Title
Essays in Development Economics

Author
Park, David Sungho

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ESSAYS IN DEVELOPMENT ECONOMICS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECONOMICS

by

David Sungho Park

June 2022

The Dissertation of David Sungho Park
is approved:

Professor Jonathan Robinson, Chair

Professor Alan Spearot

Professor Natalia Lazzati

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Essays in Development Economics

by

David Sungho Park

This dissertation contains three essays broadly related to evaluating program effects in developing countries and survey methodology.

In Chapter 1, joint work with Naresh Kumar, we evaluate the impact of a multifaceted female empowerment program on reducing intimate partner violence (IPV) in urban Liberia. We ran a randomized controlled trial (RCT) in partnership with the Liberian Red Cross. The program intervention includes intensive psychosocial therapy and vocational skills training throughout a full year. About 12 months after program completion, we find the program significantly reduced the proportion of women who experienced emotional, physical, and sexual IPV by 10-26 percentage points (from control bases of 24-62 percent). While there are multiple pathways through which IPV could be impacted, one channel is that the business training was highly effective: labor supply increased by 37 percent and expenditure by 49 percent. One focus of the program is psychological empowerment, and we find positive but statistically insignificant effects on distress and happiness indices. We also find improvements in social norms around IPV: perceived justifiability of IPV reduced by 0.3 standard deviations.

In Chapter 2, joint work with Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson and Alan Spearot, we study another important issue in this topic, which is the accurate measurement of IPV. Women may under-report intimate partner violence (IPV) due to several social and psychological factors. We conducted a measurement experiment in rural Liberia and

Malawi in which women were asked IPV questions via self-interviewing (SI) or face-to-face interviewing (FTFI). About a third of women incorrectly answered basic screening questions in SI, and SI generates placebo effects on innocuous questions even for those who "pass" screening. Because the probability of responding "yes" to any specific IPV question is less than 50 percent, and that IPV is typically reported as an index (reporting yes to at least one question), such misunderstanding increases IPV reporting. In Malawi, we find that SI increases the reported incidence of any type of IPV by 13 percentage points on a base of 20 percent; in Liberia, we find an insignificant increase of 4 percentage points on a base of 38 percent. Our results suggest SI may spuriously increase reported IPV rates.

In Chapter 3, joint work with Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson and Alan Spearot, we quantify survey fatigue by randomizing the order of questions in in-person surveys (lasting 2.5 hours on average) fielded in an evaluation of cash transfers in rural Liberia and Malawi. An additional hour of survey time increases the probability that a respondent skips a question by 10-64 percent. Because skips are more common, the total monetary value of aggregated categories such as assets or expenditures declines as the survey goes on, and this effect is sizeable for some categories: for example, an extra hour of survey time lowers food expenditures by 25 percent. Evidence from a similar experiment within high-frequency phone surveys shows that the results are not driven by the respondents deliberately choosing to skip questions in order to hasten the end of the survey, suggesting that cognitive burden is the key driver of survey fatigue.

Acknowledgments

First and foremost, I would like to thank my advisor, Jonathan Robinson, who has tirelessly guided me throughout the program. I wouldn't be here today without the incredibly generous hours he spent on me. Jon's excitement about and commitment to research have inspired me every moment in the journey. For their continuous support from the onset of each project, I am indebted to other members in my advising committee: Alan Spearot, Natalia Lazzati, and Ajay Shenoy. Without their warm encouragement and practical guidance, this dissertation wouldn't have been possible. For collaboration and helpful comments, I am grateful to my coauthors: Shilpa Aggarwal, Jenny Aker, Dahyeong Jeong, and Naresh Kumar.

I would also like to thank other faculty members at UCSC who offered constructive feedback on my research including the papers in this dissertation. I am grateful to George Bulman, Jesse Cunha, Carlos Dobkin, Robert Fairlie, Laura Giuliano, Gueyon Kim, Kenneth Kletzer, Michael Leung, Jessie Li, Justin Marion, Julian Martinez-Iriarte, Brenda Samaniego de la Parra, Dong Wei, Jeremy West, and Ariel Zucker. I would also like to thank Susan Leach, Sandra Reebie, and other UCSC staff for the administrative support throughout the PhD program.

I was fortunate to accompany the PhD program with many inspiring colleagues at UCSC: Prateek Arora, Valentina Brailovskaya, Jaehoon Choi, Dongwan Choo, Mai Hakamada, Rolly Kapoor, Luka Kocic, Jianan Liu, Bryan Pratt, Haedong Rho, Liam Rose, Monica Shandal, Youngwoo Song, Zhaoqi Wang, Brett Williams, Guanghong Xu, and Shuchen Zhao. I was not only emotionally supported by every friendly interaction, but I also learned a lot by discussing research and sharing feedback.

All papers in this dissertation were based on field work in Africa, and it would not have

been possible without the help of many people at different organizations. For data collection in the field, I am utterly grateful to the enumerators and other staff members at IPA Liberia and IPA Malawi. I would also like to thank the partner organizations, the Liberian Red Cross and GiveDirectly, for their collaboration. The projects wouldn't have been possible without the financial support from the USAID, the UK Foreign, Commonwealth Development Office, awarded through the J-PAL Crime and Violence Initiative, UC Institute on Global Conflict and Cooperation, IPA IPV Initiative, UCSC Economics Department, and UCSC Blum Center.

I couldn't be more grateful to my parents, Taeyoon Park and Changhyun Lee, for their love, encouragement, and unconditional support throughout my life. I also thank my brother, Stephen, for his continuous support. I'm also indebted to my parents-in-law and sister-in-law, who have been ever supportive.

Last and certainly not least, I'm extremely grateful to my wife, Hyojeong Kim, who is always my best supporter and companion. This dissertation, not to mention my other research, would not have been possibly completed without her support. I thank her for being there for me throughout the PhD program and the job market. I'm also grateful for our 2-year-old daughter, Lael, who has been a blessing ever since she came to our family.

Chapter 1

Reducing Intimate Partner Violence: Evidence from a Multifaceted Female Empowerment Program in Urban Liberia

1.1 Introduction

Intimate partner violence (IPV) is a serious public health problem which affects hundreds of millions of women globally. Worldwide, one in three women has experienced some form of physical or sexual IPV in their lifetime (WHO 2021; K. M. Devries et al. 2013). IPV is associated with many negative physical (Smith et al. 2017) and mental (Bacchus et al. 2018) health outcomes.[1] Moreover, IPV inflicts considerable economic costs on both survivors and society (C. Peterson et al. 2018).

There have been many policy discussions around how to effectively prevent or respond to IPV, and public health professionals recommend that a problem like IPV be targeted in multiple directions at the same time (Ranganathan et al. 2021). This is because IPV is a complex problem caused by a variety of psychological, social, and economic factors. The public health literature on IPV has been centered around the "ecological" framework (Heise 1998), where violence is conceptualized by an interaction of individual, interpersonal, and sociocultural factors. There is no single cause of violence, thus both IPV prevention and response require an intervention that addresses multiple underlying drivers.[2]

To study the effectiveness of a holistic approach to reducing IPV, we partner with the Liberian Red Cross to conduct a randomized controlled trial of a multifaceted female empowerment program in Monrovia, Liberia. The baseline prevalence of IPV is very high in Liberia. In the most recent Liberia Demographic and Health Survey (DHS) 2019-2020, 35 percent of partnered women of age 15-49 reported to have experienced physical or sexual IPV in the 12 past

[1] According to the U.S. Centers for Disease Control and Prevention (CDC), about 35% of female IPV survivors experience some form of physical injury related to IPV (Smith et al. 2017). In our study sample, about 25% of physical/sexual IPV survivors report a physical injury as a direct effect of the male partner's action of IPV.

[2] A "prevention" intervention is both to prevent violence for individuals who experienced violence earlier and to reduce the reoccurence of violence for those who already have. Note the difference from a "response" intervention, which targets at reducing revictimization of a survivor or recidivism of a perpetrator (Mary Ellsberg et al. 2015).

months. This is particularly high even compared to other African countries, a geographic region which itself is notorious for high prevalence of IPV (about 26% on average from countries where DHS data is available). There could be many explanations why IPV is highly prevalent in today's Liberia, including poverty (being one of the poorest countries in the world[3]). Yet one possible factor is the civil war that took place in 1989-2003, during which violence against civilian women and girls was weaponized (Omanyondo 2005). Research suggests that one of the hidden costs of such brutal civil war may be a persisting, permissive environment of violence in everyday lives (Steenkamp 2005).[4]

Since 2009, the Liberia National Red Cross Society (LNRCS) has run a female empowerment program targeted at marginalized women in informal settlements of Monrovia, where most of the internally displaced population fled for safety during the civil war. The program goal is to empower women economically and psychosocially so that they can self-sustain their lives and protect themselves from abuse. The program has two major components. The first is aimed at psychosocial empowerment, and includes daily group counseling sessions and cognitive behavior therapy focused on relationships with their spouses and other family member or community members. The second is to improve economic livelihoods through vocational skills and business training centered around helping beneficiaries set up and manage a small business. The program is very intensive: participants attend meetings 4-5 hours every day during the 12-month period. The total number of hours in the program is about 1,200, far more than most other programs.

Access to the program was randomized, and treatment was stratified by baseline character-

[3] CIA World Factbook.

[4] Sub-Saharan African countries with histories of internal conflict have 11%p ($p < 0.01$) higher physical or sexual IPV prevalence than countries with those (base=21%), based on authors' country-level analysis with data from K. M. Devries et al. (2013).

istics, including whether having experienced physical or sexual IPV past year. After conducting a baseline survey and randomizing the sample into treatment and control, the treatment group was invited to the program. While the original study design was to pool three cohorts (each including 400 women), due to COVID disruptions and related funding problems, our implementing partner Red Cross has been able to enroll only one cohort. This paper includes only one cohort of the sample with about 400 women.

The primary outcome of our study is the prevalence of IPV. To measure IPV, we administered the WHO's Violence Against Women module, which is a standardized questionnaire that has been extensively used and vetted by large-scale, multi-country surveys like the DHS. The module consists of 20 questions, each describing a specific IPV incidence (e.g., "Did your man ever slap you or throw something at you that could hurt you in the past 12 months?").[5] To construct our primary outcomes, responses to each yes/no question are indexed into a binary variable for each of the four categories: controlling behavior, emotional IPV, physical IPV, and sexual IPV.[6] In addition, for each IPV question, conditional on an affirmative response, a followup question is asked about how frequent such episode happened: (a) one or two times; (b) three to five times; or (c) more than five times. For each IPV category, we construct a summary index incorporating responses to these frequency questions.[7]

We have three main findings. First, we find that the intervention has sizable effects on IPV. Twelve months after program completion, it significantly reduces past-year emotional IPV by 23 percentage points (from a control base of 62 percent) and physical IPV by 26 percentage points

[5] See Appendix 2.E for full description of the IPV questionnaire.

[6] For example, Controlling Behavior Index equals to one if the respondent said yes to at least one question under the category.

[7] For each IPV categories, responses to frequency questions are standardized into a z-score using inverse covariance weighting (Michael L. Anderson 2008b).

(from 45 percent in the control). The effects on sexual IPV is 10 percentage points reduction (but insignificant). The effect sizes we find are very large compared to previous findings. For example, the cash transfer literature find effect sizes of 5-11 percentage point reductions in physical IPV (Buller et al. 2018). We also asked a set of questions for norms around IPV (e.g. "Is a husband justified in hitting or beating his wife if she burns the food?") and find that the program reduced justifiability of physical or sexual IPV by 0.3 standard deviations. This provides suggestive evidence for the change in social norms as one of the explanations for IPV reduction.

Second, we find significant improvements in economic livelihoods. Monthly expenditure increased by about $12 US from a control base of $25 (or about 49 percent). While we find no significant increase in our measure of monthly income, our survey module on expenditure is more comprehensive and contains a more exhaustive list of items, so that it could be a better measure of economic welfare (Deaton 1997). We also find the program increased labor supply on self employment by about 22 hours a month from a control base of 38 hours (or about 57 percent). This is not surprising given that the focus of the business training component of the program is on self-owned business. We find modest evidence for crowding out of labor hours from other sources, and the total labor supply hours increases by 19 hours a month (insignificant) from 51 hours in control.

Third, we find positive but statistically insignificant improvements in psychological distress and happiness. To measure distress, we use the Hopkins Symptom Checklist 10-questionnaire (HSCL-10) and construct a 1-4 scale. We find the program reduced the HSCL-10 distress index by 0.02 points (insignificant) on a control base of 2.01. For happiness we construct a summary

index from responses to the Happiness and Well-being questions in the World Values Survey,[8] and we find an effect of 0.07 standard deviation (insignificant). These results are surprising, considering that one of the major components of the program intervention is psychosocial therapy.

Recently there have been a lot of impact evaluations where IPV is an outcome. The majority of these are about cash transfers, which have increased in popularity for poverty alleviation programs. The empirical evidence shows that transfers targeted to female lead to reduction in IPV (Angelucci 2008; Hidrobo and Fernald 2013; Bobonis et al. 2013; Hidrobo et al. 2016; Haushofer et al. 2019; Roy et al. 2019),[9] and these tend to show real but modest effects in the order of about 5-11 percentage points for physical IPV (Buller et al. 2018). In a companion project in rural Liberia and Malawi (Aggarwal et al. 2020), preliminary results show unconditional cash transfers reduced proportion of women experiencing physical IPV by 2-5 percentage points (but significant only when samples are pooled).

Some studies evaluate the effect of business training programs coupled with cash transfers (Green et al. 2015; Blattman et al. 2016), but find insignificant effects on IPV.[10] While these studies are similar to ours in that they work with a marginalized population and the intervention includes business training, the intervention in our study is much more intensive. For example, about 400 hours throughout the program are spent solely on vocational skills and business

[8] Similarly to our frequency-integrated IPV indices, responses to each question are standardized into a z-score using inverse covariance weighting (Michael L. Anderson 2008b).

[9] Haushofer et al. (2019) find that IPV against women is reduced both when the cash transfers are targeted to the husband and the wife. Also, some studies find that the transfers to women lead to higher IPV for subgroups who face stronger social norms for gender roles (Angelucci 2008) or where women have the same as or higher education level than the men (Hidrobo and Fernald 2013), but overall there is less evidence that cash transfer programs increase IPV.

[10] Blattman et al. (2016) work with marginalized, war-affected women in Northern Uganda, and Green et al. (2015) extend the experiment by involving male partners, but either find no significant effects on IPV.

training, whereas in the other two studies program hours add up to about 100 hours.[11] More importantly, our intervention also includes psychosocial therapy.

In this vein, a closer study to ours is by Bandiera et al. (2020), who find that a multifaceted vocational and life skills training program to adolescent girls in Uganda decreased sex against their will, which is one form of sexual IPV. In addition to the similarities in aiming at economic empowerment, the life skills training component is similar to the psychosocial therapy in our study in that it addresses topics like conflict resolution and violence against women. However, the focus is more on sexual and reproductive health, whereas our intervention involves more intensive group counseling and cognitive behavioral therapy. The therapy sessions in our study also involve the female participants' partners and children.

This paper is also related to a growing literature on studying the effects of cognitive behavioral therapy (CBT) in developing countries. Blattman et al. (2017) find CBT coupled with $200 cash grant reduces violence committed by young men who were criminally engaged at baseline in Monrovia. Yet they find no effects on perpetrating IPV in particular. Another study in rural Kenya (Haushofer et al. 2020) finds that psychotherapy and $1,000 cash combined improve psychological wellbeing as well as economic outcomes like consumption. Instead of cash transfers, our intervention combines business training with CBT program, and we find strong evidence for improved economic livelihoods but modest effects on psychological wellbeing. This is surprising also in that the intensity of our CBT is stronger than the two other studies. The program in Blattman et al. (2017) consisted of 3 weekly sessions over 8 weeks and that in Haushofer et al.

[11] In the WINGS program evaluated by Blattman et al. (2016) and Green et al. (2015), the study sample received 4 days of training, 4-5 follow-up visits, and 3 days of self-group training (i.e., up to 96 hours total). Our intervention is unusually intensive even compared to the numerous business training programs or "graduation" programs that have been extensively tested in development economics. For example, the ILO's SIYB program (de Mel et al. 2014) included training for 7 or 9 days for 7 hours a day (i.e., 49 or 63 hours total).

(2020) 1 weekly session over 5 weeks, whereas our program involved 4-5 weekly sessions over 6 months. A recent paper by Barker et al. (2021) studies the standalone effect of CBT and finds significant improvements in mental health as well as downstream economic outcomes 3 months after the intervention.

The paper proceeds as follows. Section 1.2 describes the context and experiment and data collection. Section 3.3 presents the main results. Section 1.4 discusses possible threats to validity. Section 3.4 concludes.

1.2 Setting, Study Design, and Data

1.2.1 Context and Setting

This study was conducted in the capital city of Monrovia in Liberia, where IPV is highly prevalent. In the Liberia Demographic and Health Survey (DHS) 2019-2020, 35% of ever-partnered women of age 15-49 reported to have experienced physical or sexual IPV in the past 12 months, whereas the corresponding averages for Asian, Latin American and other African countries where DHS data is available are respectively 16%, 12%, and 26%. The study population targeted by the Red Cross reports much higher levels of IPV: in our baseline, we find that 51% of women report physical or sexual IPV in the past year.

There are numerous explanations for the high IPV prevalence in today's Liberia, including poverty.[12] Yet another contributing factor likely are the civil wars that took place in Liberia between 1989-1996 and 1999-2003 and killed around 250,000 people, amounting to approx-

[12]Liberia is one of the poorest countries in the world (CIA World Factbook) with weak institutions, and many lack access to formal education and sustainable economic activities. For example, per one of the UN's Millennium Development Goals, the net primary education enrollment in Liberia was 37% in 2016, while the average of Sub-Saharan African countries was 78% (UNESCO Institute for Statistics).

imately 10% of the population of the country then, and displaced more than another million. During the war, violence against civilians, especially women and girls, was systematically mobilized as a "weapon of war" to terrify and subdue communities. A WHO report documents that 2 in 3 Liberian women experienced sexual violence during the civil war (Omanyondo 2005).[13] Research suggests that these attitudes towards violence, once entrenched, may persist (Steenkamp 2005).[14]

1.2.2 Women Training and Integration (WIN) Program

The core intervention of this paper is a multifaceted female empowerment program called the Women Training and Integration (WIN) Program, which has been administered by the Liberian Red Cross since 2009. The program targets vulnerable women in informal settlements of Monrovia. Table 1.A2 lists the selection criteria for the WIN program. To qualify, an applicant must belong to a minimum of three groups. LNRCS has a thorough process of selecting beneficiaries. They review the application packets carefully, pay visits to the communities, and interview friends or neighbors to verify the reported information in the applications.

The program's main objective is to improve the participants' livelihoods in multiple dimensions. Specifically, the program aims at the following: 1. To economically empower women so that they can self-sustain themselves and their families; 2. To psychologically empower women so that they can better protect themselves from abuse; 3. To help establish and maintain positive relations with their families and communities; 4. To improve knowledge about and thus access to health care and psychological services.

[13] Also see Domingo et al. (2013), Jones et al. (2014), and Women (2013).

[14] Steenkamp (2005) suggests that a prolonged exposure to violence can give rise to a "culture of violence," which can be defined as "the system of norms, values, or attitudes which allow, make possible or even stimulate the use of violence to resolve any conflict or relation with another person" (Moser and Winton 2002).

The WIN program is very intensive and requires a 12-month commitment from participants, who need to be present at the WIN program center for 4-5 hours a day (either in a morning or afternoon session) for 5 days a week during the 12-month period.

The program has two major components. The first is psychosocial therapy, which includes one-to-one and group counseling sessions, thematic group discussions, cognitive behavioral therapy sessions, stress management, family/couple therapy, mediation, and conflict resolution. These aim to heal war-related trauma, reduce traumatic stress disorder, mediate family conflict situations, support coping mechanisms, build self-confidence, and promote social interaction and peaceful coexistence within their familes as well as communities.

The second is the vocational skills and business training. LNRCS offers three options for vocational skills: baking/catering, hairdressing/cosmetology, and tailoring. A participant can choose only one skill, and for those who do not have any preference, LNRCS assigns them one based on capacity constraints. The business training module provides training on handling day-to-day aspects of business, such as client interactions, sales-purchase bookkeeping, and inventory management. At the end of the program, the beneficiaries also receive business startup kits and cash grants to assist setting up their own businesses. However, due to financial constraints and COVID-related disruptions, LNRCS was not able to provide the business capital grants and cash grants for the cohort included in this paper.

The WIN program also includes several other components. The program provides routine health care check-ups and HIV/AIDS awareness and testing sessions in LNRCS's in-house clinic. Child care services are also provided when the beneficiary is at the program center. The adult literacy module targets unschooled participants and trains them in basic arithmetic, and English reading and writing skills. The curriculum is aligned with the Ministry of Education's

Alternative Learning Curriculum.

Experimental Design

The sampling frame is the pool of women who voluntarily applied to the program but selected by LNRCS through its need-based screening process. That is, our sample can be characterized by women who are disadvantaged enough for LNRCS to consider them as eligible for the program but at the same time are willing to improve their lives and have high enough agency to apply to such a program.

Several months before program start for every cohort, LNRCS advertises the program in target communities to encourage eligible women to apply. In February 2019 (for the first cohort of this study), LNRCS received about 600-700 applications in total, and after background checks and verification of the applicants' information, it shared with us a list of 450 eligible applicants divided into the "main" list of 400 and a "backup" list of 50 ranked in the order of eligibility status determined by LNRCS. In conducting the baseline survey, for those we couldn't reach after numerous attempts, we drew from the backup list in order. At the end, we enrolled 395 respondents for the study and conducted baseline in April 2019,[15] and randomly assigned 198 to treatment and 197 to control.

Treatment is stratified at two background characteristics collected in the baseline survey: (a) whether having experienced physical or sexual IPV in the past 12 months, and (b) having been affected by the civil war or having family members who have.[16]

Every woman in the treatment group was invited to the program, but some couldn't be

[15] We had completed full interviews with 400 women, but LNRCS later decided to drop anyone under 17 from the sample due to potential conflict with school enrollment.

[16] Instances include: relocation, becoming disabled/amputated, family members being killed/dead.

reached or couldn't participate in the program for other reasons, and 152 women ultimately enrolled. Moreover, due to an administrative error, 2 people from the control group were invited and joined the program. For analysis, we report both intent-to-treat (ITT) and treatment-on-treated (TOT) estimates.

Our study has been significantly affected by COVID-19 disruptions. The full design was to conduct the experiment over three cohorts for about 1,200 women, each cohort including 400.[17] The first cohort of the study was enrolled in April 2019 and the program implementation ended in March 2020, right before the government lockdowns in Liberia. However, in compliance with government restrictions on in-person activities, our partner LNRCS suspended enrollment for the second cohort. While the government restrictions have been lifted since late 2020, due to financial difficulties, as of this writing, LNRCS hasn't yet been able to resume the program, and thus this paper includes only one cohort of the sample.

1.2.3 Data Collection

We conducted the baseline survey in April 2019, and the endline in April 2021, which was about 12 months after program completion. Our primary outcome is IPV but the survey also included questions on labor supply, income, expenditure, psychological well-being, social norms around IPV, transfers, and savings.

We used the WHO's Violence Against Women module[18] to measure IPV outcomes. The module consists of a group of questions each describing an IPV-related episode, providing the respondents with multiple opportunities to report violence. These binary questions are later

[17] Such pooled design was due to LNRCS's operational constraints which allow serving up to 200 beneficiaries at a time.

[18] https://www.who.int/gender/violence/who_multicountry_study/Annex3-Annex4.pdf.

grouped into: controlling behavior, emotional, physical or sexual IPV. For all questions, we restrict the recall period to the past 12 months prior to the survey date. Appendix 2.E provides a more comprehensive description of the questionnaire.

1.2.4 IPV Measurement and Safety Protocols

We instituted WHO's ethics protocol for IPV research (WHO 2016). Study protocols have been reviewed and approved by the institutional review boards (IRBs) of the University of California, Santa Cruz and the University of Liberia, which is the relevant entity in Liberia. Second, we used the WHO's Violence Against Women module, which has been employed in multiple contexts and become a "gold standard" for IPV measurement. Third, we hired only female enumerators and provided special training both to safely conduct the interviews and to be prepared emotionally for the work. Fourth, as for the full survey itself, the survey was conducted privately without presence of anyone else than the enumerator and the respondent. Particularly for the IPV module, enumerators were trained to change questions to non-sensitive subjects in the event the survey is interrupted or eavesdropped by a third party. Fifth, while at the beginning of the whole survey respondents went through an informed consent procedure including information for the IPV, we reiterated informed consent right before the IPV module. Sixth, we prepared an information sheet that lists the services available for women experiencing IPV, including contact information for organizations where they can get help. This list was provided to every respondent who went through the IPV questionnaire, regardless of whether they reported any IPV experience.

1.2.5 Baseline Summary Statistics

Table 1.1 presents baseline summary statistics. The average age of women in the control group is about 29 years. They completed 7 years of education, on average, and about two-third of our sample have completed only primary school, while only 25% women have completed secondary school.

For the IPV questions, we restrict the sample to those who are currently partnered or have had an intimate partner 12 months prior to the survey, and the mean for this indicator at baseline was 92%.[19]

In Panel B we find that our sample had minimal access to her own income source or labor force participation. Only 11% report to have any job, and 25% are self-employed. The average income is a mere $8 dollars per month, with many zeros in the extensive margin. The mean for spouse's income is twice as large ($19). While our measures of income might not be exhaustive itself, the mean differences suggest that the women in our sample were not financially independent at baseline.

