CSAE Working Paper WPS/2015-20

Measuring attitudes regarding female genital mutilation through a list experiment[1]

Elisabetta De Cao, University of Oxford
Centre for Health Service Economics & Organisation
University of Oxford, United Kingdom

Corresponding author: elisabetta.decao@phc.ox.ac.uk

Clemens Lutz, University of Groningen
Department of Innovation Management & Strategy
University of Groningen, The Netherlands

Abstract

Understanding the attitudes toward Female Genital Mutilation (FGM) is crucial for policy, but challenging because it is a sensitive topic. This paper uses a list experiment to elicit truthful answers about FGM support in Ethiopia. A regression analysis is developed to analyze the data. To test the social desirability bias, list experiment results are compared with direct question results revealing that uneducated respondents underreport their support by 16% (p-value=0.013), while women treated by a NGO intervention by 12% (p-value=0.060). Potential bias in responses to direct questions should be considered when measuring sensitive outcomes, especially in the context of policy evaluations.

Key words: Female genital mutilation; female circumcision; list experiment; sensitive survey questions; Ethiopia.

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

[1]For supervising the data collection we thank IFPRI, Getaw Tadesse and Samson Jemaneh. For comments we thank Robert Lensink, Rob Alessie, Carol Propper, 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.

1 Introduction

Female genital mutilation (FGM) or female genital cut or female circumcision includes all procedures that alter or cause injure to the female genital organs. They are mainly carried out on young girls. FGM is recognized as an extreme form of discrimination and violence against women. Worldwide about 140 million girls and women are living with the consequences of FGM.[2] The WHO estimates that in Africa more than 3 million girls are at risk for FGM annually (WHO, 2012).

Generating knowledge about the causes and consequences of FGM and planning effective policy interventions is a way to eliminate FGM. There is a vast anthropological and ethnographical literature on the existence of FGM.[3] Economists have recently started to study FGM both from a theoretical point of view (Chesnokova and Vaithianathan, 2010; Coyne and Coyne, 2014) and empirically looking in particular at the determinants of FGM (Naguib, 2012; Ouedraogo and Koissy-Kpein, 2012; Molitor, 2014; Bellemare et al., 2015; Wagner, 2015) or at the effect of laws or program interventions against FGM (Camilotti, 2015b) .

Remarkably, quantitative research regarding FGM attitudes considers as main outcomes variables based on direct survey questions about the support for the continuation of the FGM practice.[4] 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?" However, eliciting truthful answers in surveys is challenging, especially when studying sensitive issues

[2]A review on the health consequences of FGM can be found in Makhlouf Obermeyer (2005).

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

[4]Also the FGM status, having experienced FGM or not, is considered but it is mostly self-reported, otherwise a gynecological examination would be necessary.

such as attitudes toward FGM. If asked directly, individuals may lie or refuse to answer. The dependent variable might therefore be affected by non-random measurement error that leads to biased results.[5]

When asking questions about a sensitive issue, different survey methods exist to cope with the problem of bias in self-reported answers.

New qualitative solutions have been proposed by Blattman et al. (2015) to study the direction and magnitude of the survey measurement error in the dependent variable when evaluating interventions implemented in Liberia to reduce violence and crime. Blattman et al. (2015) use different qualitative techniques to validate survey responses in relation to different behaviors (theft, drug use, homelessness, gambling and expenditures) and find different results in terms of underreporting depending on the sensitive behavior considered.

Quantitative survey methods include the randomized response technique[6] and the endorsement experiment[7]. A third method used in this paper is called *list experiment*. The idea behind a list experiment, also called 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 to indicate the total number of items with which they agree. The respondents are randomly divided in a control and 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

[5]Self-reported health status and outcomes have been found to be affected by under-reporting when they focus on sensitive topics related to sexual and reproductive health (Schroder et al., 2003; Glynn et al., 2011).

[6]The randomized response technique developed by Warner (1965) consists in asking the respondent to use a randomization device (dice, coin flip, etc) whose outcome is unknown to the interviewer. The randomization device determines the type of question (or response) the respondent answers (gives). By introducing random noise, this technique guarantees the anonymity and the respondent may be more willing to reveal the truth.

[7]In an endorsement experiment, respondents are randomly selected and asked their opinion toward a policy endorsed by a socially sensitive actor of interest (Bullock et al., 2011).

item. The difference in the total number of items between control and treatment group identifies the proportion of people in the population that agrees with the sensitive item. The list experiment technique has been mainly 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 used to study sexual risk behavior (LaBrie and Earleywine, 2000) or abortion (Moseson et al., 2015). More recently it is also 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 get truthful answers on topics such as condom use, number of sexual partners, unfaithfulness, and attitude changes with respect to the social acceptability of these behaviors (Jamison et al., 2013; Chong et al., 2013). A recent paper by Coffman et al. (2013) estimates the magnitude of anti-gay sentiment showing that it is largely underestimated when a list experiment is used to elicit truthful answers.

Surprisingly, the aforementioned economic literature considers a difference-in-means estimator to analyze the list experiment (see for example, Karlan and Zinman, 2012; Chong et al., 2013).[8] This however does not allow the identification of the relationship between preferences over the sensitive item and the respondent's characteristics. Moreover, the effect of social pressure on the answers given to direct sensitive questions may differ among groups in the population. Regressions can instead be used to study both the list experiment and the difference between indirect and direct questioning which determines the misreporting due to *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 their support

[8] 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.

towards FGM. This paper goal is to determine the true perceptions about FGM 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 a NGO intervention to improve knowledge about sexual and reproductive health and provide health services is implemented. Using the fact that FGM is formally banned in Ethiopia, but still undergone, makes the topic a very sensitive one.

