Inter-American Development Bank

Systematic Bias in Sensitive Health Behaviors and Its Impact on Treatment Effects:

An Application to Violence against Women

Jorge M. Agüero
Verónica Frisancho

Department of Research and Chief Economist

DISCUSSION PAPER Nº IDB-DP-511

April 2017

Systematic Bias in Sensitive Health Behaviors and Its Impact on Treatment Effects:

An Application to Violence against Women

Jorge M. Agüero*
Verónica Frisancho**

* University of Connecticut
** Inter-American Development Bank

April 2017

IDB
Inter-American Development Bank

Abstract[*]

Violence against women takes place mainly in the private sphere and is perpetrated by people close to the victim. These features can introduce large biases into its reporting in specialized surveys as well as to the authorities. We test for the existence of measurement error in the reporting of such violence using experimental methods in Peru, a country with several specialized surveys but one lacking reliable administrative data. We ask women to report past experiences of violent acts by randomly assigning them one of two questionnaires, one that replicates current surveys and another that relies on list experiments to provide a more private setting. We find no significant reporting bias on average. However, we uncover strong evidence of non-random measurement error by education level. For highly educated women, an increase in privacy leads to higher reporting of violence, while no change is observed for the less educated. The increase is large enough to reverse the education gradient in violence. We discuss how non-classical error in the outcome variable affects the estimation of the role of risk factors on violence. In particular, randomized controlled trials underperform instrumental variables estimates and, under certain conditions, the former could lead to even larger biases compared to cross-sectional studies.

JEL classifications: B41, C83
Keywords: Violence, Women, Education, Experimental methods, Peru

[*] Agüero: University of Connecticut, Department of Economics and El Instituto; e-mail: jorge.aguero@uconn.edu. Frisancho: Inter-American Development Bank, Research Department; e-mail: vfrisancho@iadb.org.

1 Introduction

Violence against women has been identified as a major public health problem (e.g., Krug et al. [2002]; Bott et al. [2012]; Klugman et al. [2014]) and preventing such violence is now a key target of the 2015 Sustainable Development Goals ratified by 194 United Nations Member States. In economics, a growing number of studies has tried to identify the main drivers of this violence in order to design policies that can reduce its incidence in both developed (e.g, Aizer [2011]; Lindo et al. [2015]) and developing countries (e.g., Green et al. [2015]; La Mattina et al. [2017]).

Two features distinguish violence against women from other types of violence. First, violence against women is perpetrated by people they know, mainly, their partners or spouses. Men, on the other hand, experience violence mostly perpetrated by strangers due to crime, civil unrest, or terrorism, among other scenarios (Krug et al. [2002]). Second, violence against women tends to be invisible as much of it happens behind closed doors and in the privacy of the home, while other forms of violence, including wars and riots, are more visible and frequently broadcasted on television and other media (WHO, 2002).

These two features generate a large potential for reporting error in the measurement of the prevalence of violence against women (e.g., DeKeseredy and Schwartz [1998]; Ellsberg et al. [2001]; Kishor [2005]; Aizer [2010]). On one hand, these potential biases challenge current estimates of the prevalence of violence against women and cross-country comparisons as explored by Watts and Zimmerman [2002] and more recently by Abrahams et al. [2014]. Such biases could also affect, for example, the optimal allocation of public health budgets (e.g., Klugman et al. [2014]). On the other hand, the characteristics of these biases may raise concerns about the validity of the results identified in the literature exploring the drivers of violence. For most studies that focus on the identification of the risk factors associated with violence, the existence of measurement error in the outcome variable is not a major concern. It is a well-known econometric result that while random measurement error in a risk factor leads to attenuation bias, similar *classical* errors in the outcome variable do not

bias the estimated effects. However, when the error in the outcome variable is non-classical, identifying the causal effect of risk factors is not possible even if exogenous variation in the variable of interest is available.

In this paper, we measure the bias in reporting of violence against women using experimental methods. We compare prevalence rates of violence from the widely used Demographic and Health Surveys (DHS) to the rates obtained in our experimental approach that increases the level of privacy provided. In particular, we randomize the questionnaire applied to each participant. In the control group, we follow the DHS guidelines and participants were asked directly about nine specific events regarding psychological, physical and sexual violence. These results are compared against rates obtained from list experiments.

To provide a higher level of privacy, list experiments do not ask directly about sensitive events, but rather present respondents with a list of events and ask them to mention how many of those statements are true (e.g., **?**; Glynn [2013]; **?**). To be precise, in the control group, our questionnaire contains the nine DHS-type questions about violence plus nine sets of four non-sensitive statements (e.g., "Has you cell phone ever been stolen?"). Each respondent is asked to provide the surveyor the *number* of statements that hold true, without making any reference to which of them are true. In the treatment group, the questions about violence are added to the list of non-sensitive statements as the fifth statement (e.g., "Has your partner ever pulled your hair?"). The treatment group is asked to provide the number of statements that hold true, without telling the surveyor which of them are true. Randomization guarantees that the difference between the average number of statements that are true across treatment and control groups will capture the prevalence rate of the sensitive statement while protecting the privacy of the respondent. We apply these methodology to a sample adult women who are clients of a microcredit organization in several impoverished districts of Lima, Peru.

Our first result is that, on average, there are no significant differences in reporting across direct and indirect methods in all nine measures of violence. However, as our second result,

we find that the error varies with the level of education: women with complete tertiary education report higher rates of violence when we increase the privacy level of the interview compared to the traditional survey questions used in the literature. The increase is large enough to reverse the education gradient from negative when using traditional survey data –*more* education was associated with *less* violence– to positive under our experiment. We argue that this difference in reporting may be coming from other costs of being exposed, such as stigma costs, rather than reflecting better performance in answering the list experiment questions. For instance, if highly educated women are more able to understand the experiment, then we should expect *less* differences between treatments among these women and larger differences for the less educated. We found exactly the opposite patterns, thus, eliminating the "better understanding" mechanism as a possible explanation of our findings. Furthermore, college-educated women are exposed to diverse information flows during tertiary education, they are also more aware of deviations from broader social norms that severely sanction violence against women. This awareness is likely to induce underreporting due to stigma motives among the most educated in less private environments [Bharadwaj et al., 2015].

We discuss how our findings affect the existing and growing literature identifying the risk factors related to violence against women. We show that in the presence of non-classical measurement error in the outcome variable, randomized control trials (RCT) will underperform. RCTs provide credible sources of variation in risk factors and eliminate biases from omitted variables. However, when the risk factor (e.g., income or education) is correlated with the source of measurement error, a new source of bias appears reversing the original gains from randomization. Papers using instrumental variables (IV) will suffer the least from the biases introduced by non-random measurement error. In addition to reducing the problems arising from omitted variable bias, studies using valid instruments will be able to avoid the correlation between the risk factor and the measurement error because IV estimation depends on the correlation between the *instrument* and such error, which is less likely

to occur. Interestingly, we show that when the bias from omitted variables has the opposite sign compared to the relationship between the measurement error and the risk factor, cross-sectional estimates could *outperform* RCT approaches. We provide guidelines on how to avoid the limitations of RCTs in the presence of non-classical measurement error in the outcome.

Our results are not limited to the case of violence against women and have an ample application for all research questions where the dependent variable suffers from non-random measurement error. In that sense, our paper speaks to a large literature that has considered error-in-variables in income (Bound et al. [2001]) as well as health outcomes (Butler et al. [1987]). Furthermore, our experimental approach provides researchers with a simple strategy to test for measurement error, classical or not, in contexts where administrative records are not available. In that sense, our methodology complements recent alternative approaches that focus on qualitative methods to investigate the extent of measurement error (e.g., Blattman et al. [2016]).

The paper is divided in five sections including this introduction. The second section reviews the literature on misreporting when sensitive information is gathered. The third section provides details about the design of the experiment we conduct, describes the data and the sample, goes over the estimation strategy of the bias in measurement, and presents the results. The fourth section discusses the implications of these results when trying to identify the drivers of violence against women. The last section concludes.

