Pre-analysis Plan
Manuela Angelucci and Rachel Heath
September 24, 2018

1. Introduction

In this project, we seek to explore (i) whether and how a women's economic empowerment program improves women's socioeconomic outcomes and bargaining power and (ii) whether teaching men about the importance of women's empowerment increases these effects. We are partnering with Women for Women International (WfWI) to conduct a randomized control trial of their women's economic empowerment program in eastern Democratic Republic of the Congo. Within the treatment group, a random group of husbands will also receive a men's engagement program.

2. Research Design

2.1 Intervention(s)

The key intervention is the WfWI core integrated twelve-month training program. Throughout the core program, women learn about the value of their work in the family and local economy, basic business skills, health and hygiene practices, their role in decision-making, women's rights, and the benefits of working together in a group for social and economic purposes. The training is delivered to groups of 25 women at a time. Participants also receive a monthly stipend ($10 USD). Additionally, women also choose a vocational track to be trained in to develop and pursue a strong income-generating occupational activity, and are provided with referrals for health and other financial services. Through this year-long program, WfWI aims to achieve the following four primary outcomes: women earn and save money; women improve health and wellbeing; women influence decisions in the home and community; and women connect to networks for support.

In addition, half of the women in the treatment arm have been randomly assigned to receive Men's Engagement Programming (MEP). This will take the form of the woman's male spouse, partner, or other household member participating in 4 months of men's discussion groups. In these groups, men will discuss topics including women's economic empowerment, domestic violence, women's health, and more. Couples who are identified to be at high risk for domestic violence will receive an additional 4 facilitated discussion sessions where both husband and wife are present.

2.2 Selection of Sample and Assignment to treatment

2000 women were screened and identified as as eligible for WfWI programming following normal programmatic protocols, with additional women screened to be replacements. Specifically, the WfWI M&E team members go to the pre-identified local communities and explain the program and criteria to the local chiefs (socially and economically marginalized women, aged 18-55), who draw up a list of potential women in their villages and communities. The women are then individually screened by the WfWI's M&E team on

eligibility criteria which determine their social and/or economic vulnerabilities (e.g. husband passed away, single earner in household, unable to afford school fees for children). Women are then given an explanation of the program and asked to consider the commitment to participate fully and actively in all aspects of the program for a full year, work to earn an income, and save a portion of the cash stipend. In addition to meeting the inclusion criteria, and a willingness to participate in the program, the women must be receiving support from their family to attend the training, demonstrate their ability to participate in the programming without interference from a spouse and/or family, and be of adequate health to attend the 12-month training program. The women were from the following communities in South Kivu, Democratic Republic of Congo: Kamanyola, Nyangezi, Mumosho, and Ciheraoni-Luciga

The 2000 women then received the baseline survey, which took place from July 23 to August 16, 2017. Each enumerator conducted 2 to 4 surveys per day, based on the lists of women pre-screened by the WfWI M&E staff. The full team of enumerators conducted surveys in the 4 selected communities in the following order: Kamanyola (600 surveys), Luciga (600 surveys), Nyangezi (400 surveys), and Mumosho (400 surveys). An additional 39 women were surveyed to act as replacements.

1000 of these women were then selected to receive treatment. We grouped the 2000 eligible into 80 clusters of 25 and assigned to a control (C) and treatment group (T) in equal proportions. Then, among the 40 clusters assigned to treatment, we cross-randomized 20 of these clusters into the MEP group.

The year-long treatment then began August 2017 (600 women) and October 2017 (400 women). The women who began treatment in August 2017 had all finished the baseline survey before the beginning of treatment.

2.3 Hypotheses

Our main hypothesis is that the core treatment and MEP changed women's well-being. Namely:

- Consumption (both household and woman's consumption)
- Mental and physical health (including intimate partner violence)
- Bargaining power

We will specifically measure each outcome accordingly:

Consumption

Respondents are asked about both household and own consumption of 22 different food items/groups in the past seven days. The food items range from grains such as maize and rice; roots or tubers such as cassava and sweet potatoes; vegetables; fruits; meat; eggs; fish; beans or legumes; milk; oil or ghee; sweets or processed foods; to condiments, drinks, or restaurant food.