The contributions of this paper are three. First, it focuses on a new list experiment designed to measure attitudes regarding FGM in one of the areas where FGM prevalence is among the highest. Second, the most recent regression techniques developed to analyze the list experiment and the social desirability bias are used. This allows to determine the existence and magnitude of systematic reporting measurement error of the true outcome. Third, the list experiment is used to study if respondents targeted by a NGO intervention are more or less likely to misreport their attitudes.

The main results are the following. Firstly, the list experiment shows that educated women support female circumcision less than uneducated women. Secondly, the social desirability bias is the greatest among uneducated women, they underreport by 16% (p-value=0.013) their true beliefs. Thirdly, we find that women targeted by the NGO intervention have a stronger incentive to lie about their FGM support (11%, p-value=0.060).

These results confirm the relevance of potential bias in responses to direct sensitive questions and are important to keep in mind when evaluating a program intervention where the outcome of interest is sensitive and the survey error is potentially correlated with the program treatment, leading to biased conclusions about the effectiveness of the program.

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 FGM and its limitations. Section 4 describes the list experiment and the social desirability bias results. In Section 5, we present robustness checks. 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. 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 (9 percent); the percentage of births delivered in health facility is less than 10 percent; full vaccination coverage among children age 12-23 months is 9 percent; 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 2000, Afar counted 2 hospitals, 14 Health Centers and 112 Health Posts serving the entire population (`http://www.moh.gov.et`).

The last DHS estimates indicate that Ethiopia is one of the countries with the highest FGM prevalence, about 74.3%, and up to 91.6% in the Afar region (Central Statistical Agency Ethiopia and ORC Macro, 2006).[9]

In 2004 the Ethiopian government introduced the Criminal Code Proclamation No. 414/2004, that criminalizes harmful traditional practices among which FGM. The Proclamation became law in 2005.[10] In December 2012, the United Nations General Assembly unanimously passed Resolution 67/146, condemning FGM and related harmful practices and urging member states to take measures to accelerate its elimination.[11] Even though

[9]Rahlenbeck et al. (2010) explore factors influencing attitudes towards the practice of FGM in Ethiopia. Religion is often used as a justification, even if there is no doctrinal basis for this practice in Islam, Christianity or Judaism. FGM is traditionally believed to ensure hygiene and preserve a girl's chastity and fertility. Hence, the practice is considered beneficial to girls, but also as a prerequisite for a honorable marriage.

[10]The sanctions include imprisonment that ranges from 3 months to 3 years and a fine of no less than Birr 500 to 10,000 or both imprisonment and fine. In the case of infibulation the penalty is higher with prison term of 3 to 10 years. (Ras-Work, 2009)

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

FGM is formally banned in Ethiopia, the practice still exists making discussions about it very sensitive and secretive.

Since 2011 a NGO program has been working in some areas of Afar to provide comprehensive sexuality education programs and health services. In October 2012, we collected data in the region. 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), and some non-beneficiaries without access to any of the NGO activities from a different area (zone 1). Figure 1 shows the map of Afar with the different zones highlighted. Since the NGO program mainly targets young people and women of reproductive age, our survey consists of women/mothers aged between 15 and 49 (n=631), and unmarried girls (n=217) aged between 15 and 24, for a total of 848 respondents. See the Online Appendix for details about the NGO intervention and data collection.

The information covered in the questionnaire concerns: the socio-economic back-ground of the respondent, access to sexual and reproductive health services, knowledge about sexual and reproductive health services, attitudes towards sexual and reproductive health practices and FGM, use of sexual and reproductive health services, intentions to use sexual and reproductive health services, household water supply and sanitation.[12]

Table 1 reports the descriptive statistics of the main variables. The survey contains individuals that were exposed to the NGO's program (67%). Most of the respondents are Muslim (95%) and of Afar ethnicity (78%). Not many of the respondents have ever participated in any sexual and reproductive health education or training program 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, coun-

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

seling on pregnancy/child care/contraceptives, medical treatment, condoms, contraceptives) easily accessible is 2.6 (maximum 5). About 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 few years of elementary school, and 21% with higher levels of education (elementary (13%), secondary (5%) or tertiary education (3%)).

The respondent's characteristics considered in our empirical analysis are: 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 the elementary education), NGO targeted status (dummy equal to one if targeted by the NGO intervention). Other variables are excluded because they do not present high variability in the sample. For example, we do not include being a mother or religion because 94% of the ever married women have kids, and 95% of the sample is composed by Muslims.

3 Methodology

3.1 Standard list experiment design

In order to measure the true perception about female circumcision we added to the survey a list experiment. The list experiment or unmatched count technique works by aggregating the sensitive item with a list of other non-sensitive items (Miller, 1984). The survey sample is composed by N respondents, that are randomly divided in 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 have to tell the interviewer how many of the listed items they agree on, but not which items. The treatment group respondents instead receive the same list of non-sensitive, yes/no items plus a sensitive, yes/no item ($J + 1$ in total), where $j = J + 1$ is the sensitive one.

The sensitive item measures the sensitive topic. As for the control group respondents, the treatment group respondents have to tell the interviewer the number of items they agree on.