2 Misreporting in Sensitive Survey Questions

There is an extensive literature showing that measurement error in survey data is not random but rather correlated with an array of risk factors. For example, Gottschalk and Huynh [2010] show that this source of bias in earnings leads to large biases in the measurement of inequality in the US. O'Neill and de Gaer [2004] explore the role of non-classical measurement error on

distributional analysis, focusing on unemployment and income. Several authors have shown that this problem is not limited to income data but it is also observed in health variables (see Bound et al. [2001] for a review). Butler et al. [1987] shows evidence of non-classical error in the measurement of arthritis. Johnston et al. [2009] finds a similar pattern in hypertension reporting. O'Neill [2012] expands that to the case of body mass index (BMI) and shows that such errors lead to an overestimation of the relationship between BMI and both income and education. More recently, Bharadwaj et al. [2015] compares survey and administrative data on mental health medication and finds that underreporting in the former is correlated with age, gender, and ethnicity.

The problem of non-classical measurement error is particularly worrisome whenever personal information is requested from the person interviewed. Biases can be introduced in the case of traumatic events (i.e. domestic violence) or sensitive issues (i.e., racism or physical appearance) but are also frequent in the case of wages or income data.

As discussed in the introduction, several scholars have argued that measures of violence against women could be subjected to reporting error (e.g., DeKeseredy and Schwartz [1998]; Ellsberg et al. [2001]; Kishor [2005]; Aizer [2010]). Ours, is the first paper that tries to measure misreporting in this outcome. In principle, the nature of violence against women imposes several costs in the case of being exposed as a victim. These costs prevent women to truthfully report their previous experience of violence whenever confidentiality levels are not high enough. First, there is an emotional cost that the woman may face due to her attachment to the offender and the potential sanctions (social or legal) that he may face. Second, women may also fear the potential loss of their partners' economic support if her status as a victim is revealed. Third, if exposed, the woman also faces the risk of retaliation due to an escalation of violence against them or their children. Finally, women may fear stigmatization, either from intrinsic or extrinsic sources [**?**].

Most likely, the costs of being exposed may affect women differently depending on their characteristics. These variation in costs could generate non-random misreporting patterns in

face-to-face surveys. Even when ethical and privacy protocols are enforced, the respondents still perceive certain risk of being exposed ex-post. Moreover, the fact that they have to reveal sensitive information to a person they do not know may further discourage them from reporting truthfully.

Also, note that unlike other health outcomes, administrative data cannot provide a benchmark for the "true" measure of violence against women, especially in developing countries. First, using surveys from 24 countries in the DHS program, Palermo et al. [2014] shows that only forty percent of women who experienced such violence told someone about it. Furthermore, only 7 percent of them made a formal report that would be captured in administrative data (e.g., by the police or by medical or social services). Second, the authors show that women reporting violence are not a random sample of the population and that reporting instead depends on women's socioeconomic characteristics such as age, marital status, education, and urban location. Third, as we will show below, even these surveys are subjected to measurement error. That is, the rates of violence inferred from DHS-type surveys might underestimate the *true* values in the population, and could further increase the bias in administrative data compared to the estimates made by Palermo et al. [2014].

Thus, providing higher levels of confidentiality through indirect techniques is thus intended to reduce the biases in typical surveys [**?**]. The benefits of list experiments may be very sizeable in the case of violence against women since they keep the report of violence completely anonymous. It is worth mentioning that a limitation of such approach is that the prevalence of violence is only obtained as an aggregate rate.[1]

[1]See **?** for more details.

3 Measuring Reporting Bias in Violence Against Women

3.1 List Experiments: Design

List experiments have been traditionally used to gather opinions and/or record behavior related to inherently sensitive issues which are more prone to underreport. The basic design of a list experiment will feature a control group (C), who is only given a list of S neutral statements, and a treatment group (T), who receives the same list of $S+1$ statements, where the last one refers to sensitive issue. Both groups are asked to provide the *number* of statements that hold true, without indicating the ones that are in fact true. Comparison between the average number of true statements across both groups yields the prevalence rate of the sensitive statement while providing greater levels of confidentiality than when asked directly about the prevalence.

Let d_{is} denote individual i's response to statement s, where d_{is} takes the value of one when affirmative and zero otherwise. The number of responses that hold true for individual i will thus be given by $\sum_s^S d_{is}$ in the control group and $\sum_s^{S+1} d_{is}$ in the treatment group. Random assignment of the treatment at the individual level implies that:

$$E_i\left(\sum_s^S d_{is}|T\right) = E_i\left(\sum_s^S d_{is}|C\right)$$

That is, the control group provides the counterfactual of the number of true statements if the treatment group were to receive only S statements. In our setting, the sensitive items correspond to a particular act of physical violence such as kicking or slapping. Thus, the prevalence rate of a given form of physical violence can be measured as:

$$\rho = E_i\left[\left(\sum_s^{S+1} d_{is}|T\right) - \left(\sum_s^S d_{is}|C\right)\right] \tag{1}$$

where $E_i\left(\sum_s^S d_{is}|C\right)$ approximates $E_i\left(\sum_s^S d_{is}|T\right)$.

For the list experiments to effectively protect respondents' privacy while providing a good estimator of the prevalence rate, the selection of neutral statements is crucial. In particular, designing the list of statements has to take into account the trade-off between protecting the respondent and reducing the variability of the responses. On one hand, we would like to avoid a neutral list in which a very large share of the population is likely to respond $\sum_s^S d_{is} = S$, i.e. ceiling effect, since the respondent would no longer be protected. A similar situation occurs when the list contains low-prevalence items (i.e., $\sum_s^S d_{is} \approx 0$) that may deter the respondent to answer honestly.

On the other hand, a list that avoids the two problems stated above will most likely introduce greater variability in the responses, which could then increase the variance of the estimator. Glynn [2013] provides some guidance in the development of lists so as to maximize the level of protection while sacrificing little variance. He shows that introducing negative correlation between the responses to the neutral items in the list limits the variability of the responses while minimizing the likelihood of ceiling effects. Sub-Section 3.2 provides details on the efforts we undertake to minimize extreme values in the sets of statements used while maintaining low levels of variability in the responses.

Even if the design of the instrument to implement the list experiments if flawless, these methodology poses two important limitations. First, the training of surveyors and respondents' adequate understanding of these type of questions is fundamental. If the woman interviewed is not aware of the greater levels of confidentiality provided and/or if she is overwhelmed with the mechanics of the experiment, additional measurement biases can be introduced. Sub-section 3.2 provides details on the the strategies we followed to minimize this issue.

Second, the nature of the list experiments in itself does not allow the researcher to link prevalence rates to other respondents' characteristics. The anonymity provided to the respondent limits the usage of the methodology to provide only aggregate information. However, with large enough sample sizes one can measure prevalence rates by sub-samples as we

do here (see Sub-section 3.4) and learn more about the correlation between violence and risk factors.

3.2 Sample Description and Data

The population of interest for our study is composed by adult women (aged 18 and above), in Lima, who receive microloans from the Adventist Development and Relief Agency (ADRA), a international non-governmental organization (NGO) running a village banking program in peri-urban and rural areas in Peru. ADRA's clients are microentrepreneurs from the most impoverished districts in Lima such as San Juan de Lurigancho, Villa Maria del Triunfo, Villa El Salvador, Ventanilla, Huaycan, and Los Olivos.

From the total pool of 1873 clients in 112 village banks in ADRA's microcredit program in Lima, we first drop all under-aged clients as well as all women above 65. This leaves us with a remaining universe of 1776 clients. We then draw 6 banks at random and exclude them from the study to be able to rely on them for the piloting of the instruments, which leaves us with 1690 clients in 106 banks. Finally, we work with all banks with monthly meetings scheduled during July 2015 which restricts the universe of interest to 1562 women in 98 village banks. We targeted this universe and were able to interview 1223 women between July 1st and August 25th, 2015.

