Misreporting of Intimate Partner Violence in Developing Countries: Measurement Error and New Strategies to Minimize it

June 21, 2018

Abstract

A growing literature seeks to identify policies that could reduce intimate partner violence. In the absence of reliable administrative records, this violence is often measured using self-reported data from health surveys. We conduct an experiment comparing data from such surveys against a methodology that provides greater privacy to the respondent. We identify non-classical measurement error in health surveys: college-educated women underreport physical and sexual violence while there is no bias among the less educated. We provide a low-cost solution to correct the bias in the estimation of causal effects under non-classical measurement error in the dependent variable.

Keywords: Non-classical measurement error, treatment effects, sensitive health behaviors, intimate partner violence, list experiments, survey data, direct elicitation

JEL-codes: I12, C83, C21.

1 Introduction

Despite the threat of measurement error in survey responses, much of the empirical work in economics relies on self-reported data. Due to different reasons, ranging from random mistakes, limited attention, and lack of recollection to behavioral biases or stigma, respondents may give inaccurate answers. Misreporting is expected to be even more worrisome whenever the respondent faces questions about sensitive topics such as personal earnings, crime activity, drug use, discrimination, physical appearance, or intimate partner violence (IPV).

In recent years, concerns about measurement error have been increasingly addressed in the literature. An important share of studies has made use of administrative records to directly measure and characterize misreporting in sensitive topics such as voting [Rosenfeld et al., 2016], mental health conditions [Bharadwaj et al., 2015], or personal earnings [Gottschalk and Huynh, 2010]. In developed countries, administrative records have also facilitated the analysis of the consequences of domestic violence as in the case of [Aizer, 2010]. However, data from these sources are likely to reflect self-selection into reporting that may not be random. Even worse, in developing countries, administrative records on IPV are neither well kept nor systematically collected.

Using an indirect questioning technique, this paper measures and characterizes misreporting when dealing with a sensitive topic and proposes an alternative way to minimize the bias introduced by measurement error in the estimation of treatment effects. We focus on the measurement of women's experiences of physical and sexual violence by their partners, due to its saliency as a public health issue, the urgency to generate accurate data on its prevalence to guide policy efforts, and the fact that in developing countries, as shown below, these data come from self-reported health surveys.

Our focus on IPV is extremely timely as a growing number of studies try to identify the main drivers of this phenomenon, particularly in the developing world [e.g., Angelucci, 2008; Hidrobo and Fernald, 2013; Haushofer and Shapiro, 2013; Bobonis et al., 2013; Hidrobo et al., 2016], and the impact of programs intended to reduce its prevalence [World Health Organization, 2009]. Several scholars have argued that measures of violence against women could be subject to reporting error [e.g., DeKeseredy and Schwartz, 1998; Ellsberg et al., 2001; Kishor, 2005; Aizer, 2010, 2011], but little is known about the magnitude and the characteristics of misreporting of IPV data.

From a policy perspective, the use of inaccurate self-reported data is extremely problematic as it may directly impact the design of preventive and corrective policies. If misreporting is random, treatment effects for a given risk factor can be accurately estimated by relying on

exogenous variation introduced in the variable of interest through a randomized controlled trial (RCT) or a quasi experimental approaches such as instrumental variables (IV). However, the presence of non-classical measurement error in the dependent variable biases the estimates of treatment effects even when an exogenous variation in the variable of interest is available.

We advance this literature by focusing on the misreporting of physical and sexual IPV, estimated using direct questions as applied by the Demographic and Health Surveys (DHS), a global project that is the main source of IPV data [Klugman et al., 2014]. These questions have been included in 122 surveys covering 61 developing countries, so assessing possible biases in this approach to measure IPV is extremely relevant for a vast set of countries worldwide.[1] We compare rates from such a design against an alternative method that uses indirect questions in the form of list experiments.

Our study focuses on a sample of microcredit female clients from several impoverished peri urban districts in Lima, Peru, among which we randomize two questionnaires. The control group receives seven direct questions that the DHS uses to measure the prevalence of physical and sexual IPV. In addition, the control group receives seven lists of four neutral statements and is asked to provide the *number* of statements that hold true in each list, but not the actual occurrence of each statement. In turn, the treatment group does not answer the direct questions on IPV. It only receives seven lists of five statements each, where the first four are identical to those in the list provided to the control and the last one refers to a specific act of physical or sexual violence. Again, respondents only answer *how many* statements are true. Randomization at the individual level guarantees that the average number of neutral statements that holds true is equal across treatment and control groups. Thus, the prevalence rate of a given act of physical or sexual violence can be estimated as the difference in the average number of statements that holds true for each list across treatment arms.

We find no significant differences in reporting of physical and sexual violence across direct and indirect methods. However, we find that the reporting error varies with the level of education: women with completed tertiary education report higher rates of violence under the list experiments than under the direct method while there is no significant difference among less educated women. The increased report of violent episodes among more educated women under list experiments is large enough to reverse the negative education gradient that emerges when prevalence rates are measured through direct questions.

We argue that our results have ample applications in settings where the dependent vari-

[1]The World Health Organization has conducted similar surveys about IPV but applied them in a smaller set of countries. For simplicity, we refer to both surveys as *DHS-type*.

able is likely to suffer from non-random measurement error [Bound et al., 2001; Butler et al., 1987] and where administrative records are not an alternative data source. More specifically, we argue that our study contributes to the literature on measurement error and IPV along several dimensions. First, we accurately quantify and qualify measurement error in IPV data relative to the best available direct questioning method. We devoted special care to develop the instruments and protocols that allowed us to compare the report under list experiments to the one obtained using current best practices with the direct method. Another key contribution of our study is that we try to minimize other potential biases in the self-report of data. The use of a large sample size allows us to have a treatment group that does not answer the direct IPV questions. This provides a much needed level of privacy and improves on previous papers using list experiments, in general, where the lists were applied *after* the direct questions with the possible contamination of the treatment group [e.g., Karlan and Zinman, 2012; Peterman et al., 2017; De Cao and Lutz, forthcoming].

Second, although previous studies have relied upon list experiments to measure prevalence rates of risky or socially undesirable attitudes or behaviors,[2] we add to the scarce body of work measuring misreporting in the case of victimization. The costs of self-exposing as a victim rather than a criminal or drug abuser may be quite different and have varying effects on the nature of misreporting.

Third, our paper reviews the implications of systematic misreporting on the estimation of causal effects and shows that common strategies to deal with endogeneity biases cannot address the bias induced by non-classical measurement error. We show that RCTs and (valid) IVs still yield biased treatment effects in the presence of non-classical measurement error in the outcome variable. In fact, relative to RCTs and IVs, cross-sectional estimates may provide *less* biased estimates when the sign of the bias from omitted variables is opposite to that of the relationship between measurement error and the risk factor.

Finally, our experimental approach provides researchers with a simple and inexpensive strategy to test for measurement error in contexts where fieldwork is being conducted. Furthermore, our approach allows researchers to correct their treatment effect estimates. These contributions are particularly valuable for the case of IPV, since there are no previous efforts trying to quantify the severity and patterns of underreporting in such sensitive behavior nor the implications that misreporting has on the estimation of treatment effects.

[2]Recent applications of list experiments include, for example, Karlan and Zinman [2012] to measure loan proceeds from microfinance loans, McKenzie and Siegel [2013] to elicit illegal migration rates, Coffman et al. [2013] to measure the size of LGBT population and anti-gay sentiment, Imai et al. [2014] to examine vote-selling, and Rosenfeld et al. [2016] to study anti-abortion support.