To formalize, we use the same notation as in Imai (2011) and Blair and Imai (2012). Let us define $Z_{ij}(t)$ a dummy variable that indicates the respondent i's preference for the jth control item ($j = 1, \ldots, J$) under the treatment status $t = 0, 1$. The respondent i's answer to the sensitive item for the treatment group is indicated as $Z_{i,J+1}(1)$. Z^*_{ij} corresponds to the respondent i's truthful answer to the jth item where $j = 1, \ldots, J+1$. The potential answer respondent i would give under the control or treatment group, is 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.

This design relies on three important assumptions (Imai, 2011). The first assumption is the *randomization of the treatment* meaning that the sample is randomly divided in control and treatment groups and it implies that potential and truthful responses are jointly independent of the treatment variable, hence, for any respondent $i = 1, \ldots, N$, the following needs to hold: $\{\{Z_{ij}(0), Z_{ij}(1)\}_{j=1}^{J}, Z_{i,J+1}(1)\} \perp T_i$. The second assumption called *no design effect* implies that the addition of the sensitive item does not change the sum of affirmative answers to the control 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 third assumption is called *no liars* and it implies that the respondents truthfully reply to the sensitive item, for each $i = 1, \ldots, N$, we have $Z_{i,J+1}(1) = Z^*_{i,J+1}$. 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.

If these three assumptions hold, then the unbiased estimate of the population proportion of those that agree on the sensitive item can be 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 and it characterizes each respondent's type ($2 \times (J+1)$ types in total).

Imai (2011) proposes new multivariate regression estimators, that also rely on the assumptions of no design effect and no liars, to analyze 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[13] (see also Holbrook and Krosnick, 2010):

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

where $E(\epsilon_i | X_i, T_i) = 0$, and (γ, δ) are unknown parameters.[14] Being the treatment T 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 the sensitive item affirmatively.

In this paper we analyze 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.[15] This technique is easy to interpret, but rarely

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

[14]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).

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

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 used in the context of reproductive health, hence we believe this is a further contribution.

3.2 Our list experiment

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

I want you to give me a secretive 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 shouldn't also 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 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 female circumcision:

4. *A girl should be circumcised*

As control items we selected items related to sexual and reproductive health knowledge, as well as, family planning issues. There is no need to assume that the answers to the control items are truthful, but they need not to be influenced by the presence of the sensitive item to the list. Given that the Ethiopian law prohibits FGM, people are expected to be less prone in revealing 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 so called ceiling and

floor effects are avoided (no liar effect) (Kuklinski et al., 1997).[16] Ceiling effects occur when a respondent would honestly respond "yes" to all nonsensitive items, and in the treatment group the respondent no longer has the protection to honestly report her/his 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 the items to cover his/her identity.

Table 2 reports the observed data from the list experiment. As we said the list experiment has three non-sensitive items and one sensitive item. The treatment group counts 438 respondents, the control 400. A total of 10 respondents did not answer the list experiment question, 5 in each group. We observe that the responses are well distributed and there are few responses in the extreme cases (0 and 3 for the control group, and 0 and 4 for the treatment group). Having many responses in the extreme cases can indicate the presence of ceiling effects, or floor effects.

3.3 Limitations to our list experiment

We formally test for violations of the three key assumptions of the list experiment in the Robustness section 5, and show that the design was done properly. However, we believe there are some limitations to take into consideration.

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

[16]The common advise is that the list of control items should not be too short to avoid the ceiling and floor effects (Kuklinski et al., 1997), and many empirical examples use a 3-items or 4-items 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 randomized to avoid ordering effects.

to be very informative. However, the prevalence estimates are based on self-reported FGM 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 FGM attitudes. This certainly calls for further research on this promising method. In particular it is important 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. One could have added endorsement experiments, randomized responses or qualitative techniques to validate the list experiment.[17] We have decided not to do that to avoid complicating and lengthening too much the questionnaire. As far as 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 adjusted to make our list experiment feasible before starting the data collection.

Replicability of the approach 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 FGM could and should be replicated in different contexts. Interviewers can be easily trained to ask list experiments and a pilot survey can be done to test the feasibility of the list experiment. The cost of adding list experiments to the surveys is very low, they are simply extra questions that require the randomization of the questionnaire.

[17]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 run in Afghanistan.

4 Results

4.1 Results of the list experiment

Table 3 reports the results from the list experiment, using the difference in-means estimator, commonly used to analyze the list experiment. The results indicate that 39.2% (SE=0.047) is the estimated proportion of women who agree with the sensitive item "a girl should be circumcised".

In addition to knowing the overall proportion of women that agree with FGM, it is interesting to know what type of respondent is more in favor of FGM. One can do the difference-in-means estimator separately in each subgroup, as commonly done (some examples are Kuklinski et al., 1997; McKenzie and Siegel, 2013), but 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 analyze the list experiment (Equation 1). Table 4 presents the model results. The interesting estimated coefficients are reported in the top of Table 4 (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 41.2% (SE=0.147) less likely to be in favor of circumcision even after controlling for other individuals' characteristics.

We present in Figure 2 a comparison of the difference-in-means and linear model results considering education as the main variable. Figure 2 is based on the fitted model presented in Table 4 and on the model without covariates (diff-in-means). Figure 2 presents the estimated proportions of uneducated (circle) and educated people (triangle) who agree that "a girl should be circumcised". The 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 the other covariates to their

observed values.[18] The solid lines correspond to the 95% asymptotic confidence intervals. The model without covariates (diff-in-means) does not present significantly different effects between uneducated and educated women, while the linear regression model confirms the significant difference in attitudes between educated and not educated women. In particular, 47% (SE=0.059) of the uneducated women agree with FGM, while 6% (SE=0.120) is the proportion of educated women in the multivariate linear model.