Consumption is measured by the total value of all the food items consumed by the household in the past seven days. This value is the sum of two parts: one, the value of food items that were purchased, which is measured by the household's expenditure. And two, the value of food items that were either grown by the household or were received as gifts, which is the product of the quantity consumed and the regional median price.

Respondents are also asked about their previous night's consumption, listing quantities consumed from the previous list of 22 food items/groups.

Mental and physical health

Three self-reported scales will be used to measure respondents' mental health: the GAD-7, the PHQ-9, and the locus of control scale.

The Generalized Anxiety Disorder (GAD-7) Scale (Spitzer, Kroenke, Williams, & et al, 2006) is a set of seven questions used to assess the respondent's level of generalized anxiety. Respondents are asked, over the past two weeks, how often they have been bothered by the following problems:

1. Feeling nervous, anxious, or on edge
2. Not being able to stop or control worrying
3. Worrying too much about different things
4. Having trouble relaxing
5. Being so restless that it is hard to sit still
6. Becoming easily annoyed or irritable
7. Feeling afraid as if something awful might happen

Scores of 0, 1, 2 or 3 are given for experiencing the above "Not at all", for "Several days", for "More than half the days" and for "Nearly every day", respectively. The scores are then totaled and presented from 0 to 21. Scores of 5, 10 and 15 represent cut-off points for mild, moderate and severe anxiety, respectively.

The Patient Health Questionnaire (PHQ-9) scale is a set of nine questions designed to evaluate the presence and severity of depression in respondents (Kroenke, Spitzer, & Williams, 2001). Respondents are asked, over the past two weeks, how often they have been bothered by the following problems:

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could notice. Or the opposite – being so fidgety or restless
9. Thoughts that you would be better off dead or hurting yourself in some way

Scores of 0, 1, 2, or 3 are given for experiencing the above "Not at all", for "Several days", for "More than half the days" and for "Nearly every day", respectively. The scores are then totaled and presented from 0 to 27. The final score is used, together with the responses from a tenth question ("How difficult have those problems made it for you to do your work, take care of things at home, or get along with other people?"), as a screening tool for clinical depression. The cutoffs for mild, moderate, and severe depression vary across cultures and subpopulation.

Respondents who have any question left unanswered or answered with "Don't Know" (around 10.8 percent of the respondents at baseline) will be excluded from the PHQ-9 calculation and assigned a missing value for the PHQ-9 score. If the missing values do not vary systematically by treatment arm, we will also create a version of the index that sums up all non-missing data. If this index is balanced at baseline, we will continue to use also this version. If not, we will drop it from the analysis.

In order to have a more inclusive measurement, we will also calculate a PHQ-8 scale, which leaves out the seventh question of the PHQ-9: "Trouble concentrating on things, such as reading the newspaper or watching television." This question had the highest incidence of missing values in the baseline and brought back 9.8 percent of respondents into consideration.

Locus of control: Each agreement with the first phrase (except for the second phrase for the last question) connotes an internal locus of control.

L1. FIRST PHRASE: What happens to me is my own doing. SECOND PHRASE: Sometimes I feel that I don't have enough control over the direction my life is taking.
L2. FIRST PHRASE: When I make plans, I am almost certain that I can make them work. SECOND PHRASE: It is not always wise to plan too far ahead, because many things turn out to be a matter of good or bad fortune anyhow.
L3. FIRST PHRASE: In my case, getting what I want has little or nothing to do with luck. SECOND PHRASE: Many times we might just as well decide what to do by flipping a coin.
L4. FIRST PHRASE: Many times I feel that I have little influence over the things that happen to me SECOND PHRASE: It is impossible for me to believe that chance or luck plays an important role in my life.

We will give one point for each reply that indicates internal locus of control and sum up the scores from each individual question.

Intimate partner violence (IPV). At endline, we will ask respondents whether they have been beaten, hit, or forced to have sexual intercourse in the past 30 days or twelve months. We will code each occurrence of IPV as 1 and sum up the total number of occurrences.

Activities of daily living. We ask (a) Can you currently do vigorous activities like running, lifting heavy objects, and carrying water? (b) Can you currently do moderate activities like working in the fields, sweeping, washing an infant, or walking 5 kilometers? (c) How much physical pain have you experienced in the past month? For each, we convert responses into binary variables where an answer of 1 means more difficulty (i.e., is there some difficulty or not possible? Is this activity not possible?), and construct an index out of these binary variables.