2 Misreporting in sensitive survey questions: the case of intimate partner violence

There is growing consensus about the best practices on how to ask questions about IPV. In developing countries, the most widely method to measure IPV follows the Conflict Tactic Scale (CTS), originally developed by Straus [1979]. This methodology was later modified and expanded by the World Health Organization (WHO, 1997) and has been implemented in 122 DHS surveys covering 61 developing countries. The modified CTS, as used by the DHS, follows a multiple-question approach about physical and sexual violence that provides participants with several opportunities to respond about victimization [Kishor and Johnson, 2004]. This method focuses on specific and objective acts of violence and is thus less likely to be affected by different perceptions about what constitutes violence (see Ellsberg and Heise [1999] and Bender [2017] for an extensive discussion).

The recommendations put forward by the WHO also cover rigorous implementation protocols to guarantee the safety and wellbeing of the participants. These ethical and privacy protocols try to provide adequate conditions to protect the respondent from emotional pain or further experiences of IPV, as well as to guarantee a safe environment in which she can feel at ease to share her experience. However, despite the progress made in terms of these ethical and privacy protocols, the sensitivity of the topic can make respondents reluctant to self-identify as a victim, potentially leading to misreporting.

In particular, two features of intimate partner violence generate large potential for error in the measurement of prevalence rates: it is usually perpetrated by people known to the victims, mainly their partners or ex-partners, and it tends to be invisible as much of it happens behind closed doors and in the privacy of the home.

These features introduce very large costs to self-identify as a victim. 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, a woman may also fear the potential loss of her partner's economic support if her status as a victim is revealed. Third, if exposed, she also faces the risk of retaliation through an escalation of violence against her or her children. Finally, women may fear stigmatization, either from intrinsic or extrinsic sources [Overstreet and Quinn, 2013]. Since the costs of being exposed are very likely to be heterogeneous, privacy concerns may differentially prevent women from truthfully reporting their previous experience of violence, leading to systematic misreporting.

Unlike other health outcomes or risky behaviors, administrative records do not always provide a benchmark for the measurement of IPV prevalence rates. Administrative records from the police or health establishments will capture a non-random sample of the true IPV

cases and, most likely, only the most extreme incidents.[3] Although a few reports may come from third parties, the bulk of the records rely upon the victim's decision to approach the authorities, which in turn depends on the costs of exposure she faces. Indeed, the cost may become even higher due to fear or distrust of the authority herself, which is more severe in developing contexts. Indeed, Palermo et al. [2014] shows that only seven percent of women who experienced domestic violence made a formal report that would be captured in administrative data (e.g., police, medical, or social services) in developing countries. Moreover, the authors also show that reporting depends on women's socioeconomic characteristics such as age, marital status, education, and urban location.

Ellsberg et al. [2001] argues that when greater privacy measures are enforced, higher prevalence rates of IPV are measured relative to the DHS methodology. However, although the authors compare surveys that vary in terms of privacy levels, they cannot isolate this effect on prevalence rates since the data also comes from different years and different samples of women. Moreover, as much as security and privacy protocols can be improved upon, direct methods cannot get rid of the remaining exposure cost that is derived from the expectation that the respondent may have about her report being shared with others. Our approach provides a rigorous way to minimize the costs of being exposed as a victim and/or exposing the aggressor. We rely on list experiments, which provide full anonymity to respondents, to measure and characterize misreporting in the prevalence of physical and sexual violence as committed by the women's last partner. We provide a significant contribution by establishing a benchmark, characterizing misreporting, and putting forth a strategy to correct for potential biases in the estimation of treatment effects.[4]

Two recent studies are closely related to our paper: Joseph et al. [2017] and Peterman et al. [2017]. They both rely on list experiments to measure prevalence rates of physical violence. Their contribution is valuable but they have several limitations. First, Joseph et al. [2017] measures prevalence rates at the household level, asking anyone who opens the door

[3] Jamison et al. [2013] and Moseson et al. [2015] provide additional examples where administrative records cannot provide a benchmark in developing countries. Also, while we argue that administrative records, mainly in developing countries, are not reliable to measure IPV prevalence in general, they could still be useful in certain contexts and subpopulations and when focusing on violent incidents for which reporting is not a choice, as in the case of hospital admissions for assault, especially in richer economies [Aizer, 2011].

[4] Alternative methods include qualitative approaches as in Blattman et al. [2016]. The authors combine surveying with ethnographic techniques to uncover misreporting. The method requires that the surveyor team stays for longer periods of time in the field, increasing the costs of data collection. Moreover, since it does not provide additional anonymity to the respondents, the success of the method depends heavily on the surveyors' ability to make the respondent feel safe and comfortable to truthfully report her answers or behavior. Surveyors training then becomes crucial, adding to the cost of the fieldwork and making it a technique that is hard to scale up. Other indirect questioning techniques such as endorsement experiments or randomized response techniques are often used in the political science literature and recent papers have adapted them to measure health outcomes such as abortions [Lara et al., 2006].

about the violence experienced by women in the household. Thus, it may be the case that the respondent does not know about the IPV experience of all the women in the household or that he himself is the perpetrator. Second, the sensitive statement in the lists is quite general and it is based on the single-question generic approach (*Has at least one woman member of your household faced physical aggression from her husband anytime during her life?*), greatly departing from the well-established multiple-question WHO guidelines for the measurement of violence. The same holds for Peterman et al. [2017], who targets women as respondents but uses a general sensitive statement to measure physical violence (*In the last 12 months, have you ever been slapped, punched, kicked, or physically harmed by your partner?*). Finally, neither Joseph et al. [2017] nor Peterman et al. [2017] are able to measure misreporting relative to the *best available* direct reporting method. The former did not include a direct question equivalent to the sensitive item in the control questionnaire while the latter asks the same individual the direct question on violence *before* the indirect question. This could bias both reports since the respondent is no longer protected by the list experiment.

Our design overcomes all these limitations by (i) focusing on women as respondents, (ii) following the WHO guidelines for direct questions as well as their privacy and safety protocols throughout the application of the questionnaire, (iii) asking the direct questions only to the control group, and (iv) comparing the prevalence rates obtained from the indirect method to the ones that come from the DHS direct method.[5]

3 Measuring reporting bias in intimate partner violence

3.1 List experiments: design

List experiments have been traditionally used to gather opinions and/or record behaviors related to inherently sensitive issues that are more prone to underreporting. The basic design of a list experiment features 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 statements plus one, where the last one refers to a sensitive issue. Both groups are asked to provide the *number* of statements that hold true, without indicating which ones are in fact true.

Let $d_{is} = 1$ if, for individual i, the sth statement is true and zero otherwise. In a list experiment, this is not directly observed. Instead, we observe the number of responses

[5]Recently, De Cao and Lutz [forthcoming] used a list experiment to measure attitudes towards female genital cutting in the Afar province of Ethiopia but they share some of the limitations previously mentioned for the studies on IPV such as the framing issues that emerge by asking the direct questions to both the treatment and control groups.

that hold true for each i denoted as $\sum_s^S d_{is}$ when she belongs to the control group and $\sum_s^{S+1} d_{is}$ if she is in the treatment group. Under the "no design effects" assumption (i.e., the inclusion of the sensitive statement does not distort the answers to the neutral statements in the treatment group)[Blair and Imai, 2012], random assignment of the treatment at the individual level implies that:

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

The control group serves as a counterfactual for the treatment group, yielding the average number of neutral statements that hold true if the treatment were only given the first S statements of the list. Additionally imposing the "no liars" assumption (i.e., respondents give truthful answers to the sensitive statement), the prevalence rate of the sensitive statement is measured as:

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

We apply this methodology to measure prevalence rates of physical and sexual IPV. In the direct questions and in the list experiment, we ask if the respondent ever experienced these acts as perpetrated by her *last partner*. Following the DHS protocol, the last partner refers to the current partner for women in a relationship at the time of the survey or to her previous partner if she is currently single but had at least one partner in the past.

