Pre analysis plan: Effects of information about norms on beliefs, attitudes, and behavior regarding FGM in Ethiopia

Matthew Gichohi, Andreas Kotsadam, Mette Løvgren, Charlotte Ringdal, and Espen Villanger

Abstract

We study the effects of providing information about community level attitudes on Female Genital Mutilation (FGM) on beliefs, attitudes, and behavior. In this plan we describe the hypotheses to be tested and how they will be tested. The description includes how the variables are coded, how we will deal with missing values, and the specification of the estimation equations. We also conduct a power analysis which suggests that we are able to identify relatively small effects. All deviations from the plan will be highlighted in the final paper.

1. Introduction

Female genital mutilation (FGM), the practice of cutting or removing girls' genitalia for non-medical reasons, is a human rights violation with profound negative consequences (Unicef, 2013). The practice is very common, with more than 200 million women and girls alive today having been cut and over 3 million girls at risk of being cut each year (WHO, 2020). In Ethiopia, several studies have used Demographic and Health Survey (DHS) data showing that while the FGM prevalence has declined since the year 2000 it was still as high as 65% in 2016 (Alemu, 2021; Azeze et al., 2020; Muche, 2020).

Every year, millions of dollars are spent on addressing the practice (Engelsma et al., 2020), yet it seems hard to change. Many of the interventions targeting FGM focus on changing norms in a population by influencing individual's attitudes. These programs include community mobilization, awareness rising programs and social marketing campaigns using various forms of media (Ellsberg et al., 2015). However, even if campaigns are successful in changing attitudes, that does not imply that behavior will be changed, especially if decisions are influenced by perceptions about others' attitudes and expectations.

Using data from a baseline survey in five regions of Ethiopia, we know that (i) even if parents oppose FGM, many still cut or intend to cut their daughters, (ii) the perceived social norm about FGM is strong, and parents tend to overestimate the support of FGM in their community, and (iii) perceived social norms and misperceptions are highly correlated with girls' FGM status.

Information provision experiments, where individuals are randomly given information about a topic, have become increasingly popular in the past decade (Haaland et al., 2023). A recent study in Saudi Arabia, showed how this kind of experiments can be used to study the causal effect of perceived social norms on behavior (Bursztyn et al., 2020). We plan to conduct an information provision experiment, where we randomly tell respondents what the attitudes in a nearby community concerning FGM was at baseline.

In this plan we describe the experiment in detail and how we will analyze the data from the experiment.

2. The field experiment

2.1. The timeline

In collaboration with Norwegian Church Aid (NCA) and Save the Children (SC), we collected baseline data from 1635 mothers and 1510 fathers across five regions in Ethiopia (Somali, Afar, Sidama, Amhara, and Oromia) in the early spring of 2021. Overall, we have collected the FGM status of 3361 girls. Two more data collections are planned: a midline in 2023 and an endline in 2024.

2.2. The experiment

At baseline, we asked all mothers and fathers *"A girl in our village should be circumcised"*. They could answer "Agree" or "Disagree". We also asked them what share of men and women they believed disagreed with the statement.

Our main intervention consists of randomly telling people how many individuals in a nearby community actually oppose FGM. For ethical reasons we avoid telling people that fewer individuals than they thought at baseline oppose the harmful practice, that less individuals oppose the practice than what is true in their own communities, and we never tell people that a majority support FGM.

Within all regions, we assign individuals in the experimental sample to one of three conditions:

1) 1/3 is allocated to get information about the share of women in a nearby community that oppose FGM at baseline.

2) 1/3 get information about the share of men in a nearby community that oppose FGM at baseline.

3) 1/3 of the experimental sample get no information in 2023.

In one of the regions, Afar, we are only able to assign information about the share of women in a nearby community that oppose FGM since the share of men supporting FGM was always above 50 percent. In Afar, we therefore allocate two thirds of the sample to 1 and one third to 3.

In practice, we calculate the shares of disapproval in each kebele by gender and we assign each individual the highest gender specific share of disapproval in the region that is not 100 percent. We chose not to say that 100 percent disapproves of FGM, as we expect that respondents will find it hard to believe that no-one approved of FGM in a community. Any kebele with higher disapproval rates than the highest rate is deleted from the experimental sample, as is any individual believing that the opposition is higher than the highest positive share.

