Sensitive survey questions: Measuring attitudes regarding female genital cutting through a list experiment*

Elisabetta De Cao† and Clemens Lutz††

†University of Oxford, ††University of Groningen

Conditionally Accepted at Oxford Bulletin of Economics and Statistics

Abstract

Potential bias in survey responses is higher if sensitive outcomes are measured. This study analyses attitudes towards Female Genital Cutting (FGC) in Ethiopia. A list experiment is designed to elicit truthful answers about FGC support and compares these outcomes with the answers given to

*For supervising the data collection we thank IFPRI, Getaw Tadesse and Samson Jemaneh. For comments we thank Robert Lensink, Rob Alessie, Carol Propper, Franco Peracchi, Aljar Meester, Bryn Rosenfeld, Viola Angelini, Andreas Rauch, Petros Milionis, Mariko Klasing, as well as seminar participants at the 2015 RES Women's Committee Mentoring Meetings, 2015 RES conference, 2015 CSAE conference, the 2013 IFP conference in Addis Ababa, at the PEG seminar series at the University of Groningen, the University of Wageningen for useful comments. All errors are our own. This research has been financed by The Netherlands Organisation for Scientific Research/Science for Global Development programmes (NWO/WOTRO), grant number: W 07.72.2011.115.

a direct question. Our results confirm that the average bias is substantial as answers to direct questions underestimate the FGC support by about ten percentage points. Moreover, our results provide suggestive but not statistically significant evidence that this bias is more pronounced among uneducated women and women targeted by an NGO intervention but not randomly assigned.

Key words: Female genital cutting; list experiment; attitudes; measurement; sensitive survey questions; Ethiopia.

JEL-Classification: I15; O10; C13; C83.

Word count: 12,500

1 Introduction

Eliciting honest answers in surveys is challenging, especially when studying sensitive issues. If asked directly, individuals may falsify or refuse to answer certain questions. The dependent sensitive variable, therefore, might be affected by a non-random measurement error that leads to biased results. Self-reported health status and outcomes have been determined as being affected by underreporting when, for example, they focus on sensitive topics related to sexual and reproductive health (Schroder et al., 2003; Glynn et al., 2011). When asking questions about a sensitive issue, different survey methods exist for coping with the problem of bias in self-reported answers.

New qualitative solutions have been proposed by Blattman et al. (2016) to study the direction and magnitude of the survey measurement error in the de-

pendent variable when evaluating interventions implemented in Liberia to reduce violence and crime. Blattman et al. (2016) use qualitative techniques to validate survey responses in relation to different behaviours (theft, drug use, homelessness, gambling, and expenditures) and ascertain different results in terms of underreporting depending on the sensitive behaviour that is being considered.[1]

Quantitative survey methods include the randomised response technique[2] and the endorsement experiment.[3] A third method used in this paper is called *list experiment*. The concept of a list experiment, also referred to as an *item count* or *unmatched count technique*, is that, if a sensitive question is asked indirectly, the respondent may reveal a truthful response. The method presents respondents with a list of items and asks them to indicate the total number of items with which they agree. The respondents are randomly divided into either a control or a treatment group. The control group respondents receive a list of non-sensitive items. The treatment group respondents receive the same list of non-sensitive items plus one sensitive item. The proportion of the respondents who agree with the sensitive

[1]Blattman et al. (2016) randomly selected a subsample of the respondents to validate survey responses. The goal was for the validators to determine if the respondent had engaged in any of the measured behaviour by meeting a few times with the individual with the goal of developing a rapport and gain trust. Then, by engaging in casual conversation, the validators raised indirect questions (by telling stories or scenarios) about the behaviours.

[2]The randomised response technique (RRT) consists of asking the respondent to use a randomisation device (dice, coin flip, etc) whose outcome is unknown to the interviewer. By introducing random noise, the RRT guarantees the anonymity, and the respondent may be more willing to reveal the truth. See Warner (1965) for further details.

[3]In an endorsement experiment, respondents are randomly assigned to a treatment group and asked to express their opinion toward a policy endorsed by a specific actor whose support level needs to be measured. These responses are then compared with those from a control group of respondents that answered an identical question without the endorsement. See Bullock et al. (2011) for further details.

item is estimated computing the difference in the mean response between those two groups. This technique has mainly been used in political science to understand voters' attitudes and racial attitudes (e.g., Kuklinski et al., 1997; Redlawsk et al., 2010). It has also been utilised to study sexually risky behaviour (LaBrie and Earleywine, 2000) and abortion (Moseson et al., 2015). More recently, it has also been applied in economics to study sensitive issues. In micro-finance, for example, Karlan and Zinman (2012) used a list experiment to understand how people spend their loan proceeds, showing that direct elicitation underreports the non-enterprise uses of loan proceeds. In reproductive health, list experiments have been developed to obtain truthful answers on topics such as condom use, number of sexual partners, unfaithfulness, and attitude changes with respect to the social acceptability of these behaviours (Jamison et al., 2013; Chong et al., 2013). De Cao et al. (2017) employ a series of list experiments to study if community conversations contribute to a change in social values, beliefs, and attitudes regarding harmful traditional practices against women in Ethiopia. A paper by Coffman et al. (2013) estimates the magnitude of anti-gay sentiment showing that it is generally underestimated when a list experiment is used to elicit truthful answers.

Surprisingly, the aforementioned economic literature considers a difference-in-means estimator to analyse the list experiment (see, for example, Karlan and Zinman, 2012; Chong et al., 2013).[4] This, however, does not allow the identification of the relationship between preferences over the sensitive item and the

[4]An exception is the paper by Coffman et al. (2013) that uses a regression approach to study the social desirability bias. We improve on that by considering heterogenous effects across a different set of respondents' characteristics.

respondent's characteristics. Moreover, the effect of social pressure on the answers provided to direct sensitive questions may differ among groups in the population. Regressions can instead be used to study how the probability of affirmatively answering the sensitive question varies as a function of respondents' characteristics, and also which respondents are likely to answer sensitive questions differently, depending on whether asked directly or indirectly through a list experiment. In the list experiment literature, the difference between responses to direct and indirect questioning has been interpreted as a measure of *social desirability bias* (Corstange, 2009; Holbrook and Krosnick, 2010; Imai, 2011; Blair and Imai, 2012).

In this paper, we design a list experiment to indirectly ask respondents about their support towards female genital cutting (FGC). FGC or female genital mutilation or female circumcision includes all procedures that alter or cause injury to the female genital organs. The procedure is primarily performed on young females. FGC is recognised as an extreme form of discrimination and violence against women. Worldwide, approximately 140 million girls and women are living with the consequences of it.[5] The WHO estimates that, annually in Africa, more than three million young females are at risk for FGC (WHO, 2012).

There is extensive anthropological literature on the existence of FGC.[6] Economists have recently began studying FGC from both a theoretical perspective (Ches-

[5] A review on the health consequences of FGC can be found in Obermeyer (2005).

[6] For an extensive review, see Shell-Duncan and Hernlund (2000). Theories about the nature of FGC as a social convention have been developed by Mackie and LeJeune (2009) and tested by Shell-Duncan et al. (2011), and recently disproved by Efferson et al. (2015).

nokova and Vaithianathan, 2010; Coyne and Coyne, 2014) and empirically investigating, in particular, the determinants and consequences of FGC (Naguib, 2012; Ouedraogo and Koissy-Kpein, 2012; Molitor, 2014; Bellemare et al., 2015; Wagner, 2015) or the effect of laws or program interventions against FGC (Camilotti, 2016).

Remarkably, quantitative research regarding FGC attitudes uses direct survey questions to gauge the support for the continuation of its practice. In the Demographic and Health Survey (DHS) (Yoder and Khan, 2008), for example, the question about perceptions towards female circumcision is a direct question: "Do you think that female circumcision should be continued, or should it be stopped?" We contend that this type of direct questioning may lead to misreporting due to social desirability bias. In our Ethiopian context, this means that formal institutions prohibit FGC and, therefore, place pressure on the respondents not to disclose their support based on the informal 'cultural' institutions they respect (Oliver, 1991).

This paper's goal is to determine the true perceptions about FGC by identifying if and which respondents misreport their perceptions. The analysis is based on new data collected in the region of Afar in Ethiopia where an NGO intervention is implemented. The NGO program aims at strengthening the health system and sexual and reproductive health knowledge. Its primary objective is to change the behaviour of households through information dissemination and behavioural change campaigns. FGC is formally banned in Ethiopia but still occurs, which makes the topic a very sensitive one.

The contributions of this paper are threefold. First, it focuses on a new list ex-

periment designed to measure attitudes regarding FGC in one of the areas where its prevalence is among the highest. Second, the most recent regression techniques developed to analyse the list experiment and the social desirability bias are used. This allows determining the existence and magnitude of systematic reporting measurement error of the true outcome. Third, the list experiment is employed to study if respondents targeted by a NGO intervention are more or less likely to misreport their attitudes. Since the intervention is not random, we run our regressions after a propensity score matching analysis.