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 $\sum_s^S d_{is} = S$, a *ceiling effect*, since the respondent would no longer be protected. Lists that are too short will also tend to generate ceiling effects [Glynn, 2013].[6] Similarly, we want to avoid lists that contain 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 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

[6] Although the literature on list experiments has not identified an optimal number of neutral statements, a large share of the studies that rely upon these data include neutral lists of four statements.

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. In Section 3.2 we provide details on the efforts we undertook to minimize extreme values in the sets of statements used while maintaining low levels of variability in the responses.

The reduction of misreporting obtained from providing full anonymity through the list experiments comes at the cost of foregoing individual-level data on IPV. However, with large enough sample sizes one can measure prevalence rates by sub-samples and still identify correlations between the experience of violence and other variables.

3.2 Sample description and data

The population of interest for our study are adult women, living in Lima, who receive microloans from the Adventist Development and Relief Agency (ADRA), an international non-governmental organization running a village banking program in Peru's peri-urban and rural areas. ADRA's clients in Lima are microentrepreneurs from the most impoverished districts in the city.

From the total pool of 1873 clients in 112 village banks in ADRA's program in Lima, we first drop all clients under age 18 as well as all women above 65. This leaves us with a universe of 1776 clients. We draw 6 banks at random and exclude them from the study to be able to pilot the instruments with their members. The remaining universe is comprised by 1690 clients in 106 banks. Finally, we work with all banks with monthly meetings scheduled during July 2015, which restricts the population of interest to 1562 women in 98 village banks. We targeted this restricted universe and were able to interview 1223 women between July 1st and August 25th, 2015.[7]

Randomization of the treatment was done at the individual level and was conducted by the surveyor after obtaining the informed consent of the woman. The questionnaire was implemented via tablets. Due to some initial complications with the software, we drop a few surveys which were incorrectly assigned to answer the list experiment questions from both treatment arms and we are left with a sample of 1078 valid surveys. 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.[8]

[7]Most of the non-response is due to clients who decide to drop out of the village banking program.

[8]The baseline violence prevalence rates in the area studied were obtained from the Peruvian DHS survey. We focused 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.

Table 1 confirms that the randomization was successful. There is only a small significant difference in the share of women who are household heads (at the 5 percent level).

Table 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	0.959	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	0.070	1078
	(0.464)	[0.029]**	
Works	0.730	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	0.450	-0.019	1078
	(0.498)	[0.030]	
Higher education	0.175	0.039	1078
	(0.380)	[0.024]	
Number of children	2.953	0.022	1075
	(1.712)	[0.096]	
Number of children under 12 under her care	0.844	0.028	1059
	(1.077)	[0.065]	
Memory test: % words remembered right after	0.850	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 (US$)	1552.664	8.921	1025
	(1178.413)	[72.065]	
Average savings balance in past 4 cycles (US$)	791.688	77.259	1025
	(861.449)	[63.958]	
High loan size, savings balance, and tenure	0.161	0.023	1078
	(0.368)	[0.023]	
Emotional IPV			
Humiliates her in public	0.379	0.004	1076

Continued on next page

	Control	(T-C)	N
	(0.486)	[0.030]	
Calls her ignorant or idiot	0.361	-0.025	1074
	(0.481)	[0.029]	
Calls her lazy, useless, or sleepy	0.274	-0.009	1076
	(0.446)	[0.027]	
Threatened to harm her or someone close to her	0.159	-0.015	1076
	(0.366)	[0.022]	
Threatened to leave, take children, or cut off financial support	0.328	-0.009	1076
	(0.470)	[0.029]	
Survey Application			
Interruption by men	0.045	-0.000	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]	

NOTE: Differences between control and treatment group are obtained from regressing each variable on the treatment dummy. Standard errors in parenthesis.

The implementation of list experiments requires careful preparation in terms of the development of the instrument, the training of surveyors, and the provision of tools to ensure respondents' adequate understanding of this type of questions. With this in mind, we dedicated special attention to (i) the design of the instrument, (ii) the selection and training of surveyors, and (iii) the application of the instrument.

To develop the questionnaires, we piloted 41 neutral 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 using the same time frame used in the DHS questions and in the sensitive questions of the list experiment. These prevalence rates were useful in two ways. On one hand, they measured the adequacy of the statements for our particular setting. Statements with prevalence rates too close to zero were discarded. 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.[9]

Compared to other studies using list experiments, a key advantage of our paper is a large sample size, which allows us to have separate questionnaires for the treatment and

[9]Table A.1 in Appendix A shows the prevalence rates of the 34 statements we kept for the list experiments, after removing those with very small prevalence rates. Two statements used in the final instrument were not tested in the pilot. Table A.2 reports the correlation of prevalence rates in each set of statements grouped together. Based on the correlation of responses across pairs of statements in the pilot data, 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 optimal negative correlations and checked the correlation in the full list was still negative.

control groups.[10] This reduces potential biases that may be introduced when asking the same respondent both the direct and indirect questions as done in Karlan and Zinman [2012] and Joseph et al. [2017].

The structure of the questionnaire for the treatment and control groups is shown in Table 2. Both surveys start with general questions on demographics and a memory test (modules 1-3). The control group answered to a questionnaire that had the module of direct questions on physical and sexual IPV (module 5A) presented before the list experiments section (5B). Both modules were located right after the direct questions on emotional violence (module 4). In the treatment group, only the list experiment questions with the added sensitive statement were provided in module 5B, asked also after the emotional violence module. As shown in Table 1, both groups are balanced in their rates of emotional violence. [11]

Table 2: Structure of the questionnaire

Module	**Control**	**Treatment**
1	Consent form and introduction	
2	Demographics	
3	Memory test	
4	Direct questions about emotional violence	
5A	Direct questions about physical and sexual violence	
5B	Lists (4 items) with neutral statements	Lists (5 items) with indirect questions about physical and sexual violence
6	Satisfaction with ADRA	

One may argue that the inclusion of the direct questions on physical and sexual IPV in the control group could have biased the responses to the rest of the questions in the survey, including answers to the lists of neutral statements. For instance, it could be that the mention of such a sensitive subject made the respondent relive or remember painful experiences and that this feeling lingered throughout the rest of the questionnaire, interfering with the thinking process to arrive to her answers. If that were the case, then answers to all other non-sensitive questions that followed would be affected. However, in Table 3 we test for differences in the answers and item non-response rates to the last module across treatment

[10] See sample instruments in Appendix B.

[11] This was done via direct questions and as a way to introduce the topic of violence, following the DHS questionnaire.