4.2 Social desirability bias

To assess the impact of sensitivity on responses, we compare the attitudes toward FGM 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 support towards FGM measured with the list experiment is closer to the real unobserved support than the support obtained with the direct question. The second is that the measurement error in the direct survey and list experiment data go in the same direction. This implies that the difference between the indirect and the direct question proxy the underreporting of the true support.

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

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

In the survey, we asked: "Do you agree on the following statement? A girl should be circumcised."[19] The possible answers 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 dichotomized the survey question as follows: totally agree and somehow agree correspond to 1 (yes), and the remaining ones to 0 (no). If asked directly, about 30% (SE=0.016) of the women agree upon the fact that a girl should be circumcised.[20] The proportion obtained using the difference-in-means estimator is 39.2% (SE=0.047), hence the difference is 9.2% (SE=0.049), statistically different from zero at 10% level (see the no covariates results in Table 6).

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 analyze the responses to the direct question. Table 5 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 the opposite.[21] In particular, Table 5 shows that, holding all other variables fixed, the probability that members of the other ethnic minority groups are in favor of the sensitive question is 12.3% (SE=0.036) lower than for Afar people. Being exposed to the NGO program is also significant and negatively affects the outcome variable. In particular, the probability for targeted people of being in favor of FGM is 11.8% (SE=0.034) lower than the probability for not-targeted people.

[19]We asked the direct question both to the control and the treatment group. The questionnaire was extensive and it included many different questions about sexual and reproductive health, and we believe that it did not affect the response to the list experiment.

[20]Note that if we dichotomize the direct question considering as 1 also the respondents who replied neither agree nor disagree, we obtain 35.6% (SE=0.016) as proportion of women that are in favor of the sensitive issue.

[21]Note that in Table 6 we consider the same sample used for the list experiment analysis in Table 4, keeping only the observations for which the list experiment question is not missing.

However, in the list experiment regression analysis (Table 4), we find that being targeted by the NGO program is not significant. Unfortunately, we cannot test if the difference in the two effects (NGO effect when the list experiment or the direct question are used) is statistically significant, because the list experiment generates only aggregate information.[22] Instead, we can study the social desirability bias in the NGO targeted and not-targeted groups.

Table 6 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 6 includes also the results for the model without covariates and the model with covariates (age, ethnic group, marital status, education and being targeted by the NGO). We also report the estimated proportions for different groups by controlling for all the other covariates. The differences between the indirect and direct questions are always positive (except for the educated people) and statistically significant at the 10% level in the no covariates and covariates models, as well as for the unmarried group, the other ethnic minorities group and the NGO targeted group. The difference is instead highly statistically significant and positive for the uneducated group. Therefore, it seems that the group that underreports the most is the group of uneducated people where the direct question produces a 31% (SE=0.019) of the women in favor of circumcision, compared to 47% (SE=0.059) obtained through the list experiment (difference=16%; p-value=0.013). Interestingly, when the direct question is considered, never married women underreport their support towards FGM by 27% (p-value=0.065), while women targeted by the NGO intervention underreport their support towards FGM by 12% (p-value=0.060).

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

5 Robustness

5.1 Test the assumptions for a good list experiment

In this subsection we carefully test for potential violations of the three key assumptions of the list experiments.

The first assumption is the *randomization of the treatment*. Table 7 provides sample means for the main variables in the treatment group and the control group. Comparing the means allows us to see that the randomization of the list experiment (control group and treatment group) was successful given that all important respondent's characteristics do not significantly differ between the two groups.

The second assumption is called *design effects* and it happens when the inclusion of a sensitive item affects some respondents' answers to control items. 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 $z = 0, 1$. 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 the proportions are negative, then there are design effects, while if only some are negative, it is important to understand if they are negative by chance. Table 8 reports the estimated proportion of each respondent type. They are all positive, hence, the assumption of no design effects holds. Blair and Imai (2012) propose a statistical test for detecting the design effects.[23] By applying the test to our list experiment, we fail to reject the null hypothesis of no design effects.[24] This is an indication that respondents answered in a

[23]For details about the no design effect test see Blair and Imai (2012, pages 64-65).

[24]Test statistics=1.738; test statistics/2=0.869 > alpha/2=0.05/2=0.025, then we cannot reject the null hypothesis.

truthful way to the sensitive item.

The third possible problem is the violation of the assumption *no liars*. There can be two types of liars: liars that give the answer $Y_i = J$ if assigned to the treatment condition even if the truthful answer would be $Y_i = J + 1$, affirmative for both sensitive and control items (ceiling effects); and liars that give the answer $Y_i = 0$ if assigned to the treatment condition even if the truthful answer is affirmative only for the sensitive item (floor effects). Since both types of lies lower the observed mean response of the treatment, the presence of ceiling and/or floor effects lead to the underestimation of the population proportion of those who agree with the sensitive item. As we can see from Table 2, the responses are well distributed and there are few responses in the extreme cases (0 and 3 for the control group, and 0 and 4 for the treatment group).

