Pre-analysis plan.
Violence against women: men's behaviors and opinions in Mali.

Olivia Bertelli*† Thomas Calvo† Emmanuelle Lavallée† Marion Mercier†

Sandrine Mesplé-Somps†

June 21, 2023

Abstract

Gender-based violence (GBV) is still widespread across the world. Most of the literature focuses on the victims of GBV, mostly women, but rarely on the perpetrators, mostly men. This project studies men's behaviors and opinions related to violence against women in Mali, one of the countries with the highest rates of GBV in the world. In collaboration with a local scientific partner, the GREAT, we conduct a List Experiment (LE) over a sample of 1,200 men in Bamako, the capital city. Half of the sample is randomly assigned to a treatment group and the other half to a control group. To the former, we administer a questionnaire containing eight lists of items, each one containing a different sensitive item about a violence-related behavior or opinion. To the latter, we administer the same lists but without the sensitive items and we ask, instead, the sensitive questions in a standard direct form. The result of the List Experiment will provide reliable measures of the proportion of violent men and of men in favor of various forms of GBV (*e.g.* physical and sexual violence, genital cutting, child marriage). The comparison between the prevalence rates measured with the LE and with the Direct Questioning (DQ) will provide the size of the reporting bias present in our sample. We will then explore heterogeneity of response bias across individual characteristics.

*Corresponding author. Email: olivia.bertelli@dauphine.psl.eu

†DIAL-LEDa, UMR CNRS 8007 & IRD 260, Université Paris-Dauphine, PSL Research University.

Introduction

This plan outlines a survey experiment about men's behaviors and opinions on violence against women conducted in Bamako in the 2022 Fall. Section 1 describes the research question and the context. Section 2 presents the sample selection, the respondents replacement protocol and the randomization design. Section 3 enumerates the hypotheses to be tested and primary outcomes, and discusses the causal chains involved. Section 4 outlines the empirical specifications for data analysis, and Section 5 provides details on the robustness checks we will perform.

1 The setting

This research project studies the prevalence of behaviors and opinions related to violence against women among the male population of Bamako using two survey techniques: a List Experiment (LE) and a Direct Question (DQ) technique. By comparing the prevalence rates obtained with the two survey techniques, we explore whether responses provided under a standard direct question technique are biased.

In a list experiment, respondents are given a list of statements and are asked to declare how many of these statements they experienced (or agree with), without specifying which ones. Respondents are randomly allocated to two groups. In the control group respondents are asked about a list of J so-called "baseline" items, while in the treatment group respondents are asked about a list of $J + 1$ items, the additional item being the sensitive item that is the research object. Given the random allocation of respondents into the two groups, the prevalence rates of the baseline items should be equal across groups. The difference between the average number of items declared in the treatment group and in the control group thus provides an estimate of the prevalence of the sensitive item in the sample.

Our survey contains eight list experiments, each one presenting a sensitive item about men's behaviors and opinions regarding violence against women. Each sensitive item is also administrated to the control group under the form of a direct "Yes/No" question. The lists and sensitive direct questions are reported in Table 1. In each list the sensitive item appears in italics in Table 1. Sensitive direct questions are administrated only to the control group.

This study explores multiple GBV dimensions. Four items concern the perpetration of violence (verbal violence (LE 1), physical violence (LE 2), sexual violence (LE 3), and economic control (LE 4)) and the other four concern opinions about GBV (child marriage (LE 5), female genital cutting (LE 6 and LE 7), and physical intimate partner violence (LE 8)).

The survey questionnaire contained also a preliminary section with few socio-demographic characteristics of respondents administrated before the LE. The data contain information about respondents age, the highest educational level achieved, relationship to the household head, marital status, co-habitation with wife (or wives if polygamous marriage), presence of adult women in the household, mother-tongue. These characteristics will be part of the empirical analysis exploring heterogeneity in response biases, as discussed in Section 4.

Table 1: Lists and questions about gender-based violence behaviors and opinions.