The baseline prevalence of IPV is very high. About 59% women reported having experienced emotional IPV, while the figure for the more severe form of IPV (physical or sexual) is slightly smaller (51%). This rate much higher than the national average reported in the Liberia DHS surveys, where the corresponding figures are 35% and 35% respectively in the 2019-2020 report. There could be two possible explanations. One is that our sample was selected by Red Cross in a way to be characterized as vulnerable, and one eligibility criterion was having experienced domestic abuse (Table 1.A2). Another is that the different survey tool between

[19] We later show in Table 1.A3 that this indicator is slightly unbalanced between treatment and control at endline (statistically insignificant), and also report the Lee (2009) bounds results in Table 1.A5.

our baseline and Liberia DHS 2019-2020. While our study uses the identical questionnaire to the DHS's Domestic Violence Module, at our baseline IPV was measured solely in audio computer-assisted self interviewing (ACASI), and DHS data are measured via traditional face-to-face interviewing (FTFI). In light of the findings in Section 1.4 and from our sister project in rural Liberia and Malawi (Park et al. 2021), the reported differences could be due to differing measurement modality, either through enhanced confidentiality or increased measurement error. Yet, the control group's IPV rates at our endline measured in FTFI only are still high–62% for emotional IPV, 45% for physical IPV, and 23% for sexual IPV.

Table 1.1: Baseline Summary Statistics and Randomization Check

	(1) Control Mean [SD]	(2) Treatment - Control
Panel A: Demographics		
Age	28.98	1.36*
	[7.29]	(0.73)
Years of education	7.27	0.45
	[4.11]	(0.40)
=1 if completed primary school	0.66	0.06
		(0.05)
=1 if completed secondary school	0.25	0.01
		(0.04)
=1 if currently partnered or had partner past year	0.92	-0.00
		(0.03)
Panel B: Self income and labor supply		
=1 if has own income source	0.34	0.06
		(0.05)
=1 if operated own business	0.25	0.04
		(0.04)
=1 if had any other temporary/permanent job	0.11	0.01
		(0.03)
Total income (USD)	8.38	3.36
	[27.57]	(3.09)
Panel C: Household economic well being		
Spouse's income (USD)	19.06	2.11
	[39.56]	(4.05)
Per capita expenditure (monthly, USD)	26.76	1.65
	[25.54]	(2.63)
Net value of physical assets (USD)	316.32	80.88
	[1,282.83]	(133.55)
Panel D: Intimate partner violence		
=1 if experienced the following (past 12 months):		
Controlling behavior	0.83	0.03
		(0.04)
Emotional IPV	0.59	0.00
		(0.05)
Physical IPV	0.50	-0.01
		(0.05)
Sexual IPV	0.16	0.03
		(0.04)
Physical or sexual IPV	0.51	-0.01
		(0.05)
Emotional, Physical or Sexual IPV	0.67	-0.02
		(0.05)

Note: Observations = 395.

1.2.6 Attrition Balance

In Table 1.A3, we check balance for two compliance measures: column (1) shows whether we were able to reach the respondent and complete the endline survey itself, and column (2) refers to whether she was eligible for the IPV section at endline. Given our IPV questions have a recall period of 12 months, we administered the IPV module only to those who are currently partnered or have been so in the past 12 months. Since the IPV analysis is indeed constrained to only those who went through the IPV questionnaire at all, it is necessary to check for any differential attrition in partner status. In addition, given that often in developed countries, IPV survivors are encouraged to leave the violent partner, this is also a meaningful outcome that shows how women in our study select in or out of a relationship.

For the endline survey, we were able to successfully track 359 women (91% of the baseline sample), and the attrition rate is balanced between treatment and control. We use IPV questions with a recall period of 12 months, thus we administer the IPV module to those who currently has an intimate partner or had one within 12 months prior to the survey date. Among the 359 we tracked for endline, 314 were eligible for the IPV module, and as in column (2) of Table 1.A3, we find a 2 percentage point difference between treatment and control in this partner status. While this difference is not statistically significant, we also report the Lee (2009) bound estimates for the effects on IPV outcomes in Table 1.A5.

1.3 Results

1.3.1 Effects on IPV

In this section, we examine the WIN program effects on our primary IPV outcomes. We run the following regression:

$$Y_i = \beta WIN_i + \gamma Y_{0i} + \boldsymbol{X}_{ic}'\boldsymbol{\theta} + \phi_s + \varepsilon_i, \tag{1.1}$$

where Y_i is the outcome of interest for individual i, WIN_i treatment status instrumented with original assignment, Y_{0i} baseline measurment of the outcome, $\boldsymbol{X}_i'$ a vector of individual characteristics chosen by post-double selection LASSO, and ϕ_s strata fixed effects. The coefficient of interest is β, which is the treatment-on-treated (TOT) estimates for the effects of the female empowerment program. We also report the reduced-form effects of the randomized treatment assignment. Due to problems we discuss further in Section 1.4, we exclude the random subsample for whom IPV was measured in self-interviewing modules.

The results for IPV are presented in Table 1.2. Emotional violence decreased by 23 percentage points and physical violence by 26 points from control bases of 62 percent and 45 percent, respectively.[20] The effect sizes we find are very large in comparison to the previous literature. Lighter-touch though similar interventions have shown to have null to modest effects on IPV (Green et al. 2015; Blattman et al. 2016; Bandiera et al. 2020). The cash transfer literature finds that physical violence reduces by 0-11 percentage points during the period the female receives

[20] In Table 1.A5, we show the Lee (2009) bounds results based on the difference in partner status found in Table 1.A3. For emotional IPV, the lower bound becomes statistically insignificant, but the magnitude remains fairly large with the t-statistic well greater than 1. For physical IPV, the lower bound shows a slightly smaller magnitude but remains to be statistically significant.

the transfers (Buller et al. 2018).

Table 1.2: Program Effects on IPV Indices

	(1) Controlling Behavior	(2) Emotional Violence	(3) Physical Violence	(4) Sexual Violence
Panel A. ITT				
WIN treatment	-0.02	-0.17**	-0.19***	-0.07
	(0.06)	(0.07)	(0.07)	(0.06)
Control mean	0.80	0.62	0.45	0.24
Observations	169	169	169	169
Panel B. TOT				
WIN treatment	-0.03	-0.23**	-0.26***	-0.10
	(0.09)	(0.10)	(0.10)	(0.09)
Control mean	0.80	0.62	0.45	0.24
Observations	169	169	169	169

Note: Recall period is past 12 months prior to the survey. In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, strata fixed effects, and control for ACASI vs. FTFI measurement of IPV. Standard errors in parentheses.

We next look into social norms around physical and sexual IPV. Social norms related to the acceptability of IPV has been one of the widely targeted pathways in the public health literature (Ranganathan et al. 2021). In the "social ecology" framework (Heise 1998), the dynamics between a couple are embedded in many other interpersonal relationships and the community, thus social norms around IPV is a crucial driver of IPV.

To measure social norms related to IPV acceptability, we asked relevant survey questions such as: "In your opinion, is a husband justified in hitting or beating his wife if she argues with him?" We had seven such questions and asked again each referring to what the respondent believes about the community: e.g. "In your community, is it usual for husbands to hit or beat

the wife if she argues with him?" We summarize the responses to these binary questions into a z-score per Michael L. Anderson (2008b).

Table 1.3 presents our findings on social norms around IPV. When the responses to each question are indexed, we find that justifiability of physical or sexual IPV decreases by 0.3 standard deviations. This suggests that the program did reduce the acceptability of physical or sexual IPV among the program beneficiaries and that this might have been a pathway to the reduction in actual IPV experience.

However, it's also noteworthy that most women in the control group as well report that violence is not justified in any of the given situations. Neglecting the children is where the most women said violence is justifiable in the control group (12%). Also arguing with the husband and going out without telling the husband have relatively high rates of acceptability (8% and 7% respectively). Yet, the program closes this gap, to make those cases not acceptable as excuses for violence.

In Table 1.A6, we report how women responded to similar questions but referring to what she thinks of others in her community. We find that the control means are evidently higher. One explanation is that providing affirmative responses to such questions might involve stigma or embarrassment so that when the question is directed towards others instead of the respondent herself, she might be more likely to truthfully report her belief.

Table 1.3: Program Effects on Perceived Justifiability of Physical/Sexual IPV

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	=1 if husband is justified to beat/hit wife when she:						=1 if husband is justified to force sex	Z-score
	Argues w/ husband	Goes out w/o telling	Doesn't care children	Burns food	Financial pressure	Refuses sex		
Panel A. ITT								
WIN treatment	-0.05*	-0.03	-0.08***	-0.01	-0.02	-0.01	0.01	-0.20**
	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.09)
Control mean	0.08	0.07	0.12	0.03	0.03	0.02	0.02	-0.03
Observations	359	359	359	359	359	359	359	359
Panel B. TOT								
WIN treatment	-0.06*	-0.05	-0.10***	-0.01	-0.03	-0.01	0.01	-0.26**
	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.12)
Control mean	0.08	0.07	0.12	0.03	0.03	0.02	0.02	-0.03
Observations	359	359	359	359	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment. and include strata fixed effects. Standard errors in parentheses.

1.3.2 Effects on Economic Livelihoods

Improving women's economic opportunities have been long argued as a key strategy to reducing IPV. For example, in a household bargaining model from the economics literature, increasing the wife's economic opportunities outside of the household could heighten her "threat point" and thus the husband would less likely to perpetrate violence in order to keep her in the relationship. On the other hand, if the husband's motivations are "intrumental" (e.g. to extract resources from the wife) or "backlash" (e.g. to re-assert dominance), then economically empowering the wife could lead to more IPV.[21]

In Table 1.4, we look at labor supply outcomes. We find that the program increases labor

[21] See Buller et al. (2018) for discussion of the pathways and review of related cash transfer studies.

hours for self employment by 22 hours a month (or 57 percent), while the extensive margin is not statistically distinguishable from zero. Considering the economic empowerment component of the WIN program focuses on vocational skills and business training for small businesses, this finding is not surprising. The null effect of the extensive margin is also consistent with the fact that, for the cohort we're evaluating, Red Cross was not able to provide business capital grants at the end of the program.

We check whether there was any crowding out from other sources, but we find no significant effects on either casual labor or other income sources. While it's marginally insignificant, we also find a sizeable increase in total labor hours.

In addition to the pathways discussed above, labor supply could have incapitation effects. That is, spending more time on her own business or occupation, which is likely outside of the household or intimate relationship, leads to less time spent with her partner and thus leads to a mechanical reduction in IPV.

Table 1.4: Program Effects on Labor Supply

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Self employment		Casual labor		Other income		Total	
	=1 if any	hours	=1 if any	hours	=1 if any	hours	=1 if any	hours
Panel A. ITT								
WIN treatment	0.04	16.50*	-0.03	1.33	-0.05	-3.51	-0.03	14.32
	(0.05)	(9.72)	(0.03)	(1.67)	(0.03)	(4.56)	(0.05)	(10.30)
Control mean	0.46	38.38	0.08	1.34	0.12	11.36	0.63	51.08
Observations	359	359	359	359	359	359	359	359
Panel B. TOT								
WIN treatment	0.06	21.88*	-0.04	1.77	-0.06	-4.65	-0.04	19.00
	(0.07)	(12.87)	(0.03)	(2.20)	(0.04)	(6.02)	(0.07)	(13.60)
Control mean	0.46	38.38	0.08	1.34	0.12	11.36	0.63	51.08
Observations	359	359	359	359	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include strata fixed effects. Standard errors in parentheses.

In Table 1.5, we examine how the program affected other economic outcomes. Results show that the program increased expenditure by 49 percent. The effect sizes are surprisingly large. In Table 1.A7, we show effects by expenditure categories, and we see that the effects are mostly driven by expenses on food items and nondurables. While we find no significant effects on income, our survey questions for income might not be as exhaustive as in the expenditure section to capture many income sources. Thus expenditure is our preferred measure for economic welfare.

Table 1.5: Program Effects on Economic Outcomes

	(1) Expenditure	(2) Income	(3) Food Security	(4) Net Wealth
Panel A. ITT				
WIN treatment	9.10***	-1.17	0.06	80.25
	(2.79)	(4.11)	(0.11)	(101.98)
Control mean	24.81	21.71	-0.00	453.37
Observations	359	359	359	359
Panel B. TOT				
WIN treatment	12.07***	-1.55	0.08	106.44
	(3.78)	(5.41)	(0.14)	(134.46)
Control mean	24.81	21.71	-0.00	453.37
Observations	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include strata fixed effects. Standard errors in parentheses.

1.3.3 Effects on Psychological Wellbeing

Psychological wellbeing is also a primary outcome of the program, given that counseling is one of the key "response" interventions recommended by public health experts (Ghandour et al. 2015), suggesting that IPV victimization is correlated with mental health disorders (Karen M. Devries et al. 2013; Fulu et al. 2013; Machisa et al. 2017; Trevillion et al. 2012).

We use two main outcomes. First is the distress index from the 10-question Hopkins Symptom Checklist (HSCL-10). HSCL is generally used in clinical and epidemiological settings to measure psychological distress with a fairly straightforward set of 10 questions, such as "In the past 7 days, how often were you blaming yourself for things?" Respondents choose an option among "Not at all," "A little," "Quite a bit," and "Extremely," and we add up the responses by

the assigned numeric codes. Second, we construct a happiness index using the Happiness and Well-being questions from the World Values Survey. An example question is: "In a 1 to 10 scale, how much freedom of choice and control you feel you have over the way your life turns out?" Responses to such five questions are standardized to a z-score per Michael L. Anderson (2008b).

In Table 1.6, we find rather modest effects. Both outcomes go in the expected direction, a reduction in distress and an increase in happiness, but the magnitudes are small and not statistically significant. These are indeed surprising, considering the program heavily focuses on psychological therapy sessions. Yet, the endline was 12 months after program completion, and it is possible that the effects quickly dissipated within the year. Blattman et al. (2017) and Haushofer et al. (2020) find similar results where the effect of psychotherapy sessions show significant improvement psychological wellbeing in the short term, but no effect after one year since the last therapy session.

Table 1.6: Program Effects on Psychosocial Wellbeing

	(1) Distress Index (HSCL-10)[a]	(2) Happiness Index (z-score)[b]
Panel A. ITT		
WIN treatment	-0.01	0.06
	(0.05)	(0.10)
Control mean	2.01	0.00
Observations	359	359
Panel B. TOT		
WIN treatment	-0.02	0.07
	(0.07)	(0.14)
Control mean	2.01	0.00
Observations	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include strata fixed effects. Standard errors in parentheses.
[a] 10-question Hopkins Symptom Checklist (HSCL-10).
[b] Happiness and Well-being questions from the World Values Survey, standardized per Michael L. Anderson (2008b).

1.4 Threats to Validity

1.4.1 IPV Measurement Error

A possible threat to validity of our analysis comes from the fact that our outcomes are measured by survey responses. In particular, the IPV outcomes are constructed from what women in our sample self report in our surveys, and this might lead to several concerns. In this section we address each of them.

Underreporting of IPV in surveys

It is widely concerned that IPV is underreported possibly due to factors like social taboos, feeling of shame, emotional pain, and fear of retribution (WHO 2012; Garcia-Moreno et al. 2013). However, in a professionally administered survey, these factors are likely mitigated owing to the fact that the respondent goes through an informed consent procedure where confidentiality of what she reports is assured and often the enumerator has no reason to interact with the respondent again. Yet even with underreported *levels* of IPV, these do not necessarily introduce bias to treatment effects in an impact evaluation setting, because the (nonclassical) measurement error is canceled out by taking the *differences* between treatment and control.

However, one might be concerned that the true levels of IPV become different between treatment and control (e.g. lower in the treatment if the intervention was effective), and even if the probability of IPV being underreported is contant, this could attenuate the treatment effect.[22] We cannot directly test this in this paper's setting, because the underreporting propensity is unlikely to be the same between treatment and control (discussed more in following points). Instead, in a companion project where we evaluate the effect of unconditional cash transfers in rural Liberia and Malawi (Park et al. 2021), we introduce an alternative survey tool that could alleviate social desirability bias (as we do in this paper too, and explained more below), and we find no differential cash effects on IPV between survey modes. This finding suggests that underreporting of IPV itself does not bias the treatment effects at least when the measurement error is not correlated with treatment (like unconditional cash transfers).

[22] Assume the true prevalence of IPV is $(y-\beta)$ in treatment and y in control, and that the proportion of people who truthfully report IPV is $p<1$ (constant between treatment and control). Then the estimated treatment effect based on reported IPV rates are $-p\cdot\beta$, which is smaller in magnitude than the true treatment effect β.

Experimenter demand effects

Nonetheless, the analysis in this paper could be threatened by differing IPV reporting behavior between treatment and control. One possiblity is experimenter demand effects. Given the intervention involves psychotherapy for relationships with spouses or intimate partners, the respondents in the treatment group might believe that the researchers expect them to have a better marital relationship and experience less IPV, and thus feel pressure to underreport IPV. This would overestimate the treatment effects.

Research suggests that experimenter demand effects are modest in many settings even when the researchers made the research hypothesis salient to the study sample (de Quidt et al. 2018; Dhar et al. 2018; Mummolo and E. Peterson 2019). Moreover, our endline survey was conducted by an independent survey firm that the respondents had no reason to associate with the program implementer. Also the timing of the endline was 12 months after the program had ended, so it is less likely that reporting behavior at endline was driven by the treatment.

However, to address this issue more rigorously, we cross-randomized an IPV measurement experiment at endline, where respondents answered IPV questions in either self interviewing (SI) or conventional face-to-face interviewing (FTFI). Whereas under FTFI the enumerator asks each question and the respondent responds verbally, in SI women listen to pre-recorded questions through earphones and make choices on a touchscreen by herself.[23] The main difference is that the SI module allows the respondents to report their responses anonymously to the human enumerator, which could minimize social desirability bias in IPV reporting and thus experimenter demand effects (i.e. the difference in social desirability bias between treatment

[23] We use one type of SI called audio computer-assisted self interviewing (ACASI) (Figure 2.A3). The ACASI module and the experimental design are almost identical to those of our sister project (Park et al. 2021), where we study effects of SI on IPV reporting in rural Liberia and rural Malawi.

and control).[24]

In Table 1.B5, we see the treatment effects are smaller when IPV was measured in SI, which would suggest that our main analysis based on FTFI might be driven by experimenter demand effects. However, the attenuation could be explained by measurement error introduced by the SI survey tool, which we extensively document in Park et al. (2021). If the respondent doesn't fully understand how to use the tool, she'd be making mistakes when choosing responses (classical measurement error). Since the mean of an individual yes/no IPV question is typically below 0.5, such measurement error would *increase* the rate (biased towards 0.5), and this could attenuate the treatment effect estimate in SI.[25]

In fact, a significant portion of our sample seems to be making mistakes under SI. In Table 1.B2, we find sizeable differences in how people report to a set of innocuous questions between FTFI and SI. For example, while everyone in the control under FTFI said "yes" to the questions "Did it rain in your community last year?" and "Did you sleep at all past week?", only 82% in the control group and 90% in the treatment group did so under SI. Overall, among seven questions, five of them indicate statistical significance when SI effects are pooled. Except for one question, we don't find evidence that either the treatment or control group is making less mistakes. Assuming that these questions are truly innocuous and respondents have no other reason to differentially report by FTFI and SI, the results altogether suggest that many are mak-

[24] While the original intent of SI is to minimize underreporting by protecting the respondents from feeling shame or discomfort, it is also possible on the other hand that the respondent could feel more comfortable sharing unfortunate experiences with a human being. Conducting the IPV module is typically considered a conversation, and often respondents seek counseling from the human enumerator (M. Ellsberg et al. 2001).

[25] Suppose the reported IPV rates under FTFI are $(y-\beta)$ for treatment and y for control. Under SI, assume there are two types: p fully understand the module and respond in the same way she would have under FTFI, and $(1-p)$ make mistakes under SI and randomly choose between yes and no. Then the reported rates under SI are $p\cdot(y-\beta)+(1-p)\cdot 0.5$ for the treatment and $p\cdot y+(1-p)\cdot 0.5$ for the control, and taking the difference, the estimated treatment effect under SI is $-p\cdot\beta$. This is smaller in magnitude than that under FTFI, $-\beta$, and the difference is determined by how many people don't understand the SI tool $(1-p)$.

ing mistakes in SI and the attenuation in shown Table 1.B5 is not necessarily explained by experimenter demand effects.

Enhanced sensitization of IPV

It's also possible that IPV reporting behavior is correlated with treatment in the other direction. While the treatment group becomes more sensitized of their IPV experience and more likely to truthfully report IPV, the control group might not be sensitized enough and remain underreporting IPV. This would *underestimate* the treatment effect. One could have such concern given that we find treatment effects in perceived justifiability of IPV in Table 1.3. However, it's noteworthy even among the control group, a vast majority thinks violence is not justified. One deviation is for the situation where the wife neglects the children; 13% reported that physical violence can be justified in this case, whereas the means for other questions are 3-9%. Yet, at least from what's reported, our study sample overall appears to be a context where already violence is not justified in most cases. However, even if IPV reporting behavior is significantly affected by this factor, the main results we find on IPV would be the lower bounds of the true effect.

Control group pretending to look worse

One might be concerned that the control group reports higher rates of IPV in order to look more disadvantaged. This might be plausible because our sampling frame were women who had voluntarily applied to the program for consideration. Even though this was more than two years prior to our endline, it's possible that they are still willing to be eligible for future program enrollments. However, as explained earlier, respondents had virtually no reason to link our enumeration team to the program or Red Cross. In the informed consent form we administer

at the beginning of every survey, we make it clear that no personal or identifiable information will be shared with any party, including the government or any non-government organizations. Therefore, it's unlikely that anyone in our study sample believes what she reports to us could affect her prospects for any program.

1.4.2 Incapacitation Effect

Another type of concern is that IPV experience might be reduced in the treatment group mechanically because they spend more time in the program. This could be especially concerning since the treatment group had to attend the program center 4-5 hours a day, which amounts to at least 20-25 hours a week physically away from the spouse.[26] However, our endline survey was conducted about 12 months after the program had ended, and we have no outcomes measured for more than 12 months prior to the survey. Therefore, the outcomes do not capture anything that happened while the program was running. Yet, *after* the program, as we find in Table 1.4, treatment group worked more outside of the household (and away from her partner), and it's possible this was one of the mechanisms through which IPV was reduced.

1.5 Conclusion

Our randomized evaluation of a multifaceted female empowerment program finds that it considerably reduces emotional and physical IPV experienced by women, restricting the analysis to IPV outcomes measured in a conventional setting. We also find sizeable effects on labor supply and expenditure. After 12 months since the program, we find small insignificant effects

[26] While some of controlling behavior and emotional IPV can be perpetrated remotely (e.g. over the phone), physical and sexual IPV do require physical contact.

on psychological wellbeing.

These findings suggest that a holistic approach to IPV prevention is effective. This is consistent with the public health literature on IPV emphasizing that the multi-level factors of IPV are important in designing interventions. One caveat of this study is that we cannot quantify the marginal benefit of a single program component. We leave this to future research.

Appendix 1.A

Figure 1.A1: Study Timeline and COVID-19 Disruptions

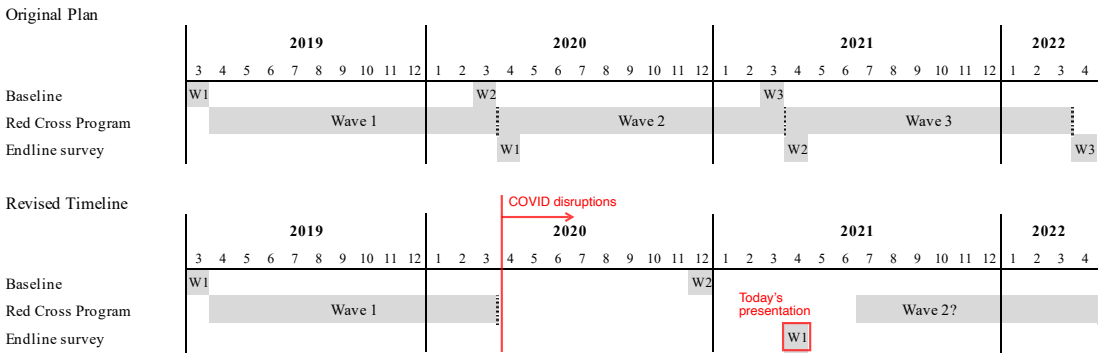


Figure 1.A2: Self Interviewing (SI) Survey Module

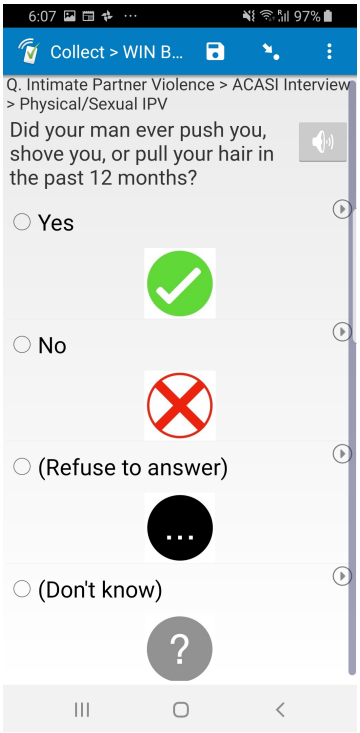


Table 1.A1: WIN Program Components

Program Component	Description
Psychological support	One-to-one and group counselling, stress management, family/couple therapy
Literacy classes	Reading and writing curriculum by Ministry of Education
Child care	During program participation
Medical checkups	Free primary medical check-ups at Red Cross clinic
Vocational skills training	Baking, cosmetology, and tailoring
Entrepreneurship training	Financial literacy, business planning/management, etc.
Business start-up capital	250 USD worth of capital along with 30 USD cash grant

Table 1.A2: Selection Criteria of WIN Program

1. Ex-combatant	5. Single mother/self-supported
2. Previous commercial sex worker	6. Illiterate
3. Victims of rape/domestic violence	7. Economically vulnerable
4. Witness of extreme violence	8. Drug user

Table 1.A3: Attrition Balance

	(1)	(2)
	=1 if completed endline survey	=1 if completed IPV survey at endline[a]
WIN treatment	0.00	-0.02
	(0.03)	(0.04)
Control mean	0.91	0.81
Overall mean	0.91	0.79
Observations	395	395

Note: Regressions include strata fixed effects. Standard errors in parentheses.
[a] IPV questionnaire is administered to only those who are currently married or has an intimate partner, or have been so in the 12 months prior to the survey.