Clearly, the implementation of list experiments requires careful preparation both in terms of the training of surveyors as well as in terms of respondents' adequate understanding of these type of questions. With this in mind, we dedicated special attention to the design of the instrument, the selection and training of surveyors, and the application of the instrument.

First, we piloted the non-sensitive statements in a small sample of ADRA's clients who were not part of the experimental sample. We came up with a list of 41 statements and asked 31 individuals to provide a yes/no answer in order to measure the prevalence rates of each statement. The questions were framed without a time horizon in order to be in line with the sensitive items on violence which intended to measure prevalence rates in a woman's

lifetime.

The prevalence rates of the non-sensitive statements were useful in two ways. On one hand, they measured the adequacy of the statements for the particular setting we were working in. Statements with prevalence rates too close to zero were discarded later. On the other hand, the prevalence rates helped us decide how to group the statements in sets of four in order to minimize ceiling effects and reduce the variance of the estimator [Glynn, 2013].[2] Table A.2 in the appendix shows the prevalence rates of the 34 statements we kept for the list experiments, after removing those with very small prevalence rates.[3] Table A.3 in the appendix reports the correlation of prevalence rates in each set of statements grouped together.

Second, we carefully selected a team of female surveyors with previous experience on the topics of gender and gender biased violence. We then asked them to go through a three-day training workshop and selected the top performers after evaluating them during the practice sessions. The workshop itself included a sensitization session provided by a local NGO, *Centro de la Mujer Peruana Flora Tristán*, which works on gender issues and women's empowerment.

Third, we tried to minimize the chances for misunderstanding or confusion when applying the instrument by providing the respondents with visual aids during the interview. Depending on the randomization outcome, the surveyor provided each respondent with a printed copy of the list experiment questions. This allowed respondents to follow the list of statements read to them and helped them remember the number of positive answers as they went along the list.

Randomization of the treatment was done at the individual level. The questionnaire was implemented via tablets. Due to some initial complications with the software, we drop a few

[2]Based on the collected data on the correlation of responses across pairs of statements, we developed an algorithm that tried to induce negative correlation within the list of non-sensitive statements. First, we chose a grouping that minimized correlation between pairs of statements. Second, we grouped pairs of statements based on this optimal negative correlations and checked the correlation in the full list was still negative.

[3]Two statements used in the final instrument were not tested in the pilot.

surveys which were incorrectly assigned to answer the list experiment questions from both treatment arms and are left with a sample of 1078 valid surveys.[4] According to our power calculations, this sample was large enough to detect an effect as small as 0.03 percentage points between the treatment and control groups.[5]

A key advantage of our paper compared to similar studies is our large sample size, which allows us to have separate questionnaires for the treatment and control groups. This allows us to reduce potential biases that may be introduced when asking the same respondent both the direct and the list experiment questions as done in **?**. Although we are able to explore differential misreporting by characteristics of the respondent, our study was not designed to be able to identify the costs that are driving the results.

Table A.1 in the Appendix confirms that the randomization was successful. There is only a small significant difference in the share of women that are household heads across treatment arms (at the 5% level). In any case, all our estimates include a full set of controls which includes a dichotomic variable indicating if the woman is the household head.

Table 1 reports the prevalence rates of ever experiencing different violent acts as collected by regular DHS surveys. Prevalence of emotional violence against women was collected for all the experimental sample while only the control group answered the questions related to physical violence. In the survey, we designed 9 direct questions and their corresponding list experiment version to indirectly measure prevalence rates of the following physical violence acts as inflicted by their actual or past partners: having her hair pulled; being pushed, shaken, or having something thrown at her; being slapped or having her arm twisted; being punched or something that may have hurt her; being kicked or dragged; being strangled o

[4]During the first three weeks of fieldwork, the randomization process was done by an offline version of the online platform we used to collect the data. Due to some complications with the software, which led some respondents to answer the two versions of the survey, we asked the surveyors to randomize using a pair of marbles from different colors during the rest of the fieldwork.

[5]Using the national DHS survey in Peru, we define the initial violence prevalence rates in the area studied. We decide to focus on one of the least frequently reported acts of violence, forced to have sexual relationships. Initial prevalence rate is set at 0.05 with a standard deviation of 0.2. With the randomization conducted at the individual level, a minimum detectable effect of 0.03 percentage points, a significance level of 10% and power of 0.8, the minimum sample size required was estimated at 550 per treatment arm.

burnt; being threatened with a knife, gun, or other weapon; being forced to have sex; and being forced to perform sex acts she does not approve of.

In general, prevalence rates are shockingly high in the context studied. Almost 80% of the women in our sample have ever experienced any type of violence, either emotional or physical. Prevalence rates for any type of emotional violence are about 0.64, close to the 0.62 prevalence rate reported for any type of physical violent act. Not only are prevalence rates high but those who are victims of violent acts tend to suffer from it quite often as reported in the last column of Table 1.

Table 1: Prevalence rates of violence against women (VAW)

	All Sample		Sample w/violence	
	N	Prevalence rate	N	High frequency
VAW	560	0.78		
Emotional VAW	1078	0.64		
Humiliate	1076	0.38	407	0.32
Insult	1074	0.35	373	0.33
Call lazy	1076	0.27	290	0.28
Threatens to harm	1076	0.15	162	0.38
Threatens to leave	1076	0.32	345	0.32
Physical VAW	560	0.62	.	.
Pull hair	560	0.31	170	0.24
Push	559	0.46	252	0.19
Slap	559	0.26	147	0.25
Punch	559	0.22	123	0.27
Kick	558	0.15	81	0.37
Strangle	560	0.06	30	0.33
Knife	560	0.06	32	0.22
Forced sex	559	0.23	127	0.36
Unapproved sex practices	558	0.09	51	0.37

NOTE: The prevalence of VAW is measured as the prevalence rate of any type of violence, emotional or physical. Similarly, the prevalence of emotional (physical) VAW is measured as the prevalence of any type of emotional (physical) aggression. The last column reports the share of women who reported experiencing a given violent incident with high frequency.

3.3 Estimation

A common source of data on violence against women are surveys from the DHS program. These surveys take into account all the ethical guidelines recommended by the WHO to measure violence prevalence rates by having an enumerator ask face-to-face questions about whether the respondent has experienced a list of violent incidents. Let the reported prevalence rates under DHS methods be denoted as p.

Let D_i be equal to the number of statements that hold true for an individual i, where $D_i = \sum_s^S d_{is}$ whenever i is assigned to the control group and $D_i = \sum_s^{S+1} d_{is}$ if i belongs to the treatment group. Let T_i denote the treatment assignment to the list experiment. The difference-in-means estimator approximates the prevalence rate of the sensitive statement included in the set of $S + 1$ sentences provided to the treatment group:

$$D_i = \alpha + \rho T_i + \xi_i \tag{2}$$

In the end, we are interested in estimating the level of misreport as measured by $(\rho - p)$ and testing whether this difference is positive and statistically significant. Since the control and treatment groups are, on average, equivalent in terms of their real prevalence rates, $(\rho - p)$ signals the existence of underreporting.

The model estimated with list experiments data can be further extended to capture prevalence rates for different sub-samples as defined by x_i:

$$D_i = \alpha + \rho T_i + \gamma x_i + \zeta T_i \cdot x_i + \xi_i \tag{3}$$

The term $(\rho + \zeta)$ captures the prevalence rate measured by experimental methods among individuals with $x_i = 1$ while ρ will measure the prevalence rate for those with $x_i = 0$. Again, we can compare these prevalence rates to their counterpart measure obtained through direct reporting, p, conditional on x_i.

3.4 Results

Although we execute the nine list experiments to measure prevalence rates of physical violence against women, we decide to analyze the data coming from only seven of these experiments. We drop the data for being pushed, shaken, or having something thrown at and being forced to have sex. Despite our efforts to group non-sensitive statements in a way that minimized ceiling effects and reduced the variance of the estimator, we envision some issues in the lists used in these two cases (see Appendix B for more details).