The effect of the core treatment on each of these measures is theoretically ambiguous. If the program increases total household resources or bargaining power, then we would expected higher women's consumption, mental well-being (Haushofer and Shapiro 2017), and relationship quality (Heath, Hidrobo, and Roy 2018). However, any backlash from men against the women's economic empowerment could lower women's consumption and well-being (Angelucci 2008; Heath 2014).

Similarly, the additional impact of the MEP is also theoretically ambiguous. Previous research has suggested that intrahousehold outcomes may not be efficient (Duflo and Udry 2004, Robinson 2012), which can lead to inefficiently low labor supply (Heath and Tan 2014). If so, there is evidence that noticing an inefficiency can change behavior (Beaman et al 2014; Hanna et al 2014). Alternatively, men's engagement could potentially decrease backlash against women's empowerment if it can successfully change their perceptions of masculinity, as in Cook et al (2015). But it could also backfire and increase resentment.

Women's bargaining power

- Women's participation in household decisions.

We asked the respondents who makes the decision regarding the following household issues: the respondent's employment status (i.e. whether or not the respondent can work outside the home), large household purchases, the respondent's contraceptive usage, seeking medical care for the respondent, and seeking medical care for the respondents' children.

We will sort responses into three categories: (1) the respondent does not make the decision; (2) the respondent makes the decision with others; and (3) The respondent makes the decision alone.

We will award one point if the respondent made the decision alone or with others. We will then construct a decision-making index following Anderson (2008).

In the baseline, almost 70% of respondents said contraceptive usage was either "Not Applicable" to their household or that no decision was made concerning this issue. Only 35% of these respondents are unmarried. If the high incidence of missing values continues in the follow-up surveys, we will exclude the contraceptive usage question from the decision-making index. As a result, the index will range from 0 to 5.

- Modified dictator game

We will use as a continuous variable the amount of money a woman passes to her husband in a modified dictator game in which we double the amount of money passed to her husband/partner, net of the amount the woman passes to a random husband/partner.

- Engel curves.

We will estimate Engel curves. If the treatment changes such curves, then we can conclude that the treatment affected women's bargaining power.

We also plan a separate analysis that focuses specifically on measuring women's bargaining power. This analysis will compare our three key ways of measuring women's bargaining power (1) the decision-making index (2) the amount of money passed to the husband in the modified dictator game and (3) the Engel curves. We will study both how these measures correlate in the cross-section, and how their treatment effects correlate. We will also consider how these three measures correlate with women's leisure, both cross sectionally and in terms of their treatment effects.

As channels for these main outcomes, we will also examine

- Total household income
  the sum of all earnings earned by working members in the respondent's household (as defined from the roster at baseline) from the past 7 days, including in-kind payments.
- Self employment
  - A dummy for whether the respondent is self employed
- Respondent's Income
  - the sum of her earnings from her job, or jobs (up to two), from the past 7 days, including in-kind payments. For wages, the woman was asked her earnings period and earnings. For self-employment, the woman was asked about earnings and expenses. Expenses about "any tools or equipment for your business that required that you made a large purchase. By large purchase, I mean a purchase of something you expect to use in your business for one month or more." were asked about separately. For these, the woman was also asked the expected number of months the item would be useful, so its total cost will be divided by lifespan to calculate the weekly cost.
  - To shed further light on the operation of the woman's enterprise, we will also look separately at revenues, variable costs, and fixed costs (as measured by the question just mentioned about "large purchases").
- Time use
  - Labor supply: total hours worked in up to two main jobs (in-kind, wage, and non-wage labor) in the past 7 days (form labor supply module) and in the previous 24 hours (from time use module)

- Non-market work: hours of non-market work in the previous 24 hours (from time use module)
- Leisure: hours of leisure, excluding sleep in the previous 24 hours (from time use module)
- Sleep: hours of sleep in the previous 24 hours (from time use module)
- Childcare: hours of child care, if applicable, in the previous 24 hours (from time use module).[1]
- Multitasking: hours in which the respondent undertook more than one activity concurrently in the previous 24 hours (from time use module)