The primary results are the following. Firstly, the list experiment shows that, on average, approximately 42% (SE=0.049) of the respondents support FGC. Secondly, the fraction of educated women supporting it is 0.12 (SE=0.134) compared to that of uneducated women which is 0.48 (SE=0.061). Thirdly, the social desirability bias is significant. If asked directly, about 32% (SE=0.017) of the women agree upon the fact that a girl should be circumcised. Hence, the difference is ten percentage points (SE=0.051) if compared with the outcome of the list experiment. Fourthly, we find suggestive evidence that the social desirability bias is the greatest among uneducated women; they underreport their true beliefs by 15 percentage points (SE=0.064). Fifthly, we find suggestive evidence that women targeted by the NGO intervention have a stronger incentive to be dishonest about their FGC support, and underreport by 12 percentage points (SE=0.068). Once we adjust the p-values for simple multiple testing procedures, the fourth and fifth results are no longer statistically significant. However, we stress the importance of further research on the potentially high social desirability bias with larger samples

to avoid type II errors.

In general, we conclude that the results confirm the relevance of potential bias in responses to direct sensitive questions. This is important to keep in mind when the outcome of interest is sensitive in order to avoid inaccurate conclusions.

The paper is structured as follows. In Section 2, we present the new data collected in Afar, Ethiopia. In Section 3, we describe the list experiment technique, the design of our list experiment about FGC, and how to measure the social desirability bias. Section 4 describes the list experiment and the social desirability bias results. In Section 5, we present robustness checks and discuss the limitations of our list experiment. Finally, Section 6 concludes. An Online Appendix is reported at the end of the paper.

2 Data

In this paper, we focus on the Afar region, one of the most remote and poorest regions in Ethiopia.[7] According to the 2011 Ethiopian DHS, in the Afar region, 57 percent of the population is in the lowest wealth quintile; 75 percent of women have no education and only 19 percent are likely to be currently employed; the use of any modern contraceptive methods is the lowest in the country (nine percent); the percentage of births delivered in a health facility is less than ten percent; full vaccination coverage among children age 12-23 months is nine percent;

[7]The results of our study might depend on the specific context raising the question of external validity. Since this is the first list experiment on FGC, we believe it could and should be replicated in different settings to be validated.

40 percent of the children are underweight; and the under-five child mortality is 127/1,000 (Central Statistical Agency Ethiopia and ICF International, 2012). According to the Afar Regional Health Bureau, in 2002, Afar counted four hospitals, 28 health centres, and 251 health posts serving the entire population estimated to be 1,494,199 (`http://www.moh.gov.et`). The last DHS estimates that the FGC prevalence is about 74.3% in Ethiopia, and 91.6% in the Afar region (Central Statistical Agency Ethiopia and ORC Macro, 2006). We note that these high prevalence rates of FGC are not unique for Ethiopia as similar figures are observed for countries such as Burkina Faso, Gambia, Guinea, Mali, and Sierra Leone (Bellemare et al., 2015). These estimates, however, are based on self-reported FGC status and, therefore, subject to bias.

In 2004, the Ethiopian Government introduced the Criminal Code Proclamation No. 414/2004 that criminalises harmful traditional practices among which FGC is included. The Proclamation became law in 2005.[8] In December 2012, the United Nations General Assembly unanimously passed Resolution 67/146, condemning FGC and related harmful practices and urging Member States to take measures to accelerate its elimination.[9] The fact that FGC is formally banned in Ethiopia and people are aware of the institutional attitude towards FGC, creates an elevated risk of underreporting it (Camilotti, 2016).

Since 2011, an NGO program has been working in some areas of Afar to

[8]The sanctions include imprisonment that ranges from three months to three years and a fine of no less than Birr 500 to 10,000 or both imprisonment and a fine. In the event of infibulation, the penalty is higher with a prison term of three to ten years (Ras-Work, 2009).

[9]`http://www.un.org/sg/statements/index.asp?nid=6529`

pro- vide comprehensive sexuality education programs and health services. The project strives to improve the sexual and reproductive health situation of Afar by increasing access to and enhancing utilisation of health services at a community level. A substantial number of activities were performed to realise the targeted goals: e.g., training and support for health workers and health promotors within the communities, renovation and equipment of a number of health facilities, and strengthening of comprehensive sexuality education at school.

In October 2012, we aggregated data in the region. Since the NGO intervention was not random, we paid particular attention to sampling subjects who were the most possibly similar in terms of their observable characteristics. The Online Appendix reports details about the NGO intervention, the data collection, and the comparability across people in the targeted and non-targeted areas (balance tests). For our survey, we used a multi-stage stratified sampling method in which strata were defined by zones representing different target groups and villages. In particular, we selected some of the NGO beneficiaries from areas where the intervention was implemented (Zones 3 and 5 of Afar) and several non-beneficiaries without access to any of the NGO activities from a different area (Zone 1 of Afar). Since the NGO program mainly targets young females and women of reproductive age, our survey consists of a total of 848 women aged between 15 and 49.

The information addressed in the questionnaire focuses on access to, knowledge about, attitudes towards, and use of sexual and reproductive health services. Next to this, we collected information about the socio-economic background of

the respondent, household water supply, and sanitation.[10]

Table 1 reports the descriptive statistics of the main variables. The survey focused on individuals that were exposed to the NGO's program (67%). Most of the respondents are Muslim (95%) and of Afar ethnicity (78%). Very few of the respondents have ever participated in any sexual and reproductive health education or training programs in the previous two years (24%) while the average number of health service providers available in the area (e.g., traditional health services, community health promoters, health extension worker, health centre) is 2.5 (maximum 4), and the average number of health services (e.g., pregnancy test, counselling on pregnancy/child care/contraceptives, medical treatment, condoms, contraceptives) that are easily accessible is 2.6 (maximum 5). Approximately 72% of the respondents are mothers and 77% are or have been married (this includes widows and divorced women). The level of education is very low with 62% of the sample being illiterate, 5% with adult education, 11% with a few years of elementary school, and 21% with higher levels of education (elementary (13%), secondary (5%), or tertiary education (3%)).

The respondent's characteristics that are considered in our empirical analysis include: women's age, marital status (dummy equal to one if ever being married), ethnicity (dummy equal to one if the woman belongs to one of the ethnic minority group), education level (dummy equal to one if the woman has at least completed elementary education), religion (dummy equal to one if the woman is Christian),

[10]The questionnaire does not contain a direct question about the FGC status, because it was considered too delicate.

motherhood (dummy equal to one if the woman has ever had any children), and the NGO targeted status (dummy equal to one if targeted by the NGO intervention).

3 Methodology

Standard list experiment design

In order to measure the true perception about FGC, we added a list experiment to the survey. The list experiment technique works by aggregating the sensitive item with a list of others that are non-sensitive (Miller, 1984). The survey sample is composed by N respondents, that are randomly divided into two groups: treatment and control. $T_i = 1$ ($T_i = 0$) implies that the respondent i belongs to the treatment (control) group. The control group respondents receive a list of J non-sensitive, yes/no items, and they must indicate to the interviewer how many of the listed items they agree on, but not which items. The treatment group respondents receive the same list of non-sensitive, yes/no items plus a sensitive, yes/no item ($J+1$ in total). The sensitive item measures the sensitive topic. As for the control group respondents, the treatment group respondents must tell the interviewer the number of items they agree on.

To formalise, we use the same notation as in Imai (2011). Let us define Z_{ij}^* as the respondent i's truthful preference to the jth item where $j = 1, \ldots, J+1$. Let us suppose that each respondent possesses a latent potential response to each non-sensitive item $j = 1, \ldots, J$ which may depend on the individual's treatment status T. Then $Z_{ij}(T)$ is equal to 1 if the answer is positive or 0 otherwise. $Z_{ij}(1) = 1$ in-

dicates, for example, that the respondent i's latent answer to the jth non-sensitive item is positive under the treatment condition. We also have that $Z_{i,J+1}(1)$ is the respondent i's latent answer to the sensitive item if in the treatment group. Now the potential responses respondent i would give under the control or treatment group are, respectively: $Y_i(0) = \sum_{j=1}^{J} Z_{ij}(0)$ or $Y_i(1) = \sum_{j=1}^{J+1} Z_{ij}(1)$. Finally, $Y_i = Y_i(T_i)$ represents the observed response, and X_i the vector of observed covariates for respondent i.

Crucial to the list experiment design is the *randomisation of the treatment* meaning that the sample is randomly divided into control and treatment groups and for any respondent $i = 1, \ldots, N$, the following needs to be valid: $\{\{Z_{ij}(0), Z_{ij}(1)\}_{j=1}^{J}, Z_{i,J+1}(1)\} \perp T_i$. This design then relies on two important assumptions (Imai, 2011; Blair and Imai, 2012). The first assumption referred to as *no design effect* implies that the addition of the sensitive item does not change the sum of affirmative answers to the non-sensitive items, hence, for each $i = 1, \ldots, N$, we have $\sum_{j=1}^{J} Z_{ij}(0) = \sum_{j=1}^{J} Z_{ij}(1)$. The second assumption is called *no liars* (also termed as ceiling and floor effects), and it implies that the respondents truthfully reply to the sensitive item, for each $i = 1, \ldots, N$, we have $Z_{i,J+1}(1) = Z^*_{i,J+1}$. Ceiling effects occur when a respondent in the treatment group would honestly respond "yes" to all nonsensitive items, losing the protection to honestly report the individual's response to the sensitive item. Floor effects instead occur when the respondent in the treatment group, whose truthful answer is affirmative only for the sensitive item, replies negatively to all of the items in order to conceal the respondent's identity. It has been shown that the presence of ceiling or floor ef-

fects leads to underestimation of the true support for the sensitive item (Blair and Imai, 2012). The assumption about the answers to the non-sensitive items is only that they are not influenced by the addition of the sensitive item, therefore, they do not necessarily need to be truthful. In the Robustness Section, we test if the randomisation is done well and for potential violations of the no design as well as no liars assumptions.