5.2 List experiment by education level and by NGO targeted status

The respondent's education level revails to be a critical characteristic. Can it be that educated women understand the mechanism behind the list experiment and trick their answers? We test this possibility by analysing the results of the list experiment by education level. Figure 3 reports the distribution of the items for the illiterate and the educated group. Responses in the two groups are well distributed with few cases in the extremes. This is an indication of no liars effect. We also formally test the presence of design effects by applying the Blair and Imai (2012) design statistical test to the subsample of educated and illiterate women, and we fail to reject the null hypothesis of no design effects.[25]

We do the same analysis to see if NGO targeted women understand the mechanism behind the list experiment. One could think that perhaps people targeted by the NGO program

[25]Test statistic for the illiterate group=1.000/2 $>$ α/2=0.05/2=0.025, then fail to reject the null of no design effects. Test statistic for the educated group 0.158/2= 0.079 $>$ α/2=0.05/2, then fail to reject the null of no design effects. (See Blair and Imai, 2012, pages 64-65).

perceived differently the list experiment because all topics touched by the items (sensitive and non-sensitive) were covered by the NGO's activities leaving only the consensus towards FGM the sensitive one because illegal. Figure 3 reports also the distribution of the items for the NGO targeted and not-targeted women. There seems to be no liar effects. Design effects are also not detected when the Blair and Imai (2012)'s design statistical test is applied to the subsample of targeted and not-targeted women.[26]

6 Conclusions

Measuring attitudes towards FGM is critical to understand who to target most, but difficult because it is a sensitive topic. This paper uses new data collected in Ethiopia, one of the countries with the highest FGM prevalence and where FGM is formally prohibited, but still a widespread advocated custom in the local culture. A list experiment is designed to elicit truthful answers about FGM support. The results of the list experiment are compared to the results of a direct question to study underreporting due to social desirability bias or the systematic reporting measurement error of the true FGM support. The goal of the paper is indeed to understand if and who are the people that misreport their true believes about FGM support, and which is the magnitude of the underreporting. Overall, the results show that underreporting can be substantial, in particular for the uneducated respondents.

Our results indicate that when asking a direct question about FGM, 30% (SE=0.016) of the women are in favor 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 favor of FGM is much higher, 39.2% (SE=0.047) leading to a difference of 9.2% only slightly statistically significant (p-value=0.061).

[26]Test statistic for the not-targeted group=$1.041/2 > \alpha/2=0.05/2=0.025$, then fail to reject the null of no design effects. Test statistic for the targeted group $1.000/2= 0.079 > \alpha/2=0.05/2$, then fail to reject the null of no design effects (See Blair and Imai, 2012, pages 64-65).

A regression analysis, based on a new statistical approach developed to analyze the list experiments (Imai, 2011), shows that some respondent's characteristics seem relevant in explaining the consent towards female circumcision. 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 favor of FGM (-41.2%, p-value=0.004) compared to the illiterate 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 16% (p-value=0.013). This indicates that illiterate women seem to be less willing to share publicly their real attitudes concerning FGM support. The educational level may affect incentives in being cut. If being cut increases the chances to get a better husband (Chesnokova and Vaithianathan, 2010), we may argue that uneducated women have more to lose if they do not favor the practice, while educated women have better chances in the job market and depend less on marriage (Ouedraogo and Koissy-Kpein, 2012; Molitor, 2014). This can also be reflected by the never married women who underreport their support towards FGM by 27% (p-value=0.065).

Another interesting result concerns the NGO effect. The positive effect of the NGO program on reducing the support for the practice seems to disappear when the sensitive information is asked indirectly via the list experiment. An obvious question is: is the NGO effect measured with the list experiment different enough from the one measured with the direct question that we can say with confidence that the NGO intervention may not have actually changed people's attitudes, but rather how respondents report it (social desirability bias)? We can partially answer this question, because in this paper we can only estimate the underreporting due to social desirability bias. By comparing the estimates obtained with the direct and indirect questioning, we find that women targeted by the NGO intervention underreport their support towards FGM when the list experiment is used to ask their opinion instead of the standard direct question, but this is only statistically significant at the 10%

level (the difference is 12%, p-value=0.060). The intervention focuses on the dissemination of sexual and reproductive health knowledge and a change in traditional attitudes. It is well possible that the respondents in the treated areas conform to the expectations of those who provided the program treatment. The NGO campaign aims at changing the local FGM customs and this may increase the social pressure around FGM resulting in a stronger incentive to reveal a biased answer.

We cannot claim that the NGO intervention is not working in changing people's attitudes and therefore behaviors. Our analysis is not an impact evaluation of the NGO program.[27] To evaluate the intervention we would need to have pre- and post-intervention data on NGO targeted and not-targeted individuals.[28] Then it would be possible to study if the NGO treatment effect is measured with error by comparing the answers to a standard survey question (direct question) with the ones obtained with a 'validation' survey question (e.g., list experiment) (Blattman et al., 2015). This would tell us if the conclusions about the impact of the program are affected by how the sensitive outcome is measured. Moreover, it would be even more useful to evaluate the intervention after a long period of time so that actual behavior (e.g., cutting young girls) can be observed and compared with different measurements of attitudes. We leave this for future research.

Lack of empirical evidence on the support towards FGM, and most importantly lack of understanding of how biased direct questions can be, make our study the further step to a

[27]There are only few studies that evaluate programs against FGM. Camilotti (2015a), for example, looks at the effect of the program of the NGO Tostan in Senegal that aims at reducing FGM and she finds that "dissemination of information on the negative effects of FGC does not correspond to the end of the practice". Camilotti (2015b) instead finds, using data from Senegal, that laws against FGM and awareness campaigns result in a reduction of the age at cutting. Discussions with the staff of the NGO considered in our study, revealed that the age at cutting is decreasing over time. Since the NGO considered for our project organizes similar activities, such as public meetings, it remains to investigate if they are effective or not.