Table 3: Difference in Responses and Item Non-Responses to the Last Module Across Treatment Arms

	Control	(T-C)	N
Differences in answers			
Satisfied with training	0.813	0.008	1077
	(0.391)	(0.024)	
Satisfied with family talks	0.834	0.014	1076
	(0.373)	(0.022)	
Satisfied with sports events	0.592	-0.025	1076
	(0.492)	(0.030)	
Satisfied with loans	0.871	-0.007	1076
	(0.335)	(0.021)	
Likely to stay in VB	0.793	-0.024	1068
	(0.405)	(0.025)	
Likely to recommend ADRA to others	0.953	-0.025	1076
	(0.211)	(0.014)	
Likely to assume role in VB committee	0.494	0.031	1073
	(0.500)	(0.031)	
Differences in item nonresponse			
Satisfied with training	0.000	0.002	1078
	(0.000)	(0.002)	
Satisfied with family talks	0.002	0.000	1078
	(0.042)	(0.003)	
Satisfied with sports events	0.002	0.000	1078
	(0.042)	(0.003)	
Satisfied with loans	0.000	0.004	1078
	(0.000)	(0.003)	
Likely to stay in VB	0.007	0.004	1078
	(0.084)	(0.006)	
Likely to recommend ADRA to others	0.002	0.000	1078
	(0.042)	(0.003)	
Likely to assume role in VB committee	0.005	-0.001	1078
	(0.073)	(0.004)	

NOTE: Differences between control and treatment group are obtained from regressing each variable on the treatment dummy. Standard errors in parenthesis.

arms (module 6). The eight questions in this module refer to client's satisfaction with ADRA. In only one case the answers across treatment and control groups differ significantly but just at the 10% level. Item non-response rates are also similar and in only one out of the eight cases the treatment group is statistically less likely to respond. We acknowledge that this test is imperfect since the treatment group was differentially exposed to the IPV questions through the list experiments. For future extensions, we suggest to randomize the order of

the direct and indirect questions on IPV in the control questionnaire.

Although we execute nine list experiments to measure prevalence rates of physical and sexual IPV, 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 faced some issues in the lists used in these two cases (see Appendix C for more details). For the remaining lists, we applied the test proposed by Blair and Imai [2012] where the null hypothesis is "no design effect". In all cases, we fail to reject the null at the 5% confidence level (results available upon request).

To implement the survey, we carefully selected a team of female surveyors with previous experience on the topics of gender and gender biased violence. They all attended a three-day training workshop and only the top performers in the practice sessions were recruited. The workshop itself included a sensitization session provided by a local organization, *Centro de la Mujer Peruana Flora Tristán*, which works on gender issues and women's empowerment.

Table 4: Prevalence rates of IPV

	N	Prevalence rate
Emotional IPV	1078	0.64
Humiliate	1076	0.38
Insult	1074	0.35
Called Lazy	1076	0.27
Threatens to harm	1076	0.15
Threatens to Leave	1076	0.32
Physical and sexual IPV	560	0.46
Pull hair	560	0.31
Slap	559	0.26
Punch	559	0.22
Kick	558	0.15
Strangle	560	0.06
Knife	560	0.06
Unapproved Sex practices	558	0.09
IPV	560	0.72

NOTE: The prevalence of emotional (physical and sexual) IPV is measured as the prevalence of any type of emotional (physical and sexual) aggression. Accordingly, the prevalence of IPV is measured as the prevalence rate of any type of violence; emotional, physical and/or sexual.

To minimize the chances for misunderstanding or confusion when applying the instrument, we provided respondents with visual aids for module 5B (list experiments). 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. We also tried to minimize potential biases in responses due to fear of having their individual answers revealed to ADRA. As shown in Appendix B.1, the consent form clearly stated that individual answers were not going to be shared with anyone outside the research team, which excluded ADRA. Moreover, surveyors reminded the respondent about the confidentiality of their answers at the beginning of module 4.

Table 4 reports shockingly high prevalence rates of ever experiencing violent acts inflicted by the woman's last partner as collected by DHS-type direct questions. About 72% of the women in our sample have ever experienced any type of violence, either emotional or physical or sexual. Prevalence for any type of emotional violence is 64% while it amounts to 46% for any type of physical or sexual violent act.

3.3 Estimation

Let T_i denote the treatment assignment to the list experiment. Also, let D_i be equal to the number of statements that hold true for 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. The difference-in-means estimator ρ approximates the prevalence rate of the sensitive statement included in the list provided to the treatment group:

$$D_i = \alpha + \rho T_i + \xi_i \tag{1}$$

Furthermore, let the reported prevalence rates under the direct questions be denoted as p. We are interested in estimating the level of misreport between the list experiment and the direct questions as measured by $(\rho - p)$ and in testing whether this difference is positive and statistically significant. Since the control and treatment groups are, on average, equivalent in terms of their true prevalence rates, $\rho - p > 0$ signals the existence of underreporting in DHS-type questions.

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

$$D_i = \alpha + \rho T_i + \gamma x_i + \zeta (T_i \cdot x_i) + \xi_i \tag{2}$$

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 $x_i = 0$.[12]

[12]These are the multivariate regression estimators obtained under linearity in x_i and $(T_i \cdot x_i)$ as proposed

Again, we can compare these prevalence rates to their counterpart measure obtained through direct reporting, p, conditional on x_i.

3.4 Results

Table 5 presents the prevalence rates using indirect and direct reporting methods for physical and sexual IPV. The last column measures the gap between ρ and p for each act of IPV while the last two rows report the results from a joint test of significance of this gap for all acts of violence analyzed.

Table 5: Difference in estimated prevalence rates of physical and sexual IPV

Violent act	List experiments (ρ)	Direct reporting (p)	($\rho - p$)
Pull hair	0.425	0.311	0.114
	(0.059)	(0.020)	(0.062)*
Slap	0.179	0.265	-0.086
	(0.065)	(0.019)	(0.067)
Punch	0.192	0.224	-0.032
	(0.068)	(0.018)	(0.071)
Kick	0.142	0.145	-0.003
	(0.064)	(0.015)	(0.067)
Strangle	-0.002	0.055	-0.057
	(0.062)	(0.010)	(0.063)
Knife	0.055	0.057	-0.002
	(0.064)	(0.010)	(0.065)
Sex acts	0.099	0.095	0.004
	(0.066)	(0.012)	(0.068)
Joint test			
χ^2		8.315	
Prob $> \chi^2$		0.306	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Significance levels: * 10%; ** 5%; *** 1%.

On average, the results in Table 5 suggest that direct questions used in DHS-type surveys do not introduce a bias in measuring the prevalence of violence when compared to experimental methods that provide more anonymity or privacy to the respondent. For six out of seven acts of physical violence, the prevalence rates obtained through experimental methods do not significantly differ from those measured using direct DHS-type questions. Indeed, we

in Blair and Imai [2012].

cannot reject the joint test that the seven gaps are zero, providing little evidence to suspect of average reporting biases.[13]

The lack of a significant difference in prevalence rates across reporting methods presented in Table 5 does not rule out the potential for misreporting among specific groups. More vulnerable groups with higher costs of being exposed could be more likely to truthfully report violence under the indirect method due to the provision of full confidentiality. We explore such potential outcomes relying on Equation (2).

Table 6: Difference in estimated prevalence rates of physical and sexual IPV by education level

	Less than tertiary education			Tertiary education		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.379	0.340	0.039	0.593	0.173	0.419
	(0.064)	(0.022)	(0.068)	(0.097)	(0.038)	(0.104)***
Slap	0.043	0.293	-0.249	0.674	0.133	0.541
	(0.070)	(0.021)	(0.073)***	(0.102)	(0.034)	(0.107)***
Punch	0.091	0.247	-0.156	0.560	0.112	0.448
	(0.074)	(0.020)	(0.077)**	(0.110)	(0.032)	(0.114)***
Kick	0.122	0.163	-0.040	0.215	0.062	0.153
	(0.071)	(0.017)	(0.073)	(0.107)	(0.024)	(0.110)
Strangle	-0.046	0.061	-0.107	0.162	0.031	0.131
	(0.068)	(0.011)	(0.069)	(0.096)	(0.017)	(0.097)
Knife	-0.042	0.058	-0.101	0.411	0.051	0.360
	(0.069)	(0.011)	(0.070)	(0.104)	(0.022)	(0.106)***
Sex acts	0.031	0.104	-0.073	0.349	0.051	0.298
	(0.073)	(0.014)	(0.075)	(0.098)	(0.022)	(0.100)***
Joint test						
χ^2		10.033			21.632	
Prob $> \chi^2$		0.187			0.003	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by education level are obtained from the model in Equation (2).