If the randomisation is well done and these two assumptions hold, the unbiased estimate of the population proportion of those that agree on the sensitive item can be subsequently computed using a difference-in-means estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - T_i) Y_i,$$

where $N_1 = \sum_{i=1}^{N} T_i$ is the size of the treatment group and $N_0 = N - N_1$ is the size of the control group. The joint distribution of $(Y_i(0), Z^*_{i,J+1})$ can be identified.

Imai (2011) proposes new multivariate regression estimators, that also rely on the assumptions of no design effect and no liars, to analyse the relationship between preferences over the sensitive item and the respondent's characteristics. One of the estimators reduces to a linear regression with interaction terms[11] (see also Holbrook and Krosnick, 2010):

$$Y_i = X_i^T \gamma + T_i X_i^T \delta + \epsilon_i, \tag{1}$$

[11]This is like modeling heterogeneity in the treatment effects.

where $E(\epsilon_i|X_i, T_i) = 0$, and (γ, δ) are unknown parameters.[12] X_i includes an intercept. Being that treatment T is randomly assigned, we can estimate (γ, δ) using ordinary least squares while we compute heteroskedasticity-consistent standard errors to account for the difference in the variance of error term between the treatment and control groups. The parameters of interests are included in the vector δ, and they indicate which respondent's characteristics (X's) explain the variation in answering affirmatively the sensitive item. It is important to note that δ needs to be interpreted as associations.

In this paper, we analyse the list experiment using the difference-in-means estimator to estimate the overall proportion of respondents that agree on the sensitive item. We then apply the linear regression estimator to study the different preferences over the sensitive item and the main respondent's characteristics.[13] This technique is easy to interpret but rarely used in the empirical research of the list experiment. Moreover, to the best of our knowledge, it is the first time that this approach is being used in the context of reproductive health, hence, we believe this is a further contribution.

[12]The estimator in this case is a nonlinear least squares estimator: $Y_i = f(X_i, \gamma) + T_i g(X_i, \delta) + \epsilon_i$, where $f(x, \gamma)$ and $g(x, \delta)$ represent the regression models for the conditional expectations of the control and sensitive items given the covariates. If X_i contains only an intercept, the difference-in-means estimator is obtained. If linearity is assumed for the two sub-models $f(x, \gamma) = x^T\gamma$ and $g(x, \delta) = x^T\delta$, then the estimator reduces to a linear regression with interaction terms. For further details about the different estimators, see Imai (2011).

[13]This estimator more efficiently estimates the relationships between the sensitive item and respondents' characteristics compared to a subgroup analysis.

Our list experiment

In our survey, the control group was presented with the following question:

I want you to give me a secret answer for the following statements. I will give you 3(4) stones and you have to hold them in your right hand. Keep your hands (both) on your back side. If you agree on the statement I will soon be reading to you, you transfer one stone to your left hand behind you (I will not see it, you also should not tell me) but, if you don't agree, do not transfer any stone. At the end, I would like to know the total number of statements you agreed on. Now, I will read the statements:

1. *HIV can be transmitted through witchcraft or other supernatural means.*
2. *It is acceptable to use contraceptives to avoid pregnancy.*
3. *In a marriage, both partners should decide on how many children they should have.*

For the treatment group, we asked an identical question, but with an extra item, a sensitive item, concerning FGC:

4. *A girl should be circumcised.*[14]

As non-sensitive items, we selected items related to sexual and reproductive health knowledge as well as family planning issues. Although items 1-3 seem controversial in the local setting, they are relevant in our FGC experiment as non-sensitive items considering the illegality of the sensitive one. There is no need to

[14]In our questionnaire, we use the term female circumcision when asking about FGC. In Afar, the word 'selot' is used. It can indicate both female circumcision or female genital mutilation. This term 'selot' has been used by our enumerators. In the paper, we have opted for the general term of "female genital cutting".

assume that the answers to the non-sensitive items are truthful, however, they need not be influenced by the presence of the sensitive item in the list. Given that the Ethiopian law prohibits FGC, people are expected to be less prone to reveal their true belief about the fourth item that is indeed considered as a sensitive issue. The choice of the non-sensitive items needs to be such that the ceiling and floor effects are avoided (Kuklinski et al., 1997).[15] We then formulated the three non-sensitive statements in a way that three yes or no outcomes would be an exception.[16]

Social desirability bias

To assess the impact of sensitivity on responses, we compare the attitudes toward FGC that are measured when the question is asked directly and when it is asked indirectly via the list experiment. Two assumptions are made. The first is that the real support towards FGC is measured with the list experiment.The second is that the measurement error in the direct survey and list experiment data follow the same direction. The difference between the indirect and the direct question, therefore, is a measure of how much the true support for female circumcision is underreported.

As in Blair and Imai (2012), we define $Z_{i,J+1}(0)$ as the respondent i's potential

[15]The common advise is that the list of non-sensitive items should not be too short to avoid the ceiling and floor effects (Kuklinski et al., 1997), and many empirical examples use a 3-item or 4-item list (Kuklinski et al., 1997; McKenzie and Siegel, 2013; Coffman et al., 2013). The sensitive item is often the last one, however, the order of the items can be randomised to avoid ordering effects.

[16]For example, respondents that agree with the more modern statements 2 and 3 are expected to disagree with the more traditional statement 1. Similarly, more 'traditional' households are expected to agree with statement 1 and disapprove statements 2 and 3.

answer to the sensitive item when asked directly. Since the social desirability bias can also vary across respondents as a function of their characteristics, it is defined as:

$$S(x) = Pr(Z^*_{i,J+1} = 1|X_i = x) - Pr(Z_{i,J+1}(0) = 1|X_i = x), \quad \text{for any} \quad x \in \chi.$$

The first term can be estimated using the linear regression estimator, Equation 1. The second term can be estimated by regressing (using, for example, a logistic regression or a linear probability model) the observed value of $Z_{i,J+1}(0)$ on X_i.

In the survey, we asked: "Do you agree on the following statement? A girl should be circumcised." In our survey, the direct question is asked both to the control and treatment groups (leading to $Z_{i,J+1}$), hence, we consider the answer to the direct question in the entire sample to compare it with the list experiment results. The list experiment question was a component of a larger survey. The questions related to FGC were included in one of the sections of the questionnaire. First, the interviewees were asked to indicate to what degree they agreed with ten statements. Several issues were raised regarding sexual and reproductive health. "A girl should be circumcised" was the fifth and only statement regarding circumcision. Subsequently, the list experiment was run. We have no reason to believe that the respondents were cognisant of this design and that the format influenced the list experiment results as so many different issues were dealt with during the interview. See also the Robustness section for further discussion.

The possible answers to the direct question were totally agree (200 answers),

somehow agree (52), neither agree nor disagree (50), somehow disagree (35), and totally disagree (511). In order to make this correspond with the yes/no scale used for each item in the list experiment, we dichotomised the survey question as follows: totally agree and somehow agree correspond to 1 (yes) and somehow disagree and totally disagree to 0 (no). We do not consider the respondents who answered 'neither agree nor disagree' in our analysis because they could not be straightforwardly classified in the yes or no category.[17]

4 Results

Results of the list experiment

The NGO intervention is not random, therefore, we run all of our analyses after a propensity score matching (PSM) analysis. The idea behind PSM is to build a statistical comparison group that is based on a model of the probability of participating in the program, using observable characteristics. The main drawback of PSM is that it relies on the conditional independence assumption meaning that individuals' unobservable characteristics do not affect participation. To participate, a woman must be living in the geographical area targeted by the NGO. Following the PSM analysis, we discard 21 observations that present a poor match (they are

[17]Separate analyses were performed by assigning zero to also the neutral category, and results were similar.

outside the common support).[18][19]

Table 2 reports the observed data from the list experiment. As we previously stated, the list experiment has three non-sensitive items and one sensitive item. The treatment group consists of 399 respondents, the control 371. A total of seven respondents did not answer the list experiment question, three in the control group and four in the treatment group. We observe that the responses are well distributed, and there are only a few responses in the extreme cases (0 and 3 for the control group). Having many responses in the extreme cases can indicate the presence of, respectively, floor and/or ceiling effects. We will discuss the no liar assumption in the Robustness section.