Our main goal is to measure if there are statistically significant differences in the report of violence across direct and experimental data collection methods. A positive gap between ρ and p would suggest that there is underreporting and that greater confidentiality levels reduce this effect.

Table 2 presents the estimated differences between indirect and direct reporting of ever experiencing different forms of intimate partner physical violence. In general, the results suggest that direct questions used in health surveys do not seem to introduce a bias in measuring the prevalence of violence when compared to experimental and more indirect methods that seek to offer greater levels of confidentiality. For six out of seven acts of physical violence, the prevalence rates obtained through experimental methods do not significantly differ from those measured using regular DHS-type questions. Only in the case of the prevalence rate of having their hair pulled, there is a positive and significant difference favoring underreporting via direct methods. However, the joint test that the seven gaps are different from zero is rejected, providing little evidence to suspect of reporting biases on average.

Although the average treatment effects are not different from zero, it may well be the case that certain vulnerable groups are more likely to report violence more accurately due to the increased confidentiality provided by the list experiments. We next explore such potential outcomes relying on (3). In particular, we try to measure heterogeneous effects depending on the level of economic and social empowerment of the respondent since we expect the costs of truthfully reporting to vary by it. For example, more economically empowered women may

Table 2: Difference in estimated prevalence rates of physical violence against women

Violent act	List experiments (ρ)	Direct reporting (p)	($\rho - p$)	
Pull hair	0.418	0.311	0.107	*
Slap	0.170	0.265	-0.094	
Punch	0.174	0.224	-0.049	
Kick	0.126	0.145	-0.019	
Strangle	-0.022	0.055	-0.077	
Knife	0.046	0.057	-0.011	
Sex acts	0.052	0.095	-0.043	
Joint test				
χ^2		8.12		
Prob $> \chi^2$		0.322		

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

be more likely to face lower costs of being exposed due to fear of their partners' economic support.

At first glance, we only find evidence of misreporting among the most educated women in the sample. Among those with completed tertiary education, Table 3 shows that there are large positive gaps between prevalence rates reported under indirect and direct methods. In turn, there are not many important differences across methods when focusing on the groups of less educated women.

The measured bias among the most educated women is large enough to reverse the education gradient in violence. Direct reporting seems to produce a negative correlation between education level and prevalence rates (see panel a in Figure 1). However, once provided greater levels of confidentiality the experimental method generates sizeable increases in the report, especially among the more educated. In fact, women with completed tertiary education are more likely than less educated women to report having been victims violence when list experiments are used (see panel b in Figure 1).

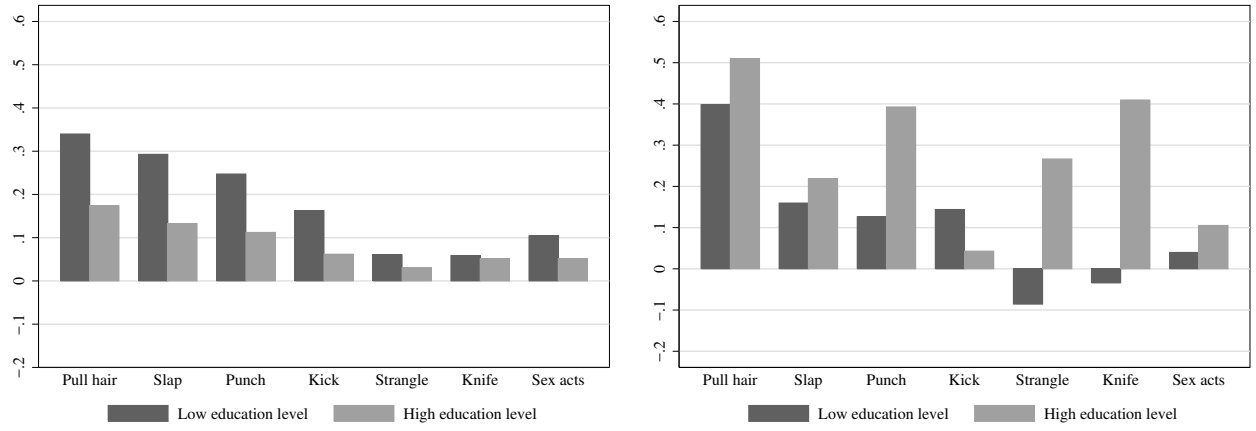
Surprisingly, no other measure of empowerment (or the lack of it) seems to be correlated

Table 3: Difference in estimated prevalence rates of physical violence against women by education level

	Less than tertiary education				Tertiary education			
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$	
Pull hair	0.398	0.340	0.058		0.510	0.173	0.336	**
Slap	0.160	0.293	-0.133	*	0.219	0.133	0.086	
Punch	0.126	0.247	-0.121		0.393	0.112	0.281	*
Kick	0.144	0.163	-0.019		0.043	0.062	-0.019	
Strangle	-0.086	0.061	-0.146	**	0.267	0.031	0.236	*
Knife	-0.034	0.058	-0.093		0.410	0.051	0.359	***
Sex acts	0.040	0.104	-0.065		0.105	0.051	0.054	
Joint test								
χ^2		10.62				22.02		
Prob $> \chi^2$		0.156				0.003		

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Figure 1: Physical Violence Prevalence Rates by Reporting Method and Education Level



(a) Direct reporting (p)

(b) Indirect reporting (ρ)

NOTE: High education level is defined as completed tertiary education.

with significant biases in the report of violence via direct survey questions. Table 4 reports the joint significance tests that the bias in the seven acts of physical violence reported is different from zero by sub-samples. While some modest differences emerge in the sub-samples

of single women and those with low loan size or low levels of savings in ADRA, these do not seem to follow a pattern as in the case of education. Table A.5 in Appendix A shows that even though the biases are jointly and significantly different from zero, no specific bias among single women is statistically significant. Moreover, the differences identified by the client's standing in ADRA do not follow a specific pattern and are only significantly different from zero for one or two acts of violence (see Tables A.10 and A.11 in Appendix A). The evidence is thus telling us that there is something particular about education, unrelated to empowerment, which is generating higher costs of truthfully reporting being victims of intimate partner violence when asked directly.

We rule out that the effect among more educated women is capturing a better understanding of the list experiment questions. If we focus on other variables that may proxy better understanding of the list experiments, no significant biases are identified. Indirect methods do not generate important differences in the report neither among women who speak Spanish as their mother tongue nor among those who have better memory to remember their answers to all statements provided (see Tables A.6 and A.7 in Appendix A).

4 Implications

Our results can be summarized in two statements. First, on average, there does not seem to be significant differences in reporting across methods with varying levels of confidentiality. Second, significant underreporting is observed among women with complete tertiary education. In this section we discuss how our findings affect studies that seek to evaluate the impact of a risk factor such as education level on violence against women.