- Risk-taking behavior
  - Whether the respondent picks option 1 in the following scenario: I have a task that I would like you to do. I am going to ask you to sort these beans so that they are in piles in different colors. I will pay you 1000 Francs for this task regardless of the amount of you sort. You will not keep the beans at the end. You now have to make a choice. Option 1: You have to sort this medium size bag of beans for me. (Show respondent the medium size). Option 2: I will place one piece of paper in one hand that says "small". I will place one piece of paper in my hand that says "big". You will not know which hand holds which piece of paper. I want you to choose one of my hands. Whichever hand you choose, you will have to sort the amount that is written on the piece of paper. There is a chance here you will sort the very small amount, but there is also a chance you will have to sort the very big amount.
  - If the respondent says she would like to be the second woman in the following scenario: "Now I will tell you about two women and I would like you to pick which woman you would rather be. Please respond according to your way of thinking and not what you think most people would say. There are no right or wrong answers. The first woman sells potatoes at the local market. She earns 1500 francs every day, very reliably. The second woman also sells potatoes at the local market. She has at a different location in the market which is sometimes very busy, but sometimes has no customers. On half the days she earns no money and has no customers, while on other days she earns 5000 francs."
  - An ordinal outcome to the question "In general, are you someone who is willing to take risks or avoids taking risks?"

- Men's use of controlling behavior
  - At endline, we will ask respondents whether they have been (a) insulted, yelled at or threatened, (b) prevent you from visiting or speaking to your family or friends, or try to prevent you from seeking medical care, (c) try to take your income, control how you spent money that you earned, or get angry because of how you spent or saved your income by a partner in the past 30 days or twelve months. We will convert each

[1] In this setting, childcare may not be perceived as a separate activity, if women carry their babies with them throughout the day. We will inspect the baseline data to understand how women classify childcare.

question into two binary outcomes, where the first captures whether the outcome had happened in the last 12 months (which includes 30 days), and the second whether the outcome had happened in the last 30 days.

2.4 Data Collection

Please use this section to provide details on pilot data and *prospective* data that you will collect after pre-results acceptance of your research design.

2.4.1 Sample Size and Power

Our data consist of a sample of 2000 participants, grouped into 80 clusters of 25 and assigned to a control (C) and treatment group (T1) in equal proportions. The sample is interviewed three times, once at baseline, once at endline, and once one year after program graduation. Our main goal is to measure the impact of the treatment, T1. Among the treatment group (T1), we further cross-randomize a second treatment (T2), assigning it to 20 of the 40 groups in the T1 arm. Our secondary goal is to measure the impact of the second treatment, T2.