The results from the list experiment using the difference in-means estimator indicate that 41.6% (SE=0.049) is the estimated prevalence of women who agree

[18]To apply PSM, we proceed in the following way: 1) We estimate a logit model of program participation; 2) We use predicted values from estimation to generate the propensity score for all NGO targeted and NGO non-targeted group members; 3) We restrict samples to ensure common support; 4) We match each NGO targeted unit to a NGO non-targeted unit with a similar propensity score; 5) We run balancing tests on the set of variables that we want to balance; 6) If the balancing tests are not good enough to improve it, we need to change the matching technique and/or the propensity score model by adding new covariates, interaction terms, or higher order covariates. Our final propensity score model is a logit where the dependent variable is a dummy equal to 1 if the woman is in the NGO targeted group and 0 otherwise, and the covariates are age, age squared, ever married, belonging to an ethnic minority group, being educated, being Christian, and being a mother. The matching technique is the radius with the caliper equal to 0.01 which corresponds to the maximum distance between non-targeted units. The balancing tests are done on the variables age, ever married, belonging to an ethnic minority group, being educated, being Christian, and being a mother. As a balancing test, we consider the absolute standardised bias (ASB) that is a measure of the average imbalance in each covariate X existing between NGO targeted and NGO non-targeted units.

[19]Similar results are obtained without the exclusion of the 71 observations (50 respondents who reply neutrally to the direct question and 21 observations who are poor matches after PSM). Also, the list experiment randomisation and the balancing tests of the NGO intervention are not compromised by the exclusion of such observations.

with the sensitive item "a girl should be circumcised".[20]

In addition to knowing the overall proportion of women that agree with FGC, it is interesting to know what type of respondent is more in favour of it. The difference-in-means estimator can be completed separately in each subgroup, which is a common occurrence (some examples are Kuklinski et al., 1997; McKenzie and Siegel, 2013), however, this leads to a small number of respondents at the subgroup level and to an increase in the standard errors.

We instead apply the linear regression model developed to analyse the list experiment (Equation 1). Table 3 presents the model results. The interesting estimated coefficients are reported in the top of Table 3 (Sensitive item), and they correspond to $\hat{\delta}$ (Equation 1). The results show that the coefficient for the education variable in the model for the sensitive item (treatment status=1) is negative, and it is statistically significantly different from zero with a p-value below 1%. This implies that, on average, educated women are 36.4 percentage points (SE=0.158) less likely to be in favour of circumcision even after controlling for other individuals' characteristics.

We present a comparison of the difference-in-means and linear model results considering education as the main variable in Figure 1. This Figure is based on the fitted model presented in Table 3 and on the model without covariates (diff-in-means). Figure 1 also presents the estimated proportions of uneducated (circle) and educated people (triangle) who agree that "a girl should be circumcised". The

[20]The mean in the control group is 1.898 (SE=0.033), while the mean in the treatment group is 2.313 (SE=0.035).

difference between those proportions is also shown (diamond). To obtain the estimated proportion for each subgroup in the models with covariates, we computed the predicted probability by setting all of the other covariates to their observed values.[21] The solid lines correspond to the 95% asymptotic confidence intervals. The model without covariates (diff-in-means) shows a difference of 14 percentage points between uneducated and educated women which is not statistically significant, while this difference increases to 36 percentage points and is statistically significant when the linear regression model is used. In particular, 48% (SE=0.061) of the uneducated women agree with FGC while 12% (SE=0.134) is the probability of educated women in the multivariate linear model.

Results on the social desirability bias

In this section, we compare the attitudes regarding FGC measured via the list experiment and the direct question. This also cannot be interpreted in a causal way.

If asked directly, approximately 32% (SE=0.017) of the women agree that a girl should be circumcised. The proportion obtained using the difference-in-means estimator is 42% (SE=0.049), hence, the difference is ten percentage points (SE=0.051) which is statistically different from zero at a 10% level (see the no covariates results in Table 5).

[21] In the case of the list experiment, by keeping a particular X constant and all of the other covariates to their observed values, we estimate the difference in the Y predictions between the control and the treatment group.

Since also the answer to the direct sensitive question might vary as a function of respondent's characteristics, we apply a linear probability model to analyse the responses to the direct question. Table 4 reports the results of the regression where the dependent variable is a dummy variable where 1 corresponds to agreeing that a girl should be circumcised and 0 indicates the opposite.[22] In particular, Table 4 demonstrates that, holding all other variables fixed, being one year older is associated with an increase in the probability of being in favour of FGC by 0.6 percentage points. The probability that members of the other ethnic minority groups are in favour of the sensitive question is instead associated with 15 percentage points (SE=0.042) less than for Afar people. Moreover, being targeted by the NGO is associated with a decrease in the probability of being in favour of FGC by 14 percentage points (SE=0.037). In contrast to the results presented for the list experiment in Table 3, education does not appear to have an effect if a direct question is asked.

This raises the question of whether the potential social desirability bias plays a role in explaining the differences in the NGO, education, or other minorities effect found between Table 3 and Table 4. Unfortunately, we cannot test if those differences are statistically significant because the list experiment only generates aggregate information.[23] Instead, we can study the social desirability bias for the

[22]Note that, in Table 5, we consider the same sample used for the list experiment analysis in Table 3, keeping only the observations for which the list experiment question is not missing.

[23]One possibility to measure the difference between a direct question and list experiment at an individual level has been proposed by Coffman et al. (2013). The idea is to directly ask each respondent all of the non-sensitive questions and then compare the total sum of the answers to the direct questions (non-sensitive and sensitive) with the list experiment answer. Under truthful

different sub-groups.

Table 5 shows the differences in estimated proportions of respondents answering the sensitive question if the direct or indirect question is used. In particular, we use the linear model to predict answers to the list experiment and the linear probability model to predict answers to the direct question. Table 5 also includes the results for the models with and without covariates (age, ethnic group, marital status, education and being targeted by the NGO). The differences between the indirect and direct questions are positive and statistically significant at the 10% level in the no covariates and covariates models.

We then calculate the social desirability bias by examining the estimated proportions for different groups by controlling for all of the other covariates. The difference is highly statistically significant and positive for the uneducated group and significant at the 10% level for the NGO targeted group (see the column of the unadjusted p-value). Therefore, it seems that the group that underreports the most is the group of uneducated people where the direct question results in 33% (SE=0.020) of the women in favour of circumcision compared to 48% (SE=0.061) obtained through the list experiment (difference=15 percentage points; p-value=0.024). Interestingly, when the direct question is considered, women targeted by the NGO intervention underreport their support towards FGC by 12 percentage points (p-value=0.064). To account for the multiplicity of the comparisons, we adjusted the p-values for simple multiple testing using Bonferroni (1935). The last column (9) of Table 5 reports the respective p-values. The statistical significance of the edu-

reporting, the expected number should be the same.

cated and NGO targeted groups disappear, and this is most likely due to the lack of power of our sample. These two results, therefore, are suggestive.

5 Robustness

Test the key elements for a good list experiment

In this subsection, we test if the randomisation of the list experiment is done well and for potential violations of the two key assumptions of the list experiments.

The first thing to verify is the *randomisation of the treatment*. Table 6 provides sample means for the main variables in the treatment group and the control group. Comparing the means allows us to see that the randomisation of the list experiment (control group and treatment group) was successful given that all important respondents' characteristics do not significantly differ between the two groups.

The first assumption is called *design effects* which occurs when the inclusion of a sensitive item affects some respondents' answers to non-sensitive items. Blair and Imai (2012) developed a statistical test to detect violation of the design effects assumption. The null hypothesis of the test indicates no design effects as indicated by the following:

$$H_0 = \begin{cases} Pr(Y_i \leq y | T_i == 0) \geq Pr(Y_i \leq y | T_i == 1) \text{ for all } y = 0, \ldots, J-1 \text{ and} \\ Pr(Y_i \leq y | T_i == 1) \geq Pr(Y_i \leq y - 1 | T_i == 0) \text{ for all } y = 0, \ldots, J \end{cases} \quad (2)$$

Equivalently, H_0 is $\pi_{yt} \geq 0$ for all y and T, where π_{yt} are the proportions of

respondent types. The population proportion of each respondent type is defined as $\pi_{yz} = Pr(Y_i(0) = y, Z^*_{i,J+1} = z)$ for $y = 0, \ldots, J$ and $z = 0, 1$ (z=0 indicates that the respondent answers no to the sensitive item; z=1 indicates the opposite). The π_{yz} is identified for all $y = 0, \ldots, J$ as:

$$\pi_{y1} = Pr(Y_i \leq y | T_i = 0) - Pr(Y_i \leq y | T_i = 1),$$

$$\pi_{y0} = Pr(Y_i \leq y | T_i = 1) - Pr(Y_i \leq y - 1 | T_i = 0).$$

If all of the proportions are positive, then H_0 cannot be rejected; if all proportions are negative, then H_0 is rejected; while, if at least one is negative, it is important to understand if it is negative by chance.[24] Table 7 reports the estimated proportion of each respondent type. They are all positive, hence, the assumption of no design effects is valid. This is an indication that the inclusion of the sensitive item did not change the responses to the non-sensitive items.