The exact text that we provide the respondents is the following:

"*Last time, we visited many people in addition to you in several regions of Ethiopia. All of them had children below the age of 13. One of the questions we asked them was whether they agree or disagree with the following statement "A girl in our village should be circumcised". In one of the communities close to your,* ***X*** *out of 100* ***women/men*** *said that they disagreed. This means that most of the* ***women/men*** *we interviewed in that community does not think that a girl should be circumcised*

One third are told about attitudes of women and one third about attitudes of men. X varies from 57 to 96.

The ethical considerations imply that we drop 732 individuals (23 percent) of the initial sample so that the experimental sample consists of 2,413 individuals, 1251 mothers and 1162 fathers. In particular, our restrictions imply that we drop individuals in the most progressive communities and that we drop more progressive individuals. In Figure 1 we see that around 50 percent in the sample that we keep agreed at baseline that girls should be cut, whereas very few of the individuals in the sample that we drop agreed. In Figure 2 we similarly show that we mostly drop individuals living in the most progressive communities. This restriction lowers the external validity but the sample becomes more relevant as we want to treat individuals that are most likely to be affected by a progressive treatment.

Figure 1

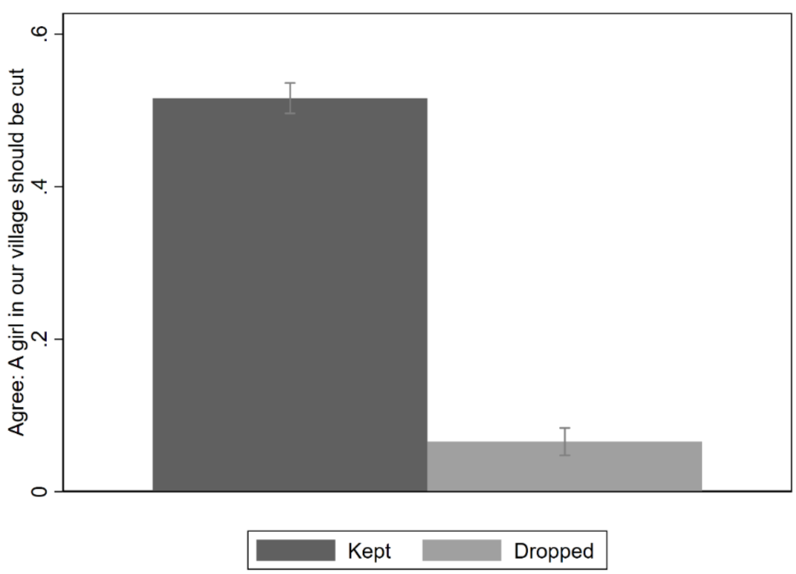
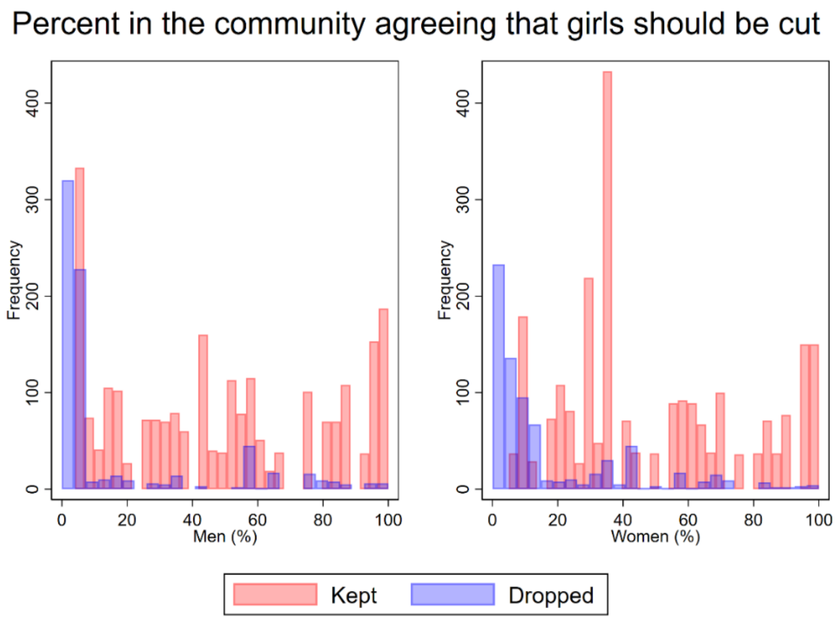


Figure 2



We expect some attrition from baseline to endline and we will replace missing households with replacement households. We will also have a treatment for replacement households but it will be less precise since we do not have baseline data for such respondents. The treatments for the replacement households is therefore:

"*Last time, we visited many people in addition to you in several regions of Ethiopia. All of them had children below the age of 13. One of the questions we asked them was whether they agree or disagree with the following statement "A girl in our village should be circumcised". In one of the communities close to your, a sizeable majority of* ***women/men*** *said that they disagreed. This means that most of the* ***women/men*** *we interviewed in that community does not think that a girl should be circumcised*

One third are told about attitudes of women and one third about attitudes of men. Again, one third of the replacement sample get no such information.