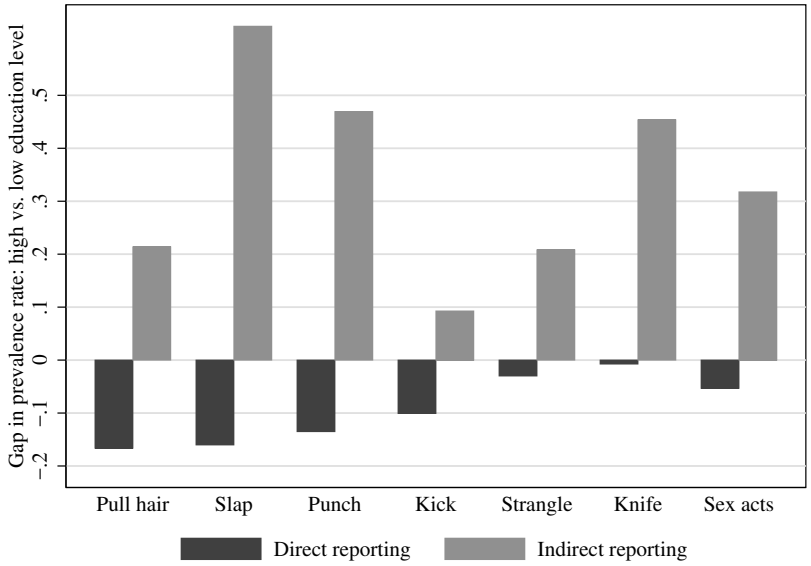
Note that, although we are able to explore differential misreporting by characteristics of the respondent, our study was not designed to identify the forces that are driving the results. In other words, we slice the data in different ways to check if systematic misreporting is

[13]A note on non-response rates is worth including here. In the control group, the non-response rate for the IPV module with the direct questions is 5.4%. List experiments do not lead to a big difference in that respect: the non-response rate for the module with list experiments is 3.9% in the treatment group and close to null in the control group.

identified for the case of physical and sexual IPV. Since the costs of exposure are likely to vary by the level of economic and social empowerment, marital status, and the number of children of the victim, we designed the survey instrument to be able to test for differences by these characteristics. However, we remain agnostic as to how these costs vary according to the observable characteristics of the woman. For example, more economically empowered women may be more likely to report truthfully since they do not fear the loss of economic support of their partner; but they may also face a greater burden from stigmatization which would make them more likely to underreport.

Even though we looked at gap across sub-samples defined by several observable characteristics, we only find evidence of misreporting among the most educated women in the sample. Table 6 shows that there are large positive gaps in the prevalence rates reported under indirect and direct methods in the group of women with complete tertiary education. The joint significance test of the gaps confirms that there is systematic misreporting in this group but not among less educated women in the sample.

Figure 1: Gap in IPV Prevalence Rates across Education Levels by Reporting Method



NOTE: The gap reported in each bar is the difference in prevalence rates across the groups of women with high and low education. High-education level is defined as completed tertiary education.

Interestingly, the measured bias among the most educated women is large enough to reverse the education gradient in violence. Figure 1 plots the difference in prevalence rates by education level for each reporting method. Under direct methods (darker bars), the difference between high and low education groups is negative for all acts of violence, implying that prevalence rates are higher for the least educated women. This negative correlation

between education level and prevalence rates reverses once indirect methods are used: the gap in prevalence rates across education levels turns positive for all acts of physical and sexual IPV. Once the costs of being exposed are minimized, women with (complete) tertiary education exhibit higher prevalence rates of physical and sexual IPV than less educated women.

Table 7: Joint significance test of $(\rho - p)$: Heterogeneous effects

Characteristic	χ^2	Prob $> \chi^2$
Age		
<50	6.610	0.471
50+	7.070	0.422
Marital status		
Single	7.765	0.354
Married	4.226	0.753
Education level		
Less than tertiary	10.033	0.187
Completed tertiary	21.632	0.003
Mother tongue		
Spanish	13.017	0.072
Other language	5.033	0.656
Memory test		
Low score	3.325	0.853
High score	9.061	0.248
Household head		
Not the head	8.436	0.296
Head	3.046	0.881
Employment		
Does not work	4.712	0.695
Works	7.288	0.400
Standing in ADRA		
Young client	11.178	0.131
Mature client	5.519	0.597

NOTE: Joint test that the seven biases are different from zero. See Table 6 for details about the regressions. Mature clients are those with loan size and savings balance above the 75th percentile and a tenure greater than two loan cycles.

Table 7 reports the joint significance tests that the bias in all measured acts of physical and sexual IPV is different from zero by sub-samples. Surprisingly, no other measure of empowerment is correlated with significant biases in the report of violence at the 95% confidence level.

We argue that the effect among more educated women is not capturing a better un-

derstanding of the list experiment questions since there are no significant biases for other characteristics that may proxy better understanding of the methodology (see Tables A.5, for language, and A.6 for the memory test in Appendix A). As mentioned above, putting forward an explanation for why education level is the main characteristic that generates systematic misreporting in our sample goes beyond the scope of this paper. Our goal is to use this case study to highlight potential problematic patterns of non-random misreporting in survey data. With the data collected in our survey and the lack of random variation in the characteristics of respondents, we cannot fully pin down the underlying sources of misreporting among more educated women. Nevertheless, below we try to provide an explanation for the differential importance of the costs of exposure by education level and present some suggestive evidence along those lines.

In most policy forums, women empowerment is considered as a tool to reduce the prevalence of IPV [e.g., Klugman et al., 2014]. However, both theoretical and empirical work show that the relationship between empowerment indicators such as education and the probability of being a victim of IPV is ambiguous. On one hand, greater access to information among more educated women may change their attitudes towards social and gender norms, which can make them less tolerant of male dominance and violent behavior at home. Moreover, under assortative matching, women with more years of schooling are more likely to find partners who are also more educated and exposed to more equal social and gender norms.

On the other hand, greater returns and better access to job market opportunities among highly educated women may lead to different equilibria within the household. Intra-household bargaining models predict that, as long as education increases their outside option, more educated women should see violence experience reduced when compared to less educated ones [Farmer and Tiefenthaler, 1996]. However, instrumental theories of IPV highlight the use of violence by men in order to control resources at home [Eswaran and Malhotra, 2011]. Depending on the context, this backlash effect may undo the positive effects of empowerment through education on IPV. Indeed, this backlash effect is the only channel that could explain the positive relationship between education level and IPV as identified through *indirect* methods.

Now, what makes it more costly for highly educated women in our sample to expose their partners's violence? There is no reason to believe that emotional attachment should be differential across education levels. In addition, more educated women should fear *less* the potential loss of their partners' economic support. We speculate that both stigma concerns and fear of retaliation could be greater burdens when reporting directly among the more educated. Exposure to more equal gender social norms increases the costs imposed by

stigma.[14] Moreover, the backlash effect can make fear of retaliation more intense [e.g., Macmillan and Gartner, 1999].