The second possible problem is the violation of the assumption *no liars*. This assumption is violated in the event of ceiling or floor effects. Ceiling effects occur when some respondents in the treatment group give the answer $Y_i = J$ even if the truthful answer would be $Y_i = J + 1$, affirmative for both sensitive and control items. The problem of $Y_i = J + 1$ as an answer is that it would reveal

[24] If this is the case, one can formally apply the Blair and Imai (2012) test. The testing procedure consists of first conducting a separate hypothesis test for each of the two stochastic dominance relationships in equation (2) and later using the Bonferroni correction to combine the results. Two p-values based on the two different statistical tests of stochastic dominance relationships are computed, and the null is rejected if and only if the minimum of these two p-values is less than $\alpha/2$ where α is the desired size of the test. For further details about the test, see Blair and Imai (2012, pages 64-65).

the respondent's support for the sensitive issue. Due to the design of our list experiment, floor effects are expected to play a minor role. $Y_i = 0$ simply reveals that an individual does not agree with the sensitive item, and $Y_i = 1$ does not reveal whether the 1 (yes) concerns the sensitive or one of the non-sensitive items. We cannot test this assumption but, as we can see from Table 2, the responses are well distributed, and there are approximately 16% of responses in the extreme cases (0 and 3 for the control group). About 15% of respondents in the control group say yes to all three non sensitive items. If this is a ceiling effect, than our estimate of the proportion of women in favour with female circumcision would be an underestimation of the real proportion.[25]

List experiment by education level and by NGO targeted status

The respondent's education level is revealed to be a critical characteristic. Can it be that educated women understand the mechanism behind the list experiment and manipulate their answers? Even if it is not a formal test, we can analyse the results of the list experiment by education level. Figure 2 reports the distribution

[25]For this group of respondents, a ceiling effect may bias the results for those who are aware of the risk that a yes to all four items may disclose their real attitude concerning the sensitive issue. In accordance with the results of our list experiment, we may assume that approximately 40% of the respondents is in favour of FGC. This implies that a ceiling effect may bias the answers of 40% of 15% = 6% of the respondents. In Table 2, we also observe that 3% of the respondents are not aware of the ceiling effect, or have no problem with showing their support (a yes for all four items). We have run a regression of an indicator variable equal to 1 if the respondent in the control group reports "3" on individual characteristics (age, marital status, ethnicity, motherhood, religion, and NGT targeted status). We find no statistically significant effect of any of those characteristics on the outcome. This further suggests that ceiling effects are not a major concern in our list experiment.

of the items for the uneducated and educated groups. Responses in the two groups are well distributed with only a small number of cases in the extremes. This is an indication that ceiling and floor effects should not be a problem in our list experiment. We can test the presence of design effects by applying the Blair and Imai (2012) design statistical test to the subsample of educated and uneducated women, and we fail to reject the null hypothesis of no design effects.[26] We do the same analysis to determine if NGO targeted women understand the mechanism behind the list experiment (Figure 2). Similar conclusions to those above can be drawn.[27]

Limitations to our list experiment

Even if our analysis suggests that the list experiment design was conducted properly, we believe there are a number of limitations to take into consideration.

Woman circumcision status We do not know if the female respondents have been circumcised themselves. If experiencing this procedure is highly correlated with the support towards the practice, then this might be driving the results. The latest DHS FGC prevalence was estimated to be 91.6% in 2005 for the Afar region (Central Statistical Agency Ethiopia and ORC Macro, 2006). If most of the

[26] All of the population proportions for the uneducated group are positive, therefore, the null hypothesis of no design effects cannot be rejected. In the educated group, instead, there is one proportion that is negative. We then apply the test of Blair and Imai (2012) to verify that this is due to chance and not to a design effect. The test statistic for the educated group is 0.158/2= 0.079 $> \alpha$/2=0.05/2, then we fail to reject the null of no design effects. (See Footnote 23 and Blair and Imai, 2012, pages 64-65).

[27] All of the population proportions for the targeted and not-targeted groups are positive, therefore, the null hypothesis of no design effects cannot be rejected.

women are still now circumcised, than the FGC status is not going to be very informative. However, the prevalence estimates are based on self-reported FGC status and, therefore, subject to bias. A medical examination would be necessary to confirm the actual status. However, this would seriously impair the willingness to participate in the survey.

External validity This study is the first to apply a list experiment to measure FGC attitudes. This certainly calls for further research on this promising method. It is of particular importance to address the external validity. Do the results depend on how the list experiment was phrased (e.g., type of non-sensitive items, total number of items, order of the items) and the context in which the data were collected? In our study, we did not add other survey methods that can confirm our results. Endorsement experiments, randomised responses, or qualitative techniques could have been added to validate the list experiment.[28] We have decided not to do so in order to avoid over complicating and lengthening the questionnaire. Regarding the content of the non-sensitive items, there is still no clear consensus on how to design them. The different items used in the list experiment were discussed with the field officers and adjusted where needed.

The results of the list experiment depend on the specific context. Is it possible to replicate this list experiment in other countries or surveys? We believe that list experiments on FGC could and should be replicated in different contexts. Interviewers can be easily trained to use them, and a pilot survey can be done to test the

[28]Blair et al. (2014), for example, compare the results of list experiments with those of endorsement experiments to validate the list experiment in a study conducted in Afghanistan.

feasibility of the list experiment. The cost of adding list experiments to the surveys is very low as they are simply extra questions that require the randomisation of the questionnaire.

6 Conclusions

Measuring attitudes towards FGC is critical for understanding the support for these traditional customs but are also difficult because it is a sensitive topic. This paper uses new data collected in Ethiopia, one of the countries with the highest FGC prevalence and where FGC is formally prohibited but is still a widespread advocated custom in the local culture. A list experiment is designed to elicit truthful answers about FGC support. The results of the list experiment are compared to the results of a direct question in order to study underreporting due to social desirability bias or the systematic reporting measurement error of the true FGC support. The goal of the paper is to understand if and who the people are that misreport their true beliefs about FGC support and what is the magnitude of the underreporting. Overall, the results indicate that underreporting can be substantial, in particular for the uneducated respondents. As extant literature on FGC uses direct questioning about support, we consider this result an important contribution to the debate. Moreover, uneducated women are then most likely to support FGC which suggests that NGOs should target them more.

Our results indicate that, when asking a direct question about FGC, 31.8% (SE=0.017) of the women are in favour of the practice. If, instead, we take into

consideration the question's sensitivity by asking it indirectly, we find that the overall proportion of women in favour of FGC is much higher, 41.6% (SE=0.049), leading to a difference of 9.8 percentage points (p-value=0.057).

A regression analysis based on a new statistical approach developed to analyse list experiments (Imai, 2011) shows that some respondents' characteristics seem relevant in explaining the consent for FGC. In particular, the women's education turned out to be the most critical variable in explaining differences in attitudes. Firstly, the list experiment shows that educated women are less in favour of FGC (-36.4 percentage points, p-value=0.021) compared to the uneducated women. Secondly, when the results of the list experiment are compared with the results obtained with the direct question to test the social desirability bias, we find that uneducated respondents underreport their attitudes by 14.5 percentage points (unadjusted p-value=0.024). Once we adjust the p-values for multiple hypothesis testing, this result is no longer statistically significant. However, we contend that our results provide suggestive evidence for the claim that uneducated women appear to be less willing to publicly share their genuine attitudes concerning FGC support. The educational level may affect incentives for undergoing the procedure. If it increases the opportunities to attract a better husband (Chesnokova and Vaithianathan, 2010), we may argue that uneducated women have more to lose if they do not favour the practice, while educated women have better chances in the job market and depend less on marriage (Ouedraogo and Koissy-Kpein, 2012; Molitor, 2014). We believe that studies that investigate the causal impact of education on FGC behaviour and attitudes could be crucial in this context (De Cao

and La Mattina, 2017).

Another interesting result, which is also to be treated with caution (see multiple hypothesis testing in Table 5), concerns a potential NGO effect. By comparing the estimates obtained with the direct and indirect questioning, we find suggestive evidence that women targeted by the NGO intervention more strongly underreport their support towards FGC. We cannot claim that this is due to the NGO intervention or alternatively that the NGO intervention is not working in changing people's attitudes because it was not random, and there might still be selection bias that we could not account for.

Lack of empirical evidence on the support towards FGC and, most importantly, lack of understanding of how biased direct questions can be make our study a further step to a future line of research that aims at focusing more on how to measure sensitive outcomes, and we believe this is especially important in the context of policy impact evaluations (e.g., Blattman et al., 2016).

References

Bellemare, M.F., Novak, L., Steinmetz, T.L., 2015. All in the family: Explaining the persistence of female genital cutting in West Africa. Journal of Development Economics 116, 252–265.

Blair, G., Imai, K., 2012. Statistical analysis of list experiments. Political Analysis 20, 47–77.

Blair, G., Imai, K., Lyall, J., 2014. Comparing and combining list and endorsement experiments: Evidence from Afghanistan. American Journal of Political Science 58, 1043–1063.

Blattman, C., Jamison, J.C., Koroknay-Palicz, T., Rodrigues, K., Sheridan, M., 2016. Measuring the measurement error: A method to qualitatively validate survey data. Journal of Development Economics 120, 99–112.

Bonferroni, C.E., 1935. Il calcolo delle assicurazioni su gruppi di teste. Tipografia del Senato , 13–60.

Bullock, W., Imai, K., Shapiro, J.N., 2011. Statistical analysis of endorsement experiments: Measuring support for militant groups in Pakistan. Political Analysis 19, 363–384.