	Treatment group's List (LE)	Control group's Direct questions (DQ)
1.	*If married:* - I rarely get angry. - It is the man who should have the last word at home. *- Sometimes my words make my wife cry.* - Women should be able to access the same jobs as men.	*Do your words ever make your wife cry?*
2.	- My father always helped me when I needed it. *- I have already hit one or more women.* - Men should comfort babies as much as women do. - A woman should always go out accompanied.	*Have you ever hit one or more women?*
3.	- I know people who are in favor of abortion. - No one ever divorced in my family. - A man and a woman should decide together about their contraception. *- I have already forced a woman to have sex with me.*	*Have you ever forced a woman to have sex with you (while she didn't want to)?*
4.	- My neighbors have financial problems - It's up to the women to do the housework *- It's up to the man to decide how his wife's money should be spent* - I know people with HIV	*Is it up to the husband to decide how his wife's money should be spent?*
5.	- I trust people in my neighborhood - I prefer to eat aside from the women in my household *- It's better for a girl to get married before the age of 18* - If I were to get married today, I would choose a woman as much as educated as me	*Is it better for a girl to get married before the age of 18?*
6.	- I recently attended a wedding of a relative - We often argue in my family *- A good wife is an excised wife* - I trust women in politics	*Is a good wife an excised wife?*
7.	- Boys must participate in household chores *- If I were to have a daughter today, I would have her circumcised* - I know one or more unfaithful men -I know one or more people who had Covid	*If you had a daughter today, would you have her circumcised?*
8.	- My parents always got along - In the evening, women are not safe in the street - Corporal punishment should never be used in school *- A wife who does not respect her husband deserves to be hit*	*Does a wife who doesn't respect her husband deserve to be hit?*

2 Pilot, sample, replacement protocol and randomization

2.1 Pilot

We first conducted a pilot survey to explore the variation of baseline items and to assess the validity of the lists of items. The success of LE rests on the assumption of no design, ceiling and floor effects which greatly depends on the choice of baseline items (Blair and Imai 2012; Glynn 2013). Baseline items should neither be too common or too uncommon, nor too innocuous or too sensitive (Chuang et al. 2021). Besides, at least one baseline item should be negatively correlated to at least another baseline item and to the sensitive item.

Respondents were randomly allocated to one type of questionnaire (out of two). One included only direct questions (Pilot Questionnaire 1), while the other one included the eight lists of the treatment group (Pilot Questionnaire 2). In Questionnaire 1, all baseline and sensitive items were administrated under the form of a direct "Yes/No" question. We also included additional baseline items in a direct form to test for possible alternative baseline items candidates. Based on respondents' answers, we identify items reporting very little variation (*i.e.,* generating only 0s or 1s) and items with many missing answers (*i.e.,* very sensitive items). We replaced either types of items with new ones or swapped baseline items across lists. We will compute pairwise correlations across items to further discuss the validity of the final lists. Furthermore, answers provided to Questionnaire 2 allowed us to assess the acceptance and level of understanding of our lists.

The face-to-face pilot survey was conducted by the GREAT (*Groupe de recherche en économie appliquée et théorique*) in Bamako in October 2022. Using the same sampling method as for the main survey, but in different enumeration areas, fifty adult men were interviewed. Among them, forty answered the Questionnaire 1 and ten the Questionnaire 2.

2.2 Sample selection

The face-to-face survey was carried out in Bamako in November 2022 among 1,200 adult men. The survey was conducted by a team of local male enumerators. The survey took place in 150 enumeration areas (EA) of the six *communes* of Bamako. The EAs were randomly selected from the list of EAs established by the National Statistics Office (INSTAT) during the mapping phase of the last General Population Census (2022). Within each EA, we randomly determined a starting point from which a random walk protocol was implemented to select the households to be surveyed. Every 5 doors an enumerator selected a household. In each household, enumerators randomly selected one adult male to be surveyed in the following way. They first listed the names of all adult men speaking Bambara or French and permanently living in the household and assigned a number next to each name. Then, they asked a household member to draw a card from a set of numbered cards to determine the person to be surveyed.

Selected individual had the opportunity to reschedule the interview to another date. Those refusing the interview were replaced by other persons living in other households selected in the

same EA with the same random walk protocol.

2.3 Randomization

Each enumerator had to interview 100 individuals, half of whom were assigned to the treated group and half to the control group. All enumerators alternated the control and treatment questionnaires. The first questionnaire was a control one if the enumerator ID number was even, while it was a treatment questionnaire if the enumerator ID number was odd.

Half of the sample was randomly assigned to the control and the other half to the treatment group. The difference between the two groups consists in the types of lists administrated. The 8 lists with the sensitive items were administered to the treatment group. In the control group the same 8 lists but without any sensitive items were administered. The sensitive items were administered instead in a standard direct question form.