Table 1.A4: Program Effects on Frequency-integrated IPV Indices

	(1)	(2)	(3)	(4)
	Frequency-integrated Indices[a]			
	Emotional IPV	Physical IPV	Sexual IPV	Any IPV
Panel A. ITT				
WIN treatment	-0.15	-0.30***	-0.18	-0.35***
	(0.14)	(0.11)	(0.12)	(0.11)
Control mean	0.00	-0.00	-0.00	0.00
Observations	169	169	169	169
Panel B. TOT				
WIN treatment	-0.20	-0.42**	-0.25	-0.48***
	(0.19)	(0.16)	(0.17)	(0.16)
Control mean	0.00	-0.00	-0.00	0.00
Observations	169	169	169	169

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, strata fixed effects, and control for ACASI vs. FTFI measurement of IPV. Standard errors in parentheses.

Table 1.A5: Program Effects on IPV Indices - Lee Bounds

	(1)	(2)	(3)	(4)	(5)	(6)
	Emotional IPV			Physical IPV		
	Baseline	Lower Bound	Upper Bound	Baseline	Lower Bound	Upper Bound
WIN treatment	-0.20**	-0.15	-0.26***	-0.22**	-0.16*	-0.25***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Control mean	0.62	0.59	0.68	0.45	0.41	0.49
Observations	169	162	162	169	162	162
	Sexual IPV			Any IPV		
	Baseline	Lower Bound	Upper Bound	Baseline	Lower Bound	Upper Bound
WIN treatment	-0.10	0.00	-0.11	-0.18**	-0.14	-0.23**
	(0.08)	(0.08)	(0.09)	(0.09)	(0.10)	(0.09)
Control mean	0.24	0.17	0.26	0.66	0.63	0.72
Observations	169	162	162	169	162	162

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, strata fixed effects, and control for ACASI vs. FTFI measurement of IPV. Standard errors in parentheses.

Table 1.A6: Program Effects on Perceived Others' Justifiability of Physical/Sexual IPV

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	=1 if husband is justified to beat/hit wife when she:						=1 if husband is justified to force sex	Z-score
	Argues w/ husband	Goes out w/o telling	Doesn't care children	Burns food	Financial pressure	Refuses sex		
Panel A. ITT								
WIN treatment	-0.04	-0.07	-0.11***	-0.09***	-0.02	-0.07*	-0.07**	-0.22**
	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.09)
Control mean	0.30	0.30	0.27	0.17	0.13	0.16	0.14	-0.02
Observations	359	359	359	359	359	359	359	359
Panel B. TOT								
WIN treatment	-0.05	-0.09	-0.15***	-0.12***	-0.03	-0.09*	-0.09**	-0.29**
	(0.06)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.04)	(0.13)
Control mean	0.30	0.30	0.27	0.17	0.13	0.16	0.14	-0.02
Observations	359	359	359	359	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment. and include strata fixed effects. Standard errors in parentheses.

Table 1.A7: Program Effects on Expenditure Items

	(1) Food	(2) Nondurables	(3) Clothes	(4) Education	(5) Health	(6) Religious contributions	(7) Family events	(8) Nonmedical emergency
Panel A. ITT								
WIN treatment	3.74**	4.67	1.17	0.69	0.42	0.33	-0.54	0.10
	(1.65)	(2.99)	(1.87)	(2.20)	(1.43)	(0.64)	(1.52)	(0.13)
Control mean	10.05	27.06	6.54	15.15	6.07	2.99	5.07	0.11
Observations	359	359	359	359	359	359	359	359
Panel B. TOT								
WIN treatment	4.96**	6.19	1.55	0.92	0.56	0.44	-0.72	0.14
	(2.21)	(4.00)	(2.47)	(2.90)	(1.88)	(0.85)	(2.01)	(0.16)
Control mean	10.05	27.06	6.54	15.15	6.07	2.99	5.07	0.11
Observations	359	359	359	359	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, and strata fixed effects. Standard errors in parentheses.

Table 1.A8: Program Effects on Income

	(1)	(2)	(3)	(4)
	Respondent			Spouse's income
	Self employment	Casual labor	Other job	
Panel A. ITT				
WIN treatment	3.63	-1.25	-3.55	-0.99
	(3.63)	(0.80)	(2.23)	(5.79)
Control mean	12.40	1.91	7.40	33.44
Observations	359	359	359	359
Panel B. TOT				
WIN treatment	4.82	-1.66	-4.71	-1.32
	(4.79)	(1.06)	(2.95)	(7.63)
Control mean	12.40	1.91	7.40	33.44
Observations	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, and strata fixed effects. Standard errors in parentheses.

Table 1.A9: Program Effects on Assets

	(1) Business capital	(2) Durables	(3) Livestock	(4) Savings	(5) Debt
Panel A. ITT					
WIN treatment	5.90	63.95	0.31	13.87	3.79
	(16.14)	(90.02)	(9.16)	(17.01)	(3.44)
Control mean	44.19	361.22	23.00	30.46	5.49
Observations	359	359	359	359	359
Panel B. TOT					
WIN treatment	7.83	84.82	0.42	18.40	5.02
	(21.24)	(118.73)	(12.07)	(22.38)	(4.53)
Control mean	44.19	361.22	23.00	30.46	5.49
Observations	359	359	359	359	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, and strata fixed effects. Standard errors in parentheses.

Table 1.A10: Program Effects on Interpersonal Transfers

	(1)	(2)	(3)	(4)
	Transfers sent		Transfers received	
	Spouse	Non-spouse	Spouse	Non-spouse
Panel A. ITT				
WIN treatment	-0.22	-1.53	2.59	1.68
	(0.48)	(1.48)	(4.52)	(2.89)
Control mean	1.40	6.41	37.40	8.15
Observations	278	359	278	359
Panel B. TOT				
WIN treatment	-0.28	-2.03	3.33	2.23
	(0.61)	(1.95)	(5.77)	(3.80)
Control mean	1.40	6.41	37.40	8.15
Observations	278	359	278	359

Note: In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, and strata fixed effects. Standard errors in parentheses.

Appendix 1.B: Possible Threats to Validity

Table 1.B1: SI Screening

	(1) Mean (=1 if yes)
Are you a woman?	0.98
Do you live in [the county/district where the survey is being conducted]?	0.97
In the past week, did you sleep, during day or night?	0.97
In the past year, did it rain in your village one time or more?	0.96
=1 if yes to all questions	0.90
=1 if yes to woman and rain questions	0.98
Observations	303

Note: These four questions were asked in SI to everyone included in SI measurement experiment.

Table 1.B2: SI Effects on Placebo Questions, by WIN treatment status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Questions for which answer should be yes:				Questions for which answer could be yes/no:				
			Index						
	Rain	Sleep	%(yes)	=1 if yes to all	Farm work	Market	Int'l travel	Rice	Meat
SI × WIN control (β)	-0.07**	-0.14***	-0.11***	-0.18***	0.11**	0.08	-0.04	-0.10***	-0.16*
	(0.03)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.08)
SI × WIN treatment (γ)	-0.04	-0.09**	-0.07**	-0.10**	-0.02	0.07	0.01	-0.13***	-0.13
	(0.03)	(0.04)	(0.03)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.08)
WIN	-0.01	-0.03	-0.02	-0.05*	0.03	0.02	-0.02	-0.03	0.01
	(0.02)	(0.02)	(0.01)	(0.03)	(0.04)	(0.06)	(0.04)	(0.02)	(0.08)
FTFI × WIN control mean	1.00	1.00	1.00	1.00	0.04	0.84	0.06	1.00	0.56
p-value ($\beta = \gamma$)	0.609	0.361	0.356	0.241	0.053	0.890	0.334	0.617	0.737
Observations	298	298	298	298	298	298	298	298	298
Post-estimation calculation									
Pooled SI effects	-0.06	-0.11	-0.09	-0.14	0.05	0.07	-0.01	-0.11	-0.14
p-value	0.010	0.000	0.000	0.000	0.139	0.049	0.659	0.000	0.016

Note: Regressions include individual controls (including all variables in Table 1.B6). "Screen Pass" is defined by selecting "yes" to all questions in Table 1.B1. Standard errors in parentheses.

Table 1.B3: SI Effects on IPV Questions, by WIN treatment status

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to individual question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
SI × WIN control (β)	0.01	-0.03	-0.06	0.06	-0.01
	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)
SI × WIN treatment (γ)	0.11***	0.04	0.06	0.05	0.07**
	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)
WIN	-0.11***	-0.08	-0.12***	-0.05	-0.10**
	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)
FTFI × WIN control mean	0.37	0.38	0.22	0.16	0.29
p-value ($\beta = \gamma$)	0.097	0.409	0.057	0.947	0.142
Number of individuals	298	298	297	298	298
Observations	2,056	1,184	1,776	889	5,905
Post-estimation calculation					
Pooled SI effects	0.06	0.01	0.00	0.06	0.03
p-value	0.046	0.845	0.963	0.112	0.255

Note: Observations at respondent-question level. See Table 1.B4 for index-level results. Regressions include question-level fixed effects. Standard errors clustered at individual level in parentheses.

Table 1.B4: SI Effects on IPV Indices, by WIN treatment status

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *any* question in the following category:				Any IPV
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
SI × WIN control (β)	0.09	-0.04	-0.14*	0.09	-0.02
	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)
SI × WIN treatment (γ)	0.18***	0.09	0.01	0.12*	0.10
	(0.06)	(0.08)	(0.07)	(0.07)	(0.08)
WIN	-0.02	-0.16**	-0.21***	-0.08	-0.15*
	(0.07)	(0.08)	(0.08)	(0.07)	(0.08)
FTFI × WIN control mean	0.77	0.63	0.47	0.24	0.67
p-value ($\beta = \gamma$)	0.290	0.271	0.164	0.762	0.301
Observations	298	298	298	298	298
Post-estimation calculation					
Pooled SI effects	0.14	0.03	-0.07	0.11	0.04
p-value	0.002	0.629	0.207	0.037	0.478

Note: See Table 1.B3 for question-level results.

Table 1.B5: Program Effects and SI Effects on IPV Indices - TOT - Screen Pass only

	(1)	(2)	(3)	(4)	(5)
	=1 if experienced any instance of the following category:				Any IPV
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. ITT					
WIN × FTFI (γ)	-0.01	-0.15**	-0.20***	-0.07	-0.14*
	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)
WIN × SI (β)	0.07	-0.03	-0.06	-0.04	-0.04
	(0.05)	(0.08)	(0.07)	(0.08)	(0.08)
Non-WIN × FTFI mean	0.77	0.63	0.47	0.24	0.67
Non-WIN × SI mean	0.84	0.54	0.29	0.31	0.60
p-value (β = γ)	0.315	0.280	0.177	0.755	0.339
Observations	298	298	298	298	298
Post-estimation calculation					
Pooled program effects	0.03	-0.10	-0.13	-0.06	-0.09
p-value	0.508	0.083	0.014	0.248	0.108
Panel B. TOT					
WIN × FTFI (γ)	0.00	-0.25**	-0.28***	-0.10	-0.22**
	(0.09)	(0.10)	(0.10)	(0.09)	(0.10)
WIN × SI (β)	0.12*	0.01	-0.03	-0.04	0.00
	(0.06)	(0.10)	(0.09)	(0.09)	(0.10)
Non-WIN × FTFI mean	0.77	0.63	0.47	0.24	0.67
Non-WIN × SI mean	0.84	0.54	0.29	0.31	0.60
p-value (β = γ)	0.296	0.074	0.060	0.584	0.111
Observations	298	298	298	298	298
Post-estimation calculation					
Pooled program effects	0.06	-0.13	-0.16	-0.07	-0.11
p-value	0.293	0.079	0.020	0.254	0.108

Note: Sample includes only those who passed screening, i.e. those who selected "yes" to all questions in Table 1.B1. In Panel B, regressions are TOT estimates, where the treatment indicator is instrumented with the original assignment to treatment, and include baseline measurement of outcome, and strata fixed effects. Standard errors in parentheses.

Table 1.B6: SI Randomization Check

	(1) Control Mean [SD]	(2) Treatment - Control
Panel A. Demographics		
=1 if currently married or has partner	0.88	0.02
		(0.04)
Age	30.44	1.74**
	[6.79]	(0.83)
Number of household members	5.06	0.72**
	[2.71]	(0.35)
Panel B. Education and digital literacy		
Years of education	8.28	-0.52
	[4.14]	(0.47)
=1 if able to write/read in English	0.84	0.03
		(0.04)
=1 if has access to mobile phone	0.89	-0.00
		(0.04)
Panel C. Household wealth		
Food security index (z-score)	0.00	-0.05
	[1.00]	(0.12)
Total expenditure (monthly)	124.08	3.06
	[83.00]	(10.21)
Net value of durables, livestock, and financial asset	421.43	121.86
	[828.44]	(108.39)
Non-agricultural income (monthly)	21.45	3.96
	[38.75]	(4.77)
Panel D. Empowerment-related outcomes		
=1 if has her own income source	0.60	0.04
		(0.06)
Number of children	2.35	0.47**
	[1.68]	(0.20)
Observations	303	

Table 1.B7: Post-SI Survey of Technical Difficulties Self-reported by Respondents

	(1)
	Mean (=1 if yes)
Was the audio loud enough to hear?	0.99
Was the audio speaking speed okay?	0.98
Was it easy for you to remember the meaning of pictures?	0.97
Was it easy for you to choose answers on the screen?	0.97
Was it easy for you to move between questions on the screen?	0.97
Observations	145

Note: Questions were asked only to those in the SI treatment group (i.e., the FTFI group did not get these questions).

Appendix 1.C: Survey instrument

Controlling behavior

1. Did your man ever try to keep you from seeing your friends in the past 12 months?
2. Did your man ever try to stop you from meeting or speaking to your family of birth in the past 12 months?
3. Did your man ever need to know where you are all the time in the past 12 months?
4. Did your man ever stop talking to you or treat you with no interest in the past 12 months?
5. Did your man ever get angry if you speak with another man in the past 12 months?
6. Did your man often think that you are unfaithful in the past 12 months?
7. In the past 12 months, did your man ever expect you to ask for his approval before you go to a health clinic or hospital?

Emotional IPV[27]

1. Did your man ever insult you or make you feel bad about yourself in the past 12 months?
2. Did your man ever make you feel small in front of other people in the past 12 months?
3. Did your man ever mean to scare you (for example, by the way he looked at you, by yelling and bursting things) in the past 12 months?
4. Did your man ever threaten to hurt you or someone you care about in the past 12 months?

[27] For each IPV question, if the answer is "yes", a follow-up question about frequency appears, asking whether it happened (i) one or two times, (ii) three to five times, or (iii) more than five times.

Physical IPV[29]

1. Did your man ever slap you or throw something at you that could hurt you in the past 12 months?
2. Did your man ever push you, shove you, or pull your hair in the past 12 months?
3. Did your man ever hit you with his hand or with something else that could hurt you in the past 12 months?
4. Did your man ever kick you, drag you or beat you up in the past 12 months?
5. Did your man ever mean to choke or burn you in the past 12 months?
6. Did your man ever threaten to use or actually use a gun, knife or other weapon against you in the past 12 months?

Sexual IPV[29]

1. Did your man ever physically force you to do man and woman business when you did not want to in the past 12 months?
2. Did you ever do man and woman business when you did not want to because you were afraid of what your man might do in the past 12 months?
3. In the past 12 months, while doing man and woman business, did your man ever force you to do something that made you feel small or bad about yourself?

Non-sensitive placebo questions

1. Did it rain in your village one time or more in the past year?

2. Did you do any farm work in the past year?

3. Did you sleep in the past week, during day or night?

4. Did you go to the market in the past week?

5. Did you travel outside of Liberia in the past week?

6. Will you, or anyone in your household, eat any rice next week, one time or more?

7. Will you, or anyone in your household, eat any type of meat next week, one time or more?

Chapter 2

Private but Misunderstood? Evidence on Measuring Intimate Partner Violence via Self-Interviewing in Rural Liberia and Malawi

2.1 Introduction

Intimate Partner Violence (IPV) is a pressing global public health and policy problem, but measuring its true prevalence is challenging because factors including social taboos, emotional pain, fear of retribution, or feelings of shame or embarrassment cause women to hesitate in reporting IPV to friends or family, as well as to physicians or to law enforcement officials (WHO 2012; Garcia-Moreno et al. 2013). Spurred by the lack of systematic data on IPV and recognizing its epidemiological nature, organizations such as the WHO began to run large-scale, multi-country surveys to measure the prevalence of IPV in the 1990s (WHO 1996).[1] These surveys reveal that nearly a third of ever-partnered women have experienced physical or sexual IPV during their lifetime (K. M. Devries et al. 2013).

Many public health professionals worry that the true rate of IPV may be higher, and that women may be understating their IPV experience even in surveys. It remains unclear if this is the case. On the one hand, some of the stigmas that drive under-reporting may be mitigated by the confidentiality afforded by a professionally done survey (as articulated in a consent form, for example), and by the fact that the surveyor is unlikely to be known by the survey respondent or her partner, or to have reason to interact with the respondent again. The survey setting also differs critically from that in normal life because the survey *directly asks* about IPV, rather than leaving the onus of initiating the conversation to the woman herself.[2] On the other hand, some of the same stigmas may still apply; for example, the victim may feel ashamed about her situation, hesitate to confide in another individual, or be scared of being overheard (despite

[1] For example, the WHO multi-country study on women's health and domestic violence was initiated in 1997 and the DHS Program started collecting information on IPV in 1990 (the first IPV module was fielded as part of the standard DHS in Colombia).

[2] In fact, the medical literature has identified one of the key measurement approaches for IPV to simply ask the person. The WHO also recommends direct questioning as the "gold standard" method of measuring IPV. See: https://apps.who.int/iris/bitstream/handle/10665/85239/9789241564625_eng.pdf.

survey precautions to guard against this).

To address some of these concerns, an alternative approach that is widely recommended is the use of confidential self-interviewing (SI). In this approach, women self-administer IPV questions privately, which ensures that their answers are shielded even from the enumerator.[3] In this paper, we evaluate one such interviewing technique which is known as Audio Computer Assisted Self-Interviewing (ACASI). In ACASI, respondents listen to pre-recorded questions via headphones and respond using a touchscreen (in this case, a tablet).[4] The enumerator has no interaction with the respondent during this part of the survey, other than to explain the module at the beginning, and to be available in case the respondent seeks clarification.[5]

The intent of ACASI is that it will destigmatize IPV reporting, which is expected to lead to an increase in reporting. However, there are two other factors which may muddy the waters (especially when benchmarked against a professionally-administered survey with a trained enumerator). One, self-interviewing lacks any human element, and it is conceivable that respondents may actually be more likely to report sensitive behaviors to a human interviewer since the respondent may perceive the enumerator to be empathetic or build a rapport with her over the course of the survey.[6] If this channel is present, ACASI will actually understate IPV.

A second factor, which is the focus of our paper, is that self-interviewing requires the re-

[3]The answers are not fully anonymized, however, since researchers have access to this data later on, but instructions during this part of the survey, and the consent form, clearly indicate that this data will be kept securely, so any risk of data breach is remote. Further, the researchers would have no reason to interact directly with the respondent outside of this research setting.

[4]Another reason to recommend ACASI is that it is virtually impossible for the interview to be overheard by anyone in close vicinity. However, this is not relevant in our experiment (or in any survey which uses best practice face-to-face interviewing), since the survey is always conducted privately.

[5]In our survey protocol, the respondent could pause the module to ask questions, and the enumerator could help her to resume from where she left off.

[6]Indeed, Mary Ellsberg et al. (2001) compile anecdotes from debriefings of IPV survey enumerators in Nicaragua recounting how they were moved or distressed by the respondents' IPV experiences, and some even reported that respondents sought their counsel during or at the end of the IPV module.

spondent to understand the questions on her own, and to use the tablet, which may not be easy. This is an especially salient concern in the case of IPV as the standard set of questions for measuring IPV has fairly complex and nuanced language, and therefore it may not be straightforward to grasp without the surveyor helping with interpretation. In almost every setting, misunderstanding will tend to cause IPV to be *over*-reported. This is because IPV is measured through a module containing 20 questions which are later indexed into 4 main categories (controlling behavior, emotional violence, physical violence, and sexual violence). Typically, the mean of each of these individual yes/no questions is well under 0.5, so a woman who does not understand the module and randomly answers yes or no will tend to bias the level of reported IPV on any given question upwards. This problem will be exacerbated in the indexing (which is set equal to 1 if the respondent reported any form of IPV).

To shed light on these various channels, we conduct a measurement experiment within surveys collected as part of an evaluation of an unconditional cash transfer program in rural Liberia and Malawi (Aggarwal et al. 2020). Women were individually randomized into whether the IPV module was asked via face-to-face interviewing (FTFI) or over self-interviewing (SI). Baseline IPV rates differ dramatically across the two samples: the proportion of women experiencing any type of IPV over the past year is 20% in Malawi but 38% in Liberia; as such, we opt to present all results separately by country.

We have three main findings. First, we check for respondents' understanding of the tool through 5 non-sensitive screening questions, for which the answer should universally be yes. These were administered to *all* respondents through SI, irrespective of the modality through which they were asked the IPV module. These questions are (1) Are you a woman?, (2) Do you live in the [location where the survey is being conducted]?, (3) Has it rained in your village

in the past year?, (4) Have you slept at all in the past week?, and (5) Have you heard of the Coronavirus? Altogether, we find that about a third of the women in each country sample do not seem to understand the screening questions, as evidenced by not answering "yes" to all of them.[7] Even the responses to the most basic questions on gender and location are not unanimously affirmative, with 2-5% of the women making errors on the gender question and nearly 10% doing so on location. In total, we find that 38% of women in Malawi and 30% in Liberia do not "pass" these screening questions.

Second, after screening, further questions were randomized to be administered either by FTFI or ACASI. As part of the IPV module, we included a further set of innocuous "placebo" questions; since these are administered either by FTFI or ACASI, we can estimate placebo treatment effects. The placebo includes 4 questions for which the answer could be yes or no: (1) Did you do any farm work in the past year?; (2) Did you go to the market in the past week?; (3) Will you, or anyone in your household, eat any [rice/maize] next week, one time or more?; and (4) Will you, or anyone in your household, eat any type of meat next week, one time or more? If ACASI is accurate, we should find no ACASI effects on these placebo questions (at least for those who cleared screening). Yet, we find placebo effects even for those who passed screening; surprisingly, these placebo effects are similarly sized for those who pass screening and who do not. We interpret these results as suggesting that even among those who get screened in, many do not understand the questions.[8]

[7] The research team found such responses even during pre-testing, and so repeatedly refined the surveys; however, such reporting remained. While we have no definitive answer to why women answer as they do, one anecdote is that women with young children or who are nursing interpret "sleep" as being about getting restful sleep. We have no explanation for the rain question.

[8] In order to allay concerns that some of those who passed screening may be answering "yes" to all questions, and did not in fact understand the module, we also conduct robustness with a 6th screening question, for which the answer should be "no" for most women: "In the past week, have you traveled outside the country?" The placebo effects are nearly identical when we include this question. However, this is imperfect since some women may in fact have traveled abroad (since parts of the study regions are close to international borders and people routinely

Third, we find that SI increases IPV reporting, but that this increase may be entirely spurious. The increase is dramatic, for all categories of IPV (i.e. controlling behavior, emotional, physical, and sexual IPV): on a given question, 7% of women in Malawi and 14% in Liberia report yes, and ACASI increases this percentage by 5 percentage points in Malawi and 3 percentage points in Liberia. As an index, the effects are even larger, at least in Malawi, where the probability of emotional, physical or sexual IPV increased by 5-10% points, on a base of 20%. In Liberia, the effects are more modest: 1-8% points (significant for only sexual IPV) on a base of 38%. Naively interpreted, the increase in IPV we document would match the narrative that women are hesitant to report IPV, and that FTFI dramatically understates prevalence. However, we know from the screening questions that at least 1/3 of women do not seem to understand ACASI, and the effect sizes for our placebo results are similar to those for IPV. Our interpretation is that ACASI is not appropriate, at least for these populations, and researchers should be extremely cautious about using ACASI.

Our paper is related to a large but as yet inconclusive literature about the effects of ACASI on measuring sensitive behaviors. Studies comparing ACASI and FTFI in a variety of contexts suggest that self-administration increases the reporting of sensitive behaviors.[9] However, since researchers typically do not have an objective measure of the underlying behavior, it is not clear whether this increase is indicative of increased truthful reporting, or miscomprehension. Falb et al. (2016) tested ACASI with adolescent girls in the DRC and in refugee camps along the

cross into the neighboring country to buy and sell wares). Therefore, we do not include this in our main set of screening questions.

[9]See Tourangeau and Yan (2007) for a review. There are also examples where FTFI is more effective in some contexts. For example, Fincher et al. (2015) find evidence that FTFI is more effective than ACASI in screening women for IPV in WIC clinics in the US. In one of the few studies set in Africa, Cullen (2020) finds that the likelihood of reporting IPV over ACASI is no higher than reporting it in standard FTFI. In a study set in the context of a syringe exchange program, Newman et al. (2002) find that ACASI increases reporting of stigmatized behaviors but decreases reporting of psychological distress.

Sudan-Ethiopia border, and report that self-reported *average* ACASI comprehension levels are only 90% for the DRC and 75% for the Sudan-Ethiopia border, a level similar to our study. Park and Kumar (2022), a concurrent study to ours in Monrovia, Liberia, find that even urban and educated women struggle with ACASI comprehension.