Camilotti, G., 2016. Interventions to stop female genital cutting and the evolution to the custom: Evidence on age at cutting in Senegal. Journal of African Economies 25, 133–158.

Central Statistical Agency Ethiopia and ICF International, 2012. Ethiopia Demographic and Health Survey 2011. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International.

Central Statistical Agency Ethiopia and ORC Macro, 2006. Ethiopia Demographic and Health Survey 2005. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ORC Macro.

Chesnokova, A.C., Vaithianathan, R., 2010. The economics of female genital cutting. The B. E. Journal of Economic Analysis & Policy 10, 64.

Chong, A., Gonzales-Navarro, M., Karlan, D., Valdivia, M., 2013. Effectiveness and spillovers of online sex education: Evidence from a randomized evaluation in Colombian public schools. NBER Working Paper No. 18776 .

Coffman, K.B., Coffman, L.C., Ericson, K.M.M., 2013. The size of the LGBT population and the magnitude of anti-gay sentiment are substantially underestimated. NBER Working Paper No. 19508 .

Corstange, D., 2009. Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. Political Analysis 17, 45–63.

Coyne, C.J., Coyne, R.L., 2014. The identity economics of female genital mutilation. The Journal of Developing Areas 48, 137–152.

De Cao, E., Huis, M., Jemaneh, S., Lensink, R., 2017. Community conversations as a strategy to change harmful traditional practices against women. Applied Economics Letters 24, 72–74.

De Cao, E., La Mattina, G., 2017. The impact of education on female genital cutting in Nigeria. Mimeo .

Efferson, C., Vogt, S., Elhadi, A., Ahmed, H.E.F., Fehr, E., 2015. Female genital cutting is not a social coordination norm. Science 349, 1446–1447.

Glynn, J.R., Kayuni, N., Banda, E., Parrott, F., Floyd, S., Francis-Chizororo, M., Nkhata, M., Tanton, C., Hemmings, J., Molesworth, A., et al., 2011. Assessing the validity of sexual behaviour reports in a whole population survey in rural Malawi. PLoS One 6, e22840.

Holbrook, A.L., Krosnick, J.A., 2010. Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly 74, 37–67.

Imai, K., 2011. Multivariate regression analysis for the item count technique. Journal of the American Statistical Association 106, 407–416.

Jamison, J., Karlan, D., Raffler, P., 2013. Mixed method evaluation of a passive mHealth sexual information texting service in Uganda. Information Technologies & International Development 9, 1–28.

Karlan, D.S., Zinman, J., 2012. List randomization for sensitive behavior: An application for measuring use of loan proceeds. Journal of Development Economics 98, 71–75.

Kuklinski, J.H., Cobb, M.D., Gilens, M., 1997. Racial attitudes and the "New South". The Journal of Politics 59, 323–349.

LaBrie, J.W., Earleywine, M., 2000. Sexual risk behaviors and alcohol: higher base rates revealed using the unmatched-count technique. Journal of Sex Research 37, 321–326.

Mackie, G., LeJeune, J., 2009. Social dynamics of abandonment of harmful practices: A new look at the theory. Special series on social norms and harmful practices. Technical Report. Innocenti Working Paper.

McKenzie, D., Siegel, M., 2013. Eliciting illegal migration rates through list randomization. Policy research working paper 6426, World Bank .

Miller, J.D., 1984. A new survey technique for studying deviant behavior. Ph.D. thesis. The George Washington University.

Molitor, V., 2014. Family Economics in Developing Countries. Ph.D. thesis. Universit$\ddot{a}$t Mannheim.

Moseson, H., Massaquoi, M., Dehlendorf, C., Bawo, L., Dahn, B., Zolia, Y., Vittinghoff, E., Hiatt, R.A., Gerdts, C., 2015. Reducing under-reporting of stigmatized health events using the list experiment: Results from a randomized, population-based study of abortion in Liberia. International Journal of Epidemiology , dyv174.

Naguib, K., 2012. The effects of social interactions on female genital mutilation: Evidence from Egypt. Working Paper, Boston University .

Obermeyer, M.C., 2005. The consequences of female circumcision for health and sexuality: an update on the evidence. Culture, Health & Sexuality 7, 443–461.

Oliver, C., 1991. Strategic responses to institutional processes. Academy of management review 16, 145–179.

Ouedraogo, S., Koissy-Kpein, S.A., 2012. An economic analysis of female genital mutilation: How the marriage market affects the household decision of excision. Unpublished Manuscript .

Ras-Work, B., 2009. Legislation to address the issue of female genital mutilation (FGM). Technical Report. United Nations.

Redlawsk, D.P., Tolbert, C.J., Franko, W., 2010. Voters, emotions, and race in 2008: Obama as the first black president. Political Research Quarterly 63, 875–889.

Schroder, K.E., Carey, M.P., Vanable, P.A., 2003. Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. Annals of Behavioral Medicine 26, 104–123.

Shell-Duncan, B., Hernlund, Y., 2000. Female "circumcision" in Africa: Culture, controversy, and change. Lynne Rienner Publishers.

Shell-Duncan, B., Wandera, K., Hernlundc, Y., Moreau, Y., 2011. Dynamics of change in the practice of female genital cutting in Senegambia: Testing predictions of social convention theory. Social Science & Medicine 73, 1275–1283.

Stichting-Gezamenlijke-Evaluaties, 2015. MFS-II Evaluations - Joint Evaluations of the Dutch Co-Financing System 2011-2015 - Country Report Ethiopia. Technical Report. Partos.

Wagner, N., 2015. Female genital cutting and long-term health consequences –

Nationally representative estimates across 13 countries. Journal of Development Studies 51, 226–246.

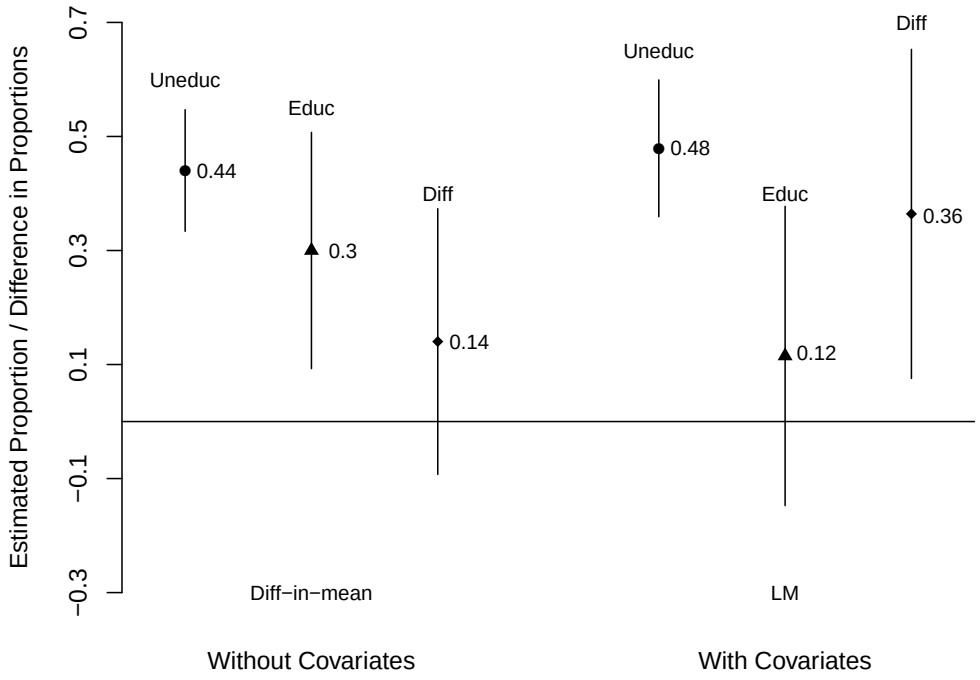
Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60, 63–69.

WHO, 2012. Female Genital Mutilation, Fact Sheet 241. Technical Report. World Health Organization.

Yoder, P.S., Khan, S., 2008. Numbers of women circumcised in Africa: The production of a total. DHS Working Paper 39.

7 Figures and tables

Figure 1: Estimated proportion of women who are in favour of FGC based on the linear regression model for the list experiment design.



Note. Predictions are based on the difference-in-means estimator when no covariates are used and on the linear regression model when covariates are considered. The solid lines correspond to a 95% confidence interval for the estimated proportions. In the linear regression model, the results are averaged over the sample distribution of covariates.

Figure 2: List experiment results by respondents' education level and by respondents' NGO targeted status.