The test of randomization balance will be conducted using the collected respondent's characteristics: age, education level, marital status, mother tongue, relationship with the household head, cohabitation with the wife or at least one of the wives and presence of adult women in the household.

We expect some individuals to refuse to participate in our survey. We will test whether survey participation is correlated to the treatment assignment. We also expect some individuals to decline to answer some questions. We will investigate whether refusal to answer to specific questions is correlated with respondents characteristics.[1] The same set of observable characteristics used for the randomization balance test will also be used in this case.

3 Hypotheses, primary outcomes and causal chain

The hypotheses at the core of this analysis are the following:

1. Questions about gender-based violence behaviors and opinions are sensitive and, when administrated in a standard Direct Question survey technique, the answers will likely suffer from response bias.
2. Response biases are not uniform in the population but vary across sub-groups.
3. Individual characteristics correlate differently with the prevalence rates of gender-based violence behaviors and opinions across interview techniques.

Our key outcomes of interest are the prevalence rates of eight gender-based violence behaviors and opinions. They will be measured in two ways, depending on the survey technique used. For the LE, the prevalence rate of each sensitive item will be retrieved based on the comparison

[1] We will not test whether refusal to answer specific questions is correlated to treatment assignment, as only respondents in the control group were given the option to skip specific direct questions.

between the counts declared by respondents in the treatment and control groups. For the DQ, the prevalence rates will correspond to the average answers to the eight Yes/No questions.

The causal chain involved in hypothesis 1 lies upon the fact that the DQ-technique, by asking respondents to fully disclose personal information, generates declarative biases that are, instead, mitigated by the LE-technique. We argue that revealing personal behaviors and opinions about violence against women is sensitive for men in Mali and thus likely to be affected by those biases when directly asked. Tourangeau and Yan (2007) find that misreporting in DQ is mostly due to respondents trying to avoid embarrassing themselves, and to avoid repercussions from third parties, which can happen in case of lack of confidentiality. Due to social desirability bias, we expect respondents to under-report opinions and behaviors that deviate from what they perceive as the prevalent social norm and to over-report socially desirable opinions and behaviours. Moreover, if respondents fear a lack of confidentiality, they may under or over-report sensitive opinions or behaviors to keep the information private, and under-report reprehensible opinions or behaviors to avoid possible retaliation from third parties. The LE-technique, by avoiding respondents to reveal any specific personal information, is deemed to avoid these sources of declarative bias by mitigating respondents potential discomfort in answering direct questions about sensitive issues.[2]

The second hypothesis we make (hyp. 2), is that response biases are not uniform in the population but vary across sub-groups. Socio-demographic characteristics are likely to influence the social norms to which individuals aim at complying; the degree to which they are attached to this norm; and the feeling of shame that they associate with certain behaviors – hence the social desirability bias. Individual characteristics might also determine their fear of a lack of confidentiality, their perception of the possible consequences of their answers being disclosed to third parties, and/or their willingness to keep their personal information private.

To explore heterogeneity in response biases we will account for the respondents observable characteristics collected in the socio-demographic module of the survey questionnaire. We will explore four main dimensions of heterogeneity across all our GBV-related outcomes: respondent education, age, the relationship with the women in the household, and ethnicity, proxied by mother tongue.

The existing literature highlights some heterogeneity of response bias on questions related to GBV based on women education. For instance, De Cao and Lutz (2018) find that support to FGM in Ethiopia is substantially under-reported, especially by uneducated women. Joseph et al. (2017) and Agüero and Frisancho (2022) show that women education is positively related to the under-reporting of intimate partner violence (IPV) in India and in Peru, respectively. Looking at available data from standard surveys with direct questions suggests a negative correlation between men education and GBV opinions or practices. For instance, 15.6% of men with secondary or more educational level justifying beating their wife if she refuses to have sex, against 27.8% men without any education, according to the Malian DHS (2019) data. However, these

[2] Other sources of declarative bias linked, for instance, to the misunderstanding of the questions or to the lack of memory can exist in both types of survey methods. It is also possible that measurement errors are amplified by the LE if respondents misunderstand the exercise.

answers could suffer from response biases. For instance, educated men might feel a stronger pressure to comply with the mainstream discourse condemning GBV than low educated men. We will use the variable on the highest level of education achieved to explore whether men's response bias on GBV-related questions varies according to their level of education.