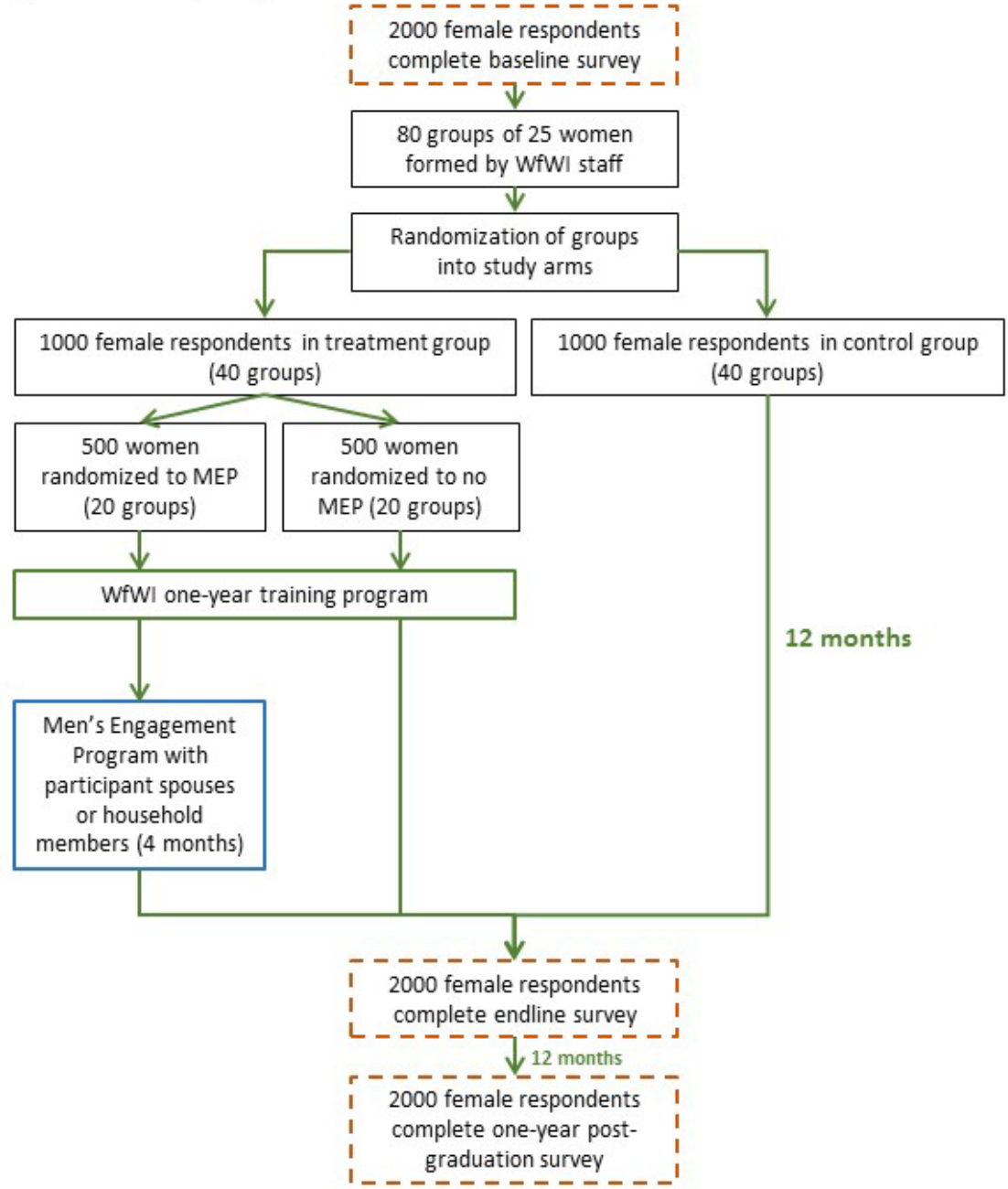
Our method consists of finding the minimum effect size we can detect in multiples of the standard deviation of the outcome, considering a significance level of 0.05 and a power of 0.9. We assume that the outcomes have a correlation coefficient of 0.5 between each of the three data waves (higher correlations let us detect a smaller minimum impact). We also assume an intra-cluster correlation of 0.1.
Under these assumptions, we can detect a minimum impact size for the main treatment, T1, of 0.2SD. Under the same assumptions, we can further detect a minimum impact size for the secondary treatment, T2, of 0.27SD. In both cases, these effect sizes are estimated using an ANCOVA estimator.

2.4.2 Surveys

The baseline survey took place from July 23 to August 16, 2017. Endline data collection begun on August 13, 2018 (though neither Angelucci nor Heath had seen the data at the time this document was prepared) and is expected to continue through November 2018.

2.5 Summary of Treatment and Surveys

Figure 1. RCT Study Design



3. Empirical Analysis

3.1 Statistical model

We are interested in measuring the causal effect of the core program (T), as well as whether this effect varies if the participants' husbands receive the men's engagement program (MEP) on some outcome Y. To do that, we will estimate the parameters of the following equation for woman *i* at time *t*:

$$Y_{it} = \beta_{0+}\beta_1 T_i \ + \beta_2 \, T_{\,i} \times \, MEP_{\,i} + \, X_i{}'\delta + \, \varepsilon_i$$

The covariates *X* include the baseline value of Y (Y_{i0}), for variables that we collected at baseline. We will also control for a quadratic in a woman's age and dummies for the community (Kamanyola, Nyangezi, Mumosho, and Ciheraoni-Luciga), as well as any variables that are significantly different between treatment and control at baseline.

The parameter β_1 identifies the average treatment effect of the women's program, while the parameter β_2 identifies whether this effect varies depending on whether the participant's husbands receive the men's program. These parameters are identified under random assignment and absent spillover effects. We will check that the randomization worked using balance tests and believe that our design minimizes spillover effects, as the treated women and their husbands are a very small fraction of the underlying community.

To improve the precision of our estimates, we will use either the above approach (ANCOVA) or instead use a DID or FD based on the degree of serial correlation in the dependent variable; we will choose an approach for all estimations based on the efficiency results in McKenzie (2012). Specifically, we will assess the serial correlation in our main outcomes of interest and decide between ANCOVA, DID, and FD based on the strategy that is optimal for the greatest number of these outcomes.