The idea of exposure to new social norms and stigma as the reason for differential underreporting could also explain the findings by De Cao and Lutz [forthcoming] regarding female genital cutting in Ethiopia. Since the practice was declared illegal, one can expect a shift in social norms when asked about support of FGC. In this context, less educated women are expected to experience a greater change since they are likely to be less exposed to information and equal gender norms before the law was passed. Accordingly, De Cao and Lutz find that it is precisely this group where underreporting is the largest. Along the same lines, the authors also find that underreporting the support of FGC practices is higher among women exposed to a campaign seeking to change households' behavior regarding sexual and reproductive health. Under the light of this evidence, we speculate that more educated women in our experiment face greater stigma costs due to the social norms they are exposed to, which leads to higher underreporting levels relative to the less educated.[15]

4 Non-classical measurement error in the outcome

Our results show that, on average, there is no evidence of misreporting of physical and sexual IPV experience. However, the provision of anonymity through list experiments exposes the presence of non-classical measurement error. More educated women underreport when using DHS-type direct questions, the current best-practice and the most common way to measure violence in applied research.

This finding has extremely important implications on the empirical literature that tries to identify the main drivers and triggers of intimate partner violence. In a context where evidence is increasingly being used to move into action in the policy arena, our results are particularly important as they show that targeting strategies and prevention and mitigation programs may be designed with the wrong parameters in mind.

4.1 The data generating process

To understand the implications of the presence of non-classical error in the measurement of an outcome, we consider a simple model. Suppose that a researcher wants to estimate β:

[14] See Lindbeck et al. [1999] for an example of how social norms and stigma are related in the case of welfare recipients.

[15] Note, however, that one cost the list experiments cannot remove is the one associated to admitting to oneself that you are experiencing violence.

$$y_i = \beta x_i + \epsilon_i \qquad i = 1, \ldots, N. \tag{3}$$

In our particular case of interest, y_i would capture a measure of IPV and x_i would represent women's education, her income, or any other "risk factor" explored in the literature. The error term ϵ_i is assumed to be iid and, for simulation purposes, distributed $N(0, 1)$. For simplicity, (3) assumes that y_i and x_i are measured in deviations from the mean and ignores the role that other variables can play in explaining violence against women.[16]

Now consider the case when y_i is measured with some noise. The researcher observes $\tilde{y_i}$ instead of the true value, y_i:

$$\tilde{y_i} = y_i + \omega_i$$

Furthermore, let x_i be measured *without* error[17] and define it as follows:

$$x_i = \gamma \epsilon_i + \tau_i$$

That is, the risk factor is correlated with ϵ_i whenever $\gamma \neq 0$, introducing endogeneity in the estimation of β. In the simulations, we assume that $\tau_i \sim N(0, \kappa)$ so that $\text{var}(\tau_i) = \kappa \text{var}(\epsilon_i)$.

Now, we model the measurement error as a mix between a classical and a non-classical component:

$$\omega_i = \phi x_i + \nu_i \tag{4}$$

where $\nu_i \sim N(0, 1)$ for our simulations.

4.2 Causal estimation under endogeneity and measurement error biases

Consider the case where x_i is correlated with ϵ_i ($\gamma \neq 0$) and measurement error is non-classical ($\phi \neq 0$). In this situation, $E(\omega_i) = 0$, which is consistent with our findings of no underreporting, on average, so the measurement error has zero mean. However, two types of biases are introduced in the estimation of β using cross-sectional data:

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

[17] See Calvi et al. [2017] for an example where x is endogenous and measured with error but where y is observed without error.

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= \beta + \frac{\text{cov}(\epsilon_i, x_i)}{\text{var}(x_i)} + \frac{\text{cov}(\omega_i, x_i)}{\text{var}(x_i)} \\ &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi \end{aligned} \tag{5}$$

In equation (5), the second term captures the endogeneity bias because $\gamma \neq 0$ but the third element (ϕ) corresponds to the non-classical measurement error bias.

4.3 Implications for current evidence

Several papers in the literature have tried to estimate (3) via ordinary least squares using only cross-sectional variation to identify the impact of risk factors on violence against women.[18] More recent papers have tried to reduce or eliminate the endogeneity bias relying on exogenous variations introduced by RCTs. For example, Hidrobo and Fernald [2013], Hidrobo et al. [2016], Haushofer and Shapiro [2013], Angelucci [2008], and Bobonis et al. [2013], among others, have explored the role of income on IPV using the random allocation of conditional cash transfers (CCTs) to women in developing countries.[19] Other studies have tried to look at the impact of social norm interventions under an experimental design (see Pronyk et al. [2006] and World Health Organization [2009]). Another common strategy to deal with endogeneity problems is the use IV techniques as in Erten and Keskin [2018], where the authors rely on a school reform in Turkey as an instrument to evaluate the impact of women's education on the prevalence of violence.

By introducing random (or exogenous) variation in x_i, these papers are able to convincingly set $\gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} = 0$. However, if x_i in itself makes women more likely to misreport violence, the bias stemming from measurement error does not go away. This is very likely to occur in the context of CCT programs since the cash transfer tends to come within a bundle of other program components that may provide the recipient with information, changes in what is socially acceptable, or changes in the costs of being exposed. The same applies to education as the increase in human capital could translate into access to more information, exposure to different social norms, better access to labor market opportunities, to name a few of the factors that may affect the report of IPV.

Thus, non-classical measurement error imposes a limit to the gains that randomization or IV provide to obtain less biased estimates of treatment effects. Since ϕ in (5) does not go

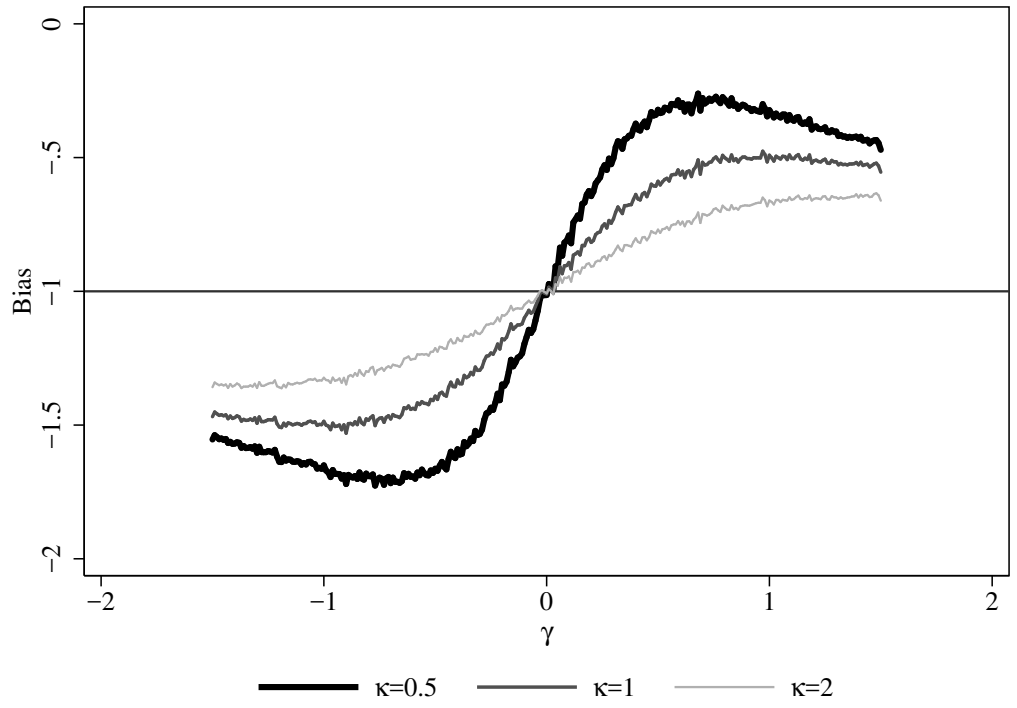
[18] See Jewkes et al. [2002], Koenig et al. [2003], Breiding et al. [2008], Fulu et al. [2013], where demographic and socioeconomic variables are considered among a long list of possible risk factors. See also Capaldi et al. [2012] for a recent review.