[28]Our study does not present such characteristics. See the Online Appendix for details about the NGO intervention and data collection.

future line of research that aims at focusing more on how to measure sensitive outcomes, and we believe this is specially important in the context of policy impact evaluations.[29]

Finally, we suggest that both quantitative survey methods, such as list experiments, endorsement experiments or randomized responses, and qualitative survey methods[30], should be added to surveys, such as the DHS for example, to measure if respondents misreport their attitudes or behaviors when the outcome of interest is sensitive.[31] They could be applied to other stigmatized health topics, such as, for example, domestic violence or sexually transmitted infections.

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., 2015. Measuring the measurement error: A method to qualitatively validate survey data. NBER Working Paper No. 21447 .

[29] In a recent paper, for example, Blattman et al. (2015) validate the results of a field experiment by using a qualitative approach to measure the measurement error affecting sensitive outcome variables.

[30] Such as the one suggested by Blattman et al. (2015).

[31] A recent paper by Rosenfeld et al. (2015) validates for the first time direct questioning, list experiment, endorsement experiment and randomized response against the official outcome, in the context of an anti-abortion referendum held during the 2011 Mississippi General Election.

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., 2015a. Changing female genital cutting. Evidence from Senegal. Mimeo .

Camilotti, G., 2015b. Interventions to stop female genital cutting and the evolution to the custom: Evidence on age at cutting in Senegal. Journal of African Economies , 1–26.

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 oc 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.

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.

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.

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

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. Universität 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 .

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 .

Rahlenbeck, S., Mekonnen, W., Melkamu, Y., 2010. Female genital cutting starts to decline among women in Oromia, Ethiopia. Reproductive BioMedicine Online 20, 867–872.

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.

Rosenfeld, B., Imai, K., Shapiro, J., 2015. An empirical validation study of popular survey methodologies for sensitive questions. American Journal of Political Science 0, 1–20.

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.

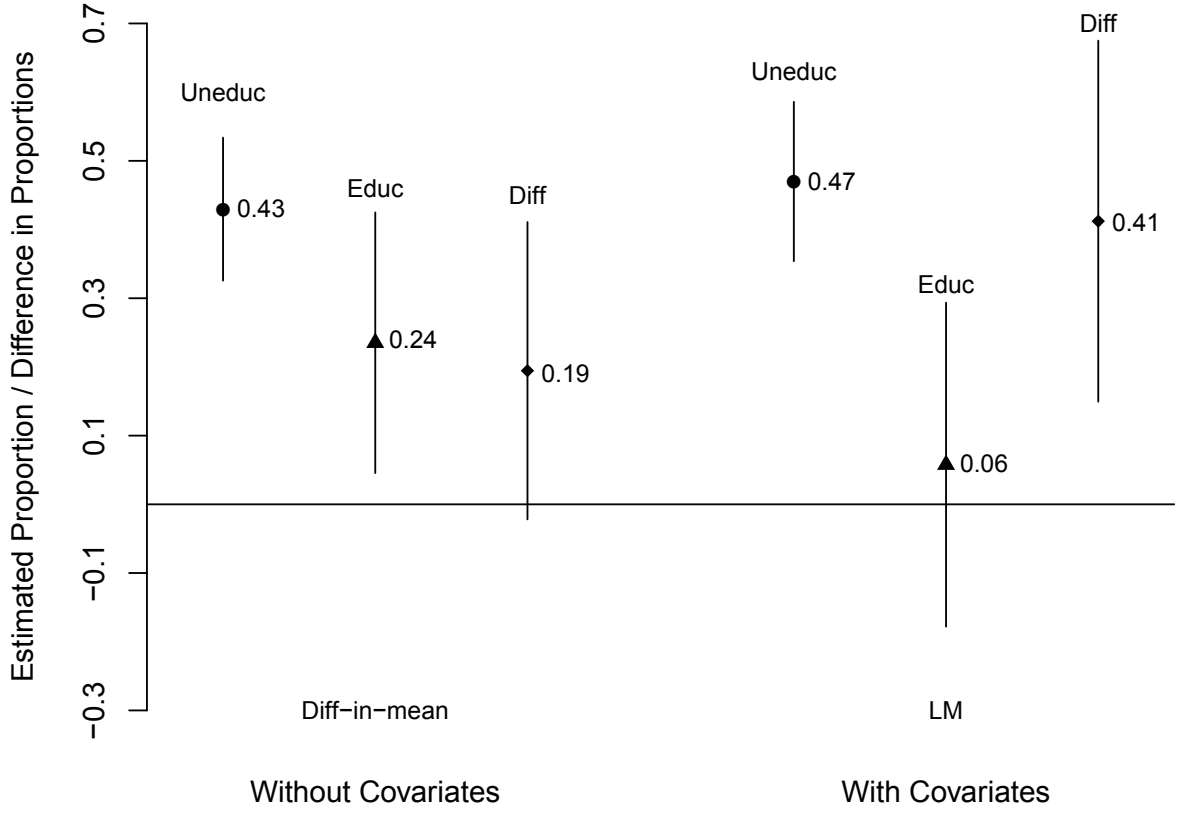
Yoder, P.S., Khan, S., 2008. Numbers of women circumcised in Africa: The production of a total. DHS Working Paper 39.

7 Figures and tables

Figure 1: Map of Afar, Ethiopia.

Note: This map shows the Afar region in Ethiopia (UN OCHA, `http://www.unocha.org`).

Figure 2: Estimated proportion of women who are in favor of FGM 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 95% confidence interval for the estimated proportions. In the linear regression model the results are averaged over the sample distribution of covariates.

Figure 3: List experiment results by respondents' education level and by respondents' NGO targeted status.