3. Data and coding of main variables

We collected baseline data in the spring of 2021. The intervention will be implemented during the second round of data collection that will take place during June of 2023.

We first collect data on demographics, prevalence of early marriage and prevalence of FGM using the household roster. Then we treat individuals. Later in the survey we ask questions about beliefs, attitudes, and planned behavior.

3.1 Primary and secondary dependent variables

Our short run outcomes Y, are:

- Decision about FGM for a hypothetical girl (Vignette): *If you have a daughter in the future, do you intend to circumcise her?* The variable is an indicator variable taking the value 1 if the respondent answers "yes", and 0 if the respondent answers "no".
- Own attitudes about FGM: *Do you agree or disagree with the following statement: A girl in our village should be circumcised*. Indicator variable taking the value 1 if respondent agrees and 0 if respondent disagrees.
- Plans for uncut daughters, where applicable: *Do you plan to circumcise [name of daughter]?* The variable is an indicator variable taking the value 1 if the respondent answers "yes" for at least one daughter, and 0 if the respondent answers "no" for all daughters.

- Beliefs about attitudes of FGM in the community: The average of the answers to the two questions: *Imagine that there are 100 women in your village. Of these, how many agree with the statement: A girl in our village should be circumcised?* Discrete variable between 0 and 100.
  And: *Imagine that there are 100 men in your village. Of these, how many agree with the statement: A girl in our village should be circumcised?* Discrete variable between 0 and 100.

These four variables are our main short run variable of interest. In addition, we will use one long-run outcome; the share of cut daughters at endline in 2024. In total, we have 5 main outcomes. If treatment has an effect in 2023 we will inform everyone in 2024 so there will be no evaluation of long term effects on other variables.

3.1.1. *Secondary outcome*: S*ocial desirability index.*

We create a *Social desirability index* based on responses on a Marlowe-Crowne module which measures a person's general tendency to give socially desirable answers. The index is based on the following questions where individuals can answer agree or disagree:

i. It is sometimes hard for me to go on with my work if I am not encouraged.
ii. I sometimes feel resentful when I don't get my way.
iii. On a few occasions, I have given up doing something because I thought too little of my ability.
iv. There have been times when I felt like rebelling against people in positions of authority even though I knew they were right.
v. No matter who I'm talking to, I'm always a good listener.
vi. There have been occasions when I took advantage of someone.
vii. I'm always willing to admit it when I make a mistake.
viii. I sometimes try to get even rather than forgive and forget.
ix. I am always courteous, even to people who are disagreeable.
x. I have never been irked when people expressed ideas very different from my own.
xi. There have been times when I was quite jealous of the good fortune of others.
xii. I am sometimes irritated by people who ask favours of me.
xiii. I have never deliberately said something that hurt someone's feelings.

We use the baseline answers to these questions and create a variable *High Social Desirability* which equals 1 if the respondents answer more than 7 of the questions in a socially desirable way. We test for treatment effect heterogeneity by interacting this variable with treatment. The worrisome pattern would be if the treatment effects were driven by individuals with a high propensity to disingenuously give socially desirable answers and vanished for those with a low such tendency.

3.2 Main independent variables

All main independent variables except treatment are from the baseline surveys. The value codings and response categories may be different from baseline to follow up but as we use ancova specifications (see below) and have general coding rules for dummy variables, this is not a problem. We will always have the strata variables (region fixed effects) as independent variables since randomization is within regions.

Our main treatment variable is: Information about high share of FGM opposition. This variable is coded as 1 if information about either women's or men's high share of opposition is given and zero otherwise. Note that this treatment is well defined also for the replacement sample.

The other main independent variables are all from the baseline survey and they are:

<u>Continuous numerical values of:</u>

- Age
- Education ("What is the highest level of education you have completed")
- Number of children
- Number of girls
- How common is it currently for girls to be circumcised in your community?
- From what you know, how does your religion view female circumcision?