Turning to other respondent characteristics, Cullen (2022) shows that women's bargaining power in the household (proxied, among other, by her age, education and marriage type) is negatively correlated with IPV experience in Nigeria. In particular, highly-educated women under-report IPV experience when directly asked. In India, Joseph et al. (2017) report that respondents from the youngest and oldest age cohorts under-report domestic violence more than those of the middle-age cohort. Building upon this literature, we expect that respondents age, marital status, type of relationship with the household head and with the women in the household might correlate with response bias. Our socio-demographic module allows us to directly observe respondents' age. Regarding the relationship with the women in the household, three dimensions can be explored: marital status (married or not, monogamous or polygamous union), the presence of at least one adult woman in the household, the relationship with the household head. We will assess whether these variables are pairwise correlated. If they are highly correlated among themselves, we will conduct the heterogeneity analysis focusing on one of them and may rely on the other ones for robustness checks. If, instead, they are not strongly correlated among themselves, we will explore the various dimensions of heterogeneity side by side.

Last, a growing literature shows the link between ethnicity and GBV. For instance, Alesina, Brioschi, and La Ferrara (2021) find that certain ancestral traits of ethnic groups, such as endogamy or virilocal residence are correlated with higher levels of IPV prevalence and acceptance. FGM practice is also shown to vary considerably across ethnic groups. Based on data from 2009, Diabate and Mesple-Somps (2019) observe that the rate of FGM among girls below the age of 14 is close to or even outreaches 90% in the Bambara, Soninke, Malinke, Senufo, Fulani and Dogon groups, while it is less than 50% among Bobo and Songhai girls. Moreover, in Ethiopia De Cao and Lutz (2018) show that the response bias concerning support to FGM differs between the Afar and the other ethnic groups. It is then likely that response bias on GBV-related opinions and behaviours is also ethnic group-specific, something that we will explore using the respondent mother-tongue as a proxy for ethnic identity.

Finally, the third hypothesis we intend to investigate is whether individual characteristics differently correlate with prevalence rates depending on the survey technique used. Heterogeneous response biases might, in fact, lead to biased estimates of the individual determinants of GBV-related behaviours and opinions when measured with the DQ-technique. For instance, opposite response biases among low-educated and high-educated respondents might show a "fake" significant correlation of education with the DQ prevalence rates, but not with the LE prevalence rates. Building upon the results regarding hyp. 2, we plan to explore hyp. 3 focusing on respondents education, age, marital status, relationship with the women in the household, and ethnicity.

4 Specifications

To test hypothesis 1, we will compare the prevalence rates computed with the LEs with those computed with the DQ-technique.

The prevalence rate of sensitive item k measured with the DQ-technique, $\hat{g}_k$, will be obtained with a regression of the direct answers on a constant:

$$Z_{k,i} = g_k, \tag{1}$$

where $Z_{k,i}$ equals 1 if individual i answers 'Yes' to sensitive question k, and 0 otherwise.

The prevalence rate of sensitive item k measured with the LE-technique, $\hat{\gamma}_k$, will be measured based on the comparison of the control and treatment groups. In particular, we will estimate the following equation:

$$Y_{i,k} = \alpha_k + \gamma_k T_i + \lambda_k W_i^{'} + \epsilon_{i,k} \tag{2}$$

where $Y_{k,i}$ denotes the response given by individual i to the list of items containing k ($Y_{k,i}$ ranges from 0 to 3 for respondents of the control group and from 0 to 4 for respondents of the treatment group), T_i is a dummy variable equal to 1 if i is in the treatment group and to 0 if i is in the control group, and $\epsilon_{i,k}$ is the error term robust to heteroskedasticity. Moreover, we will introduce a vector of enumerator fixed-effects ($W_i^{'}$) to account for potential systematic differences across enumerators.

Testing hypothesis 1 consists in testing whether $\hat{g}_k$ and $\hat{\gamma}_k$ are significantly different from each other. For each of the eight sensitive questions, the difference between the LE and DQ prevalence rates can be interpreted as the size of the declarative bias the LE-technique neutralizes, showing the added value of using this technique rather than the DQ.[3]

Equation 2 yields an accurate estimate of the prevalence of item k provided that the assignment to treatment is random. Balance tests of randomization will be performed on the set of observed individual characteristics (age, educational level, marital status, ethnic group proxied by mother tongue, cohabitation with the wife, or at least one of the wives, cohabitation with other women, relationship with the household head).