We will estimate the parameters by OLS, cluster the standard errors by group (the randomization unit) and add time and region dummies.

3.2 Statistical methods

- Missing values. We will test whether missing values differ systematically by treatment arms. For indices, our main approach will follow Anderson (2008), which sums over standardized non-missing values. For other variables, we exclude missing observations, if missing values do not vary by arm.
- For PHQ9, GAD, and locus of control indices, we will also construct this measure for individuals who answered all questions, for comparability with the literature.
- We will handle outliers by winsorizing continuous outcomes at the 5th and 95th percentiles.

3.3 Multiple outcome and multiple hypothesis testing

- We will follow Benyamini and Hochberg (1995) and control the false discovery rate.
- Indices will be constructed for key outcomes, as described in section 2.3 and 3.2

3.4 Heterogeneous outcomes

Broadly, we expect that there could be differential treatment effects based on women's ability, constraints (including her health), and beliefs. We will test for differential treatment effects based on the baseline levels of the following variables (at the median for the continuous variables) --

- hh-level per capita consumption
- literacy
- whether the woman worked outside the home
- marital status
- health (as measured by the mental health and ADL indices)
- empowerment (as measured by the decision-making index)
- beliefs about gender (as measured by her responses to questions SA15, SA16, SA17 on the survey)
- we will also regress the main outcomes of interest on baseline socio-economic variables; for each outcome, we will give one point to the top 5 variables with the highest partial R-squared. We will then estimate heterogeneous effects for the 5 variables with the highest points.

4 List of References

Anderson, Michael L. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." Journal of the American statistical Association 103.484 (2008): 1481-1495.

Angelucci, Manuela. "Love on the rocks: Domestic violence and alcohol abuse in rural Mexico." *The BE Journal of Economic Analysis & Policy* 8.1 (2008).

Beaman, Lori, Jeremy Magruder, and Jonathan Robinson. "Minding small change among small firms in Kenya." Journal of Development Economics 108 (2014): 69-86.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the royal statistical society. Series B (Methodological) (1995): 289-300.

Cook, Philip J., Kenneth Dodge, George Farkas, Roland G. Fryer Jr, Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg. "Not Too Late: Improving Academic Outcomes for Disadvantaged Youth." Institute for Policy Research Northwestern University Working Paper WP-15-01. http://www. ipr. northwestern. edu/publications/papers/2015/ipr-wp-15-01. html (2015).

Donald, Aletheia, Gayatri Koolwal, Jeannie Annan, Kathryn Falb, and Markus Goldstein. "Measuring women's agency." (2017).

Duflo, Esther, and Christopher Udry. Intrahousehold resource allocation in Cote d'Ivoire: Social norms, separate accounts and consumption choices. No. w10498. National Bureau of Economic Research, 2004.

Field, Erica, et al. On Her Account: Can Strengthening Women's Financial Control Boost Female Labor Supply?. Technical report, 2016.

Gallegos, J. V. and I. A. Gutierrez (2011). The effect of civil conflict on domestic violence: the case of peru. Unpublished manuscript.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. "Learning through noticing: Theory and evidence from a field experiment." The Quarterly Journal of Economics 129.3 (2014): 1311-1353.

Heath, Rachel, and Xu Tan. "Intrahousehold bargaining, female autonomy, and labor supply: Theory and evidence from india." University of Washington (2014).

Heath, Rachel, and Xu Tan. Worth Fighting For: Daughters Improve their Mothers' Autonomy in South Asia. Tech. rep., mimeo, 2015.

McKenzie, David. "Beyond baseline and follow-up: The case for more T in experiments." Journal of Development Economics 99.2 (2012): 210-221.

La Mattina, Giulia. "Civil conflict, domestic violence and intra-household bargaining in post-genocide Rwanda." *Journal of Development Economics*124 (2017): 168-198.

Robinson, Jonathan. "Limited insurance within the household: Evidence from a field experiment in Kenya." American Economic Journal: Applied Economics 4.4 (2012): 140-164.

Seymour, Greg, and Amber Peterman. "Understanding the Measurement of Women's Autonomy: Illustrations from Bangladesh and Ghana." (2017)