Our paper is also related to a broader recent literature about survey methodologies aimed at preserving respondent confidentiality; ACASI is only one of these.[10] Other methods rely on indirect responses, such as list experiments or randomized response techniques.[11] There is no consensus on the efficacy of these methods.[12] For example, Chuang et al. (2021) finds logical errors in a list experiment and a randomized response technique focused on sexual and reproductive behavior. A recent set of evaluations compares list experiments with FTFI interviews specifically for measuring IPV and the evidence is mixed. While Agüero and Frisancho (2021) find no difference in IPV reporting between the two methods among urban microfinance borrowers in Lima, Peru, Cullen (2020) finds that the list method leads to substantially larger reporting of IPV among rural women in Nigeria and Rwanda.

The rest of this paper is laid out as follows. Section 3.2 describes the experiment and data collection. Section 3.3 presents our main results. Section 2.4 discusses evidence on potential pathways and heterogeneity. Section 3.4 concludes.

[10] Researchers have also tried other unconventional methods to measure IPV indirectly, such as asking female community leaders, but these efforts have not been very successful (Agüero et al. 2020).

[11] In list experiments, yes/no questions about sensitive behaviors are included in a list with other innocuous binary-response questions, and subjects report the number of items for which the answer is "yes" or "no", which allows the researcher to back out the population level prevalence of a behavior without being able to identify whether a specific individual engaged in that behavior. The randomized response technique (RRT) bundles a question with a random event, such as a throw of the dice. Respondents report "yes" if either the truth is "yes" or if the die is a certain number.

[12] See Höglinger and Jann (2018) and Lensvelt-Mulders et al. (2005) for reviews.

2.2 Data and Experimental Design

2.2.1 Setting

The ACASI experiment we analyze was done as part of an endline survey for a cash transfer RCT in Liberia and Malawi (the transfers were implemented by the NGO GiveDirectly, as part of a USAID-funded study). The study takes place in Bong and Nimba counties in Liberia, and in Chiradzulu and Machinga districts in Malawi. The study includes 300 villages in each country, with half of the villages receiving cash transfers.[13] While we do not evaluate the transfers themselves in this study, one important detail is that villages were included in the study only if they fell below a population threshold (as measured in the most recent population census).[14] The reason for this is that transfers were given out universally in treatment villages, and so our partner NGO chose smaller villages to be able to preserve their liquidity.

In Liberia, we implemented the project in two waves: a smaller first wave (90 villages), which had its endline in late 2020; and a bigger second wave (210 villages), which had its endline in September-November 2021. Most of our ACASI protocols were developed, tested, and refined over the course of the Wave 1 endline. Therefore, this sample is excluded from our results, and our results for Liberia are restricted to Wave 2 only. In Malawi, all 300 villages were enrolled at once and the endline was in April-July 2021.[15]

In both countries, we attempted to enroll 10 households per village into data collection for program evaluation, though in some cases we were only able to enroll fewer households.[16]

[13] The average transfer amount was $500, and was randomized between three amounts ($250, $500, and $750). In Liberia, transfers were also randomized between being paid as "lump sum" or quarterly transfers.

[14] In Malawi, the upper threshold was 100 household per village; in Liberia, it was 125, reflecting the larger village sizes in the study region.

[15] Figure 3.A1 presents the project timeline.

[16] The total sample size for the cash transfer study is 2,715 in Liberia and 2,944 in Malawi. Yet we conducted the IPV module at endline only to women who currently have a partner or have been so in the past 12 months prior to the survey date. For the IPV measurement experiment, we further excluded those who reported to have vision or

Surveys were targeted at female heads of households (male heads were interviewed only when the female was not present, and could not be reached within a few days of the baseline study. When the male head was interviewed, the IPV module was not asked).

2.2.2 Questionnaire Design and ACASI Experiment

Measuring Intimate Partner Violence

To measure IPV, we employed WHO's standard Violence Against Women module.[17] The questionnaire includes 20 questions about experience with specific forms of violence, over a time period of 12 months prior to the survey. Following the literature, we group these questions into four categories: controlling behavior, emotional IPV, physical IPV, and sexual IPV. In conducting this module, we followed WHO's ethics protocol for IPV research (WHO 2016).[18] The IPV module was administered to all women who had an intimate partner within the 12 months preceding the survey. The sample to which the IPV module was asked is 2,998 women (1,737 in Malawi and 1,261 in Liberia).

ACASI Implementation

In ACASI, respondents listen to questions on headphones and answer questions privately on a tablet. In each country, audio readings of the questions were recorded by an enumerator who was chosen for having clear enunciation. The recorded audio files were uploaded to SurveyCTO, along with an image file containing choice options (i.e. "yes" / "no" / "refuse to

hearing impairment. We are also excluding the Liberia Wave 1 sample from analysis in this paper. As a result, the total sample size for this paper is 1,737 women in Malawi and 1,231 in Liberia.

[17] The WHO's standard questionnaire for measuring IPV, which is widely used for measuring IPV, can be found here: https://www.who.int/gender/violence/who_multicountry_study/Annex3-Annex4.pdf. Our survey module on IPV can be found in the Appendix.

[18] The protocol includes: hiring only female enumerator; training enumerators to safely conduct the interviews and to be prepared emotionally for the work; conducting all surveys privately; reiterating consent just before the IPV module; and providing all respondents with an information sheet that listed the services available for women experiencing IPV (including law enforcement and local hospitals).

answer" / "don't know"). As shown in Figure 2.A2, the resulting interface on the tablet has a speaker icon (which the respondent could touch to listen to the question) and four images (from which the respondent could choose her answer by touching the screen herself).

In the field, the enumerator explained how to take the module, and then demonstrated how to conduct the module by going through a handful of practice questions with the respondent, and making sure that she could clearly hear the audio and accurately choose the option she intends to. When the respondent felt ready to take the actual module, the enumerator handed the tablet over to the respondent for her to take the module. In order to make sure that she had complete privacy while doing so, the enumerator kept sufficient distance to be unable to see the screen but remained in the same room or vicinity to be available to answer questions. When the respondent hands back the tablet, the screen is blank so that the responses are blinded to the enumerator.

Experimental Design

In each survey round for each country, half of the sample was randomly assigned to ACASI and the other half to FTFI. However, before starting the IPV questions, *every* respondent was asked to take 5 "screening" questions via ACASI.[19] The answers to all of these questions are expected to be yes: (1) "Are you a woman?"; (2) "Do you live in [the county/district in which the survey is being conducted]?"; (3) "In the past week, did you sleep?"; (4) "In the past year, did it rain in your village?"; and (5) "Have you heard of the coronavirus?" We also asked one question that would likely be answered "no": "Have you traveled outside the country in the past week?" We do not use this in our main specifications, however, because some women

[19]These screening questions were added after piloting, when it became apparent that women were answering unexpectedly to innocuous placebo questions.

could potentially travel across borders (especially in Nimba county in Liberia, which borders Guinea, and Machinga district in Malawi, which borders Mozambique), due to which "no" is not a perfect benchmark.

After screening, women began questions in their experimental group (ACASI or FTFI). As discussed throughout the paper, this module included questions on IPV; however, it also included questions on psycho-social well-being, as well as "placebo" questions. These placebo questions were meant to be innocuous and free from any social desirability bias or stigma, and were meant to be a further tool to calibrate the effects of ACASI, and included 4 questions: (1) Did you do farm work in the past year?; (2) Did you visit the market in the past week?; (3) Will you eat maize/rice in the next week?;[20] and (4) Will you eat meat in the next week?

Other Subtreatments

To explore possible technical reasons for misunderstanding, we cross-cut multiple sub-treatments. First, we randomized whether the "yes" or "no" option would appear at the top of the screen (Figure 2.A3). This randomization was implemented in order to test whether respondents are more or less likely to pick the first option. Second, in order to examine possible learning effects in which respondents became more comfortable with the method with more experience, we randomized whether the placebo questions come before or after the IPV module.

2.2.3 Summary Statistics and Balance Check

Table 2.1 shows summary statistics by country sample, as well as the difference between the ACASI and FTFI groups. Panel A shows household demographics. Because the sample is restricted to women with an intimate partner at any point during the past 12 months, the pro-

[20]This question is about the staple food, which is maize in Malawi and rice in Liberia.

portion of women who are currently partnered is very high (97%). The average respondent is about 37-38 years old and lives in a household with 5-6 members. Panel B shows education and mobile phone ownership. Average educational attainment is 5.2 years in Malawi and only 2.4 in Liberia. Sixty-six percent of women in Malawi are literate, compared to 30% in Liberia. Mobile phone ownership is similar in the two countries, ranging from 42% in Liberia to 45% in Malawi.

Panel C shows some indicators of household income and wealth, and reveals that households are better off in Liberia than in Malawi: average total monthly household expenditures are $26 in Malawi and $66 in Liberia, or about $0.17-0.39 in per capita daily expenditures. In Malawi, the average household reports about $160 worth of assets, compared to $420 in Liberia. Most of the households in the study villages are subsistence farmers, and the average monthly non-agricultural income measured in our surveys is $8-10.

Panel D shows a few proximate indicators related to female empowerment. Forty-four percent of women in Malawi have their own income source, compared to 31% in Liberia. The age difference (in years) between husband and wife is 2.9 in Malawi and 4.1 in Liberia.

Turning to Columns 2 and 4, we find 2 outcomes for which the differences are significant at 10% in Malawi (food security and total expenditure, which are both lower in the ACASI group), and none in Liberia. While the randomization appears to show no cause for concern, we present results separately with and without controls, and find no difference in results. In any case, we control for all variables reported in Table 2.1 for the main analysis.

2.3 Results

2.3.1 Screening questions

We start by documenting responses to the five screening questions which were administered to all respondents via ACASI. Results, which are shown in Table 2.2, suggest major cause for concern. Only 95-98% report being a woman, and 91-93% report living in their county/district of residence. Even more surprisingly, only 78-86% report that they slept in the past week and 83-85% report that it rained in the past year. While we do not have a good explanation for these results, an *ex post* explanation from some of our field staff was that some women interpreted the sleep question as "getting a good night's sleep," to which some women reported no. We do not have a good explanation for the rain question, but again, an *ex post* explanation is that women may have interpreted it as meaning whether it rained "enough." The reasons for misinterpretation of these questions notwithstanding, the bottom line is that even these *simple* questions were very likely misinterpreted, raising concerns about how well the more nuanced IPV questions would be understood.

Taking the questions together, we find that only 62% of the respondents in Malawi and 70% in Liberia correctly answered all the questions. This finding alone shows that ACASI will be problematic, since presumably the other 30-38% of women will not be able to use ACASI effectively. These women are not randomly selected, and in Table 2.A1, we show that in Malawi less educated women are more likely to fail the screening (though in Liberia the correlation surprisingly goes the other way). The screening results alone make it clear that it is impossible to estimate a population level prevalence using ACASI. Moreover, the opposing directions of the correlation in Table 2.A1 suggests that it may also not possible to predict the suitability of

SI for a sample of women *ex ante*.

In any case, to shed further light on the use of SI for measuring IPV, we separate women who "passed" screening from those that did not for the next set of analyses.

2.3.2 Placebo effects

Next, we examine effect of ACASI on the innocuous placebo questions. Specifically, we run the following regression:

$$Y_{ic} = \beta SI_{ic} \times ScreenPass_{ic} + \gamma SI_{ic} \times NonPass_{ic} + \delta ScreenPass_{ic} + \boldsymbol{X}'_{ic}\boldsymbol{\theta} + \phi_c + \varepsilon_{ic}, \quad (2.1)$$

where $ScreenPass_{ic}$ is equal to 1 if individual i chose yes to all of the five questions in Table 2.2, and 0 otherwise (and $NonPass_{ic}$ is the complement). $\boldsymbol{X}$ is a vector of covariates including all variables in Table 2.1.[21] The coefficients of interest are β, which represents the SI effects for those who passed the screening, and γ, which is the SI effect for those who did not. We also present p-values for a test of equality of β and γ.

Results are presented in Table 2.3, separately for Malawi (Panel A) and Liberia (Panel B). In Malawi, for those who did pass screening, we find placebo effects on 3 of 4 questions (visiting the market, eating maize, and eating meat). These effects are large, ranging from 8-15 percentage points. For those that did not pass, there are significant effects on 2 of 4 outcomes. Surprisingly, the effects are, if anything, somewhat large for those that passed; nevertheless, we cannot reject equality for any outcome other than eating maize next week (Column 3). At the bottom of the Panel, we show the effect of SI for the average respondent (i.e. a weighted average of β and γ), and test for significance. Effects are highly significant for 3 of 4 outcomes, and economically large. The pattern is largely the same in Panel B, where 2 of 4 outcomes are

[21] Results without controls are shown in Appendix 2.B, and show essentially identical results.

significant for both those who passed and those who did not, and where equality is not rejected for any outcome.[22]

Table 2.A3 runs a version of these results with a 6th screening question: "Did you travel internationally in the past week?" This question was added to try to include a question for which the answer should be no, and so could address the concern that some of the people who pass screening simply answer yes to everything, and may in fact, not understand ACASI. However, we do not use it as primary measure because some people do travel across boundaries, especially in Nimba county, Liberia, which borders Guinea. Results are very similar with this screening definition.

Finally, in Table 2.A4 and Table 2.A5, we examine heterogeneity in placebo effects with background characteristics that might be correlated with being able to complete the module, including education, mobile phone access, literacy, and age. Our results from Malawi suggest that these background characteristics have no bearing on comprehension; while there is some suggestive evidence from Liberia these characteristics matter. While this evidence is not definitive, it is suggestive that ACASI might be more effective with educated younger women, who have had more experience with mobile phones,[23] at least in the Liberian context. However, given the contrast in findings from Liberia and Malawi, taken together, these results again underscore the near impossibility of making any *ex ante* judgments about the suitability of ACASI for any given context.

[22] In Table 2.A2, we decompose the ACASI effect on the probability of choosing "don't know" or "refuse to answer" and find small effects of farm work, market visit, and meat. However, in Malawi, we find a decline in the probability of responding in this manner to the maize question. The overall probability of choosing these options is low for the farm and market question, but 27-29% for the other questions in Malawi. For the main analysis, observations for such responses ("don't know" or "refuse to answer") are dropped.

[23] This is similar to the finding in Falb et al. (2016).

2.3.3 Implications of placebo effects on measured IPV prevalence

These results show clear placebo effects, even among those who pass screening, and strongly suggest that some of those who pass screening still do not understand the questions. These placebo effects suggest that there will be a spurious effect on IPV reporting in ACASI.

To get some rough sense of the magnitude of this problem, we assume there are 2 types of women: those that understand the question and answer correctly, and others who do not understand and who simply randomly choose yes or no. Based on Section 2.3.1, we know that at least 38% of women in Malawi and 30% in Liberia do not understand ACASI, as measured by failing screening (though misunderstanding is clearly higher than this, so this is a conservative estimate).

If the true prevalence of an IPV measure is p, then the rate under ACASI will be

$$p_{ACASI} = (1-q) * p + q * 0.5 \tag{2.2}$$

where q is the proportion of women who do not understand the module. As shown in the next section, p is about 0.07 in Malawi and 0.14 in Liberia (assuming that the rates reported in FTFI are true prevalence). Thus with a q of 0.38 in Malawi and 0.30 in Liberia, p_{ACASI} could be as high as 0.25 in Liberia and 0.23 in Malawi (an 11-16 percentage point increase). Ultimately, these effects will be even worse, because IPV is typically reported as an index (equal to 1 if a women reported *any* violence in a given category).

2.3.4 Effect of ACASI on IPV reporting

Next we show the ACASI effects on the main outcome of interest, IPV. In this section, we pool those who pass and did not pass screening (based on the placebo results).[24] We first estimate a regression at the *question* level:

$$IPV_{icq} = \beta SI_{ic} + \boldsymbol{X}'_{ic}\boldsymbol{\theta} + \phi_c + \psi_q + \varepsilon_{icq} \tag{2.3}$$

where IPV_{icq} is the binary indicator of whether individual i in country sample c responded yes to question q, and ψ_q question-level fixed effects. All other notation is the same as Equation (2.1). We report results separately for each category of IPV: controlling behavior, emotional IPV, sexual IPV, and physical IPV. In a second analysis, we estimate the same equation but for the IPV index, which is set equal to 1 if a respondent reported violence on any question in that category.

The question level results are presented in Table 2.4 for Malawi (Panel A) and Liberia (Panel B). For Malawi, all effects are statistically significant, and range between 1 (physical IPV) and 9 percentage points (controlling behavior). A specification that pools all question categories together (Column 5) finds a 5 percentage point increase in reporting, also significant. In Liberia, effects are slightly more modest, where only 3 of 5 coefficients are significant, and effect sizes range from 1-5 percentage points. However, as discussed above, these results are well within the bounds suggested by the placebo effects.

In Table 2.5, we show results at the index level. The findings are qualitatively similar to the ones for individual questions, although results here differ dramatically by country. In Malawi,

[24] Disaggregated results are shown in Table 2.A6.

ACASI increases emotional IPV by 10 percentage points (base 16%), physical IPV by 5 percentage points (base 8%), and sexual IPV by 6 percentage points (base 7%). Across all forms of IPV (not including controlling behavior), ACASI increases prevalence by 13 percentage points, a 65% increase on the base of 20%. In Liberia, effects are positive but surprisingly much more modest: the index of any form of IPV increases by only 4 percentage points (on the much higher base of 38%).[25]

To sum up our findings, we find clear evidence that ACASI dramatically increases IPV reporting (at least in one country, Malawi). While it is possible that some of this increase is indeed indicative of destigmatization, it is also entirely possible that the effects are driven purely by comprehension difficulties. Our results suggest that caution is warranted in using ACASI, at least in settings like these.

2.4 Investigation of Heterogeneity and Pathways

Debriefing: did technical problems impede understanding?

A simple hypothesis for these results is that technical problems made it hard to understand or complete the ACASI module, and therefore, a technically superior module may eliminate the purported miscomprehension. We believe that this is unlikely as before implementing these protocols, we extensively pre-tested the modules, especially after early results showed similar patterns to those reported here. We carefully tested that the audio instructions were well articulated and read at a reasonable speed, and refined the implementation over time. Nevertheless, technical difficulties could have remained.

[25] In Table 2.A7, we analyze whether ACASI increases the likelihood of respondents picking "don't know" or "refuse to answer." We find that the probability of these answers is miniscule in FTFI, and SI leads to a small increase in these being chosen.

To evaluate this, after the respondent handed back the tablet to the enumerator, she asked a handful of debriefing questions about whether the respondent had faced any technical or comprehension difficulties during the module, which we present in Appendix 2.C. As shown in Table 2.C1, only 1-2% reported technical issues; most respondents could hear the module, and felt the recordings were slow enough to understand. In Table 2.C2, we regress answers to these technical questions on passing screening. We find no correlation here, which is perhaps not surprising given the low level of technical difficulties. We find no evidence that simple technical problems were the explanation.

On the other hand, we show in Table 2.C3 that 8-12% reported comprehension difficulties with the module, in remembering which picture meant "yes" (a green check) and which meant "no" (a red cross), or in using the tablet. In Table 2.C4, we regress passing screening on these measures of self-reported comprehension. In Malawi, we see no correlation, but in Liberia we see that people who reported understanding the module were more likely to pass screening (though significantly so only for one measure). This is consistent with the idea that some people had trouble understanding how the module worked.

Subtreatments

While informative, these are only debriefing questions. To shed further light on this, we randomized several subtreatments to evaluate technical components of the module (results are presented in Appendix 2.D). First, to examine whether the location of the choice options on the screen affects reporting, we randomized the order of the yes and no options. This sub-treatment was motivated by our suspicion that when in doubt, some women may have the tendency to simply choose the first option. We start by analyzing this for the placebo questions in Table 2.D1, and

find evidence that respondents were more likely to choose yes when it appears at the top of the choice options in Malawi, although not in Liberia. Surprisingly, however, in Table 2.D2, we find no evidence of the presence of such behavior in either country when it comes to the IPV questions. We have no good explanation for why this may be the case.

Second, in order to check for the possibility that respondents may get better at understanding the module with practice, we randomized the order between the non-sensitive placebo questions and the IPV questions. Specifically, for half the sample (both FTFI and SI), the IPV questions came before the placebo ones, while for the other half, this order was reversed. For the placebo questions (Table 2.D3), we find no effect of ordering, other than for the farm work question in Malawi. However, the effect goes contrary to the expected direction as the placebo effect of SI comes about when the placebos come later (i.e., practice does not help). That said, we do not wish to make much of this lone coefficient, as the placebo effect of SI is the same for "placebos first" and for "IPV first" in all other cases. For the IPV questions, we report coefficients in (Table 2.D4), and find that IPV reporting increases for sexual IPV if the placebos come first. Overall, for Malawi, the effect of SI on the probability of answering a question "yes" is about 4 percentage points if the IPV questions come first, but 6 percentage points if the placebos come first (p-value for difference = 0.217). We find no significant effect of the ordering in Liberia either. This finding is consistent with the possibility that survey fatigue causes measurement error to increase, though it is also possible that the increase in IPV is real and that women became more familiar with the module over the course of the survey. We leave a further investigation of this channel to future work.

2.5 Conclusion

In this paper, we test the efficacy of ACASI versus FTFI in eliciting truthful responses regarding IPV from women respondents in the context of a cash transfer experiment in rural Liberia and Malawi. Our results suggest that women do not understand ACASI, as evidenced by the fact that 1/3 of women incorrectly answer basic screening questions and that, even among those who pass, we observe a strong ACASI effect on innocuous placebo questions. This lack of understanding will tend to *increase* IPV reporting, since the rate of IPV is much less than 50%. And indeed, we do find a striking increase in reported IPV in one country (Malawi). However, this result is likely entirely spurious. This is deeply concerning because measurement error goes in the same direction as destigmatization, and so what looks like a decrease in stigma could be purely fictional. SI could therefore give very misleading results.

Our results, combined with our read of the literature, suggest that there may be greater benefit from having well-trained, empathetic enumerators than from SI in the context of measuring IPV. For example, in a natural experiment in Serbia, respondents of a WHO-run IPV survey ended up getting randomly assigned to either a previously inexperienced but well-trained enumerator (training duration of 2.5 weeks) or to an experienced, professional enumerator, but with less than a day of IPV training.[26] While 21% of the women reported having experienced physical or sexual IPV to the untrained enumerators, 26% reported IPV to the trained ones (Jansen et al. 2004).

Another relevant data quality issue that we want to note from a companion study in the same setting (Jeong et al. 2021) is that time into the survey at the point at which a question is

[26]This was done in an effort to speed up the fieldwork midway through surveying after the assassination of then Prime Minister Zoran inić in March 2003.

asked appears to adversely impact response quality, a phenomenon known as survey fatigue. While fatigue would be a consideration for any survey, it may be particularly germane for IPV measurement as most surveys place the IPV module at the end - for example, the standard DHS surveys ask about domestic violence at the end; we also chose to always place the IPV module at the end, even as we randomized the location of other survey modules within the survey. While this is usually done to minimize shame or embarrassment stemming from continued interaction with the enumerator after having answered the IPV module, we reiterate our read that concerns about stigmatization from the enumerator are very likely overblown,[27] and purported remedial actions, such as SI or late placement within the survey may be opening up non-obvious channels of bias.

[27] In our study, we asked our enumerators a few debriefing questions after the IPV module and they reported that among respondents who reported any IPV incidence in FTFI, 44% in Liberia and 28% in Malawi shared more about their IPV experience with them than what was asked, suggesting that stigmatization by the enumerators may not be a big concern.

Table 2.1: Summary Statistics and Experimental Balance

	(1)	(2)	(3)	(4)
	Malawi[a]		Liberia[b]	
	FTFI	SI	FTFI	SI
	Mean [SD]	- FTFI	Mean [SD]	- FTFI
Panel A. Demographics				
=1 if currently married or has partner	0.97	-0.01	0.97	-0.00
		(0.01)		(0.01)
Age	37.97	-0.94	37.13	0.67
	[12.88]	(0.60)	[10.96]	(0.61)
Number of household members	5.03	-0.02	5.59	-0.09
	[1.78]	(0.09)	[2.27]	(0.13)
Panel B. Education and mobile phone ownership				
Years of education	5.22	0.01	2.44	-0.01
	[3.50]	(0.17)	[3.43]	(0.19)
=1 if able to write/read	0.66	-0.01	0.30	0.01
		(0.02)		(0.03)
=1 if has access to mobile phone	0.45	0.03	0.42	-0.01
		(0.02)		(0.03)
Panel C. Household wealth				
Food security index (z-score)	0.00	-0.09*	0.00	0.05
	[1.00]	(0.05)	[1.00]	(0.06)
Total expenditure (monthly)	26.03	-2.13*	65.71	-0.52
	[24.46]	(1.17)	[47.08]	(2.59)
Net value of durables, livestock, and financial asset	162.55	4.24	416.43	33.88
	[235.93]	(11.45)	[823.80]	(51.15)
Non-agricultural income (monthly)	10.27	0.96	7.84	0.85
	[16.73]	(0.81)	[20.52]	(1.13)
Panel D. Empowerment-related outcomes				
=1 if has her own income source	0.44	0.01	0.31	-0.03
		(0.02)		(0.03)
Age difference from spouse	2.94	-0.16	4.09	0.72
	[10.78]	(0.51)	[12.59]	(0.72)
Observations	1,737		1,261	

Note: Sample is restricted to women with an intimate partner over the 12 months prior to the survey, and those who do not report any vision or hearing impairments. Columns 1 and 3 present the mean for the FTFI groups, and Columns 2-4 show the difference between the ACASI and FTFI groups. Standard deviation is in square brackets in Columns 1 and 3 and standard error in parentheses in Columns 2 and 4. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.2: Self-interviewing (SI) Screening Questions

	(1)	(2)
	Mean (=1 if yes)	
	Malawi	Liberia
Questions for which answer should be yes:		
1. Are you a woman?	0.95	0.98
2. Do you live in [the county/district where the survey is being conducted]?	0.91	0.93
3. In the past week, did you sleep, during day or night?	0.78	0.86
4. In the past year, did it rain in your village one time or more?	0.83	0.85
5. Have you heard about Coronavirus?	0.93	0.94
Summary measures for "passing" screening		
=1 if YES to all five questions	0.62	0.70
Observations	1,737	1,261

Note: These five questions were asked in ACASI to everyone included in ACASI measurement experiment.