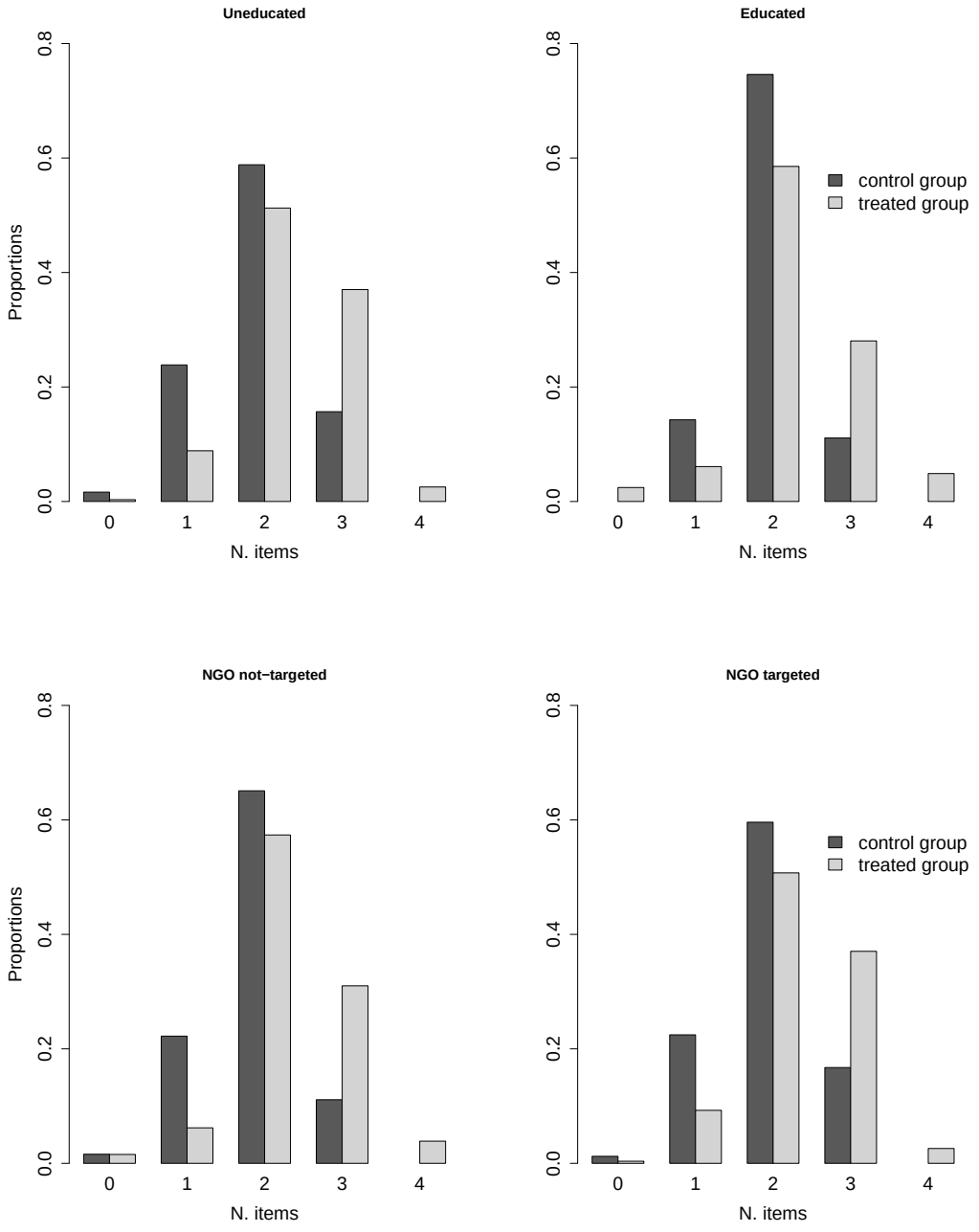


Table 1: Descriptive statistics

Variable	**N**	**Mean**	**Std. Dev.**
Age	845	28.226	9.512
Religion (1=Christian; 0=Muslim)	839	0.049	
Ethnic group (proportions)			
Afar	848	0.78	
Other ethnic minorities	848	0.22	
Areas in Afar (proportions)			
Zone 1	848	0.33	
Zone 3	848	0.34	
Zone 5	848	0.33	
Health education/training (1=yes; 0=no)	835	0.243	
Health providers available (0-4)	848	2.514	1.003
Health services accessible (0-5)	848	2.637	2.041
Having children (1=yes; 0=no)	846	0.722	
Ever being married (1=yes; 0=no)	843	0.770	
Educated† (1=yes; 0=no)	844	0.213	
Sex and HIV knowledge (0-6)*	847	4.046	1.279
NGO program target (1=yes; 0=no)	848	0.667	

Note. * This variable is the percentage of correct answers of a battery of six questions related to sexual knowledge and HIV. † includes people that have at least completed elementary school.

Table 2: Observed data from the experiment result.

	Control group		Treatment group	
Response value	**Freq.**	**Perc. (%)**	**Freq.**	**Perc. (%)**
0	5	1.35	3	0.75
1	83	22.37	33	8.27
2	228	61.46	211	52.88
3	55	14.82	140	35.09
4			12	3.01
Total	371	100	399	100

Note. The table displays the number of respondents for each value of the observed outcome variable (total number of items the respondent agrees on) and its proportions, separately for the control and the treatment group where the sensitive item is "a girl should be circumcised".

Table 3: Results of the linear regression model for the list experiment.

	Response to the list experiment question	
Variables	**Est**	**SE**
Sensitive item		
T	0.873***	(0.219)
Age×T	-0.009	(0.007)
Ever married ×T	-0.092	(0.246)
Other ethnic minorities×T	-0.063	(0.127)
Educated×T	-0.364**	(0.158)
Christian×T	0.313	(0.305)
Mother×T	-0.035	(0.230)
NGO program target×T	-0.051	(0.105)
non-sensitive items		
Age	0.001	(0.005)
Ever married	-0.017	(0.188)
Other ethnic minorities	0.126	(0.084)
Educated	0.124	(0.106)
Christian	0.119	(0.181)
Mother	0.090	(0.178)
NGO program target	0.097	(0.072)
Intercept	1.697***	(0.149)
N	751	

Note. The dependent variable is the response to the list experiment question. It is either 0,1,2,3 for the respondents in the control group or 0,1,2,3,4 in the treatment group. Estimated coefficients from the item count technique linear regression model 1 where the sensitive item is whether or not "a girl should be circumcised". T corresponds to the treatment status dummy (1 treated; 0 control). The sensitive item estimated parameters correspond to δ in equation 1. The non-sensitive item estimated parameters correspond to γ in equation 1. Robust SE Signif. codes: (*) if $p < .1$, (**) if $p < .05$, (***) if $p < .01$.

Table 4: Results of the linear probability model applied to responses to the direct question.

	"A girl should be circumcised"	
Variables	**Est**	**SE**
Age	0.006**	(0.003)
Ever married	0.047	(0.097)
Educated	-0.070	(0.050)
Other ethnic minorities	-0.152***	(0.042)
Christian	0.009	(0.085)
Mother	-0.016	(0.099)
NGO program target	-0.135***	(0.037)
Intercept	0.262***	(0.070)
N	751	
R-squared	0.063	

Note. The dependent variable is a dummy variable equal to 1 if the respondent replies yes to the statement "a girl should be circumcised". Robust standard errors are in parentheses. Signif. codes: (*) if $p < .1$, (**) if $p < .05$, (***) if $p < .01$. To be consistent with the list experiment analysis, we only consider the subsample of respondents for which the list experiment question is not missing.

Table 5: Estimated proportion of women answering the sensitive item in the affirmative way by socio-demographic characteristics, and differences between direct and indirect questioning.

	List experiment		Direct question		Differences		test	Unadj	Bonferroni
	Est	SE	Est	SE	Est	SE	Est	p-value	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No covariates	0.416	0.049	0.318	0.017	0.098	0.051	1.900	0.057*	-
Covariates	0.409	0.049	0.321	0.017	0.088	0.052	1.709	0.087*	-
Uneducated	0.479	0.061	0.334	0.020	0.145	0.064	2.252	0.024**	0.292
Educated	0.115	0.134	0.264	0.043	-0.149	0.141	-1.059	0.290	1.000
Never married	0.480	0.212	0.284	0.077	0.196	0.225	0.869	0.385	1.000
Ever married	0.389	0.074	0.332	0.028	0.057	0.079	0.723	0.470	1.000
Ethnic group Afar	0.423	0.056	0.353	0.020	0.070	0.060	1.169	0.242	1.000
Other ethnic minorities	0.360	0.112	0.201	0.035	0.159	0.118	1.351	0.177	1.000
Muslim	0.399	0.050	0.321	0.017	0.078	0.052	1.497	0.134	1.000
Christian	0.712	0.324	0.330	0.084	0.382	0.334	1.142	0.253	1.000
No mother	0.435	0.187	0.332	0.075	0.103	0.201	0.510	0.610	1.000
Mother	0.400	0.082	0.317	0.031	0.083	0.087	0.955	0.340	1.000
NGO not-targeted	0.443	0.086	0.411	0.031	0.033	0.092	0.355	0.722	1.000
NGO targeted	0.392	0.061	0.276	0.020	0.116	0.064	1.824	0.068*	0.817

Note. The sensitive item corresponds to "a girl should be circumcised". Predictions are based on the linear probability model for the direct question, and on the linear model for the indirect question. The results are averaged over the sample distribution of covariates. The last column presents Bonferroni multiple hypothesis testing to adjust the p-values of column (8). Signif. codes: (*) if $p < .1$, (**) if $p < .05$, (***) if $p < .01$.

Table 6: Tests of randomisation for the list experiment.