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

away under these methodologies, estimates of β could be still far off from the true value. In fact, OLS may yield *less* biased estimates of β whenever the sign of the correlation between x_i and ϵ_i is opposite to that of the correlation between x_i and ω_i.[20]

We conduct Monte Carlo simulations relying on the data generating process outlined in sub-section 4.1 to provide a better sense of the conditions that yield less biased estimates of β in the case of OLS when compared to RCTs and IVs. In Figure 2, we set $\phi = -1$ and plot the bias obtained by OLS for different values of κ (relative variance of the measurement error and random error) and γ (correlation between x_i and ϵ_i). First, note that if $\gamma = 0$, the only bias in the estimation of the effect of risk-factor x_i on IPV is driven by ϕ, which is shown in the horizontal line at -1. This is also true for estimates using valid IV. Second, cross-sectional studies that do not have an exogenous variation in x_i have smaller biases under OLS than RCTs (or IVs) when γ and ϕ have opposite signs. Since we set ϕ to -1, Figure 2 shows that the three lines get closer (vertically) to a zero bias when γ is positive.

Figure 2: Bias in OLS estimates ($\gamma \frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi$) with $\phi = -1$



NOTE: Simulations were conducted in a sample of 3000 observations with 100 replications. See text for details.

Moreover, $\gamma \frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi$ becomes close to zero whenever γ increases relative to ϕ, and more

[20]Note that by the nature of our experiment, where measurement error is not observed at the individual level, we can only estimate $E[\omega | x_i = x]$.

so whenever κ is smaller. Thus, the comparison of β estimates relying on RCT and observational data (from a baseline survey, for example) can be informative about the presence and potential magnitude of measurement error in IPV.

From (4), notice that ϕ is the slope of the relation between the risk factor of interest (x_i) and the measurement error in the dependent variable (ω_i). By conducting an experiment similar to ours, researchers can directly estimate $E[\omega_i|x_i = x]$ and obtain ϕ by correlating it with different values x. This will allow them to compute the bias in their estimates of β. We thus argue that the lists experiments used in our study provide an inexpensive way to directly measure ϕ and correct biased estimates from RCTs or IV methods. Based on our study's budget and sample size, the cost per women to conduct our experiment was close to US$8. For projects already conducting fieldwork, as those implementing a RCT, the marginal cost of adding the questions required to conduct list experiments is even smaller.

4.4 Non-linear measurement error

In the previous section, we consider the possibility of a linear source of non-classical measurement error as in Blattman et al. [2016]. We extend this case to consider non-linear and non-classical measurement error as the one we identify in our sample. We redefine the measurement error introduced in equation (4) as follows:

$$\omega_i = \pi_i(\phi x_i + \nu_i) + (1 - \pi_i)(\nu_i) \tag{6}$$

where $\pi_i = I[x_i > \mu_x]$ and $\mu_x = \bar{\mu}$. In this case, measurement error in the dependent variable is related to x_i in a non-linear way. As in our case study, the indicator function activates whenever the woman has completed tertiary education, i.e., has accumulated years of schooling above $\bar{\mu}$.

In this new framework, the OLS estimator of β becomes:

$$\begin{aligned} \beta_{OLS} &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi\frac{\text{cov}(x_i, \pi_i x_i)}{\text{var}(x_i)} \\ &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi E(\pi_i) \end{aligned} \tag{7}$$

Thus, when the measurement error is not linear, the bias of the OLS estimator still depends on ϕ as before but now it is also affected by the relative size of the group that generates non-classical measurement error.[21] As an example, we provide an estimate of the bias remaining when estimating treatment effects of college education on IPV using RCT or

[21] See Appendix D for derivation of Equation 7.

IV methods. Using the findings from Table 6 and the fact that 17.5 percent of the women in our sample completed college, we can estimate ϕ for a given act of IPV during a woman's experience with her last partner: the bias due to measurement error in β is 0.049 ((0.336-0.058)*0.175) in the case of having her hair pulled and 0.079 in the case of being attacked with a knife. Although we have no way to pin down the bias due to endogeneity, we provide $\hat{\beta}_{OLS}$ corresponding to education level in the case of these two acts of violence in our sample as a reference: -0.143 and 0.009 for having her hair pulled and being attacked with a knife, respectively.

5 Conclusion

Our paper uses indirect methods to measure misreporting in sensitive topics. In particular, we study the case of physical and sexual IPV as committed by the woman's last partner and rely on list experiments to provide full anonymity in its report.

We are the first to measure misreporting of IPV when using direct questions, the current best-practice and widely used in health surveys worldwide. We find that, on average, there are no significant differences in direct versus indirect reporting. Unlike previous studies using list experiments, our design avoids the contamination of the treatment group as the direct questions are asked only the to control. Furthermore, our results show that underreporting in our sample is concentrated among women with complete tertiary education, who do not fit the typical victim stereotype. This has important implications on the invisibility of violence that certain groups may suffer and the targeting efforts conducted to prevent and combat IPV. 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 victimization truthfully. 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 and greater fear of retaliation possibly related to a backlash effect.

Our contribution goes beyond our particular application to IPV. Even when (quasi) random assignment in the risk factor is introduced, non-classical measurement error in the dependent variable can still bias the estimates of treatment effects. We show that under certain conditions, randomization (and instrumental variables) could lead to even larger biases compared to cross-sectional studies. We provide a solution to correct biased causal effects under the presence of non-classical measurement error in the dependent variable. Paired with instrumental variable techniques or randomized controlled trials that deal with endogeneity biases, our approach offers the potential to estimate unbiased treatment effects.

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 could be more accepting of violence. But even in this setting we are able to find evidence of misreporting for a highly educated women. Further research should explore whether the misreporting 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 on IPV that try to estimate treatment effects with outcome variables that seem to be systematically misreported.

For studies examining the impact of risk factors on violence against women as well as for studies analyzing any other sensitive behavior in settings where administrative records are not reliable, we advocate for the inclusion of list experiment questions in the survey instruments used by researchers during data collection efforts. This will allow them to measure the magnitude of the bias in the estimated treatment effects introduced by non-classical measurement error based on the risk factor of interest.

It is worth highlighting that our design was implemented at a very low cost per woman (US$8). This implies that there are potentially important savings from this method when compared to other procedures [Blattman et al., 2016] that require intensive qualitative approaches. This opens up the possibility to replicate our design with other samples with different contextual characteristics.

References

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. [2008], 'Love on the Rocks: Domestic Violence and Alcohol Abuse in Rural Mexico', *The B.E. Journal of Economic Analysis & Policy* **8**(1), 1–43.

Bender, A. K. [2017], 'Ethics, methods, and measures in intimate partner violence research: the current state of the field', *Violence against women* **23**(11), 1382–1413.

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

Blair, G. and Imai, K. [2012], 'Statistical Analysis of List Experiments', *Political Analysis* **20**(1), 47–77.

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.

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.

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.

Calvi, R., Lewbel, A. and Tommasi, D. [2017], 'Late with mismeasured or misspecified treatment: An application to women's empowerment in india'.

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.