To understand the estimation problems arising in the presence of non-classical measurement error in outcome variables consider a simple model. Suppose that a researcher wants to estimate the impact of x_i on the outcome y_i as described by the following equation:

Table 4: Joint significance test of $(\rho - p)$: Heterogeneous effects

	χ^2	Prob $> \chi^2$
Age		
<50	4.124	0.765
50+	8.219	0.314
Civil status		
Single	13.436	0.062
Married	4.318	0.742
Education level		
Less than tertiary	10.617	0.156
Completed tertiary	22.018	0.003
Mother tongue		
Spanish	10.934	0.142
Other language	7.306	0.398
Memory test		
Low score	3.993	0.781
High score	6.598	0.472
Household head		
Not the head	8.781	0.269
Head	4.729	0.693
Employment		
Does not work	6.218	0.515
Works	6.481	0.485
Loan size in ADRA		
Low	16.087	0.024
High (p75+)	9.319	0.231
Savings balance in ADRA		
Low	12.842	0.076
High (p75+)	4.810	0.683

NOTE: Joint test that the seven biases are different from zero. See Table 3 for details about the regressions.

$$y_i = \beta x_i + \epsilon_i \qquad i = 1, \ldots, N. \tag{4}$$

In our particular case of interest, y_i would capture a measure of violence against women and x_i would represent women's education, her income or other "risk factors" explored in the literature. For simplicity, equation (4) assumes that both variables are measured in deviations from the mean and ignores the role that other variables can play in explaining violence against women.[6]

Now consider the case where variables capturing the prevalence of violence against women are prone to measurement error. In particular, we study the case where we only have data on a noisy measure of violence, $\tilde{y_i}$, instead if the true value y_i:

$$\tilde{y_i} = y_i + \nu_i \tag{5}$$

with $E(\nu_i) = 0$, that is, on average the violence if measured without error, as in our main finding. To fix some ideas, assume that the risk factor x_i is observable without error and that it is uncorrelated with ϵ_i. We will later relax the latter assumption and consider models with measurement error and omitted variables. In this simple model, ordinary least squares (OLS) estimates of β yield:

$$\hat{\beta}_{OLS} = \beta + \frac{cov(\nu_i, x_i)}{var(x_i)} \tag{6}$$

Equation (6) shows that the bias in OLS depends on the correlation between the measurement error, ν_i, and the risk factor x_i. Assuming that this correlation is zero, yields the well-known conclusion that classical measurement error in outcome variables does not bias the estimates of β and only affects the standard errors [Cameron and Trivedi, 2005]. Our goal is to compare different approaches used in the literature to estimate β and how their accuracy is affected by presence of non-random measurement error in the outcome variable.

[6]Bound et al. [1994] provide a general framework where x_i is a vector instead of a scalar.

4.1 Cross-sectional estimates

Several papers in the literature estimate (4) via ordinary least squares using only cross-sectional variation to identify the impact of risk factors on violence against women. Examples include Jewkes et al. [2002], Koenig et al. [2003], Breiding et al. [2008], Fulu et al. [2013], where demographic and socioeconomic variables are considered among a long list possible risk factors.[7] Under this research design, it is possible for the risk factor under study to be correlated with the unobservables explaining violence so that $cov(\epsilon_i, x_i)$ is not zero, due to omitted variable bias. Therefore, OLS estimates can be expressed by the following terms:

$$\hat{\beta}_{OLS} = \beta + \frac{cov(\epsilon_i, x_i)}{var(x_i)} + \frac{cov(\nu_i, x_i)}{var(x_i)} \tag{7}$$

That is, with non-classical measurement error, the omitted variable bias in OLS estimates is *amplified* when the correlation between x_i with the unobservables (ϵ_i) has the *same* sign as the correlation between x_i and ν_i, the source of misclassification. However, when these correlations have different signs, the additional bias from non-classical measurement error attenuates the overall deviation from the true β. For instance, if education creates a stigma so that women underreport violence ($cov(\nu_i, x_i) < 0$), as shown in our list experiments, but education is positively correlated with unobserved ability as expected in human capital models (e.g., Card [2001]), then the initial bias from omitted variables is *reduced* because these two sources of bias oppose each other.

4.2 Randomized controlled trials

Several papers use experimental data to identify the causal impact of different risk factors on violence against women. For example, in developing countries, Hidrobo and Fernald [2013], Hidrobo et al. [2016], Haushofer and Shapiro [2013], Angelucci et al. [2008], and Bobonis et al. [2013], among others, have used the random or near-random allocation of cash

[7] See Capaldi et al. [2012] for a recent review.

transfers to women as part of antipoverty programs to look at their effect on violence.[8] By introducing random variation in x_i, these papers are able to convincingly set $cov(\epsilon_i, x_i) = 0$. Thus, when y_i is accurately measured or when the error is classical, these studies obtain unbiased estimates of β. However, if receiving the transfer makes women more likely to misreport violence, due to the higher emotional costs that their new status might create, then $cov(\nu_i, x_i) \neq 0$ and OLS estimates in (7) are now biased. Interestingly, non-classical measurement error partially undoes the gains from randomization. This implies that, when $cov(\epsilon_i, x_i)$ and $cov(\nu_i, x_i)$ have opposite signs, cross sectional estimates may have *less* bias (in absolute terms) when compared to studies that randomize the risk factor of interest.

4.3 Instrumental variables

We now consider a third case, where researchers take advantage of the availability of an instrument z_i for x_i, as a way to account for the possibility of omitted variable bias. For example, Erten and Pinar [forthcoming] use a school reform in Turkey to evaluate the impact of women's education on the prevalence of violence. In that case, the IV estimate can be decomposed in the following terms:

$$\hat{\beta}_{IV} = \beta + \frac{cov(\epsilon_i, z_i)}{cov(x_i, z_i)} + \frac{cov(\nu_i, z_i)}{cov(x_i, z_i)} \tag{8}$$

Let us assume that the instrument is valid. In that case, it has a strong first stage ($cov(x_i, z_i) \neq 0$) and it is uncorrelated with unobservables ($cov(\epsilon_i, z_i) = 0$). This implies that the first possible source of bias in (8) is set to zero. However, and unlike RCTs, the only possible source of variation comes from the correlation between the *instrument* and the measurement error. The correlation between x_i and ν_i no longer introduces an additional bias.

For instance, consider the case where x_i is education and z_i comes from the changes in the compulsory schooling laws as in Erten and Pinar [forthcoming]. In this situation, even

[8] See also De Koker et al. [2014] for a review of RCT papers in the United States.

when educated women underreport violence, it is hard to expect that changes in the laws are correlated with the misreporting of violence; $cov(\nu_i, z_i)$ is thus zero or near zero. The use of a valid instrument could solve not only the problem of omitted variable bias but can reduce or even eliminate the bias due to non-classical measurement error in outcome variables.

The comparison of these three cases leaves us with some important recommendations for research seeking to estimate the role of risk factors on violence against women. First, papers using valid instruments, are less likely to be affected by the presence of non-classical misreporting of violence. Second, for those studies considering an experimental intervention, it is important to test whether misclassification is correlated with the policy instrument (e.g., education, income, empowerment). If that is the case, researchers could rely on the average treatment on the treated effects and use the randomization as an *instrument* rather than using the randomized variable to estimate intention to treat effects. This strategy would reduce the bias in their estimates of β.

5 Conclusion

Our paper uses experimental methods to measure reporting biases in violence against women, in a setting with low-quality administrative data. We contribute to the literature on measurement of violence against women twofold. We are the first to measure the bias in reporting violence against women by using indirect survey questions such as list experiments and comparing estimated prevalence rates to those obtained by DHS. Previous uses of the methodology have been applied to other type of sensitive issues such as sexual orientation, racism, support of weapons, among others.[9] We find that, on average, there are no significant differences in reporting across methods with varying levels of confidentiality.

Second, our findings show that underreporting in our sample is concentrated among college-educated women, who do not fit with the typical victim stereotype. This has important implications on the invisibility of violence that certain groups may suffer and the

[9]See Coffman et al. [n.d.] for a recent application.

targeting efforts conducted to prevent and combat intimate partner violence. More educated women seem to face larger costs of being exposed and thus require higher levels of privacy and confidentiality to make them feel safe enough to report their true status. Since this pattern is not identified among more empowered women as measured by other proxies, we speculate that more educated women are more prone to face higher stigma costs related with the anticipation and internalization of negative stereotypes against them.

Our paper also contributes to the strand of the literature that tries to measure the impact of individual characteristics on the prevalence of violence. We show that, even when random assignment in the risk factor is introduced, non-classical measurement error in the dependent variable biases the estimates of treatment effects. In fact, IV techniques outperform RCTs since they bypass the issue of the correlation between the risk factor and measurement error. With a valid instrument, the only possible source of bias comes from the correlation between the instrument and the measurement error, which is much less likely to exist. Moreover, under certain conditions, randomization could lead to even larger biases compared to cross-sectional studies.