  Beliefs about:
- Percentage of men in the village agree with girls FGM
- Percentage of women in the village agree with girls FGM

<u>Dummy variables for:</u>

- Employment status (=1 if having engaged in income generating activities)
- Literacy (1 if yes on "Do you know how to read and write in any language")
- Gender of respondent
- Answering "Yes" to "If you have a daughter in the future, do you intend to circumcise her?"
- Thinking that "A girl in our village should be circumcised"

For the replacement households we will add the corresponding answers to some of these questions (the demographic ones) but they are asked in the midline survey before the treatment.

For both samples we will also add the following pre-treatment questions from the midline survey:

<u>Dummy variables for:</u>

Dummy for stopped: Do you think that female circumcision should be continued, or should it be stopped? *1=continued 2=stoped 3=Do not know*

Do you think that FGM is harmful to the girl? *1=Yes 2=No*

Do you think FGM is violating the rights of the girl? *1=Yes 2=No*

Continuous numerical values of:

	To what extent do you agree that…	Response option
a	Cutting shows respect to our elders.	1.strongly agree 2. agree 3.neither agree nor disagree 4. disagree 5. strongly disagree
b	Cutting helps a girl stay a virgin until she marries.	
c	Cutting teaches girls obedience and respect.	
d	Cutting is not the right thing to do to girls in our community.	
e	Cutting marks the transition from a girl child to a woman/adult.	
f	Cutting is not part of our traditions and culture.	
g	Cutting ensures that a girl retains her femininity.	
h	Uncut girls are not pure.	

4. Empirical strategy

4.1 Individual level effects of treatment

For the analysis of individual level treatment effects, we estimate the following regression:

$$Y_{it} = \beta Treated_i + \delta X_{1i} + \varepsilon_i$$

where t=2 indicates midline follow up (immediate effects), t=3 is the longer-term effects for a planned endline taking place in 2024 (long term effects), and t=1 indicates baseline, i indexes individuals. The vector X will always include fixed effects for the regions and the main specification will include optimally chosen control variables. In particular, we will pick optimal controls from the total list of controls in section 3.2. using the double debiased LASSO procedure (Belloni et al. 2014; Ahrens et al. 2018). This is likely to increase power, especially since we have baseline values of the outcomes. In the regressions, all continuous variables will be dummy coded up until they contain at least five percent of the sample. If we have missing values on explanatory variables, we will code the variables as zero and include dummy variables controlling for missing status so that we do not lose observations. We use robust standard errors in all estimations unless otherwise stated.

4.2 Balance and attrition

To test for balance, we will regress treatment on the independent variables from the baseline (or the corresponding demographic variables for the replacement sample at midline) described in section 3.2, both individually and together, while controlling for the Strata variables (region fixed effects). We will judge whether the randomization worked by conducting an F-test of whether the control variables jointly predict treatment status. Continuous variables will not be dummy coded in the balance tests.

We will probably not manage to reach all the respondents in the final endline survey. We will check whether attrition and missing outcomes are correlated with treatment. If there are statistically significant differences in attrition or non-response between treatment and control (controlling for the strata variables), we will follow the correction proposed by Lee (2009).

5. Heterogeneity and further exploratory analyses

We will conduct exploratory analyses where we test if the effect differs when depending on whether information is given about men and when the respondent is either a woman or a man. We will also test for differences between replacement and original households. These households may be different in many ways, so the results will not simply be interpreted as responses to precise or fuzzier treatments.

We will use machine learning techniques to automate the search for heterogenous treatment effects. In particular, we will use honest causal forest estimation to credibly search for heterogeneity.

Under the assumption that all the treatment effect runs via updated beliefs we can estimate the effects of beliefs on attitudes by running IV regressions instrumenting beliefs with treatment. While the assumptions may not be met, we still believe it is a useful exercise.

We will also test whether the size of the possible updating by the treatment is important. We can do this since we use different information in different areas, and since we have baseline beliefs in the data.

6. Power calculation, reporting, and discussion of null findings

Assuming that we manage to reach a sample of 1000 individuals we can detect effects as small as 0.14 standard deviations with 80 percent power (calculation based on 2/3 being allocated to treatment). This is a relatively small effect, but that is for a continuous variable as the main outcome. If we assume that the vignette is our main outcome, the effect would have to be around 10 percentage points (assuming that around 48 percent of the control would answer yes, as in the baseline survey). If we have 5 main outcomes instead (and follow Fink et al. 2014, see below), we would need the effect to be around 13 percentage points. That is not such a big difference. We therefore propose to have 4 or 5 main outcomes. We are also likely to gain substantially powerwise by adding baseline values of the control variables.