To test hypothesis 2, we will conduct an heterogeneity analysis by interacting the treatment dummy with individual characteristics, one at a time. This will allow us to identify whether response bias varies across respondents characteristics. We plan to compute prevalence rates by education level (whether primary / secondary education was completed or not) and age (two categories, less or more than the median age of the sample). We will also explore heterogeneity depending on the respondents type of relationships with the women in the household, relying on a set of alternative variables – namely, marital status (married or not, monogamous or polygamous union), whether at least one adult woman lives in the household, and whether the

[3] By construction, LEs suffer from higher variance than the DQ-based results. The possible lack of statistical significance of the gap between the LE-based and DQ-based prevalence rates should thus be interpreted with caution.

respondent is the household head. Finally, we will compute prevalence rates by ethnic groups, differentiating relatively conservative groups (based on Diabate and Mesple-Somps (2019)) from the others.

We will follow Blair and Imai (2012) to test whether the response bias (still measured by comparing prevalence rates across the two survey methods, LE and DQ), varies across individual characteristics. The DQ prevalence rate of each of the eight sensitive questions will be estimated with the following equation:

$$Z_{i,k} = g_k + \mu_k X_i + \epsilon_{i,k} \tag{3}$$

where $Z_{i,k}$ denotes the answer given by individual i to sensitive question k and X_i is one of the dummies of interest (education, age category, type of relationship with women in the household, conservative ethnic group). The DQ prevalence rate of sensitive question k is then given by $\hat{g}_k$ for the omitted category of variable X_i, and by $\hat{g}_k + \hat{\mu}_k$ for the category under study.

The LE prevalence rate of each of the eight sensitive questions will be estimated with the following equation:

$$Y_{i,k} = \alpha_k + \beta_k X_i + \gamma_k T_i + \delta_k X_i \times T_i + \lambda_k W_i^{'} + \epsilon_{i,k} \tag{4}$$

where we interact the treatment status T_i with one of the dummies of interest. The estimated term $\hat{\gamma_k}+\hat{\delta_k}$ captures the prevalence rate as measured by LE among individuals with $X_i = 1$ while $\hat{\gamma_k}$ measures the prevalence for those with $X_i = 0$. As before, $W_i^{'}$ is the vector of enumerators fixed-effects.

The second empirical step consists in comparing $\hat{g}_k$ and $\hat{\gamma}_k$ on the one hand, and $\hat{g}_k + \hat{\mu_k}$ and $\hat{\gamma}_k + \hat{\delta_k}$ on the other hand, based on the coefficients estimated in Equations 3 and 4. Given that there are eight different item-count questions, eight regressions will be performed.

To test hypothesis 3, we will explore the correlation between individual characteristics and prevalence rates measured with DQ or LE. Individual characteristics might significantly correlate with prevalence rates measured with one technique but not the other, especially if response biases are heterogeneous across sub-groups of the population, as the test of hypothesis 2 will investigate. In other words, the results about hypothesis 3 will provide suggestive evidence about possible consequences of heterogeneous response biases in regression analysis when using a standard direct question technique. To run this analysis we will consider the terms estimated separately for each survey technique: $\hat{\mu_k}$ (Eq. 3) for the DQ and $\hat{\delta_k}$ (Eq. 4) for the LE. While $\hat{\mu_k}$ captures the correlation between respondents characteristics and the DQ-measured prevalence rates, $\hat{\delta_k}$ captures the same correlation but for the LE prevalence rates. We will perform standard linear tests on the coefficients to assess whether the correlations of respondents characteristics with the prevalence rates are statistically significantly different from one survey technique to the other.

5 Robustness checks

5.1 Alternative specifications

We will test whether results are robust to a list of control variables including respondent age, education, ethnic group, marital status and presence of female household members. The variables for the randomization balance check and for the heterogeneity analysis will be mainly dummy variables: gender, whether the individual completed primary / secondary school, whether the respondent is married or not, whether he belongs to the relatively conservative ethnic groups and whether there is at least one adult female household member.

Besides, we will combine sub-sets of questions into indexes to increase power, to limit multiple hypotheses testing and to detect a possible overall effect. We will build a "GBV opinion index" (LE4, LE5, LE6, LE7, LE8), a "GBV behavior index" (LE1, LE2, LE3) and an "FGM opinion index" (LE6 and LE7). Two versions of the indexes will be used. The first one will simply sum the number of items declared by respondents. The second one will create standardized indexes where we will place lower weight on questions with more variance by dividing the number of "Yes" items in each list by that same list's standard deviation.