Table 2.3: Effect of Self-interviewing (SI) on Placebo Questions

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
SI × Screen Pass (β)	-0.01	0.08***	0.13***	0.15***
	(0.01)	(0.03)	(0.03)	(0.03)
SI × Non-Pass (γ)	-0.03	0.09**	-0.00	0.07*
	(0.02)	(0.04)	(0.04)	(0.04)
Screen Pass	0.02	0.05	0.03	-0.01
	(0.02)	(0.03)	(0.04)	(0.03)
FTFI × Non-Pass mean	0.93	0.43	0.52	0.28
p-value (β = γ)	0.445	0.875	0.018	0.121
Observations	1,718	1,713	1,345	1,228
Effect of SI for the average respondent				
Pooled SI effects	-0.02	0.09	0.08	0.11
p-value	0.132	0.000	0.003	0.000
Panel B. Liberia				
SI × Screen Pass (β)	0.01	0.07**	-0.03**	0.00
	(0.03)	(0.03)	(0.01)	(0.03)
SI × Non-Pass (γ)	0.02	0.12**	-0.06**	-0.03
	(0.04)	(0.05)	(0.03)	(0.05)
Screen Pass	0.04	0.09**	0.03*	0.05
	(0.04)	(0.04)	(0.02)	(0.04)
FTFI × Non-Pass mean	0.77	0.61	0.95	0.65
p-value (β = γ)	0.836	0.415	0.266	0.637
Observations	1,259	1,260	1,226	1,101
Effect of SI for the average respondent				
Pooled SI effects	0.01	0.08	-0.04	-0.01
p-value	0.548	0.001	0.002	0.794

Note: Regressions are at the respondent-question level. Regressions include individual controls (including all variables in Table 2.1). "Screen Pass" is defined by selecting "yes" to all questions in Table 2.2. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.4: Effect of Self-interviewing (SI) on IPV (Individual Questions)

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *individual* question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI	0.09***	0.05***	0.01*	0.03***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
FTFI mean	0.11	0.07	0.03	0.04	0.07
Number of individuals	1,715	1,711	1,712	1,709	1,716
Observations	11,887	6,802	10,181	5,095	33,965
Panel B. Liberia					
SI	0.05***	0.01	0.01	0.03***	0.03**
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
FTFI mean	0.20	0.19	0.09	0.04	0.14
Number of individuals	1,259	1,259	1,259	1,259	1,259
Observations	8,752	5,006	7,508	3,758	25,024

Note: Regressions are at the respondent-*question* level (violence is not aggregated into indexes). See Table 2.5 for results in which IPV questions are aggregated into indices. Regressions include question-level fixed effects and individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

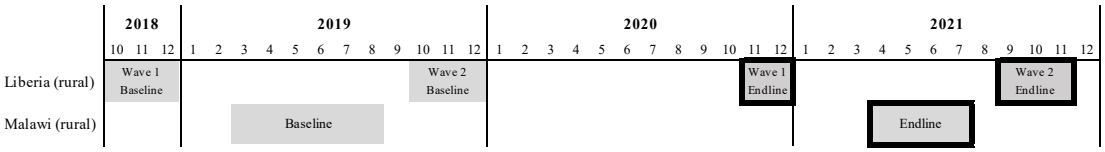
Table 2.5: Effect of Self-interviewing (SI) on IPV Indices

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *at least one* question in the following category:				Any IPV
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI	0.18***	0.10***	0.05***	0.06***	0.13***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
FTFI mean	0.38	0.16	0.08	0.07	0.20
Observations	1,737	1,737	1,737	1,737	1,737
Panel B. Liberia					
SI	0.07***	0.04	0.01	0.08***	0.04
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
FTFI mean	0.57	0.34	0.23	0.07	0.38
Observations	1,261	1,261	1,261	1,261	1,261

Note: IPV measures are indexed by category; index is set equal to 1 if the respondent answered "yes" to any question in the category. Regressions include individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Appendix 2.A: Additional Figures and Tables

Figure 2.A1: Timeline of Survey Activities



Note: Bold rectangles refer to the survey rounds where ACASI vs. FTFI randomization was implemented. Liberia Wave 1 sample is excluded from our results in this paper, as most ACASI protocols were developed, tested, and refined during Liberia's Wave 1 Endline.

Figure 2.A2: Self-interviewing Module

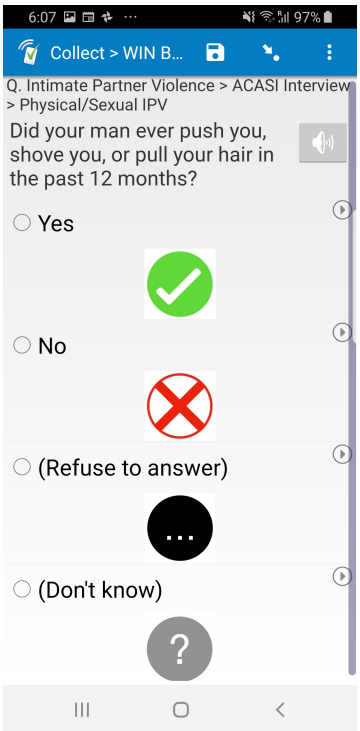
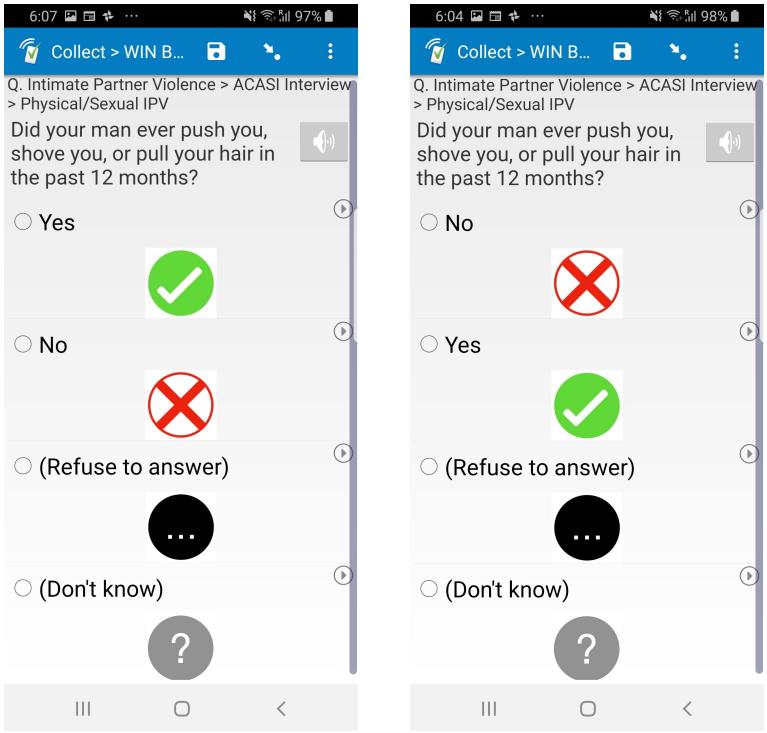


Figure 2.A3: Appearance of Module with “yes” or “no” option appearing first



Notes: Women would see either the display on the left or right.

Table 2.A1: Correlates of "Passing" ACASI Screening Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	=1 if passed SI screening[a]							
	Malawi				Liberia			
Years of education	0.011***			0.017***	-0.012***			-0.011*
	(0.003)			(0.006)	(0.004)			(0.006)
=1 if able to write/read in English		0.046*		-0.037		-0.065**		0.016
		(0.025)		(0.038)		(0.029)		(0.044)
=1 if has access to mobile phone			0.009	-0.009			-0.017	-0.028
			(0.023)	(0.024)			(0.026)	(0.027)
R-square	0.006	0.002	0.000	0.013	0.009	0.004	0.000	0.039
Overall mean of outcome	0.62	0.62	0.62	0.62	0.70	0.70	0.70	0.70
Observations	1,737	1,737	1,737	1,737	1,261	1,261	1,261	1,261

Note: Columns 1-3 and 5-7 present bivariate regressions. Columns 4 and 8 include all variables in Table 2.1, but other coefficients are not reported for space. Standard errors clustered at individual level are in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

[a] Passing threshold is choosing "yes" for all the five questions in Table 2.2.

Table 2.A2: Effect of ACASI on Choosing "Don't know" or "Refuse to answer" in Placebo Questions

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
SI	0.013***	0.013**	-0.080***	-0.000
	(0.005)	(0.005)	(0.020)	(0.022)
FTFI mean	0.005	0.007	0.266	0.294
Observations	1,737	1,737	1,737	1,737
Panel B. Liberia				
SI	0.003	0.002	0.017*	-0.011
	(0.002)	(0.002)	(0.009)	(0.019)
FTFI mean	0.000	0.000	0.020	0.132
Observations	1,261	1,261	1,261	1,261

Note: Regressions include individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.A3: Effect of Self-interviewing (SI) on Placebo Questions, Alternative Definition of Passing

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
SI × Screen Pass (β)	-0.01	0.07**	0.11***	0.10***
	(0.01)	(0.03)	(0.03)	(0.04)
SI × Non-Pass (γ)	-0.03	0.10**	-0.01	0.06
	(0.02)	(0.04)	(0.04)	(0.04)
Screen Pass	0.02	0.06*	0.02	0.05
	(0.02)	(0.03)	(0.04)	(0.04)
FTFI × Non-Pass mean	0.93	0.43	0.53	0.26
p-value (β = γ)	0.262	0.660	0.024	0.496
Observations	1,718	1,713	1,345	1,228
Effect of SI for the average respondent				
Pooled SI effects	-0.02	0.08	0.06	0.08
p-value	0.144	0.001	0.021	0.002
Panel B. Liberia				
SI × Screen Pass (β)	-0.00	0.07**	-0.02*	0.02
	(0.03)	(0.03)	(0.01)	(0.03)
SI × Non-Pass (γ)	0.05	0.10**	-0.06**	-0.03
	(0.04)	(0.04)	(0.02)	(0.05)
Screen Pass	0.04	0.05	0.02*	0.02
	(0.03)	(0.04)	(0.01)	(0.04)
FTFI × Non-Pass mean	0.78	0.64	0.96	0.68
p-value (β = γ)	0.247	0.581	0.118	0.443
Observations	1,259	1,260	1,226	1,101
Effect of SI for the average respondent				
Pooled SI effects	0.02	0.09	-0.03	-0.00
p-value	0.490	0.001	0.002	0.969

Note: Alternatively "Screen Pass" is defined by not only selecting "yes" to all questions in Table 2.2 but also choosing "no" to the question "Did you travel outside of the country in the past week?" By this alternative definition, 59% in Malawi and 62% in Liberia are in the "Screen Pass" group. Regressions are at the respondent-question level. Regressions include individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.A4: Heterogeneity in Effects of ACASI on Placebo Questions (Malawi)

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Primary education completion				
SI × Primary Educ (β)	-0.02	0.10***	0.06	0.07*
	(0.01)	(0.03)	(0.04)	(0.04)
SI × No Primary Educ (γ)	-0.01	0.07**	0.10***	0.16***
	(0.02)	(0.03)	(0.04)	(0.03)
Primary Educ	0.03	0.03	0.04	0.10*
	(0.02)	(0.05)	(0.06)	(0.05)
FTFI × No Primary Educ mean	0.93	0.41	0.52	0.22
p-value (β = γ)	0.749	0.659	0.404	0.096
Observations	1,718	1,713	1,345	1,228
Panel B. Access to mobile phone				
SI × Mobile (β)	-0.01	0.13***	0.09**	0.09**
	(0.02)	(0.03)	(0.04)	(0.04)
SI × No Mobile (γ)	-0.02	0.05	0.07*	0.14***
	(0.02)	(0.03)	(0.04)	(0.03)
Mobile	-0.02	-0.02	-0.01	0.00
	(0.02)	(0.03)	(0.04)	(0.04)
FTFI × No Mobile mean	0.95	0.46	0.53	0.26
p-value (β = γ)	0.549	0.096	0.763	0.314
Observations	1,718	1,713	1,345	1,228
Panel C. Able to read/write in English				
SI × English (β)	-0.02*	0.10***	0.06*	0.10***
	(0.01)	(0.03)	(0.03)	(0.03)
SI × No English (γ)	-0.01	0.06	0.12***	0.15***
	(0.02)	(0.04)	(0.05)	(0.04)
English	0.03	-0.04	0.01	-0.00
	(0.02)	(0.05)	(0.05)	(0.05)
FTFI × No English mean	0.92	0.40	0.51	0.23
p-value (β = γ)	0.499	0.385	0.239	0.334
Observations	1,718	1,713	1,345	1,228
Panel D. Age				
SI × Below-median Age (β)	-0.03*	0.07**	0.04	0.10***
	(0.02)	(0.03)	(0.04)	(0.04)
SI × Above-median Age (γ)	-0.00	0.11***	0.12***	0.13***
	(0.02)	(0.03)	(0.04)	(0.04)
Below-median Age	-0.00	0.05	0.05	0.09*
	(0.02)	(0.05)	(0.05)	(0.05)
FTFI × Above-median Age mean	0.94	0.41	0.52	0.27
p-value (β = γ)	0.188	0.397	0.120	0.511
Observations	1,718	1,713	1,345	1,228

Note: Regressions include individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.A5: Heterogeneity in Effects of ACASI on Placebo Questions (Liberia)

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Primary education completion				
SI × Primary Educ (β)	0.07	0.00	-0.01	-0.06
	(0.05)	(0.05)	(0.03)	(0.06)
SI × No Primary Educ (γ)	-0.00	0.11***	-0.04***	0.01
	(0.02)	(0.03)	(0.01)	(0.03)
Primary Educ	-0.11*	0.01	-0.03	-0.06
	(0.06)	(0.07)	(0.03)	(0.07)
FTFI × No Primary Educ mean	0.82	0.66	0.98	0.68
p-value (β = γ)	0.159	0.097	0.194	0.337
Observations	1,259	1,260	1,226	1,101
Panel B. Access to mobile phone				
SI × Mobile (β)	0.08**	0.05	-0.03*	0.02
	(0.03)	(0.04)	(0.02)	(0.04)
SI × No Mobile (γ)	-0.03	0.11***	-0.04***	-0.03
	(0.03)	(0.03)	(0.02)	(0.04)
Mobile	-0.07**	0.09**	0.00	-0.00
	(0.03)	(0.04)	(0.01)	(0.04)
FTFI × No Mobile mean	0.83	0.63	0.98	0.69
p-value (β = γ)	0.017	0.293	0.677	0.313
Observations	1,259	1,260	1,226	1,101
Panel C. Able to read/write in English				
SI × English (β)	0.02	0.04	-0.02	0.01
	(0.04)	(0.05)	(0.02)	(0.05)
SI × No English (γ)	0.01	0.11***	-0.04***	-0.02
	(0.03)	(0.03)	(0.01)	(0.03)
English	0.00	0.03	-0.01	0.01
	(0.04)	(0.05)	(0.02)	(0.05)
FTFI × No English mean	0.81	0.66	0.98	0.69
p-value (β = γ)	0.861	0.203	0.512	0.597
Observations	1,259	1,260	1,226	1,101
Panel D. Age				
SI × Below-median Age (β)	-0.01	0.09**	-0.02	0.01
	(0.03)	(0.04)	(0.02)	(0.04)
SI × Above-median Age (γ)	0.03	0.08**	-0.05***	-0.02
	(0.03)	(0.04)	(0.02)	(0.04)
Below-median Age	0.04	-0.03	-0.01	0.04
	(0.04)	(0.05)	(0.02)	(0.05)
FTFI × Above-median Age mean	0.82	0.68	0.98	0.70
p-value (β = γ)	0.399	0.816	0.277	0.607
Observations	1,259	1,260	1,226	1,101

Note: Regressions include individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.A6: Does the effect of ACASI differ between those who pass screening and those who don't? (Individual IPV Questions)

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *individual* question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI × Screen Pass (β)	0.08***	0.05***	0.01*	0.03***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SI × Non-Pass (γ)	0.10***	0.06***	0.01	0.03*	0.05***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Screen Pass	-0.00	-0.00	0.00	-0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
FTFI × Non-Pass mean	0.12	0.08	0.03	0.04	0.07
p-value (β = γ)	0.502	0.765	0.683	0.779	0.791
Number of individuals	1,715	1,711	1,712	1,709	1,716
Observations	11,887	6,802	10,181	5,095	33,965
Effect of SI for the average respondent					
Pooled SI effects	0.09	0.05	0.01	0.03	0.05
p-value	0.000	0.000	0.070	0.001	0.000
Panel B. Liberia					
SI × Screen Pass (β)	0.05***	0.02	0.02	0.02**	0.03**
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
SI × Non-Pass (γ)	0.06**	0.00	-0.01	0.04**	0.02
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Screen Pass	-0.02	-0.01	-0.01	-0.01	-0.01
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
FTFI × Non-Pass mean	0.23	0.21	0.11	0.06	0.16
p-value (β = γ)	0.840	0.711	0.358	0.419	0.860
Number of individuals	1,259	1,259	1,259	1,259	1,259
Observations	8,752	5,006	7,508	3,758	25,024
Effect of SI for the average respondent					
Pooled SI effects	0.05	0.01	0.01	0.03	0.03
p-value	0.000	0.479	0.434	0.005	0.013

Note: Regressions are at the respondent-*question* level (violence is not aggregated into indexes). Regressions include question-level fixed effects and individual controls (including all variables in Table 2.1). "Screen Pass" is defined by selecting "yes" to all questions in Table 2.2. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.A7: Effect of ACASI on Choosing "Don't know" or "Refuse to answer" in IPV Questions

	(1)	(2)	(3)	(4)	(5)
	=1 if don't know or refusal to *individual* question in following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI	0.011**	0.007*	0.011**	0.006**	0.012**
	(0.006)	(0.003)	(0.005)	(0.003)	(0.006)
FTFI mean	0.017	0.009	0.015	0.007	0.017
Number of individuals	1,737	1,737	1,737	1,737	1,737
Observations	12,159	12,159	12,159	12,159	34,740
Panel B. Liberia					
SI	0.007**	0.002	-0.001	0.003**	0.004
	(0.003)	(0.002)	(0.003)	(0.001)	(0.003)
FTFI mean	0.005	0.004	0.007	0.001	0.006
Number of individuals	1,261	1,261	1,261	1,261	1,261
Observations	8,827	8,827	8,827	8,827	25,220

Note: Regressions are at the respondent-*question* level (violence is not aggregated into indexes). Regressions include question-level fixed effects and individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Appendix 2.B: Main results, without controls

Table 2.B1: Effect of Self-interviewing (SI) on Placebo Questions, no individual controls

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
SI × Screen Pass (β)	-0.01	0.08***	0.11***	0.11***
	(0.01)	(0.03)	(0.03)	(0.03)
SI × Non-Pass (γ)	-0.03	0.09**	-0.02	0.04
	(0.02)	(0.04)	(0.04)	(0.04)
Screen Pass	0.02	0.07*	0.04	0.02
	(0.02)	(0.03)	(0.04)	(0.04)
FTFI × Non-Pass mean	0.93	0.43	0.52	0.28
p-value (β = γ)	0.478	0.808	0.019	0.189
Observations	1,718	1,713	1,345	1,228
Effect of SI for the average respondent				
Pooled SI effects	-0.02	0.08	0.06	0.09
p-value	0.150	0.001	0.020	0.001
Panel B. Liberia				
SI × Screen Pass (β)	0.01	0.07**	-0.03**	0.01
	(0.03)	(0.03)	(0.01)	(0.03)
SI × Non-Pass (γ)	0.04	0.11**	-0.06**	-0.03
	(0.04)	(0.05)	(0.03)	(0.05)
Screen Pass	0.05	0.08**	0.03**	0.06
	(0.04)	(0.04)	(0.02)	(0.04)
FTFI × Non-Pass mean	0.77	0.61	0.95	0.65
p-value (β = γ)	0.515	0.546	0.299	0.541
Observations	1,259	1,260	1,226	1,101
Effect of SI for the average respondent				
Pooled SI effects	0.01	0.09	-0.04	-0.00
p-value	0.500	0.001	0.002	0.948

Note: Regressions are at the respondent-question level. "Screen Pass" is defined by selecting "yes" to the first five questions in Table 2.2. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.B2: Effect of Self-interviewing (SI) on IPV Reporting in Individual Questions, no individual controls

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *individual* question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI	0.09***	0.06***	0.01*	0.03***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
FTFI mean	0.11	0.07	0.03	0.04	0.07
Number of individuals	1,715	1,711	1,712	1,709	1,716
Observations	11,887	6,802	10,181	5,095	33,965
Panel B. Liberia					
SI	0.05***	0.00	0.00	0.03***	0.02**
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
FTFI mean	0.20	0.19	0.09	0.04	0.14
Number of individuals	1,259	1,259	1,259	1,259	1,259
Observations	8,752	5,006	7,508	3,758	25,024

Note: Regressions are at the respondent-*question* level (violence is not aggregated into indexes). See Table 2.B3 for results in which IPV questions are aggregated into indices. Regressions include question-level fixed effects. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.B3: Effect of Self-interviewing (SI) on IPV Indices, no individual controls

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *at least one* question in the following category:				Any IPV
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI	0.18***	0.11***	0.05***	0.06***	0.14***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
FTFI mean	0.38	0.16	0.08	0.07	0.20
Observations	1,737	1,737	1,737	1,737	1,737
Panel B. Liberia					
SI	0.07**	0.03	0.01	0.07***	0.03
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
FTFI mean	0.57	0.34	0.23	0.07	0.38
Observations	1,261	1,261	1,261	1,261	1,261

Note: IPV measures are indexed by category; index is set equal to 1 if the respondent answered "yes" to any question in the category. "Screen Pass" is defined by selecting "yes" to all questions in Table 2.2. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Appendix 2.C: Self-reported technical and comprehension difficulties

Table 2.C1: Debriefing Survey on Technical Issues with ACASI Module

	(1)	(2)
	Mean (=1 if yes)	
	Malawi	Liberia
Was the audio loud enough to hear?	0.99	0.98
Was the audio speaking speed okay?	0.99	0.98
Observations	866	616

Note: Questions were asked only to those in the ACASI treatment group (i.e., the FTFI group did not get these questions).

Table 2.C2: Relationship between Reporting Technical Difficulties and Passing Screening

	(1)	(2)	(3)	(4)
	=1 if passed SI screening[a]			
	Malawi		Liberia	
=1 if said:				
audio loud enough to hear	0.203		-0.043	
	(0.143)		(0.127)	
audio speaking speed okay		0.016		-0.044
		(0.220)		(0.127)
R-square	0.002	0.000	0.000	0.000
Outcome mean when said no	0.42	0.60	0.75	0.75
Observations	866	867	616	615

Note: Standard errors clustered at individual level are in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.
[a] Passing threshold is choosing "yes" for the first five questions in Table 2.2.

Table 2.C3: Debriefing Survey on Comprehension Issues with ACASI Module

	(1)	(2)
	Mean (=1 if yes)	
	Malawi	Liberia
Was it easy for you to remember the meaning of pictures?	0.90	0.90
Was it easy for you to choose answers on the screen?	0.91	0.88
Was it easy for you to move between questions on the screen?	0.92	0.88
Observations	866	616

Note: Questions were asked only to those in the ACASI treatment group (i.e., the FTFI group did not get these questions).

Table 2.C4: Relationship between Reporting Comprehension and Passing Screening

	(1)	(2)	(3)	(4)	(5)	(6)
	=1 if passed SI screening[a]					
	Malawi			Liberia		
=1 if said:						
easy to remember the meaning of pictures	0.019			0.136**		
	(0.055)			(0.065)		
easy to choose answers on screen		-0.001			0.074	
		(0.057)			(0.059)	
easy to move between questions on screen			-0.003			0.074
			(0.060)			(0.058)
R-square	0.000	0.000	0.000	0.008	0.003	0.003
Outcome mean when said no	0.60	0.62	0.62	0.59	0.64	0.64
Observations	866	865	866	616	616	616

Note: Standard errors clustered at individual level are in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.
[a] Passing threshold is choosing "yes" for the first five questions in Table 2.2.