We acknowledge that the external validity of our results is limited. However, in a setting with high prevalence rates, such as the one studied here, it would have been more difficult to identify underreporting since the local social norms may be more accepting of violence. But even in this setting we are able to find evidence of misreporting for a given group. Further research should explore whether the misclassification is larger in areas with lower prevalence rates and if the heterogeneous effects vary by context. This is particularly urgent given the growing number of studies that try to identify the main drivers of this violence in order to design policies that can reduce its prevalence in both, developed (e.g, Aizer [2011] Aizer [2010], and Lindo et al. [2015]) and developing countries (e.g., Hidrobo and Fernald [2013] Hidrobo et al. [2016], **?**, Bobonis et al. [2013], and La Mattina et al. [2017]). Our results suggest that one should be cautious when interpreting these results since the presence of non-classical measurement error in the dependent variable may lead to erroneous

conclusions.

It is worth highlighting that our design was implemented at a very low cost: we were able to survey 1221 women at a cost of US$8 per woman. This means that there are potentially important savings from this methods when compared to other procedures [Blattman et al., 2016] that require intensive qualitative approaches. This opens the possibility to replicate our design with other samples with different contextual characteristics.

References

Abrahams, N., Devries, K., Watts, C., Pallitto, C., Petzold, M., Shamu, S. and García-Moreno, C. [2014], 'Worldwide prevalence of non-partner sexual violence: a systematic review', *The Lancet* **383**(9929), 1648–1654.

Aizer, A. [2010], 'The gender wage gap and domestic violence', *American Economic Review* **100**(4), 1847–59.

Aizer, A. [2011], 'Poverty, violence, and health: The impact of domestic violence during pregnancy on newborn health.', *Journal of Human Resources* **46**(3), 518–538.

Angelucci, M. et al. [2008], 'Love on the rocks: Domestic violence and alcohol abuse in rural mexico', *The BE Journal of Economic Analysis & Policy* **8**(1), 1–43.

Bharadwaj, P., Pai, M. M. and Suziedelyte, A. [2015], Mental health stigma, Technical report, National Bureau of Economic Research.

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K. and Sheridan, M. [2016], 'Measuring the measurement error: A method to qualitatively validate survey data', *Journal of Development Economics* **120**, 99 – 112.
URL: *http://www.sciencedirect.com/science/article/pii/S0304387816000122*

Bobonis, G. J., González-Brenes, M. and Castro, R. [2013], 'Public transfers and domestic violence: The roles of private information and spousal control', *American Economic Journal: Economic Policy* pp. 179–205.

Bott, S., Guedes, A., Goodwin, M. and Mendoza, J. A. [2012], *Violence Against Women in*

Latin America and the Caribbean: A comparative Analysis of Population-Based Data from 12 Countries, Washington, DC: Pan American Health Organization.

Bound, J., Brown, C., Duncan, G. J. and Rodgers, W. L. [1994], 'Evidence on the validity of cross-sectional and longitudinal labor market data', *Journal of Labor Economics* **12**(3), 345–368.

Bound, J., Brown, C. and Mathiowetz, N. [2001], 'Measurement error in survey data', *Handbook of econometrics* **5**, 3705–3843.

Breiding, M. J., Black, M. C. and Ryan, G. W. [2008], 'Prevalence and risk factors of intimate partner violence in eighteen us states/territories, 2005', *American journal of preventive medicine* **34**(2), 112–118.

Butler, J. S., Burkhauser, R. V., Mitchell, J. M. and Pincus, T. P. [1987], 'Measurement error in self-reported health variables', *The Review of Economics and Statistics* **69**(4), 644–650. **URL:** *http://www.jstor.org/stable/1935959*

Cameron, A. C. and Trivedi, P. K. [2005], *Microeconometrics: methods and applications*, Cambridge university press.

Capaldi, D. M., Knoble, N. B., Shortt, J. W. and Kim, H. K. [2012], 'A systematic review of risk factors for intimate partner violence', *Partner abuse* **3**(2), 231–280.

Card, D. [2001], 'Estimating the return to schooling: Progress on some persistent econometric problems', *Econometrica* **69**(5), 1127–1160.

Coffman, K., Coffman, L. and Marzilli Ericson Keith, year = 2013, i. . N. [n.d.], The size of the

lgbt population and the magnitude of anti-gay sentiment are substantially underestimated, Technical report.

De Koker, P., Mathews, C., Zuch, M., Bastien, S. and Mason-Jones, A. J. [2014], 'A systematic review of interventions for preventing adolescent intimate partner violence', *Journal of Adolescent Health* **54**(1), 3–13.

DeKeseredy, W. S. and Schwartz, M. D. [1998], 'Measuring the extent of woman abuse in intimate heterosexual relationships: A critique of the conflict tactics scales', *US Department of Justice Violence Against Women Grants Office Electronic Resources* .

Ellsberg, M., Heise, L., Pena, R., Agurto, S. and Winkvist, A. [2001], 'Researching domestic violence against women: methodological and ethical considerations', *Studies in family planning* **32**(1), 1–16.

Erten, B. and Pinar, K. [forthcoming], 'For Better or Worse? Education and Prevalence of Domestic Violence in Turkey', *American Economic Journal: Applied Economics* .

Fulu, E., Jewkes, R., Roselli, T., Garcia-Moreno, C. et al. [2013], 'Prevalence of and factors associated with male perpetration of intimate partner violence: findings from the un multi-country cross-sectional study on men and violence in asia and the pacific', *The lancet global health* **1**(4), e187–e207.

Glynn, A. N. [2013], 'What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment', *Public Opinion Quarterly* **77**(S1), 159–172.

Gottschalk, P. and Huynh, M. [2010], 'Are earnings inequality and mobility overstated?

the impact of nonclassical measurement error', *The Review of Economics and Statistics* **92**(2), 302–315.

Green, E. P., Blattman, C., Jamison, J. and Annan, J. [2015], 'Women's entrepreneurship and intimate partner violence: A cluster randomized trial of microenterprise assistance and partner participation in post-conflict uganda (ssm-d-14-01580r1)', *Social Science & Medicine* **133**, 177–188.

Haushofer, J. and Shapiro, J. [2013], 'Household response to income changes: Evidence from an unconditional cash transfer program in kenya'.

Hidrobo, M. and Fernald, L. [2013], 'Cash transfers and domestic violence', *Journal of Health Economics* **32**(1), 304–319.

Hidrobo, M., Peterman, A. and Heise, L. [2016], 'The effect of cash, vouchers, and food transfers on intimate partner violence: Evidence from a randomized experiment in northern ecuador', *American Economic Journal: Applied Economics* **8**(3), 284–303.

Jewkes, R., Levin, J. and Penn-Kekana, L. [2002], 'Risk factors for domestic violence: findings from a south african cross-sectional study', *Social science & medicine* **55**(9), 1603–1617.

Johnston, D. W., Propper, C. and Shields, M. A. [2009], 'Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient', *Journal of health economics* **28**(3), 540–552.

Kishor, S. [2005], 'Domestic violence measurement in the demographic and health surveys: The history and the challenges', *Division for the Advancement of Women* pp. 1–10.

Klugman, J., Hanmer, L., Twigg, S., Hasan, T., McCleary-Sills, J. and Santamaria, J. [2014], *Voice and Agency: Empowering Women and Girls for Shared Prosperity*, Washington, DC: World Bank Group.

Koenig, M. A., Ahmed, S., Hossain, M. B. and Mozumder, A. K. A. [2003], 'WomenÕs status and domestic violence in rural bangladesh: individual-and community-level effects', *Demography* **40**(2), 269–288.

Krug, E. G., Mercy, J. A., Dahlberg, L. L. and Zwi, A. B. [2002], 'The world report on violence and health', *The lancet* **360**(9339), 1083–1088.

La Mattina, G. et al. [2017], 'Civil conflict, domestic violence and intra-household bargaining in post-genocide rwanda', *Journal of Development Economics* **124**(C), 168–198.