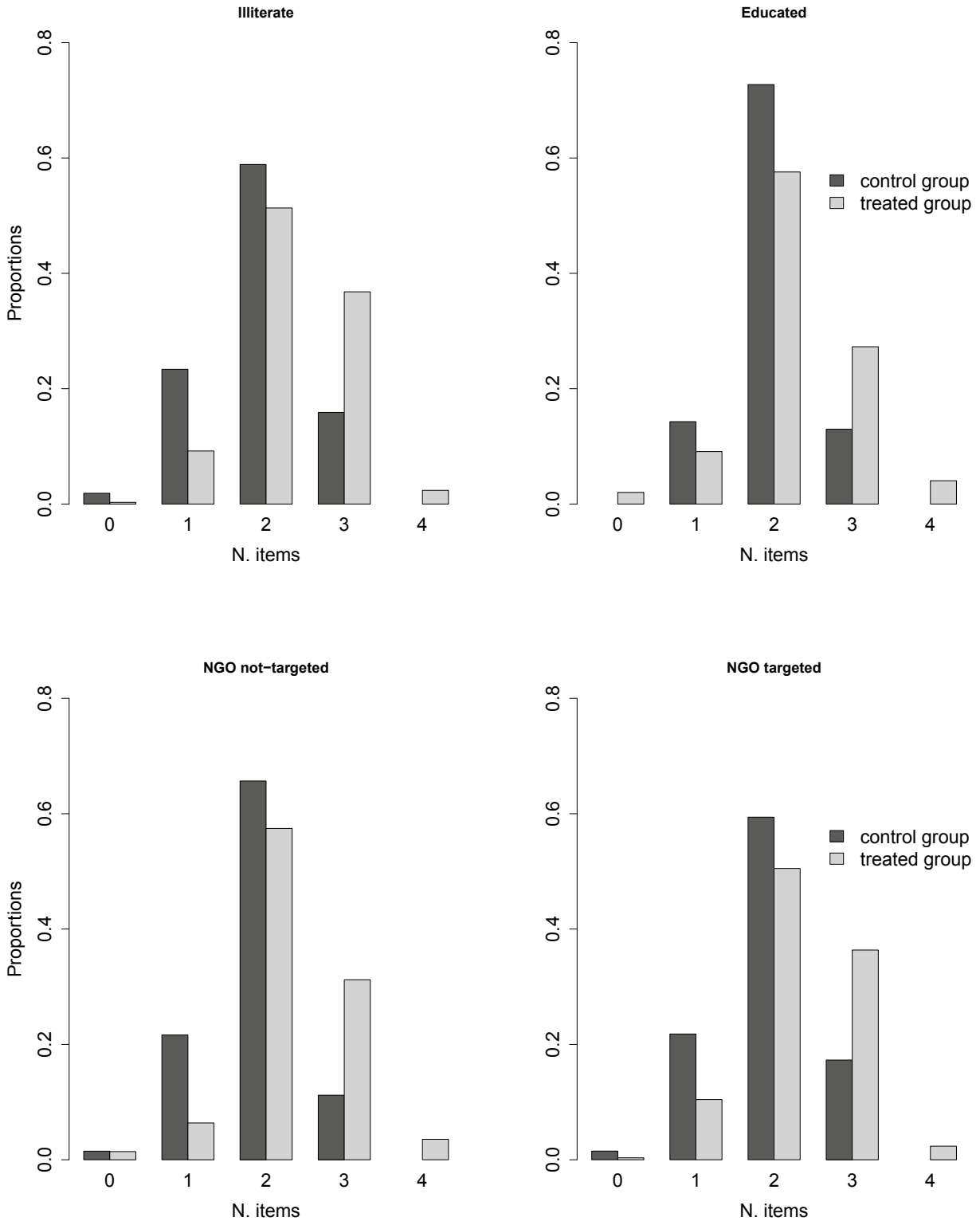


Table 1: Descriptive statistics

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 6 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	6	1.5	3	0.68
1	87	21.75	40	9.13
2	246	61.5	231	52.74
3	61	15.25	152	34.7
4			12	2.74
Total	400	100	438	100

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

Table 3: List experiment difference-in-means result.

	Control group	**Treatmen group**	**Diff-in-means estimate**
Mean	1.905	2.297	0.392***
S.E.	0.032	0.033	0.047
N	400	438	

Note. The table displays the average response to the observed outcome variable (total number of items the respondent agree on), separately for the control and the treatment group where the sensitive item is "a girl should be circumcised", and the difference-in-means estimation. Robust S.E. Signif. codes: (*) if $p < .05$, (**) if $p < .01$, (***) if $p < .001$.

Table 4: Results of the linear regression model for the list experiment.

Variables	**Est**	**SE**
Sensitive item		
T	0.813***	0.209
Age×T	-0.007	0.007
Ever married×T	-0.179	0.167
Educated×T	-0.412**	0.147
Other ethnic minorities×T	0.009	0.119
NGO program target×T	-0.022	0.100
Control items		
Intercept	1.721***	0.141
Age	-0.000	0.005
Ever married	0.097	0.117
Educated	0.138	0.097
Other ethnic minorities	0.152*	0.076
NGO program target	0.084	0.068
N	826	

Note. 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 control item estimated parameters correspond to γ in equation 1. Robust S.E. Signif. codes: (†) if $p < .1$, (*) if $p < .05$, (**) if $p < .01$, (***) if $p < .001$.

Table 5: Results of the linear probability model applied to responses to the direct question.