Appendix 2.D: Subtreatments

Table 2.D1: Effect of Ordering of Yes and No Options in ACASI on Placebo Questions

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
YES First	-0.02	0.05	0.06*	0.07*
	(0.02)	(0.03)	(0.04)	(0.04)
NO First mean	0.94	0.54	0.58	0.34
Observations	854	851	708	615
Panel B. Liberia				
YES First	0.00	0.03	0.01	-0.01
	(0.03)	(0.04)	(0.02)	(0.04)
NO First mean	0.82	0.74	0.94	0.70
Observations	615	616	595	542

Note: Includes only those who are in the ACASI group (FTFI group excluded). Regressions include country sample fixed effects and individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.D2: Effect of Ordering of Yes and No Options in ACASI on IPV Questions

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *individual* question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
YES First	-0.01	0.00	0.00	-0.00	-0.00
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
NO First mean	0.21	0.13	0.05	0.07	0.12
Number of individuals	858	854	855	852	859
Observations	5,915	3,385	5,062	2,531	16,893
Panel B. Liberia					
YES First	-0.01	-0.02	-0.02	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
NO First mean	0.25	0.19	0.10	0.08	0.17
Number of individuals	617	617	617	617	617
Observations	4,268	2,446	3,676	1,832	12,222

Note: Includes only those who are in the ACASI group (FTFI group excluded). Observations at respondent-question level. Regressions include question-level fixed effects and individual controls (including all variables in Table 2.1). Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.D3: Effect of Placebo Module Position on SI Effects for Placebo Questions

	(1) Farm work (past year)	(2) Market visit (past week)	(3) Maize/Rice (next week)	(4) Meat (next week)
Panel A. Malawi				
SI × Placebos First (β)	0.01	0.08**	0.11***	0.11***
	(0.02)	(0.03)	(0.04)	(0.04)
SI × IPV First (γ)	-0.04**	0.08***	0.05	0.12***
	(0.02)	(0.03)	(0.04)	(0.03)
Placebos First	0.00	0.00	0.01	0.05
	(0.02)	(0.03)	(0.04)	(0.03)
FTFI × IPV First mean	0.94	0.46	0.54	0.26
p-value ($\beta = \gamma$)	0.036	0.992	0.326	0.895
Observations	1,718	1,713	1,345	1,228
Panel B. Liberia				
SI × Placebos First (β)	0.03	0.10***	-0.03*	0.01
	(0.03)	(0.04)	(0.02)	(0.04)
SI × IPV First (γ)	-0.01	0.07**	-0.04**	-0.03
	(0.03)	(0.04)	(0.02)	(0.04)
Placebos First	-0.01	-0.04	-0.01	-0.04
	(0.03)	(0.04)	(0.01)	(0.04)
FTFI × IPV First mean	0.81	0.69	0.98	0.70
p-value ($\beta = \gamma$)	0.356	0.633	0.610	0.446
Observations	1,259	1,260	1,226	1,101

Note: Regressions include individual controls (including all variables in Table 2.1) and the order between IPV module and PHQ-9 module. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 2.D4: Effect of Placebo Module Position on SI Effects for IPV Questions

	(1)	(2)	(3)	(4)	(5)
	=1 if responded yes to *individual* question in the following category:				All questions pooled
	Controlling Behavior	Emotional IPV	Physical IPV	Sexual IPV	
Panel A. Malawi					
SI × Placebos First (β)	0.10***	0.05***	0.02**	0.05***	0.06***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
SI × IPV First (γ)	0.08***	0.05***	0.00	0.01	0.04***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Placebos First	0.01	0.02	-0.01	-0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
FTFI × IPV First mean	0.11	0.06	0.04	0.04	0.07
p-value ($\beta = \gamma$)	0.275	0.916	0.276	0.030	0.217
Observations	11,887	6,802	10,181	5,095	33,965
Panel B. Liberia					
SI × Placebos First (β)	0.05***	-0.00	0.00	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
SI × IPV First (γ)	0.05**	0.02	0.01	0.04**	0.03**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Placebos First	0.00	0.00	0.01	0.00	0.01
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
FTFI × IPV First mean	0.20	0.19	0.09	0.04	0.14
p-value ($\beta = \gamma$)	0.965	0.415	0.587	0.415	0.614
Observations	8,752	5,006	7,508	3,758	25,024

Note: Observations at respondent-question level. Regressions include question-level fixed effects, individual controls (including all variables in Table 2.1), and the order between IPV module and PHQ-9 module. Standard errors clustered at individual level in parentheses. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Appendix 2.E: Survey instrument

Non-sensitive placebo questions[28]

1. Did you do any farm work in the past year?

2. Did you go to the market in the past week?

3. Will you, or anyone in your household, eat any [rice/maize] next week, one time or more?

4. Will you, or anyone in your household, eat any type of meat next week, one time or more?

Controlling behavior

1. Did your man ever try to keep you from seeing your friends in the past 12 months?

2. Did your man ever try to stop you from meeting or speaking to your family of birth in the past 12 months?

3. Did your man ever need to know where you are all the time in the past 12 months?

4. Did your man ever stop talking to you or treat you with no interest in the past 12 months?

5. Did your man ever get angry if you speak with another man in the past 12 months?

6. Did your man often think that you are unfaithful in the past 12 months?

7. In the past 12 months, did your man ever expect you to ask for his approval before you go to a health clinic or hospital?

[28] Some questions asked in the ACASI Screening module were reasked later either by SI or by FTFI. For placebo effects analysis, we exclude those questions and include only the four questions listed here, which were not previously asked in the screening module.

Emotional IPV[29]

1. Did your man ever insult you or make you feel bad about yourself in the past 12 months?
2. Did your man ever make you feel small in front of other people in the past 12 months?
3. Did your man ever mean to scare you (for example, by the way he looked at you, by yelling and bursting things) in the past 12 months?
4. Did your man ever threaten to hurt you or someone you care about in the past 12 months?

Physical IPV[29]

1. Did your man ever slap you or throw something at you that could hurt you in the past 12 months?
2. Did your man ever push you, shove you, or pull your hair in the past 12 months?
3. Did your man ever hit you with his hand or with something else that could hurt you in the past 12 months?
4. Did your man ever kick you, drag you or beat you up in the past 12 months?
5. Did your man ever mean to choke or burn you in the past 12 months?
6. Did your man ever threaten to use or actually use a gun, knife or other weapon against you in the past 12 months?

Sexual IPV[29]

1. Did your man ever physically force you to do man and woman business when you did not want to in the past 12 months?

[29] For each IPV question, if the answer is "yes", a follow-up question about frequency appears, asking whether it happened (i) one or two times, (ii) three to five times, or (iii) more than five times.

2. Did you ever do man and woman business when you did not want to because you were afraid of what your man might do in the past 12 months?

3. In the past 12 months, while doing man and woman business, did your man ever force you to do something that made you feel small or bad about yourself?

Chapter 3

Exhaustive or Exhausting? Evidence on Respondent Fatigue in Long Surveys

3.1 Introduction

Many of the surveys that are administered in development economics or by multilateral agencies such as the World Bank to measure poverty or as part of evaluations are long and complicated, and require the sustained attention of a respondent for several hours. For any researcher who has observed such a survey, it is clear that some respondents disengage as the survey drags on, because they are exhausted, bored, or because their attention wanders. As a result, response quality during the later part of a long survey may suffer, a phenomenon known as survey fatigue.

While survey fatigue is well-documented in the literature,[1] until recently there has been comparatively little research to rigorously quantify its effects. In this paper, we provide such a quantification by randomizing the order in which modules appear in a long survey, generating exogenous variation in the time-into-survey when a particular question was asked. This random order of questions allows us to compare responses to the *same* question when it is asked sooner in the survey versus when it is asked later, and quantify the divergence in responses. We conduct this experiment within surveys administered at baseline and endline for a randomized evaluation of cash transfers in rural Liberia and Malawi (Aggarwal et al. 2020). These surveys were long, averaging about 2.5 hours, and the experimental randomization induced meaningful variation in the time it took to reach a specific question: the average time to reach a specific question was changed by as much as about 30 minutes as a result of the randomization.

We have two main findings. First, and consistent with other work, we find clear evidence of survey fatigue. We estimate survey fatigue separately for two ways of asking questions. The first is an "open-ended" method which we used for those questions in which there is no top code or pre-listed set of options. For example, for transfers given out, respondents were asked to

[1] For example, survey fatigue has its own entry in the Encyclopedia of Survey Research Methods.

provide the number of transfers that they gave, and could list as many or as few as they wanted. For such questions, we find that each additional hour of surveying causes a 26-64% decrease in the number of items listed. The second method, or "fixed list" method, is one in which the list of items was pre-coded. For example, in the food expenditures section, we generated a list of around 35 food items and asked about each of these separately. Survey fatigue might be reduced with this method, if the listing serves as a memory aid for those who need help with recall later in the survey as they begin to tire out. In addition, for some categories (such as food expenditure), there are minimal follow-up questions so that listing a value of zero would not reduce survey length substantially. However, we still observe survey fatigue in this method, though much less than in the prior method: for every additional hour, respondents are about 10-19% more likely to report no value for a given question. While survey fatigue appears less prevalent when using the fixed list method, we are unable to definitively attribute this to the question type, since the method is not random – it is also possible that these categories are less subject to survey fatigue.

Second, we quantify the extent to which this skipping reduces the *value* of aggregate categories such as the total value of transfers or expenditures. For any skipped question, the value of that category would be set to zero by default, and so we would expect survey fatigue to lower aggregated values. This effect might be modest if the categories that are skipped tend to be more marginal. However, the effects we find are sizeable: for example, an additional hour of survey time reduces the value of food expenditures by 25%, and has even larger effects (in percentage terms) on smaller categories (such as transfers).

This paper contributes to a recent literature that experimentally evaluates the effect of survey time on survey fatigue. Laajaj and Macours (2021) randomize the order of cognitive, non-

cognitive and technical questions in a sample of farmers in Western Kenya but, unlike us, find no effect of survey time on reporting. Two other papers were conducted contemporaneously to this study, and find similar results to ours. Ambler et al. (2021) randomize the order of a household labor supply module, where questions are asked about the labor supply of each household member, but the order in which the household members are listed was randomized. The authors find a 2% reduction in the number of activities reported when a household member is moved back by one position in the household roster. Abay et al. (2021) employ a methodology similar to ours, in which the authors randomize the placement of a dietary diversity module within a phone survey in Ethiopia. Like us, they find large effects: a 15 minute increase in survey time before the module leads to an 8-17% decline in reported dietary diversity.[2] Finally, in a similar but different design in a different context, Backor et al. (2007) conduct a web-based time-use survey in the US in which an extra question is included at a random order, creating variation in how many hours had already been asked about when a particular question appeared in the survey. Similar to these other papers, the authors find that asking about an additional hour lowers the number of activities reported in each subsequent hour by 5 percentage points.

Our experiment also furthers the literature by helping us rule out some of the explanations for *why* survey fatigue occurs. Past research suggests that survey fatigue may be driven by people deliberately choosing to not answer questions in order to expedite the end of the survey, or if people become more likely to inadvertently make mistakes as they become tired. Some researchers have also conjectured that, over time, respondents learn that answering "no" to a question often invokes a skip code that will allow them to skip a number of follow-up ques-

[2]Another related paper is Kilic and Sohnesen (2019), who find that poverty incidence differs when measured in a short or a long survey in Malawi. However, in their case, since everybody got the same long survey or the same short survey, it is not possible to disentangle the effects of survey length from those of question order, i.e., when your responses are impacted by a question being preceded by another question (see here).

tions. This behavior, known as "satisficing," has been documented in survey settings (Krosnick 1991). To examine this, we also randomized the order of modules within phone surveys that we conducted with respondents repeatedly every 2 months. These surveys took about 30-40 minutes to complete. Importantly, the randomization began more than a year after the phone surveys had started. Therefore, at the time of the phone survey experiment, we would expect that respondents were already familiar with the structure of the surveys, including the mechanics of skip patterns over time as they go through multiple rounds of the survey. If satisficing were present, fewer questions would be answered right from the outset during the later rounds of the phone surveys, and there would be no evidence of experimental survey fatigue *within* a survey round. However, our evidence is not consistent with satisficing: we find evidence of survey fatigue similar to our baseline and endline surveys in the 2-3 rounds of the phone survey. Our results suggest that fatigue is very likely driven by an increase in cognitive burden as the survey progresses.

The rest of this paper proceeds as follows. Section 3.2 explains the data and experimental design, Section 3.3 presents results, and Section 3.4 concludes.

3.2 Data and Experimental Design

3.2.1 Setting

We use data from baseline and endline surveys conducted as part of a cash transfer RCT with the NGO GiveDirectly in Liberia and Malawi. In the experiment, the treatment group received cash transfers via mobile money. The average amount of the transfer was $500; however, the amount and other implementation details were varied experimentally – see our trial registry on the AEA

website (Aggarwal et al. 2020) for more details on the design of the underlying experiment.[3]

In each country, the project took place in rural areas, with universal targeting in treatment villages (i.e. all households in treatment villages received transfers). For this reason, the total allocation to a village depends on its size; to ensure liquidity, the NGO decided to only include villages which were small. Operationally, we set a population threshold based on the most recent population census.[4] In Liberia, the study takes place in Bong and Nimba Counties; in Malawi, it takes place in Chiradzulu and Machinga Districts. In each country, the project enrolled 300 villages, with half selected for treatment.

In each village, we attempted to enroll 10 households into the survey sample.[5] We chose to target women for the study, though many questions were asked at the household level. Male heads were interviewed only when the female was not present, and would not be reachable within a few days; our sample was ultimately 76% female in Liberia and 94% in Malawi.

Two of the 10 sampled households in each village were further randomly sampled to participate in a monthly panel survey that was conducted over the phone and was designed to measure a pre-defined set of outcomes at a high frequency. While the major focus of these surveys was to measure food security, they also included questions on income, labor supply, transfers, savings, and credit. We designed these surveys such that each household was called every other month, but the 2 households in each village alternated months, such that each village provided a data point every month. The phone surveys took about 30-40 minutes to complete.

[3] In both countries, the size of the transfer was varied between $250, $500, and $750. In addition, in Liberia, cash was disbursed either as a "lump-sum" or via quarterly payments. However, even the lump sum was disbursed in increments of $250 per month, so that cash was paid out over 3 months for the largest transfer.

[4] In Malawi, the upper threshold was 100 household per village according to the 2008 national census. In Liberia, we conducted the experiment in two cohorts; the first cohort included villages that had up to 25 households in the 2008 national census, and the threshold for the second cohort was 125, reflecting the larger village sizes in the study region.

[5] It was not always possible to enroll 10 households per village. The total sample size is 2,715 in Liberia and 2,944 in Malawi

Figure 3.A1 shows the timeline of project activities.

3.2.2 Question order randomization

This experiment takes place within baseline and endline surveys which are similar to World Bank LSMS surveys and take about 2.5 hours to complete on average. The surveys contain 19 self-contained sections, including household demographics, agriculture, income, expenditures, savings, assets, labor supply, shocks, and other topics.[6] The beginning of the survey (which included household identifying information, demographics, and agriculture) and the end of the survey (which had a section on intimate partner violence, followed by the collection of household tracking information) were the same across all versions. The remaining sections were grouped into 3 modules, and the order of these 3 modules was randomized, giving us 6 versions of the survey (which we refer to as versions A-F – see Figure 3.A3). The survey software records the amount of time elapsed (since beginning) at each question, allowing us to calculate the specific time at which a question appeared in the survey.

The amount of time it takes to progress through the survey varies depending on a number of factors, including respondent and enumerator characteristics, and the details of a household's circumstance. For example, because our survey had a focus on agriculture, a household which grew multiple crops would be asked a number of questions about each one of them. Table 3.A1 shows information on the average survey duration. The baseline and endline surveys took on average 2.3 and 2.7 hours respectively in Liberia; and 3 and 2.8 hours respectively in Malawi. The standard deviation in survey time is sizeable, ranging from 0.7 to 1.1 hours. Figure 3.A4 shows a CDF of the time until completion of different points of the survey (using survey Version A only) for both countries and for both baseline and endline pooled together (i.e., for 4

[6] See Figure 3.A2 for the list of sections.

country-survey combinations). The figure shows CDFs for various quantiles in the survey time distribution (i.e. relative to completing the question which makes up the p-th percentile of the overall distribution of time to survey completion). The CDFs show that even 10% into the survey, the standard deviation of time is already over 30 minutes and that for all percentiles, there are surveys that take a large amount of time. For example, about 10% of people take over 3 hours to even get halfway through the survey (Panel C).

Finally, although not the main focus of this paper, we also randomized survey order for the final 2-3 rounds of the phone survey. In order to do this, we randomized the location of the Expenditures and Transfers sections to appear at either the very beginning or the very end of the survey, and the order between the two sections, generating 4 possible permutations of section order within the survey (Figure 3.A5). We return to this randomization in the discussion section, when we discuss possible explanations for survey fatigue.

Table 3.1 shows the effect of the randomized survey versions on the time until which the first question of each section was administered. The reported means and standard deviations at the bottom of the table are those pertaining to that section for Version A of the survey. As can be seen from this table, the module randomization introduced significant variation in the time-into-survey when a section starts. For example, looking at Column 1, we can see that the Assets section started just after the 80th minute on average for those who got Version A of the survey. However, the full range for when this section started ranges from 77th minute (version B) to 107th minute (Version F) - a difference of 30 minutes. This range of about 30 minutes is consistently observed across all sections.

We use the survey version that was used for each respondent as an instrument for the time-into-survey when a particular set of questions began to be asked of that respondent. The

first-stage F-statistics are shown at the bottom of Table 3.1, and range from 35 to almost 200.

Table 3.1: Experimental variation in time before which sections were administered

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Time into survey (minutes) at the beginning of following section:						
	Module 1				Module 2	Module 3	
	Assets	Savings	Credit	Transfers	Expenditure	Shocks	Contributions
Version B	-3.30***	-2.84**	-2.79**	-2.63*	6.39***	-12.20***	-8.83***
	(1.28)	(1.36)	(1.36)	(1.37)	(1.37)	(1.47)	(1.50)
Version C	19.21***	17.71***	17.54***	17.67***	-16.83***	-4.74***	-3.24**
	(1.27)	(1.36)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version D	23.96***	22.52***	22.24***	22.35***	-18.00***	-16.52***	-5.45***
	(1.27)	(1.36)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version E	6.61***	6.01***	5.67***	6.04***	5.79***	-25.74***	-15.79***
	(1.27)	(1.35)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version F	26.06***	24.56***	24.01***	24.07***	-8.38***	-27.57***	-14.49***
	(1.28)	(1.36)	(1.36)	(1.38)	(1.37)	(1.47)	(1.50)
Version A: Mean	80.01	93.47	93.89	95.61	109.39	125.53	134.72
Version A: SD	38.78	40.34	40.26	40.87	44.23	47.39	49.78
F-statistic: joint significance	197.30	151.42	146.55	143.33	127.92	114.25	35.05
Number of respondents	5,591	5,597	5,597	5,597	5,597	5,597	5,592
Observations	10,153	10,226	9,952	10,228	10,227	10,224	10,154

Note: The omitted category is version A. Observations include in-person baseline and endline survey data. Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

3.2.3 Respondent characteristics and randomization check

Table 3.2 presents summary statistics for several basic demographic indicators, as well as comparisons across treatment groups. We present these statistics only for those indicators which were asked before the module randomization kicked in as the variables from the later sections would by definition be imbalanced under our central hypothesis for this paper. We show the balance across versions separately for the baseline and endline surveys, but pool them across

the 2 countries. For each survey (baseline or endline), we show the mean and standard deviation (for non-binary variables) pertaining to Version A of the survey (chosen arbitrarily), followed by the p-value for the joint test of equality across all 6 versions of the survey. Panel A shows respondent characteristics. Almost 90% of the sample is female, three-quarters are married, and the average age is 41. Average years of education (for the respondent) is only 4.2, and 57% are literate in English (these last 2 variables were measured at baseline only).

Panel B shows household statistics. At baseline, the average household has 4.8 members, and 96% were engaged in farming. About 40% of the sample live in a house with a thatch roof, and 80% live in a house with a mud floor. About 77% own their dwelling and only 2% have electricity. We cannot reject equality across treatments for all of these variables.

Finally, Panel C shows the other experimental treatments. Cash was randomly given out to 50% of villages (and given that we sampled about 10 households per village, it was given, by design, to roughly 50% of the respondents). The phone surveys were administered to 20% of the respondents. As expected, the survey experiment is orthogonal to both of these treatments.

Table 3.2: Summary Statistics and Randomization Check

	Baseline Survey		Endline Survey	
	Version A (Mean/SD)	*p*-value: test of equality over 6 versions	Version A (Mean/SD)	*p*-value: test of equality over 6 versions
Panel A. Respondent Characteristics				
=1 if female	0.87	0.188	0.89	0.308
=1 if currently married or has partner	0.76	0.970	0.74	0.188
Age	40.50	0.661	40.95	0.388
	(15.20)		(14.31)	
Years of education	4.18	0.553		
	(3.75)			
=1 if can read/write in English	0.57	0.667		
	(0.50)			
Panel B. Household Characteristics				
Number of household members	4.77	0.436	4.98	0.744
	(2.11)		(2.16)	
=1 if household engaged in farming past year	0.96	0.786	0.90	0.803
=1 if thatch roof	0.40	0.206	0.24	0.780
=1 if mud/dirt floor	0.80	0.848	0.77	0.392
=1 if owns dwelling	0.77	0.844	0.77	0.840
=1 if has electricity in dwelling	0.02	0.280	0.02	0.523
Panel C. Cross-randomized groups				
Cash Treatment Group	0.53	0.216	0.51	0.914
Phone survey group	0.21	0.640	0.22	0.655
Observations	4,879		5,349	

Note: Column 1 and 3 (Version A) represent control mean with standard deviation in parentheses. Columns 2 and 4 present *p*-values from the joint test of equality of the means for all the 6 survey versions, A-F.

3.3 Results

3.3.1 Quantifying survey fatigue

We start by examining the impacts of time-into-survey on the count of items or instances reported in response to the open-ended questions (questions described in Figure 3.A6). To do this,

we run the following regression:

$$Y_{icq} = \beta Hours_{icq} + \phi_c + \varepsilon_{icq}, \tag{3.1}$$

where Y_{icq} refers to the count of items corresponding to question q reported by survey respondent i in country-survey sample c, $Hours_{icq}$ denotes elapsed time into survey (in hours) at which question q is asked to respondent i, instrumented with the randomized module order (Versions A-F) that was fielded to the respondent, ϕ_c represents country-survey fixed effects (e.g., Malawi Baseline), and ε_{icq} is the error term.

In this analysis, there is no reason to expect heterogeneity in responses based on outcomes – ex ante, we expect similar results for any question category. Therefore, to discipline our analysis, we present results exhaustively for every relevant outcome, and adjust the standard errors to account for a false discovery rate (FDR) using the procedure in Michael L Anderson (2008a). For each outcome, we present only q-values from this procedure, and statistical significance is ascertained only based on the q-values obtained after FDR-correction.

We present these results in Table 3.3. We show 5 outcomes: the number of Rotating Savings and Credit Associations (ROSCAs) and Village Savings and Loan Associations (VSLAs) that the respondent reported being part of in the savings section; the reported number of transfers received and given during the past month; and the number of credit purchases during the past month.[7] Four out of 5 of these outcomes are statistically significant at 10% (and 2 are significant at 5%), even with the FDR adjustment. The effect sizes are large: an extra hour reduces the number of items by 26-64%. Because these surveys average 2.5 hours, this implies that the

[7]For both transfers and credit purchases, some earlier survey versions included questions recalling for the past 3 months instead. Later for analysis on aggregated values, the monetary values collected from these versions are divided by 3, comparable to the past-month values.

decision to place a question at the beginning rather than the end of the survey can have a large effect.

Table 3.3: Survey time and the probability of missing responses ("Open-ended" questions)

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Hours into Survey	0.002	-0.058**	-0.074*	-0.209***	-0.095*
	[0.468]	[0.033]	[0.065]	[0.001]	[0.078]
Dependent variable: Mean	0.056	0.205	0.275	0.328	0.366
Hours into Survey: Mean	1.9	1.9	1.9	1.9	1.9
Hours into Survey: SD	1.0	1.0	1.0	1.0	1.0
Number of respondents	5,596	5,597	5,596	5,594	5,597
Observations	10,225	10,224	10,223	10,215	10,228

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. See Table 3.B2 for results by country and Table 3.C2 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Next, we investigate the impacts of elapsed survey time on choosing an item in questions asked via the fixed-list method (questions described in Figure 3.A7), and run the following regression:

$$Y_{icqj} = \beta Hours_{icq} + \phi_c + \psi_{qj} + \varepsilon_{icqj}, \tag{3.2}$$

where Y_{icqj} is a binary indicator of whether respondent i in country-survey sample c responded "yes" to having consumed/bought/experienced item j in question q of the survey, $Hours_{icq}$ elapsed time into survey (in hours) at the beginning of question q, instrumented with the ran-

domized module order (Versions A-F), ϕ_c country-survey fixed effects, ψ_{qj} question-item fixed effects, and ε_{icqj} error term. Like before, we adjust the standard errors for multiple testing, and report only the FDR-corrected q-values in our tables.

Table 3.4 presents this analysis for a set of 9 outcomes: livestock, farm tools, durable goods, savings, loans, food expenditures, non-durables expenditures, household shocks, and public goods contributions. Note that these regressions are at the question level, and so are much better powered than the previous set of outcomes: we find that 4 of 9 outcomes are significant at 5% (and even of those not significant, nearly all are negative signed). Effect sizes are more moderately measured than for the "open-ended" questions, ranging from 10-19% for the statistically significant outcomes. Nevertheless, survey fatigue is clearly evident here as well.[8,9]

[8] Please note, however, that in both Table 3.3 and Table 3.4, the effect sizes in percent terms are slightly overestimated due to the fact that the dependent variable means are calculated across all versions and are therefore, depressed due to survey duration effects. Nevertheless, the effects are large enough in an absolute sense to be economically meaningful.

[9] See Appendix 3.B and Appendix 3.C for heterogeneity in these results by country and by survey type (baseline or endline).

Table 3.4: Survey time and the probability of missing responses ("Fixed list" questions)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Hours into Survey	-0.007**	-0.004	-0.001	-0.002	0.000	-0.025***	-0.038***	-0.025***	-0.002
	[0.037]	[0.144]	[0.406]	[0.365]	[0.468]	[0.001]	[0.001]	[0.001]	[0.378]
Dependent variable: Mean	0.072	0.154	0.176	0.060	0.020	0.203	0.249	0.130	0.050
Hours into Survey: Mean	1.7	1.7	1.8	1.9	1.9	1.9	1.9	2.0	2.0
Hours into Survey: SD	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.2
Number of respondents	5,594	5,594	5,594	5,597	5,597	5,597	5,597	5,597	5,349
Observations	134,831	208,281	212,373	114,045	138,711	366,947	112,497	166,524	48,141

Note: Observations at respondent-question-item level. Each regression is an IV regression, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. See Table 3.B1 for results by country and Table 3.C1 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

3.3.2 Effect of survey fatigue on aggregated values

The prior section implies that aggregated values of categories such as expenditures or transfers will be attenuated by survey fatigue; in this section, we quantify this attenuation. We run regressions identical to Equation (3.1), except that the dependent variable is now in dollar amounts, rather than counts; in addition, results are shown for both open-ended and fixed list questions. Results are shown in Table 3.5. We find that the vast majority (9 of 11) of point estimates are negative, more than half of which (5) are significant at conventional levels despite being corrected for multiple hypothesis testing. In addition, 2 of the coefficients - those for farm tools and public goods - are marginally significant at 15% and 13% respectively. In addition to being statistically significant, the effect sizes are economically meaningful. Focusing on just the statistically significant effects, the coefficient magnitudes range from 25% of the mean (for food

expenditure) to 86% (for transfers given).