Lindo, J. M., Siminski, P. M. and Swensen, I. D. [2015], College party culture and sexual assault, Technical report, National Bureau of Economic Research.

O'Neill, D. [2012], The consequences of measurement error when estimating the impact of bmi on labour market outcomes, Technical report, IZA Discussion Paper No. 7008.

O'Neill, Donal, O. S. and de Gaer, D. V. [2004], The consequences of non-classical measurement error for distributional analysis, Technical report, National University of Ireland, Maynooth.

Palermo, T., Bleck, J. and Peterman, A. [2014], 'Tip of the iceberg: reporting and gender-based violence in developing countries', *American journal of epidemiology* **179**(5), 602–612.

Watts, C. and Zimmerman, C. [2002], 'Violence against women: global scope and magnitude', *The lancet* **359**(9313), 1232–1237.

A Additional Figures and Tables

Table A.1: Summary Statistics and Balance Check

	Control	(T-C)	N
Demographic Characteristics			
Age	43.825	0.903	1078
	[11.604]	[0.693]	
Married	0.798	-0.007	1078
	[0.402]	[0.025]	
Literate	1.959	0.002	1078
	[0.199]	[0.012]	
Spanish is not mother tongue	0.114	0.019	1078
	[0.318]	[0.020]	
Household head	0.313	0.07	1078
	[0.464]	[0.029]**	
Works	0.73	0.005	1078
	[0.444]	[0.027]	
Less than complete primary	0.109	0.017	1078
	[0.312]	[0.020]	
Primary education	0.266	-0.036	1078
	[0.442]	[0.026]	
Secondary education	0.45	-0.019	1078
	[0.498]	[0.030]	
Higher education	0.175	0.039	1078
	[0.380]	[0.024]	
Number of children	2.987	-0.013	1076
	[1.891]	[0.102]	
Number of children under 12 under her care	0.897	-0.025	1060
	[1.641]	[0.083]	
Memory test: % words remembered right after	0.85	0.026	1078
	[0.357]	[0.021]	
Memory test: % words remembered at the end	0.489	0.038	1078
	[0.500]	[0.030]	
Always lived in current locality	0.632	-0.028	1078
	[0.483]	[0.030]	
Financial Situation			
Average loan size in past 4 cycles	1552.664	8.921	1025
	[1178.413]	[72.065]	
Average savings balance in past 4 cycles	791.688	77.259	1025
	[861.449]	[63.958]	
High loan size and high savings balance	0.284	0.038	1078
	[0.451]	[0.028]	
Partner's characteristics			

Continued on next page

	Control	(T-C)	N
Jealous when speaking to other men	0.979	0.195	1077
	[7.224]	[0.488]	
Accuses her of being unfaithful	0.452	0.521	1078
	[4.196]	[0.420]	
Prevents her from visiting or being visited by friends	0.801	-0.203	1077
	[7.233]	[0.408]	
Limits contact with family	1.096	-0.511	1078
	[9.310]	[0.477]	
Wants to know where she is at all times	0.828	-0.34	1077
	[5.909]	[0.251]	
Does not trust her with money	0.428	0.374	1077
	[4.199]	[0.375]	
Humiliates her in public	0.555	0.018	1078
	[4.196]	[0.261]	
Calls her ignorant or idiot	0.538	0.37	1078
	[4.196]	[0.375]	
Calls her lazy, useless, or sleepy	0.45	0.006	1078
	[4.196]	[0.261]	
Threatened to harm her or someone close to her	0.512	-0.368	1078
	[5.913]	[0.250]	
Threatened to leave, take children, or cut off financial support	0.68	-0.362	1078
	[5.910]	[0.251]	
Survey Application			
Interruption by men	0.045	0	1078
	[0.207]	[0.013]	
Interruption by partner	0.007	-0.003	1078
	[0.084]	[0.004]	
Presence partner	0.018	-0.006	1078
	[0.133]	[0.007]	

Source: ADRA Survey 2015.

Table A.2: Prevalence rates of non-sensitive statements in the pilot

Have you ever	Mean	S.D.
made improvements to your dwelling?	0.774	0.425
traveled with your family on vacation? *	0.613	0.495
seen any soap opera? **	1.000	0.000
lost your cell phone? **	0.645	0.486
reared farm animals for consumption?	0.613	0.495
felt insecure in your neighborhood?	0.710	0.461
paid rent for the place where you live?	0.548	0.506
run out of money to cover the household's monthly expenses?	0.710	0.461
bought any high-end clothes?	0.290	0.461
been part of a Christian church?	0.484	0.508
purchased a TV with HD?	0.290	0.461
witnessed robberies in your neighborhood?	0.516	0.508
been robbed on the street?	0.516	0.508
seen *Al fondo hay sitio*? * [a/]	0.903	0.301
had to truncate your studies to care for your family?	0.742	0.445
pursued a technical degree?	0.387	0.495
read *El Comercio*? ** [b/]	0.645	0.486
helped your children with their homework?	0.968	0.180
participated in other microfinance programs?	0.645	0.486
had multiple businesses at the same time?	0.387	0.495
experienced that your business' sales are insufficient to cover your household expenses?	0.516	0.508
had insurance from ESSALUD, the armed forces or the police?	0.323	0.475
suffered from a serious medical condition that has required medical assistance?	0.677	0.475
bought expensive clothes?	0.226	0.425
traveled with your children?	0.839	0.374
played any games on your cell phone? *	0.290	0.461
visited the cathedral of Lima? **	0.677	0.475
used the subway as a means of transportation?	0.290	0.461
traveled with your friends?	0.323	0.475
participated in a committee or association in your neighborhood?	0.548	0.506
been to the movies with your family?	0.452	0.506
been out for a walk with your children?	0.968	0.180
bought new clothes for your children on important dates (Christmas, birthdays, etc.)? *	0.968	0.180
had problems with your partner because of money issues?	0.839	0.374

NOTE: * These statements are the ones in the 2nd list experiment question (push). ** These statements are the ones in the 8th list experiment question (forced sex). // [a/] *Al fondo hay sitio* is a very popular soap opera than run for several years in Peru. // [b/] *El Comercio* is one of the most read newspapers in the country, particularly in Lima.

Table A.3: Correlation of prevalence rates among non-sensitive statements

	1a	1b	1c	1d
1a	1.00			
1b	-0.29	1.00		
1c	0.12	-0.03	1.00	
1d	0.33	0.10	-0.34	1.00

	2a	2b	2c	2d
2a	1.00			
2b	-0.29	1.00		
2c	-0.08	0.23	1.00	
2d	-0.03	-0.06	-0.26	1.00

	3a	3b	3c	3d
3a	1.00			
3b	-0.29	1.00		
3c	-0.12	-0.16	1.00	
3d	0.34	-0.29	-0.35	1.00

	4a	4b	4c
4a	1.00		
4b	-0.29	1.00	
4c	0.25	-0.02	1.00

	5a	5b	5c	5d
5a	1.00			
5b	-0.37	1.00		
5c	-0.07	0.22	1.00	
5d	-0.06	-0.07	-0.37	1.00

	6a	6b	6c	6d
6a	1.00			
6b	-0.28	1.00		
6c	-0.23	-0.10	1.00	
6d	-0.05	0.14	-0.31	1.00

	7a	7b	7c	7d
7a	1.00			
7b	-0.54	1.00		
7c	0.15	0.03	1.00	
7d	0.09	-0.13	-0.28	1.00

	8a	8b	8c	8d
8a	1.00			
8b	-0.13	1.00		
8c	-	-	-	
8d	0.07	0.50	-	1.00

	9a	9b	9c
9a	1.00		
9b	-0.24	1.00	
9c	-0.04	-0.11	1.00

NOTE: Questions 4 and 9 include only 3 statements because the fourth one used in these questions did not come from the list of statements tested in the pilot. In question 8, statement c had a prevalence rate of 1.