	"A girl should be circumcised"
Variables	
Age	0.004
	(0.002)
Ever married	0.068
	(0.052)
Educated	-0.072
	(0.045)
Other ethnic minorities	-0.123***
	(0.036)
NGO program target	-0.118***
	(0.034)
Intercept	0.264***
	(0.065)
Observations	826
R-squared	0.054

Note. The dependent variable is a dummy variable whether or not a girl should be circumcised. Robust standard errors are in parentheses. Signif. codes: (†) if p < .1, (*) if p < .05, (**) if p < .01, (***) if p < .001. 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 6: 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	
	Est	SE	Est	SE	Est	SE
No covariates	0.392	0.047	0.300	0.016	0.092†	0.049
Covariates	0.384	0.047	0.300	0.016	0.083†	0.050
Uneducated	0.470	0.059	0.315	0.019	0.155*	0.062
Educated	0.058	0.120	0.243	0.038	-0.186	0.126
Never married	0.522	0.142	0.248	0.042	0.274†	0.148
Ever married	0.343	0.057	0.316	0.020	0.027	0.061
Ethnic group Afar	0.382	0.053	0.327	0.018	0.054	0.056
Other ethnic minorities	0.391	0.106	0.204	0.031	0.187†	0.110
NGO not-targeted	0.398	0.081	0.380	0.029	0.018	0.086
NGO targeted	0.376	0.058	0.262	0.018	0.115†	0.061

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. Signif. codes: (†) if $p < .1$, (*) if $p < .05$, (**) if $p < .01$, (***) if $p < .001$.

Table 7: Tests of randomization for the list experiment.

Respondent's characteristics	**Control mean**	**Treatment mean**	**T test/ chi-squared p-value**
Age	28.225	28.227	0.998
Religion (1=Christian; 0=Muslim)	0.052	0.046	0.675
Ethnic (1=Afar; 0=Other ethnic minorities)	0.778	0.785	0.784
Areas in Afar (proportions)			0.885
Zone 1	0.489	0.511	
Zone 3	0.474	0.526	
Zone 5	0.470	0.530	
Health education/training (1=yes; 0=no)	0.239	0.247	0.773
Health providers available (0-4)	2.472	2.553	0.238
Health services accessible (0-5)	2.654	2.621	0.811
Having children (1=yes; 0=no)	0.732	0.713	0.544
Ever being married (1=yes; 0=no)	0.787	0.754	0.253
Educated (1=yes; 0=no)	0.196	0.229	0.242
Sex and HIV knowledge (0-6)	4.064	4.029	0.693
Agree circumcision (1=yes; 0=no)	0.301	0.294	0.804
NGO targeted (1=yes; 0=no)	0.659	0.675	0.628
N	405	443	

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

Table 8: Design effects. Estimated respondent types for the list experiment.

y **value**	π_{y0}	**se**	π_{y1}	**se**
0	0.68%	0.004	0.82%	0.007
1	8.32%	0.016	13.43%	0.026
2	39.31%	0.031	22.19%	0.029
3	12.51%	0.020	2.74%	0.008
Total	60.82%		39.18%	
N	838			

Note. The table shows the estimated proportion (and standard error) of respondent types, $\hat{\pi}_{yz}$, characterized 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 organizations (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 marginalized groups have and enjoy their sexual and reproductive health and rights. The Alliance, in collaboration with partner organizations 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 talk about the NGO program, we refer to the UFBR program implement in Ethiopia where AMREF was the leading partner organization. The area where to intervene 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 utilization of health services at community level. Specifically the project is aiming at:

- Objective 1: Increased quality and delivery of comprehensive sexuality education
- Objective 2: Increased utilization 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 centers, 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 counseling 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, 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 endline report 04, pp. 257-381 (Stichting-Gezamenlijke-Evaluaties, 2015, `https://www.partos.nl/joint-MFSII-evaluations`). This report concerns an impact study of the project.

Appendix B Data details

Appendix B.1 Sampling strategy

Since the primary objective of the project is to change the behavior of households through information dissemination and behavioral 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 when possible from the same household.[32]

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 from Afar zones 3 and 5 individuals targeted by the intervention, while the interviewees from Afar zone 1 had no access to any of the services supported by the intervention. Zone 1 was selected taking

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

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 of woredas was not necessarily random because of the limited number of woredas 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, and Telalak from zone 5, and Mile and Chifira from zone 1 (see the map, Figure 1).

Selection of Kebeles. Kebeles are stratified in 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 based on project coverage. We selected kebeles which 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 out of the age range, we replaced the household with the neighboring household.

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

Balance tests between the NGO targeted and non-targeted groups are reported in Table B1. Some of the respondent's characteristics differ in the two groups. It is important to notice 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 both when 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	Control mean	Treatment 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 FGM (1=yes; 0=no)	0.362	0.265	0.004
List experiment answer	2.084	2.122	0.452
N	282	566	

Appendix B.2 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.

Appendix B.3 Enumerator selection and training

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

The enumerators were trained by our partner IFPRI-ESARO to make sure that the survey questions were understandable and well phrased. Adjustments to the survey were made before starting 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 around during the interview. The interview took place in an area near to the home of the interviewee. Two supervisors and 10 enumerators were hired and trained. Two teams were going to different woredas in the same zone and at the same time. Each team conducted the interviews in 3 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, in particular the Ethiopia endline report 04, pp. 257-381 (Stichting-Gezamenlijke-Evaluaties, 2015) and the survey included in this report (`https://www.partos.nl/joint-MFSII-evaluations`). This report concerns an impact study of the project. The list experiment data were collected as part of the baseline study for this evaluation.