One surprising result is that our effect sizes for reported monetary values (as shown in Table 3.5) are in some cases, much larger in percent terms than they are for the counts that were collected via the open-ended and the fixed-list questions in Table 3.3 and Table 3.4, respectively. This is especially true for some of the small categories such as transfers given, where an extra hour reduces the value by $0.59, on a base of just $0.69, or 86%, while the effect of an hour on the count in Table 3.3 is a reduction of -0.21 transfers on a base of 0.33 (or 64%). But even for a larger category like food, the percent decline in value is 25%, compared to 12% in skipping in Table 3.4. This is at odds with others in the literature, such as Ambler et al. (2021) and Abay et al. (2021), who find that respondents are likely to forget the more marginal categories as they progress through the survey. While we can only conjecture as to what may cause this, our results are consistent with recent work such as Brzozowski et al. (2017), who show that recall errors in surveys tend to not be mean zero, but are in fact, negatively correlated with true behavior - i.e., when respondents make mistakes, they tend to overstate the low values and understate the high values.

Table 3.5: Survey Fatigue and Reported Total Monetary Value of Aggregated Categories

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Hours into Survey	-13.47	-1.32	4.68	-1.47	0.73	-4.12***	-2.52***	-0.10	-0.51**	-0.59***	-0.65***
	[0.334]	[0.148]	[0.365]	[0.365]	[0.345]	[0.001]	[0.001]	[0.121]	[0.015]	[0.001]	[0.002]
Dependent variable: Mean	95.78	10.48	58.11	15.52	6.40	16.22	7.93	0.14	0.95	0.69	0.81
Hours into Survey: Mean	1.7	1.6	1.8	1.9	1.9	1.9	2.0	2.1	1.9	1.9	1.9
Hours into Survey: SD	1.0	1.3	1.0	1.0	1.0	1.3	1.3	1.6	1.0	1.0	1.0
Number of respondents	5,594	5,349	5,594	5,597	5,597	5,597	5,597	5,349	5,597	5,597	5,597
Observations	10,189	5,349	10,189	10,226	9,952	10,227	10,227	5,349	10,228	10,228	10,228

Note: All values in USD. Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. For transfers and credit purchases, some earlier survey versions included questions recalling for the past 3 months instead of past month. The monetary values collected from these versions are divided by 3, making them comparable to the past-month values. See Table 3.B3 for results by country and Table 3.C3 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

3.3.3 Are effects driven by satisficing?

In this subsection, we investigate whether the practice known as "satisficing" is likely an explanation behind the observed pattern of results. Satisficing is a term used to describe the phenomenon where respondents may be answering questions in such a way that helps them avoid follow-ups, and therefore, reduce survey length. In this case, satisficing would entail responding "no" to questions, or not listing additional values of items such as transfers, in order to avoid follow-up questions on those items. In order to check for this, we utilize our phone surveys. As mentioned in Section 3.2.1, we randomly selected 20% of our sample to participate in phone surveys, which began shortly after the baseline survey. Respondents were called once every 2 months for about 16-26 months (or 8-13 rounds).

After deciding to implement the survey order randomization into the longer in-person sur-

veys, we later decided to also randomize the order in the phone surveys. Importantly, the randomization began around the 8th round of the survey in Liberia and the 11th in Malawi, so respondents already had lots of experience with the questionnaire.[10] If satisficing is an explanation, we would therefore expect survey fatigue to be minimal in this experiment (since people would be equally able to skip questions wherever they appeared in the survey).[11] The randomization was very similar to the longer surveys, though less involved: specifically, as shown in Figure 3.A5, we varied the location of the expenditure and transfers sections within the survey.

Table 3.6: Impacts of survey time on open-ended and fixed list questions, phone surveys

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Number of distinct items reported for the following:					=1 if item is selected (not skipped):			
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases	Savings	Loans	Food expend	Non-durables
Hours into Survey	0.050	0.308	0.091	-0.246	-0.346*	0.048	-0.014	-0.103***	-0.069*
	[0.315]	[0.108]	[0.308]	[0.105]	[0.091]	[0.108]	[0.185]	[0.001]	[0.091]
Dependent variable: Mean	0.088	0.372	0.205	0.190	0.283	0.140	0.031	0.216	0.351
Hours into Survey: Mean	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Hours into Survey: SD	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Number of respondents	780	780	780	780	780	780	780	780	780
Observations	1,762	1,762	1,762	1,762	1,762	18,678	24,654	63,059	20,083

Note: For columns 1-4, observations at respondent-question-item level, and regressions include country-sample fixed effects and question-item level fixed effects. For columns 5-9, observations at respondent level, and regressions include country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

In Table 3.6, we analyze the effect of hours-into-survey on responses in the phone survey.

[10] See Figure 3.A1 for the specific survey rounds when order randomization was implemented.

[11] Another implication of survey fatigue is that the total survey time, and thus the value of categories, should decline over time as respondents learn the skip codes. However, we have no way of testing this since the number of rounds is colinear with time trends.

Columns 1-5 analyze responses to open-ended questions, and Columns 6-9 show outcomes for questions that follow the fixed list pattern. To study these, we run the same regressions as in Equation (3.1) and Equation (3.2) respectively, except that the outcomes are now drawn from the phone survey. Similarly, in Table 3.7, we show the impacts on the value of aggregated categories, a replication of the analysis that we show in Table 3.5.

Contrary to the predictions of a satisficing hypothesis, we find evidence of negative effects of survey duration on both, the counts as well as the value of objects/outcomes reported by the respondents. Taken together, these results suggest that the operative channel for survey fatigue effects is the cognitive burden imposed by long surveys, and not deliberate gaming by the respondents.

Table 3.7: Effect of survey time of total value of aggregated categories, phone surveys

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Total value of reported items for the following:						
	Savings	Loans	Food Expend	Non-durables	Transfers received	Transfers given	Credit purchases
Hours into Survey	8.56	-2.45	-9.63**	-2.62	2.68*	-0.73	-2.38*
	[0.185]	[0.333]	[0.039]	[0.185]	[0.091]	[0.153]	[0.091]
Dependent variable: Mean	10.19	8.63	13.37	7.49	1.55	0.59	1.28
Hours into Survey: Mean	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Hours into Survey: SD	0.1	0.1	0.1	0.1	0.1	0.2	0.2
Number of respondents	780	780	780	780	780	780	780
Observations	1,762	1,762	1,762	1,762	1,762	1,762	1,762

Note: Observations at respondent level, and regressions include country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

3.4 Conclusion

In this paper, we randomize the order of questions asked as part of the baseline and endline surveys of a cash transfer experiment to provide evidence on the impact of survey duration on the quality of responses elicited during the survey. Our results point to strong fatigue effects, on the order of a 10-64% reduction in the count of reported items, which leads to even bigger effects on the reported monetary values of categories that aggregate over these items.

An important implication of these results is that the effects of any program might be attenuated if effects are measured later in the survey. For example, if the effect of survey fatigue is to proportionally reduce the number of items mentioned, then treatment-control differences will become smaller (in absolute value, though not in percentages) if measured later in the survey. This effect could be magnified if there exist non-linearities, for example if there is some threshold level of cognitive load that the treatment group is more likely to encounter because they have more to report. In our case, we can examine if the effect of the cash transfer differs when outcomes are measured later in the survey. However, we find no compelling evidence of this effect in this data, perhaps because power is limited because this analysis can only be conducted on the endline and because the cash treatment requires clustering at the village level (results are shown in tables Table 3.A3 and Table 3.A4, in which we regress outcomes on cash, time into the survey, and its interaction).[12] We leave a further evaluation of this to future work.

Is there a way for these findings to inform survey design? Survey fatigue is not a recent discovery, and practitioners suggest a variety of remedies to address this concern, most of which boil down to fielding shorter surveys, or splitting surveys into multiple shorter versions. For

[12] In the regressions, we demean the hours variable, but find the interaction term is negative only for 5 of 9 outcomes, and none are significant (Table 3.A4).

example, Aggarwal et al. (2021) is an example of a multi-day baseline survey. Other strategies involve sacrificing detail in order to avoid survey fatigue, for example by splitting the survey into shorter versions, administering only one of the versions to each respondent, and imputing responses to the unasked questions (Herzog and Bachman 1981; Raghunathan and Grizzle 1995). Another strategy is to replace ordinal questions with binary ones (Dolnicar et al. 2011). However, each of these remedies comes with its own set of problems, either in terms of detail and measurement error, or in cost.

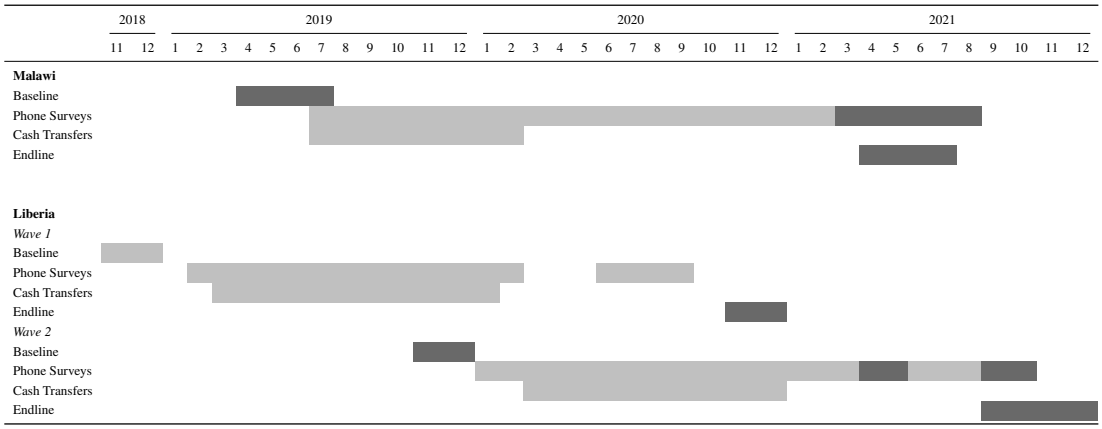
While we have no easy fixes to recommend, an obvious remedial step would be to place the most important questions (for example, those about the primary outcomes in an RCT), as early as possible in the survey. Relatedly, it may also be a good survey practice for enumerators to suggest taking a short break before they start asking important questions that are placed later in the survey. This may be an important consideration especially for interventions in which the primary outcome is sensitive (for example, intimate partner violence, which was placed at the end of these surveys for exactly this reason).[13] Researchers often choose to place such sensitive questions later in the survey to allow respondents some time to become familiar with the enumerator and with the survey, but this paper suggests that this consideration should be balanced against the risk of survey fatigue.

A final implication from this paper is that, for those working with secondary data collected via long surveys, such as the LSMS or the DHS surveys, it may be useful to recognize that cross-country comparisons or even within country comparisons across survey waves may be complicated because of varying survey duration. It may be important to design panel surveys such that outcomes are measured at similar points in the survey over waves.

[13] See Park et al. (2021) and Park and Kumar (2022) for related work on the pitfalls of measuring IPV in this and a related sample in Liberia.

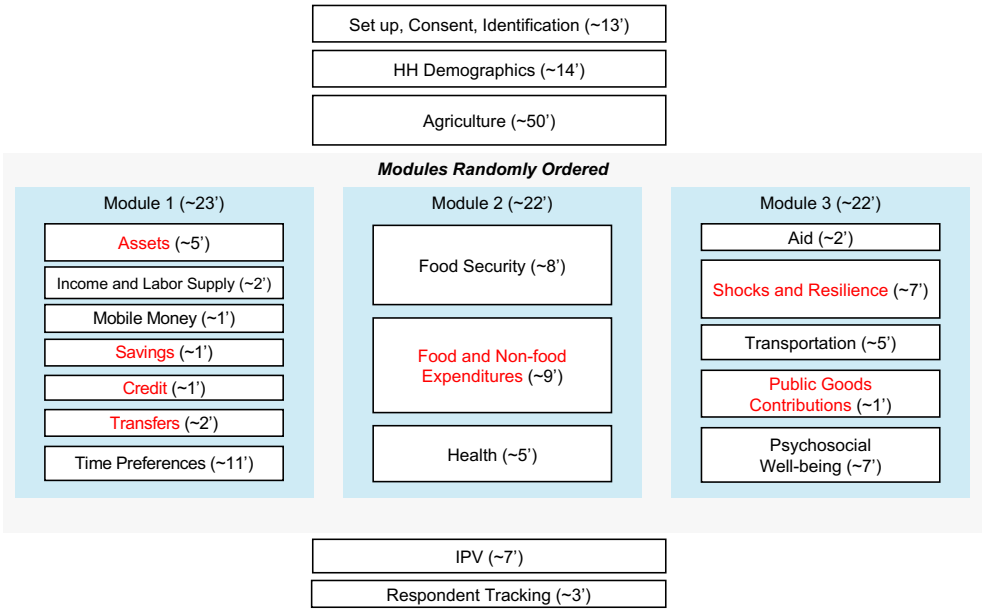
Appendix 3.A: Main Appendix Figures and Tables

Figure 3.A1: Timeline of Survey Activities



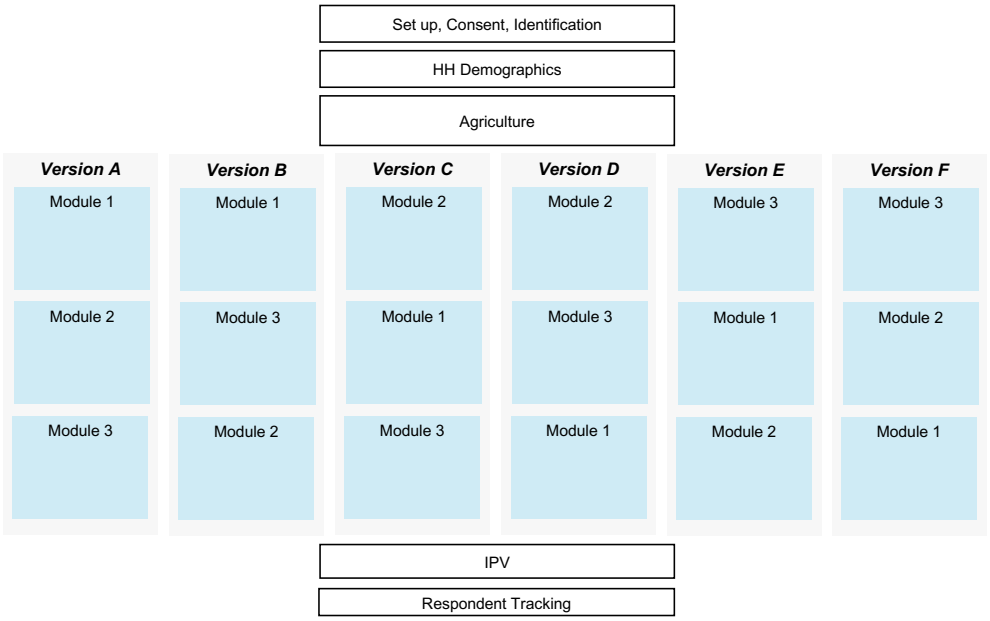
Note: Darker grey blocks indicate the survey rounds where module order randomization was conducted and thus data for which are included for analysis in this paper.

Figure 3.A2: Sections in In-person Surveys



Note: Approximate duration for each section (in minutes) are reported in parentheses. In red are the sections for which survey questions are relevant for analysis in this paper.

Figure 3.A3: Randomized Order of Modules in In-person Surveys

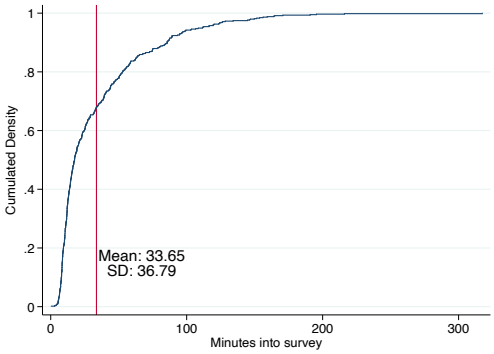


Note: A respondent is randomly provided with one among Versions A-F. The order in which the sections not included in the modules is not randomized. For every version among A-F, survey set-up, demographics, and agriculture come at the beginning, while IPV and respondent tracking are at the end.

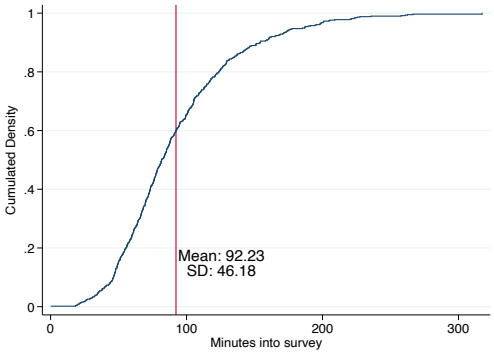
Figure 3.A4: Distribution of Survey Time

Distribution of time to reach the question where on average the survey is:

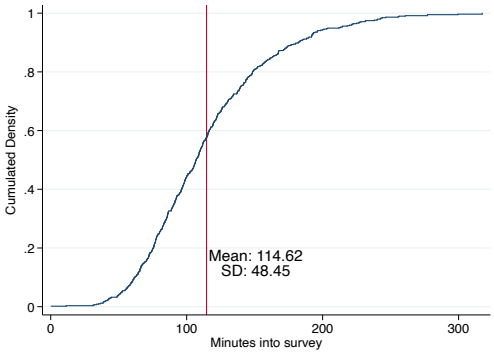
(a) 10% completed



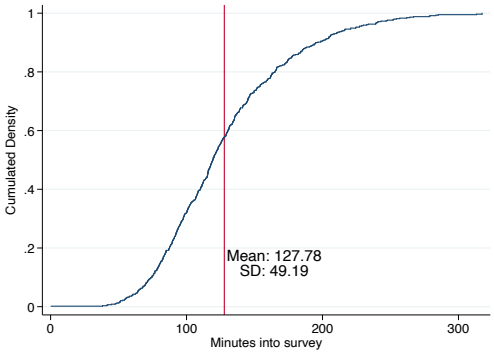
(b) 25% completed



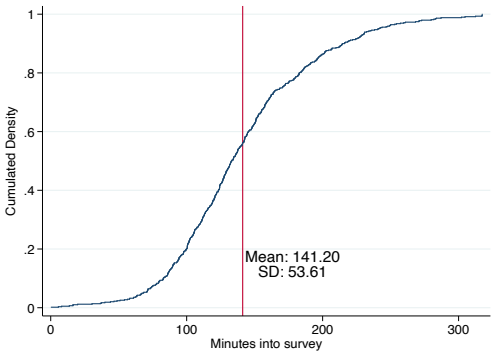
(c) Median



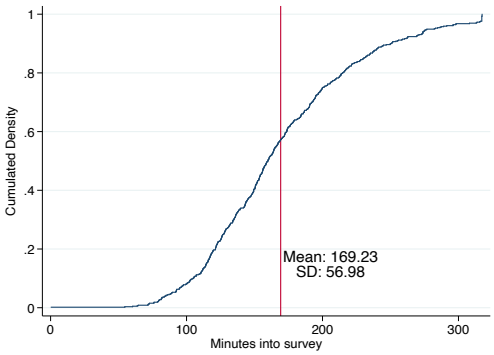
(d) 75th percentile



(e) 90th percentile

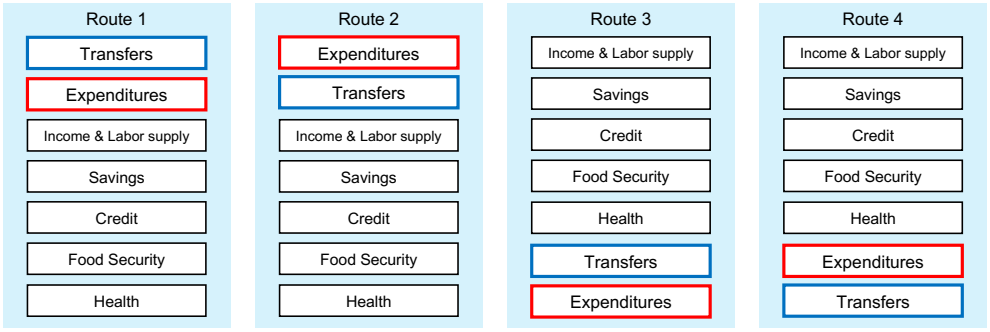


(f) Total survey length



Note: Based on Version A only.

Figure 3.A5: Randomized Order of Modules in Phone Surveys



Note: A respondent is randomly provided with one among Routes 1-4.

Figure 3.A6: Example of "Open-Ended" Question Order

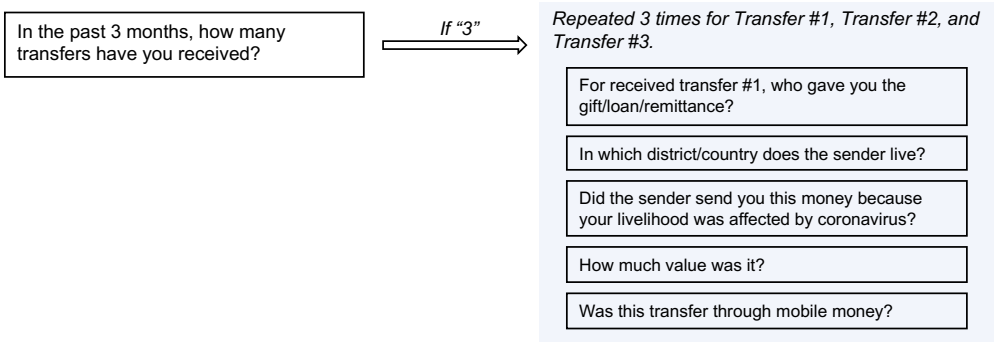


Figure 3.A7: Example of "Fixed List" Question

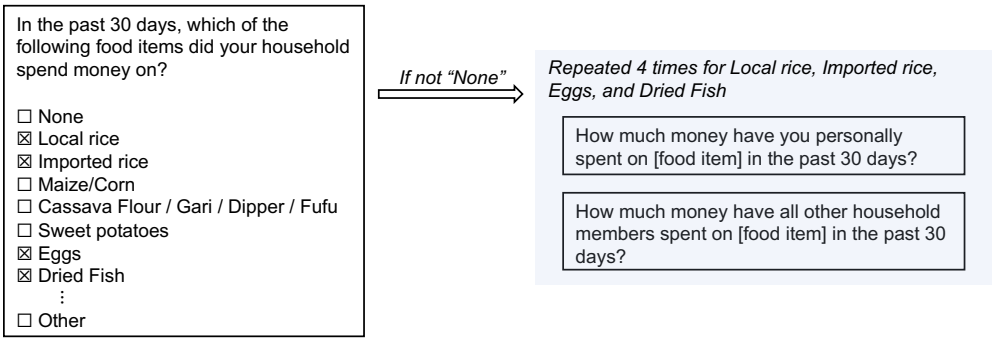


Table 3.A1: Average Duration by Survey Versions (in hours)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Survey Version						Overall
	A	B	C	D	E	F	
Panel A: Liberia							
Baseline	2.28	2.27	2.24	2.31	2.29	2.24	2.27
	(0.69)	(0.65)	(0.69)	(0.75)	(0.67)	(0.70)	(0.69)
Endline	2.73	2.64	2.74	2.68	2.72	2.77	2.71
	(1.04)	(1.05)	(1.12)	(1.02)	(1.09)	(1.16)	(1.08)
Panel B: Malawi							
Baseline	3.15	3.03	3.06	3.03	3.01	3.04	3.05
	(1.02)	(0.89)	(0.93)	(0.92)	(0.91)	(0.90)	(0.93)
Endline	2.75	2.81	2.80	2.76	2.75	2.78	2.77
	(0.80)	(0.82)	(0.81)	(0.79)	(0.82)	(0.82)	(0.81)

Note: Standard deviations in parentheses.

Table 3.A2: Experimental variation in time before sections were administered (phone surveys)

	(1)	(2)	(3)	(4)
	Time into survey (minutes) at the beginning of following section:			
	Savings	Credit	Transfers	Expenditure
Version B	-0.17	-0.08	8.66***	-1.45***
	(0.32)	(0.34)	(0.34)	(0.29)
Version C	-9.14***	-9.03***	10.48***	9.95***
	(0.31)	(0.33)	(0.34)	(0.28)
Version D	-9.66***	-9.59***	17.52***	8.53***
	(0.31)	(0.32)	(0.33)	(0.28)
Version A: Mean	15.47	16.53	3.21	4.81
Version A: SD	6.89	6.98	3.10	3.70
F-statistic: joint significance	585.88	523.70	941.79	837.01
Number of respondents	780	780	779	780
Observations	1,762	1,762	1,759	1,760

Note: Observations include only phone survey data. Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table 3.A3: The effect of survey time on the measurement of the effect of cash

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	=1 if tem is selected (not skipped)									Number of distinct items reported for the following				
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Time into Survey (hr)	0.01	-0.00	-0.01	-0.01	-0.01	-0.02	-0.08	0.00	-0.01	-0.02	-0.13	0.03	-0.15	-0.22
	[0.591]	[0.766]	[0.591]	[0.256]	[0.256]	[0.591]	[0.222]	[1.000]	[0.594]	[0.594]	[0.228]	[0.705]	[0.228]	[0.228]
Cash × Time into Survey (hr)	-0.02	-0.01	-0.02	0.00	0.01	0.01	0.08	-0.02	0.00	-0.02	0.21	-0.29	-0.05	0.16
	[0.304]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[0.276]	[1.000]	[1.000]	[1.000]	[0.276]	[0.135]	[1.000]	[1.000]
Cash	0.01***	0.01*	0.02***	0.01***	0.00	0.01*	0.02***	-0.01*	-0.00	0.01*	0.04*	0.02	0.05**	-0.02
	[0.001]	[0.072]	[0.001]	[0.001]	[0.266]	[0.087]	[0.003]	[0.072]	[0.376]	[0.087]	[0.098]	[0.152]	[0.021]	[0.178]
Control Mean	0.06	0.13	0.15	0.05	0.02	0.18	0.20	0.08	0.04	0.03	0.24	0.16	0.15	0.34
Hours into Survey: Mean	1.7	1.7	1.8	1.9	1.9	1.9	2.0	2.2	2.2	0.0	-0.0	-0.0	-0.0	0.0
Hours into Survey: SD	0.9	0.9	0.9	1.0	0.9	0.9	0.9	1.0	1.0	0.9	0.9	0.9	0.9	0.9
Observations	54,714	80,419	82,023	44,761	51,489	141,028	43,582	63,392	35,658	3,961	3,962	3,962	3,958	3,962

Note: Regressions include baseline measurement of outcome, fixed effects for cash treatment randomization strata, and country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened *q*-values (calculated from *p*-values based on standard errors clustered at village level) in brackets.