Table A.4: Difference in estimated prevalence rates of physical violence against women by age

	< 50 years old			50+ years old		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.386	0.304	0.082	0.477	0.324	0.153
Slap	0.151	0.251	-0.100	0.206	0.293	-0.087
Punch	0.185	0.213	-0.028	0.155	0.245	-0.090
Kick	0.115	0.124	-0.009	0.146	0.187	-0.041
Strangle	0.023	0.048	-0.026	-0.105	0.069	-0.174
Knife	0.007	0.048	-0.042	0.118	0.074	0.044
Sex acts	0.011	0.059	-0.048	0.127	0.166	-0.039
Joint test						
χ^2		4.12			8.22	
Prob > χ^2		0.765			0.314	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.5: Difference in estimated prevalence rates of physical violence against women by civil status

	Single			Married		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.547	0.345	0.201	0.386	0.302	0.084
Slap	0.195	0.354	-0.159	0.164	0.242	-0.078
Punch	0.144	0.336	-0.193	0.182	0.195	-0.013
Kick	0.263	0.214	0.049	0.092	0.128	-0.036
Strangle	0.039	0.133	-0.094	-0.037	0.036	-0.073
Knife	0.072	0.097	-0.025	0.039	0.047	-0.008
Sex acts	0.106	0.133	-0.026	0.038	0.085	-0.047
Joint test						
χ^2		13.44			4.32	
Prob > χ^2		0.062			0.742	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.6: Difference in estimated prevalence rates of physical violence against women by mother's tongue

	Spanish				Other language		
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$
Pull hair	0.444	0.315	0.129	*	0.239	0.281	-0.043
Slap	0.142	0.258	-0.116	*	0.368	0.317	0.050
Punch	0.138	0.216	-0.078		0.423	0.281	0.142
Kick	0.083	0.138	-0.055		0.426	0.203	0.223
Strangle	-0.048	0.054	-0.103		0.160	0.063	0.098
Knife	0.057	0.056	0.000		-0.030	0.063	-0.092
Sex acts	0.044	0.083	-0.038		0.103	0.190	-0.088
Joint test							
χ^2		10.93				7.31	
Prob $> \chi^2$		0.142				0.398	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.7: Difference in estimated prevalence rates of physical violence against women by memory

	Bad memory				Good memory		
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$
Pull hair	0.477	0.350	0.127		0.362	0.270	0.092
Slap	0.253	0.262	-0.009		0.091	0.267	-0.176 *
Punch	0.247	0.248	-0.001		0.105	0.198	-0.093
Kick	0.165	0.155	0.011		0.088	0.135	-0.047
Strangle	0.006	0.063	-0.057		-0.049	0.047	-0.096
Knife	0.061	0.073	-0.013		0.032	0.040	-0.008
Sex acts	0.137	0.116	0.021		-0.029	0.073	-0.102
Joint test							
χ^2		3.99				6.60	
Prob $> \chi^2$		0.781				0.472	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.8: Difference in estimated prevalence rates of physical violence against women by household head status

	Household head				Not the household head		
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$
Pull hair	0.455	0.275	0.180	**	0.348	0.389	-0.040
Slap	0.174	0.240	-0.066		0.163	0.320	-0.157
Punch	0.136	0.197	-0.061		0.246	0.282	-0.035
Kick	0.131	0.112	0.019		0.117	0.218	-0.102
Strangle	-0.012	0.026	-0.038		-0.041	0.120	-0.161
Knife	0.057	0.034	0.023		0.025	0.109	-0.083
Sex acts	0.024	0.065	-0.041		0.103	0.160	-0.057
Joint test							
χ^2		8.78				4.73	
Prob $> \chi^2$		0.269				0.693	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.9: Difference in estimated prevalence rates of physical violence against women by employment

	Does not work			Works		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.468	0.351	0.117	0.400	0.296	0.104
Slap	0.207	0.272	-0.065	0.157	0.262	-0.105
Punch	0.339	0.252	0.087	0.114	0.213	-0.099
Kick	0.070	0.185	-0.116	0.146	0.130	0.016
Strangle	0.014	0.086	-0.072	-0.035	0.044	-0.079
Knife	0.062	0.066	-0.004	0.040	0.054	-0.014
Sex acts	0.039	0.113	-0.073	0.056	0.088	-0.032
Joint test						
χ^2		6.22			6.48	
Prob $> \chi^2$		0.515			0.485	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.10: Difference in estimated prevalence rates of physical violence against women by loan size

	Low loan size				High loan size		
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$
Pull hair	0.431	0.313	0.118		0.379	0.317	0.062
Slap	0.135	0.294	-0.159	**	0.278	0.187	0.091
Punch	0.177	0.254	-0.078		0.167	0.138	0.028
Kick	0.057	0.161	-0.104		0.336	0.114	0.223
Strangle	-0.064	0.069	-0.133	*	0.106	0.016	0.089
Knife	0.093	0.067	0.026		-0.096	0.016	-0.112
Sex acts	-0.016	0.094	-0.110		0.257	0.098	0.160
Joint test							
χ^2		16.09				9.32	
Prob $> \chi^2$		0.024				0.231	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

Table A.11: Difference in estimated prevalence rates of physical violence against women by savings balance

	Low savings balance				High savings balance		
Violent act	ρ	p	$(\rho - p)$		ρ	p	$(\rho - p)$
Pull hair	0.429	0.308	0.121		0.384	0.333	0.051
Slap	0.134	0.286	-0.152	**	0.279	0.214	0.065
Punch	0.174	0.246	-0.073		0.177	0.167	0.010
Kick	0.057	0.155	-0.097		0.332	0.135	0.198
Strangle	-0.058	0.062	-0.120		0.086	0.040	0.046
Knife	0.040	0.062	-0.022		0.063	0.032	0.032
Sex acts	0.013	0.085	-0.072		0.167	0.127	0.040
Joint test							
χ^2		12.84				4.81	
Prob $> \chi^2$		0.076				0.683	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory.

B Ceiling Effects

Although using a very small sample (31 observations), the pilot data allows us to measure the prevalence of each non-sensitive statement before designing the list experiments. Relying on this data, we grouped statements in sets of 4 while trying to minimize ceiling effects and reduce the variance of the estimator (see sub-section 3.2). Since we had to construct 9 sets of 4 non-sensitive statements simultaneously, we relied on an algorithm that tried to minimize these two problems for the 9 sets of statements altogether. Thus, the final grouping we obtained may have been more conducive to generate ceiling effects in certain questions.

In particular, we believe that there may be a higher propensity to yield ceiling effects in the questions related to push and forced sex. Table B.1 reports some statistics on the prevalence rates of the sets of non-sensitive statements with data from the pilot. The first column reports the mean prevalence of the 4 statements, while the second and third report the standard deviation and the 75th percentile of this 4 prevalence rates. The non-sensitive statements grouped with the sensitive ones on pushing and forced sex have very high average prevalence rates and low variance. Moreover, the 75th percentile of prevalence rates for these sets of 4 statements is very high, which shows that many statements in these groups have high prevalence rates. In fact, one of the statements grouped with forced sex has a prevalence rate of 1 ("ever watched a soap opera").

In what follows, we discard the results on these two acts of violence. We focus on the acts of violence related to the other seven list experiment questions that seem more robust to biases in the instrument design.

Table B.1: Prevalence of 4 non-sensitive statements by question

	Distribution of prevalence		
Statements grouped with:	Mean	SD	p(75)
Slap	0.419	0.083	0.484
Kick	0.500	0.194	0.661
Knife	0.508	0.152	0.597
Pull Hair	0.613	0.411	0.968
Push	0.694	0.310	0.935
Strangle	0.694	0.150	0.790
Forced sex	0.742	0.173	0.839

NOTE: Columns 1-3 report means, standard deviations, and the 75th percentile for the prevalence rates of each sample of 4 non-sensitive statements. Only 3 out of the 4 statements grouped with punch and sex acts come from the pilot and are thus not reported.