5.2 Multiple hypotheses testing

Given the number of estimations we run, our results are subject to null hypotheses false rejections. Indeed, in step-1 of our analysis, we will be testing for 8 hypotheses (for the eight sensitive items) while in step-2 we will be testing for more than 80 hypotheses (8 sensitive items times the 10+ individual characteristics tested). For both steps of our analysis, we will resort to two methods to adjust the p-values for multiple hypotheses testing. We will compute q-values adjusted for false discovery rate following Benjamini, Krieger, and Yekutieli (2006) and Anderson (2008). We will also adjust p-values controlling for family-wise error rate following Romano and Wolf (2005b) and Romano and Wolf (2005a). The latter method, while more restrictive, accounts for dependence among the p-values.

5.3 Survey participation and refusal to answer

We define the rate of survey participation as the ratio between the number of respondents who accepted to participate to the survey, over the number of individuals that we reached. We will test whether survey participation is correlated with treatment status.

We will then explore the possible consequences of missing values for specific questions. We will first estimate the determinants of refusal to answer to sensitive questions under DQ. For each of the eight sensitive indicators, we will estimate a linear probability model of refusal to answer on an extensive set of individual characteristics and enumerators fixed-effects.

Second, we will explore whether refusal to answer affects the measurement of response bias. We will first test whether our main results are robust to excluding respondents who refused to

answer certain sensitive items under DQ. To do so, we will re-estimate the LE-based prevalence rate of each sensitive item (and the associated response bias) by dropping those respondents who refused to answer the corresponding direct sensitive question. Second, instead of dropping them, we will impute different DQ answers to sensitive items. In particular, we will successively assume that, had they answered, they would all have answered Yes, all have answered No, or they would have answered Yes in a proportion in line with the results of our list experiment. While relying on arbitrary hypotheses, these three exercises aim at illustrating the extent to which the response bias we measure is sensitive to refusal to answer standard direct questions.

References

Agüero, Jorge M and Veronica Frisancho (2022). "Measuring violence against women with experimental methods". *Economic Development and Cultural Change* 70.4, pp. 1565–1590.

Alesina, Alberto, Benedetta Brioschi, and Eliana La Ferrara (2021). "Violence Against Women: A Cross[U+2010]cultural Analysis for Africa." *Economica* 88.349, pp. 70 –104.

Anderson, Michael L. (2008). "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association* 103.484, pp. 1481–1495.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli (2006). "Adaptive linear step-up procedures that control the false discovery rate". *Biometrika* 93.3, pp. 491–507.

Blair, Graeme and Kosuke Imai (2012). "Statistical analysis of list experiments". *Political Analysis* 20.1, pp. 47–77.

Chuang, Erica et al. (2021). "Sex, lies, and measurement: Consistency tests for indirect response survey methods". *Journal of Development Economics* 148, p. 102582.

Cullen, Claire (2022). "Method Matters: The Underreporting of Intimate Partner Violence". *The World Bank Economic Review*.

De Cao, Elisabetta and Clemens Lutz (2018). "Sensitive survey questions: Measuring attitudes regarding female genital cutting through a list experiment". *Oxford Bulletin of Economics and Statistics* 80.5, pp. 871–892.

DHS (2019). *Enquête Démographique et de Santé au Mali 2018*. Ed. by CPS/SS-DS-PF et ICF INSTAT, p. 643.

Diabate, Idrissa and Sandrine Mesple-Somps (2019). "Female genital mutilation and migration in Mali: do return migrants transfer social norms?" *Journal of Population Economics* 32.4, pp. 1125–1170.

Glynn, Adam N (2013). "What can we learn with statistical truth serum? Design and analysis of the list experiment". *Public Opinion Quarterly* 77.S1, pp. 159–172.

Joseph, George et al. (2017). "Underreporting of Gender-Based Violence in Kerala, India".

Romano, Joseph P and Michael Wolf (2005a). "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing". *Journal of the American Statistical Association* 100.469, pp. 94–108.

Romano, Joseph P. and Michael Wolf (2005b). "Stepwise Multiple Testing as Formalized Data Snooping". *Econometrica* 73.4, pp. 1237–1282.

Tourangeau, Roger and Ting Yan (2007). "Sensitive questions in surveys." *Psychological bulletin* 133.5, p. 859.