Table 3.A4: The effect of survey time on the measurement of the effect of cash

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Expenditure		Assets					Transfers	
	Food	Nondurables	Livestock	Farm tools	Durables	Savings	Loans	Given	Received
Time into Survey (hr)	-0.55	-0.39	6.94	-2.47	-7.09	-7.71	-3.98	0.14	-1.94
	[1.000]	[1.000]	[1.000]	[0.999]	[1.000]	[1.000]	[0.999]	[1.000]	[1.000]
Cash × Time into Survey (hr)	-0.21	1.12	-45.35	0.30	36.20	-5.41	7.99	-2.70	-2.15
	[1.000]	[1.000]	[1.000]	[1.000]	[0.962]	[1.000]	[0.286]	[0.286]	[1.000]
Cash	0.19	0.27	26.00**	1.47***	21.02***	4.56***	-0.19	0.26	1.33**
	[0.197]	[0.227]	[0.034]	[0.004]	[0.001]	[0.007]	[0.545]	[0.197]	[0.019]
Control Mean	3.08	6.30	90.00	9.75	56.32	8.68	6.94	1.66	6.85
Control SD	4.90	9.37	367.73	10.66	138.21	55.59	19.14	6.58	14.53
Hours into Survey: Mean	1.9	2.0	1.7	1.7	1.8	1.9	1.9	1.9	1.9
Hours into Survey: SD	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.9	0.9
Observations	3,962	3,962	3,962	3,962	3,962	3,962	3,687	3,962	3,962

Note: Regressions include baseline measurement of outcome, fixed effects for cash treatment randomization strata, and country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values (calculated from p-values based on standard errors clustered at village level) in brackets.

Appendix 3.B: Heterogeneity by Country

Table 3.B1: Heterogeneity by Country in Fixed List Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Panel A. Liberia									
Hours into Survey	-0.012**	0.002	0.001	-0.002	0.001	-0.033***	-0.053***	-0.011	0.006
	[0.026]	[0.654]	[0.713]	[0.571]	[0.522]	[0.001]	[0.001]	[0.256]	[0.522]
Dependent variable: Mean	0.097	0.196	0.171	0.048	0.009	0.189	0.234	0.065	0.075
Hours into Survey: Mean	1.5	1.6	1.6	1.7	1.7	1.7	1.7	1.8	1.8
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.5
Number of respondents	2,653	2,653	2,653	2,653	2,653	2,653	2,653	2,653	2,566
Observations	49,511	94,521	90,020	54,012	62,397	166,537	49,511	72,016	23,094
Panel A. Malawi									
Hours into Survey	-0.002	-0.011**	-0.001	-0.002	-0.001	-0.016***	-0.025**	-0.028***	-0.014*
	[0.535]	[0.015]	[0.551]	[0.551]	[0.551]	[0.001]	[0.015]	[0.001]	[0.099]
Dependent variable: Mean	0.057	0.119	0.179	0.071	0.028	0.214	0.261	0.180	0.028
Hours into Survey: Mean	1.9	1.9	1.9	2.0	2.0	2.1	2.1	2.2	2.2
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9
Number of respondents	2,941	2,941	2,941	2,944	2,944	2,944	2,944	2,944	2,783
Observations	85,320	113,760	122,353	60,033	76,314	200,410	62,986	94,508	25,047

Note: Observations at respondent-question-item level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table 3.B2: Heterogeneity by Country in Open Ended Questions Question

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Panel A. Liberia					
Hours into Survey	-0.001	-0.036	-0.037	-0.157**	-0.180**
	[0.884]	[0.145]	[0.522]	[0.026]	[0.026]
Dependent variable: Mean	0.106	0.063	0.297	0.381	0.349
Hours into Survey: Mean	1.7	1.7	1.8	1.8	1.8
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2
Number of respondents	2,652	2,653	2,652	2,650	2,653
Observations	4,500	4,500	4,498	4,494	4,501
Panel A. Malawi					
Hours into Survey	0.004	-0.077	-0.122*	-0.240***	-0.016
	[0.551]	[0.121]	[0.072]	[0.001]	[0.551]
Dependent variable: Mean	0.017	0.316	0.258	0.285	0.380
Hours into Survey: Mean	2.0	2.0	2.1	2.1	2.1
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8
Number of respondents	2,944	2,944	2,944	2,944	2,944
Observations	5,725	5,724	5,725	5,721	5,727

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table 3.B3: Heterogeneity by country on total monetary values of aggregated categories

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Panel A. Liberia											
Hours into Survey	-22.28	0.14	-3.04	-0.47	1.06	-6.05***	-4.10**	-0.15	-0.59	-0.87**	-1.25***
	[0.557]	[0.867]	[0.713]	[0.867]	[0.522]	[0.007]	[0.013]	[0.247]	[0.145]	[0.019]	[0.007]
Dependent variable: Mean	155.11	11.12	53.59	27.39	4.62	21.42	10.59	0.28	1.50	1.23	1.39
Hours into Survey: Mean	1.5	1.5	1.6	1.7	1.8	1.7	1.8	1.9	1.8	1.8	1.8
Hours into Survey: SD	1.3	1.6	1.3	1.2	1.2	1.8	1.8	2.2	1.2	1.2	1.2
Number of respondents	2,653	2,566	2,653	2,653	2,653	2,653	2,653	2,566	2,653	2,653	2,653
Observations	4,501	2,566	4,501	4,501	4,501	4,501	4,501	2,566	4,501	4,501	4,501
Panel A. Malawi											
Hours into Survey	-4.48	-3.86*	15.58	-1.51	0.22	-2.41**	-1.08	-0.01	-0.39**	-0.26***	-0.02
	[0.551]	[0.094]	[0.147]	[0.202]	[0.551]	[0.046]	[0.111]	[0.551]	[0.024]	[0.003]	[0.551]
Dependent variable: Mean	48.83	9.89	61.68	6.18	7.87	12.13	5.83	0.02	0.52	0.26	0.36
Hours into Survey: Mean	1.9	1.8	1.9	2.0	2.0	2.1	2.1	2.3	2.1	2.1	2.1
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8	0.8
Number of respondents	2,941	2,783	2,941	2,944	2,944	2,944	2,944	2,783	2,944	2,944	2,944
Observations	5,688	2,783	5,688	5,725	5,451	5,726	5,726	2,783	5,727	5,727	5,727

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Appendix 3.C: Heterogeneity by Survey type

Table 3.C1: Heterogeneity by Survey (Baseline or Endline) in Fixed List Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Panel A. Baseline surveys									
Hours into Survey	-0.008*	-0.003	0.008*	0.005	0.002	-0.032***	-0.044***	-0.026***	
	[0.052]	[0.190]	[0.065]	[0.190]	[0.190]	[0.001]	[0.001]	[0.001]	
Dependent variable: Mean	0.076	0.166	0.191	0.069	0.022	0.227	0.289	0.194	
Hours into Survey: Mean	1.9	1.8	1.9	2.0	2.0	1.9	2.0	2.0	
Hours into Survey: SD	0.7	0.7	0.7	0.7	0.8	0.8	0.7	0.8	
Number of respondents	4,840	4,840	4,840	4,877	4,877	4,878	4,878	4,875	
Observations	64,860	98,735	102,610	52,640	67,977	174,600	53,658	80,940	
Panel A. Endline surveys									
Hours into Survey	-0.005	-0.005	-0.012*	-0.009	-0.002	-0.017**	-0.032**	-0.017	-0.002
	[0.313]	[0.316]	[0.077]	[0.148]	[0.394]	[0.012]	[0.032]	[0.196]	[0.469]
Dependent variable: Mean	0.068	0.144	0.162	0.053	0.018	0.180	0.212	0.070	0.050
Hours into Survey: Mean	1.6	1.6	1.7	1.8	1.8	1.8	1.9	2.0	2.0
Hours into Survey: SD	1.1	1.2	1.1	1.2	1.2	1.2	1.2	1.2	1.2
Number of respondents	5,349	5,349	5,349	5,349	5,073	5,349	5,349	5,349	5,349
Observations	69,971	109,546	109,763	61,405	70,734	192,347	58,839	85,584	48,141

Note: Observations at respondent-question-item level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table 3.C2: Heterogeneity by Survey (Baseline or Endline) in Open-Ended Questions

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Panel A. Baseline surveys					
Hours into Survey	0.020	-0.076**	-0.085	-0.267***	-0.043
	[0.183]	[0.035]	[0.106]	[0.001]	[0.275]
Dependent variable: Mean	0.067	0.204	0.382	0.494	0.414
Hours into Survey: Mean	2.0	2.0	2.0	2.0	2.0
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8
Number of respondents	4,877	4,875	4,874	4,870	4,879
Observations	4,877	4,875	4,874	4,870	4,879
Panel A. Endline surveys					
Hours into Survey	-0.010	-0.029	-0.061	-0.139**	-0.159*
	[0.440]	[0.384]	[0.216]	[0.032]	[0.077]
Dependent variable: Mean	0.046	0.205	0.178	0.176	0.323
Hours into Survey: Mean	1.8	1.8	1.8	1.9	1.9
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2
Number of respondents	5,348	5,349	5,349	5,345	5,349
Observations	5,348	5,349	5,349	5,345	5,349

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table 3.C3: Heterogeneity by survey type (baseline or endline) on total monetary values of aggregated categorie

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Panel A. Baseline surveys											
Hours into Survey	-15.61*		6.40	-4.45	1.78	-4.65***	-2.50***		-0.36*	-0.44**	-0.27*
	[0.065]		[0.190]	[0.183]	[0.106]	[0.001]	[0.002]		[0.056]	[0.018]	[0.060]
Dependent variable: Mean	51.24		46.06	18.16	6.60	16.93	8.31		0.83	0.71	0.54
Hours into Survey: Mean	1.8		1.8	2.0	2.0	1.9	2.0		2.0	2.0	2.0
Hours into Survey: SD	0.7		0.7	0.8	0.8	0.8	0.7		0.8	0.8	0.8
Number of respondents	4,840		4,840	4,877	4,878	4,878	4,878		4,879	4,879	4,879
Observations	4,840		4,840	4,877	4,878	4,878	4,878		4,879	4,879	4,879
Panel A. Endline surveys											
Hours into Survey	-11.42	-1.32	3.29	0.33	-0.53	-3.06	-2.06	-0.10	-0.59	-0.69*	-1.06**
	[0.480]	[0.219]	[0.502]	[0.576]	[0.480]	[0.104]	[0.127]	[0.186]	[0.167]	[0.062]	[0.041]
Dependent variable: Mean	136.08	10.48	69.01	13.11	6.21	15.57	7.57	0.14	1.05	0.67	1.06
Hours into Survey: Mean	1.6	1.6	1.7	1.8	1.8	1.9	1.9	2.1	1.8	1.9	1.9
Hours into Survey: SD	1.3	1.3	1.3	1.2	1.2	1.7	1.7	1.6	1.2	1.2	1.2
Number of respondents	5,349	5,349	5,349	5,349	5,074	5,349	5,349	5,349	5,349	5,349	5,349
Observations	5,349	5,349	5,349	5,349	5,074	5,349	5,349	5,349	5,349	5,349	5,349

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Bibliography

Abay, Kibrom A., Guush Berhane, John Hoddinott, and Kibrom Tafere Hirfrfot (2021). "Assessing response fatigue in phone surveys: Experimental evidence on dietary diversity in Ethiopia". World Bank Policy Research Working Paper No. 9636.

Aggarwal, Shilpa, Jenny Aker, Dahyeon Jeong, Naresh Kumar, David S. Park, Jonathan Robinson, and Alan Spearot (2020). "The Effect of Cash Transfers and Market Access on Households in Rural Liberia and Malawi". AEA RCT Registry.

Aggarwal, Shilpa, Rebecca Dizon-Ross, and Ariel Zucker (2021). "Incentivizing Behavior Change: The Role of Time Preferences". National Bureau of Economic Research Working Paper No. 27079.

Agüero, Jorge M., Úrsula Aldana, Erica Field, Veronica Frisancho, and Javier Romero (2020). "Is Community-Based Targeting Effective in Identifying Intimate Partner Violence?" *AEA Papers and Proceedings* 110: 605–609.

Agüero, Jorge M. and Veronicar Frisancho (2021). "Measuring Violence Against Women with Experimental Methods". *Economic Development and Cultural Change*.

Ambler, Kate, Sylvan Herskowitz, and Mywish K. Maredia (2021). "Are we done yet? Response fatigue and rural livelihoods". *Journal of Development Economics* 153: 102736.

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

— (2008b). "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association* 103 (484): 1481–1495.

Angelucci, Manuela (2008). "Love on the Rocks: Domestic Violence and Alcohol Abuse in Rural Mexico". *The B.E. Journal of Economic Analysis & Policy* 8 (1).

Bacchus, Loraine J., Meghna Ranganathan, Charlotte Watts, and Karen Devries (2018). "Recent intimate partner violence against women and health: a systematic review and meta-analysis of cohort studies". *BMJ Open* 8 (7): e019995.

Backor, Kristen, Saar Golde, and Norman Nie (2007). "Estimating Survey Fatigue in Time Use Study". SIQSS Working Paper.

Bandiera, Oriana, Niklas Buehren, Robin Burgess, Markus Goldstein, Selim Gulesci, Imran Rasul, and Munshi Sulaiman (2020). "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa". *American Economic Journal: Applied Economics* 12 (1): 210–259.

Barker, Nathan, Gharad T. Bryan, Dean Karlan, Angela Ofori-Atta, and Christopher R. Udry (2021). "Mental Health Therapy as a Core Strategy for Increasing Human Capital: Evidence from Ghana". Working Paper 29407. National Bureau of Economic Research.

Blattman, Christopher, Eric P. Green, Julian Jamison, M. Christian Lehmann, and Jeannie Annan (2016). "The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda". *American Economic Journal: Applied Economics* 8 (2): 35–64.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan (2017). "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia". *American Economic Review* 107 (4): 1165–1206.

Bobonis, Gustavo J., Melissa Gonzalez-Brenes, and Roberto Castro (2013). "Public Transfers and Domestic Violence: The Roles of Private Information and Spousal Control". *American Economic Journal: Economic Policy* 5 (1): 179–205.

Brzozowski, Matthew, Thomas F. Crossley, and Joachim K. Winter (2017). "A comparison of recall and diary food expenditure data". *Food Policy* 72: 53–61.

Buller, Ana Maria, Amber Peterman, Meghna Ranganathan, Alexandra Bleile, Melissa Hidrobo, and Lori Heise (2018). "A Mixed-Method Review of Cash Transfers and Intimate Partner Violence in Low- and Middle-Income Countries". *The World Bank Research Observer* 33 (2): 218–258.

Chuang, Erica, Pascaline Dupas, Elise Huillery, and Juliette Seban (2021). "Sex, lies, and measurement: Consistency tests for indirect response survey methods". *Journal of Development Economics* 148: 102582.

Cullen, Claire (2020). "Method Matters: Underreporting of Intimate Partner Violence in Nigeria and Rwanda". World Bank Policy Research Working Paper 9274.

Deaton, Angus (1997). *The analysis of household surveys*. The World Bank.

De Mel, Suresh, David McKenzie, and Christopher Woodruff (2014). "Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka". *Journal of Development Economics* 106: 199–210.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018). "Measuring and Bounding Experimenter Demand". *American Economic Review* 108 (11): 3266–3302.

Devries, K. M. et al. (2013). "The Global Prevalence of Intimate Partner Violence Against Women". *Science* 340 (6140): 1527–1528.

Devries, Karen M., Joelle Y. Mak, Loraine J. Bacchus, Jennifer C. Child, Gail Falder, Max Petzold, Jill Astbury, and Charlotte H. Watts (2013). "Intimate Partner Violence and Incident Depressive Symptoms and Suicide Attempts: A Systematic Review of Longitudinal Studies". *PLOS Medicine* 10 (5): e1001439.

Dhar, Diva, Tarun Jain, and Seema Jayachandran (2018). "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India". Working Paper 25331. National Bureau of Economic Research.

Dolnicar, Sara, Bettina Grün, and Friedrich Leisch (2011). "Quick, Simple and Reliable: Forced Binary Survey Questions". *International Journal of Market Research* 53 (2): 231–252.

Domingo, Pilar, Rebecca Holmes, Alina Rocha Menocal, and Nicola Jones (2013). "Assessment of the evidence of links between gender equality, peacebuilding and statebuilding". ODI (Overseas Development Institute) Report.

Ellsberg, M., L. Heise, R. Peña, S. Agurto, and A. Winkvist (2001). "Researching domestic violence against women: methodological and ethical considerations". *Studies in Family Planning* 32 (1): 1–16.

Ellsberg, Mary, Diana J. Arango, Matthew Morton, Floriza Gennari, Sveinung Kiplesund, Manuel Contreras, and Charlotte Watts (2015). "Prevention of violence against women and girls: what does the evidence say?" *The Lancet* 385 (9977): 1555–1566.

Ellsberg, Mary, Lori Heise, Rodolfo Pena, Sonia Agurto, and Anna Winkvist (2001). "Researching Domestic Violence against Women: Methodological and Ethical Considerations". *Studies in Family Planning* 32 (1): 1–16.

Falb, Kathryn et al. (2016). "Implementation of Audio-Computer Assisted Self-Interview (ACASI) among adolescent girls in humanitarian settings: feasibility, acceptability, and lessons learned." *Conflict and Health* 10 (32).

Fincher, Danielle, Kristin VanderEnde, Kia Colbert, Debra Houry, L. Shakiyla Smith, and Kathryn M. Yount (2015). "Effect of Face-to-Face Interview Versus Computer-Assisted Self-Interview on Disclosure of Intimate Partner Violence Among African American Women in WIC Clinics". *Journal of Interpersonal Violence* 30 (5): 818–838.

Fulu, Emma, Rachel Jewkes, Tim Roselli, and Claudia Garcia-Moreno (2013). "Prevalence of and factors associated with male perpetration of intimate partner violence: findings from the UN Multi-country Cross-sectional Study on Men and Violence in Asia and the Pacific". *The Lancet Global Health* 1 (4): e187–e207.

Garcia-Moreno, Claudia, Christina Pallitto, Karen Devries, Heidi Stockl, Charlotte Watts, and Naeema Abrahams (2013). *Global and Regional Estimates of Violence Against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-partner Sexual Violence*. World Health Organization.

Ghandour, Reem M., Jacquelyn C. Campbell, and Jacqueline Lloyd (2015). "Screening and Counseling for Intimate Partner Violence: A Vision for the Future". *Journal of Women's Health* 24 (1): 57–61.

Green, Eric P., Christopher Blattman, Julian Jamison, and Jeannie Annan (2015). "Women's entrepreneurship and intimate partner violence: A cluster randomized trial of microenterprise assistance and partner participation in post-conflict Uganda (SSM-D-14-01580R1)". *Social Science & Medicine* 133: 177–188.

Haushofer, Johannes, Robert Mudida, and Jeremy P. Shapiro (2020). "The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-being". Tech. rep. w28106. National Bureau of Economic Research.

Haushofer, Johannes, Charlotte Ringdal, Jeremy P Shapiro, and Xiao Yu Wang (2019). "Income Changes and Intimate Partner Violence: Evidence from Unconditional Cash Transfers in Kenya". Working Paper 25627. National Bureau of Economic Research.

Heise, Lori (1998). "Violence Against Women: An Integrated, Ecological Framework". *Violence Against Women* 4 (3): 262–290.

Herzog, A. Regula and Jerald G. Bachman (1981). "Effects of Questionnaire Length on Response Quality". *The Public Opinion Quarterly* 45 (4): 549–559.

Hidrobo, Melissa and Lia Fernald (2013). "Cash transfers and domestic violence". *Journal of Health Economics* 32 (1): 304–319.

Hidrobo, Melissa, Amber Peterman, and Lori Heise (2016). "The Effect of Cash, Vouchers, and Food Transfers on Intimate Partner Violence: Evidence from a Randomized Experiment in Northern Ecuador". *American Economic Journal: Applied Economics* 8 (3): 284–303.

Höglinger, Marc and Ben Jann (2018). "More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model". *PLOS ONE* 13 (8): e0201770.

Jansen, Henrica, Charlotte Watts, Mary Ellsberg, Lori Heise, and Claudia Garcia-Moreno (2004). "Interviewer Training in the WHO Multi-Country Study on Women's Health and Domestic Violence". *Violence against Women* 10 (7): 831–849.

Jeong, Dahyeon, Shilpa Aggarwal, Jonathan Robinson, Naresh Kumar, Alan Spearot, and David S. Park (2021). "Exhaustive or Exhausting? Evidence on Survey Fatigue in Long Surveys". Working Paper.

Jones, Nicola, Janice Cooper, Elizabeth Presler-Marshall, and David Walker (2014). "The fall-out of rape as a weapon of war". ODI (Overseas Development Institute) Report.

Kilic, Talip and Thomas Pave Sohnesen (2019). "Same Question But Different Answer: Experimental Evidence on Questionnaire Design's Impact on Poverty Measured by Proxies". *Review of Income and Wealth* 65 (1): 144–165.

Krosnick, Jon A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". *Applied Cognitive Psychology* 5 (3): 213–236.

Laajaj, Rachid and Karen Macours (2021). "Measuring skills in developing countries". *Journal of Human Resources* 56 (4): 1254–1295.

Lee, David S. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". *The Review of Economic Studies* 76 (3): 1071–1102.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas (2005). "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation". *Sociological Methods & Research* 33 (3): 319–348.

Machisa, Mercilene T., Nicola Christofides, and Rachel Jewkes (2017). "Mental ill health in structural pathways to women's experiences of intimate partner violence". *PLOS ONE* 12 (4): e0175240.

Moser, Caroline and Ailsa Winton (2002). "Violence in the Central American Region: Towards an Integrated Framework for Violence Reduction". ODI (Overseas Development Institute) Working Paper.

Mummolo, Jonathan and Erik Peterson (2019). "Demand Effects in Survey Experiments: An Empirical Assessment". *American Political Science Review* 113 (2): 517–529.

Newman, Jessica Clark, Don C. Des Jarlais, Charles F. Turner, Jay Gribble, Phillip Cooley, and Denise Paone (2002). "The Differential Effects of Face-to-Face and Computer Interview Modes". *American Journal of Public Health* 92 (2): 294–297.

Omanyondo, Marie-Claire (2005). "Sexual Gender-based Violence and Health Facility Needs Assessment". WHO Report.

Park, David Sungho, Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson, and Alan Spearot (2021). "Private but Misunderstood? Evidence on Measuring Intimate Partner Violence via Self-Interviewing in Rural Liberia and Malawi". National Bureau of Economic Research Working Paper No. 29584.

Park, David Sungho and Naresh Kumar (2022). "Reducing Intimate Partner Violence: Evidence from a Multifaceted Female Empowerment Program in Urban Liberia". Unpublished.

Peterson, Cora, Megan C. Kearns, Wendy LiKamWa McIntosh, Lianne Fuino Estefan, Christina Nicolaidis, Kathryn E. McCollister, Amy Gordon, and Curtis Florence (2018). "Lifetime Economic Burden of Intimate Partner Violence Among U.S. Adults". *American Journal of Preventive Medicine* 55 (4): 433–444.

Raghunathan, Trivellore E. and James E. Grizzle (1995). "A Split Questionnaire Survey Design". *Journal of the American Statistical Association* 90 (429): 54–63.

Ranganathan, Meghna, Lori Heise, Amber Peterman, Shalini Roy, and Melissa Hidrobo (2021). "Cross-disciplinary intersections between public health and economics in intimate partner violence research". *SSM - Population Health* 14: 100822.

Roy, Shalini, Melissa Hidrobo, John Hoddinott, and Akhter Ahmed (2019). "Transfers, Behavior Change Communication, and Intimate Partner Violence: Postprogram Evidence from Rural Bangladesh". *The Review of Economics and Statistics* 101 (5): 865–877.

Smith, S.G., J. Chen, K.C. Basile, L.K. Gilbert, M.T. Merrick, N. Patel, M. Walling, and A. Jain (2017). "The National Intimate Partner and Sexual Violence Survey (NISVS): 2010-2012 State Report". Tech. rep.

Steenkamp, Chrissie (2005). "The Legacy of War: Conceptualizing a 'Culture of Violence' to Explain Violence after Peace Accords". *The Round Table* 94 (379): 253–267.

Tourangeau, Roger and Ting Yan (2007). "Sensitive questions in surveys". *Psychological Bulletin* 133 (5): 859–883.

Trevillion, Kylee, Sian Oram, Gene Feder, and Louise M. Howard (2012). "Experiences of Domestic Violence and Mental Disorders: A Systematic Review and Meta-Analysis". *PLOS ONE* 7 (12): e51740.

WHO (1996). "Violence against women : WHO consultation, Geneva, 5-7 February, 1996".

— (2012). "Understanding and addressing violence against women: Intimate partner violence". World Health Organization Report.

— (2016). "Ethical and safety recommendations for intervention research on violence against women". Tech. rep.

— (2021). "Violence Against Women Prevalence Estimates, 2018: Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women". Tech. rep.

Women, UN (2013). "The Contribution of UN Women to Increasing Women's Leadership and Participation in Peace and Security and in Humanitarian Response". UN Women Evaluation Report.