We will adjust the p-values for the fact that we are testing the impact on 5 outcomes. We follow the recommendations of Fink, McConnell, and Vollmer, (2014) to minimize the false non-discovery rate. The main advantage of the method is that it is limiting the risk of false discoveries while only adjusting the critical values based on other true hypotheses. The false discovery rate method implies that the m p-values of the i hypotheses are ordered from low to high and that the critical value of the p-value is then p(i) = a*i/m. To illustrate, with 5 hypotheses and a significance level (a) of 0.05, the critical p-value would be 0.01 for the one with the lowest p-value (0.05* 1/5, which is the same as a Bonferroni correction). For the second hypothesis, the critical p-value is 0.02 (0.05*2/5) and for the fifth it is 0.05 (0.05*6/5). As we have 5 main hypotheses, the most significant effect would have to have a p-value of less than 0.01 in order for it to pass the 5 percent bar after multiple hypotheses correction.

Note that we apply this correction to the main dependent variables only; when discussing individual variable results within particular outcome groups, we use conventional significance levels. We use this approach because the purpose of studying individual variables within the outcome groups is to understand mechanisms, rather than to single out particular variables for general conclusions.

We may not publish all of the findings in the same paper but we will create a working paper where we present all results outlined in this plan. This working paper will then be referred to and we will mention all main outcomes studied in each published paper.

Some of our findings are likely to be null results. It is often difficult to judge whether such results are showing a meaningful lack of effect or whether they arise due to low power. To investigate if the effects are meaningful null findings we will conduct equivalence tests with two one-sided t-tests (TOST) and show how large positive and negative effects we can reject. The tests are one sided in equivalence testing as one tests whether effects are larger than a highest value and lower than a lowest value. In practice, the procedure is equivalent to presenting the bounds of a 90 percent confidence interval.

6. Ethical approval

This study was reviewed and approved by the Norwegian Centre for Research Data and the ethical committee at CMI.

7. Archive and sharing of replication data

The pre-analysis plan is archived before any follow-up data is collected. We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: https://www.socialscienceregistry.org/ on June 12. We will start the data collection June 13, 2023 and data is expected in the fall of 2023.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. (2018). PDSLASSO: Stata module for or post-selection and post-regularization OLS or IV estimation and inference.

Alemu, Addisu Alehegn, “Trends and Determinants of Female Genital Mutilation in Ethiopia: Multilevel Analysis of 2000, 2005 and 2016 Ethiopian Demographic and Health Surveys,” International Journal of Women’s Health, 2021, 13, 19.

Azeze, Gedion Asnake, Anita Williams, Hannock Tweya, Mohammed Suleiman Obsa, Taklu Marama Mokonnon, Zewde Zema Kanche, Robera Olana Fite, and Anthony D Harries, “Changing prevalence and factors associated with female genital mutilation in Ethiopia: Data from the 2000, 2005 and 2016 national demographic health surveys,” PLoS one, 2020, 15 (9).

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler. "Optimal design of experiments in the presence of interference." *Review of Economics and Statistics* 100, no. 5 (2018): 844-860.

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

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. "Misperceived social norms: Women working outside the home in Saudi Arabia." *American economic review* 110, no. 10 (2020): 2997-3029.

Engelsma, Brian, Gerry Mackie, and Brandon Merrell, “Unprogrammed abandonment of female genital mutilation/cutting,” World Development, 2020, 129

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness*, *6*(1), 44-57.

Haaland, I., Roth, C., & Wohlfart, J. (2023). Designing information provision experiments. *Journal of economic literature*, *61*(1), 3-40.

Lee, David S. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76, no. 3 (2009): 1071-1102.

Muche, Achenef, “Magnitude of Female Genital Mutilation in Ethiopia: A systematic review and Meta-analysis,” Ethiopian Journal of Reproductive Health, 2020, 12 (01), 19–19.

UNICEF, Female genital mutilation: a statistical exploration of the dynamics of change 2013,Unicef, 2013.

WHO, “Female Genital Mutilation,” https://www.who.int/news-room/fact-sheets/detail/female-genital-mutilation 2020.