Respondent's characteristics	Control mean	Treatment mean	T test/ chi-squared p-value
Age	28.346	28.328	0.979
Religion (1=Christian; 0=Muslim)	0.035	0.035	0.976
Ethnic (1=Afar; 0=Other ethnic minorities)	0.789	0.784	0.874
Zone 1 (1=Zone 1; 0=Zone 3 or 5)	0.342	0.328	0.664
Zone 3 (1=Zone 3; 0=Zone 1 or 5)	0.329	0.340	0.744
Zone 5 (1=Zone 5; 0=Zone 1 or 3)	0.329	0.333	0.914
Health education/training (1=yes; 0=no)	0.231	0.247	0.602
Health providers available (0-4)	2.473	2.551	0.276
Health services accessible (0-5)	2.644	2.603	0.777
Mother (1=yes; 0=no)	0.745	0.732	0.690
Ever being married (1=yes; 0=no)	0.788	0.767	0.489
Educated (1=yes; 0=no)	0.175	0.207	0.262
Sex and HIV knowledge (0-6)	4.051	4.045	0.948
Agree circumcision (1=yes; 0=no)	0.318	0.315	0.927
NGO targeted (1=yes; 0=no)	0.658	0.673	0.664
N	374	403	

Note. A good randomisation of the list experiment is a crucial assumption. All important characteristics do not vary between the two groups.

Table 7: Design effects. Estimated respondent types for the list experiment.

y **value**	π_{y0}	**se**	π_{y1}	**se**
0	0.75%	0.004	0.60%	0.007
1	7.76%	0.016	14.70%	0.026
2	38.19%	0.033	23.27%	0.031
3	11.82%	0.020	3.01%	0.009
Total	58.43%		41.58%	
N	770			

Note. The table shows the estimated proportion (and standard error) of respondent types, $\hat{\pi}_{yz}$, characterised by the total number of affirmative answers to the control questions, y, and the truthful answer for the sensitive item.

Online Appendix

Appendix A NGO intervention details

In 2010, five Dutch organisations (Rutgers WPF, AMREF Flying Doctors, Simavi, dance4life and Choice) formed the Sexual and Reproductive Health and Rights Alliance (SRHR Alliance). The Alliance aims at working towards a society free of poverty in which all women and men, girls and boys, and marginalised groups have and enjoy their sexual and reproductive health and rights. The Alliance, in collaboration with partner organisations in developing countries, formed the 'Unite for Body Rights (UFBR)' program, a five year program (2011 - 2015) implemented in nine countries: five in Africa (Ethiopia, Kenya, Malawi, Tanzania and Uganda) and four in Asia (Bangladesh, India, Indonesia and Pakistan).

In Ethiopia, the UFBR program is implemented by three partners: AMREF Health Africa Ethiopia, Youth Network for Sustainable Development (YNSD), and Talent Youth Association (TaYA joined the program in 2013). In our paper, when we discuss the NGO program, we are referring to the UFBR program implemented in Ethiopia where AMREF was the leading partner organisation. The intervention area was selected by AMREF in close cooperation with the government. Important criteria were the non-existence of other donors and accessibility of the area.

Generally, the project strives to improve the sexual and reproductive health situations of Afar by increasing access to health services and enhancing utilisation

of health services at a community level. Specifically, the project is aiming at:

- Objective 1: Increased quality and delivery of comprehensive sexuality education
- Objective 2: Increased utilisation and quality of sexual and reproductive health services
- Objective 3: Reduction of sexual and gender based violence (SGBV)

To improve sexual reproductive health and rights (SRHR) services in Afar, the program trains and supports health workers at three levels in the health system: health centres, rural health extension posts, and within the communities through community health promoters. Trainings address, for example, SRHR/SGBV issues, including emergency obstetric care, clean and safe delivery and referral (for traditional birth attendants), youth friendly service provision, and counselling of victims of SGBV. The program provides training and support for district and health management teams. Some health facilities are renovated and equipped.

Besides focusing on strengthening the health system, the project also focuses on strengthening comprehensive sexuality education for in and out of school youth. For this component of the project, AMREF Health Africa Ethiopia also works in close collaboration with YNSD, the partner of the Ethiopian SRHR Alliance.

For further information about the intervention, we refer to the MFS II evaluation report published on the Partos Website, in particular the Ethiopia end-line report 04, pp. 257-381 (Stichting-Gezamenlijke-Evaluaties, 2015, `www.partos.nl/`

joint-MFSII-evaluations). This report concerns an impact study of the project.

Appendix B Data details

Sampling strategy

Since the primary objective of the project is to change the behaviour of households through information dissemination and behavioural change campaigns, all households with children (10-24 years) and women of reproductive age (15/49) living in the targeted districts are defined as the "targeted group". We sampled women of reproductive age (15/49) and unmarried girls aged between 15 and 24 and from the same household when possible.[29]

We used a multi-stage stratified sampling method in which strata are defined by zones which represent different target groups, woredas and kebeles. We sampled individuals targeted by the intervention from Zones 3 and 5, while the interviewees from Zone 1 had no access to any of the services supported by the intervention. Zone 1 was selected taking into account the geographical proximity (similarity) to the treatment zones. Data from Zone 1 reflect the situation for households that do not have access to the program.

Selection of Woredas. We identified a list of intervention woredas from each zone. From this list, we selected two woredas per zone. The selection was not nec-

[29] A small sample of boys were interviewed, but were not considered in this paper.

essarily random because of the limited number of them targeted by the program and their accessibility to conduct the survey. Households were selected from the following woredas in each zone: Awash and Amibara from Zone 3, Dawe, Telalak from Zone 5, and Mile and Chifira from Zone 1.

Selection of Kebeles. Kebeles are stratified into rural and urban. In most cases, an urban kebele is the center of the woreda. Three kebeles (one urban and two rural) were selected from a woreda. We selected kebeles that were targeted by the program in Zones 3 and 5. Kebeles in Zone 1 have not been targeted by any intervention.

Selection of households or woman. Our sample concerned women within the age group of 15 to 49. In kebeles where there was a list of residents available, we used the list to sample households (35 households were selected for each kebele using a lottery method). However, in villages where there was no list, we randomly selected houses from the village. If the age of the woman in the house was outside of the age range, we replaced the household with the neighbouring one.

Sampling of the unmarried girls. When possible, we selected one girl from the family of the interviewed women (mother). We interviewed approximately 12 girls per kebele.

Balance tests between the NGO targeted and non-targeted groups are reported in Table B1. A number of the respondents' characteristics differ in the two groups. It is important to note that: 1) the NGO decided to intervene in areas that were not benefitting from other donor interventions and accessible for the NGO staff, and 2) our data were collected after the beginning of the intervention. The most

important characteristics are controlled for in the regression analysis when either the outcome is measured with a direct question or with the list experiment.

Table B1: Balance tests for NGO targeted and non-targeted groups.

Variables	NGO Non-targeted mean	NGO Targeted mean	T test/ chi-squared p-value
Respondent's characteristics			
Age	27.240	28.710	0.034
Religion (1=Christian; 0=Muslim)	0.014	0.066	0.001
Ethnic group (1=Afar; 0=Other ethnic minorities)	0.741	0.802	0.043
Health education/training (1=yes; 0=no)	0.258	0.236	0.479
Health providers available (0-4)	2.319	2.611	0.000
Health services accessible (0-5)	2.238	2.836	0.000
Having children (1=yes; 0=no)	0.711	0.728	0.600
Ever being married (1=yes; 0=no)	0.782	0.764	0.551
Educated (1=yes; 0=no)	0.207	0.216	0.760
Sex and HIV knowledge (0-6)	4.274	3.933	0.000
Outcomes			
Agree FGC (1=yes; 0=no)	0.362	0.265	0.004
List experiment answer	2.084	2.122	0.452
N	282	566	

Timing and focus of the survey

The data used for this paper were part of the baseline study and collected in August/September 2012. The survey data concerned information about the following

issues: socio-economic background of the respondent and the household; access to sexual and reproductive health services; knowledge about sexual and reproductive health services; attitudes towards sexual and reproductive health practices; use of sexual and reproductive health services; intentions to use sexual and reproductive health services; household water supply; and household sanitation.

Enumerator selection and training

Female enumerators were used to interview women and girls to make respondents more comfortable. All enumerators spoke the local language, i.e., Afar.

The enumerators were trained by our partner IFPRI-ESARO to ensure that the survey questions were understandable and well phrased. Adjustments to the survey were made before beginning the data collection.

The enumerators were supervised by supervisors who checked the questionnaires day by day during the interviewing process. Individual face-to-face interviews were conducted in a location where only the interviewer and the respondent were present. Since many questions were private, no other person was supposed to be present during the interview. The interview took place in an area near to the home of the interviewee. Two supervisors and ten enumerators were hired and trained. Two teams were traveling to different woredas in the same zone and at the same time. Each team conducted the interviews in three woredas, one from each zone.

For further information about the data collection, we refer to the MFS II evaluation report published on the Partos Website (`https://www.partos.nl/`

joint-MFSII-evaluations), in particular, the Ethiopia endline report 04, pp. 257-381 (Stichting-Gezamenlijke-Evaluaties, 2015) and the survey included in this report. This report concerns an impact study of the project. The list experiment data were collected as a component of the baseline study for this evaluation.