Coffman, K., Coffman, L. and Keith, M. [2013], The Size of the LGBT Population and the Magnitude of Anti-Gay Sentiment are Substantially Underestimated, Technical report, NBER Working Paper No. 19508.

De Cao, E. and Lutz, C. [forthcoming], 'Sensitive survey questions: Measuring attitudes regarding female genital cutting through a list experiment', *Oxford Bulletin of Economics and Statistics* (http://dx.doi.org/10.1111/obes.12228).

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. and Heise, L. [1999], 'Putting womens safety first: ethical and safety recommendations for research on domestic violence against women', *Geneva, Switzerland: World Health Organization* .

Ellsberg, M., Heise, L., Pena, R., Agurto, S. and Winkvist, A. [2001], 'Researching Ddomestic Violence Against Women: Methodological and Ethical Considerations', *Studies in Family Planning* **32**(1), 1–16.

Erten, B. and Keskin, P. [2018], 'For better or for worse?: Education and the prevalence of domestic violence in turkey', *American Economic Journal: Applied Economics* **10**(1), 64–105.

Eswaran, M. and Malhotra, N. [2011], 'Domestic Violence and Women's Autonomy in Developing Countries: Theory and Evidence', *Canadian Journal of Economics* **44**(4), 1222–1263.

Farmer, A. and Tiefenthaler, J. [1996], 'Domestic Violence: The Value of Services as Signals', *American Economic Review* **86**(2), 274–279.

Fulu, E., Jewkes, R., Roselli, T. and Garcia-Moreno, C. [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.

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.

Imai, K., Park, B. and Greene, K. [2014], 'Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models', *Political Analysis* **23**, 180–196.

Jamison, J. C., Karlan, D. and Raffler, P. [2013], 'Mixed-method evaluation of a passive mhealth sexual information texting service in uganda', *Information Technologies & International Development* **9**(3), pp–1.

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.

Joseph, G., Usman Javaid, S., Andres, L. A., Chellaraj, G., Solotaroff, J. L. and Rajan, S. I. [2017], Underreporting of Gender-Based Violence in Kerala, India: An Application of the List Randomization Method, Technical report, Policy Research Working Paper N. 8044, World Bank.

Karlan, D. and Zinman, J. [2012], 'List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds', *Journal of Development Economics* **98**, 71–75.

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.

Kishor, S. and Johnson, K. [2004], Profiling Domestic Violence: A Multi-Country Study, Technical report, Calverton, Maryland: ORC Macro.

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], 'Women's Status and Domestic Violence in Rural Bangladesh: Individual-and Community-level Effects', *Demography* **40**(2), 269–288.

Lara, D., García, S. G., Ellertson, C., Camlin, C. and Suárez, J. [2006], 'The measure of induced abortion levels in mexico using random response technique', *Sociological methods & research* **35**(2), 279–301.

Lindbeck, A., Nyberg, S. and Weibull, J. W. [1999], 'Social norms and economic incentives in the welfare state', *The Quarterly Journal of Economics* **114**(1), 1–35.

Macmillan, R. and Gartner, R. [1999], 'When she brings home the bacon: Labor-force participation and the risk of spousal violence against women', *Journal of Marriage and the Family* pp. 947–958.

McKenzie, D. and Siegel, M. [2013], Eliciting Illegal Migration Rates through List Randomization, Technical report, Policy Research Working Paper N. 6426, World Bank.

Moseson, H., Massaquoi, M., Dehlendorf, C., Bawo, L., Dahn, B., Zolia, Y., Vittinghoff, E., Hiatt, R. A. and Gerdts, C. [2015], 'Reducing under-reporting of stigmatized health events using the list experiment: results from a randomized, population-based study of abortion in liberia', *International journal of epidemiology* **44**(6), 1951–1958.

Organization, W. H. et al. [1997], 'Protocol for who multi-country study on womens health and domestic violence', *World Health Organization, Geneva, Switzerland* .

Overstreet, N. and Quinn, D. [2013], 'The Intimate Partner Violence Stigmatization Model and Barriers to Help-Seeking', *Basic Appl Soc Psych.* **35**(1), 109–122.

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.

Peterman, A., Palermo, T., Handa, S. and Seidenfeld, D. [2017], 'List randomization for soliciting experience of intimate partner violence: Application to the evaluation of Zambia's unconditional child grant program', *Health Economics Letter* pp. 1–7.

Pronyk, P., Hargreaves, J., Kim, J., Morison, L., Phetla, G., Watts, C., Busza, J. and Porter, J. [2006], 'Effect of a Structural Intervention for the Prevention of Intimate-Partner Violence and HIV in Rural South Africa: A Cluster Randomised Trial', *Lancet* **368**, 1973–83.

Rosenfeld, B., Imai, K. and Shapiro, J. N. [2016], 'An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions', *American Journal of Political Science* **60**(3), 783–802.

Straus, M. [1979], 'Measuring intrafamily conflict and violence: The conflict tactics (cs) scales', *Joornal of Marriage and tho Family, _4l,/5-88* .

World Health Organization [2009], Changing Cultural and Social Norms that Support Violence, Technical report, Series of briefings on violence prevention: the evidence.

Online appendices: not for publication

A Additional Figures and Tables

Table A.1: 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

NOTES: * 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.2: 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.3: Difference in estimated prevalence rates of physical and sexual IPV by age

	< 50 years old			50+ years old		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.491	0.304	0.188	0.304	0.324	-0.020
	(0.066)	(0.024)	(0.070)***	(0.089)	(0.034)	(0.095)
Slap	0.175	0.251	-0.075	0.186	0.293	-0.107
	(0.073)	(0.023)	(0.076)	(0.097)	(0.033)	(0.102)
Punch	0.229	0.213	0.016	0.124	0.245	-0.120
	(0.078)	(0.021)	(0.081)	(0.099)	(0.031)	(0.104)
Kick	0.092	0.124	-0.032	0.233	0.187	0.046
	(0.074)	(0.017)	(0.076)	(0.095)	(0.029)	(0.100)
Strangle	0.041	0.048	-0.007	-0.080	0.069	-0.149
	(0.073)	(0.011)	(0.073)	(0.084)	(0.019)	(0.086)*
Knife	0.039	0.048	-0.009	0.085	0.074	0.010
	(0.072)	(0.011)	(0.073)	(0.094)	(0.019)	(0.095)
Sex acts	0.029	0.059	-0.030	0.228	0.166	0.062
	(0.075)	(0.012)	(0.076)	(0.095)	(0.027)	(0.100)
Joint test						
χ^2		6.610			7.070	
Prob $> \chi^2$		0.471			0.422	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by age are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.4: Difference in estimated prevalence rates of physical and sexual IPV by marital status

	Single			Married		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.543	0.345	0.198	0.394	0.302	0.092
	(0.103)	(0.045)	(0.112)*	(0.064)	(0.022)	(0.067)
Slap	0.291	0.354	-0.063	0.150	0.242	-0.092
	(0.120)	(0.045)	(0.128)	(0.069)	(0.020)	(0.071)
Punch	0.323	0.336	-0.013	0.157	0.195	-0.038
	(0.125)	(0.044)	(0.132)	(0.073)	(0.019)	(0.075)
Kick	0.346	0.214	0.131	0.089	0.128	-0.039
	(0.121)	(0.039)	(0.127)	(0.069)	(0.016)	(0.071)
Strangle	-0.020	0.133	-0.152	0.003	0.036	-0.033
	(0.111)	(0.032)	(0.116)	(0.066)	(0.009)	(0.067)
Knife	0.164	0.097	0.067	0.027	0.047	-0.020
	(0.113)	(0.028)	(0.117)	(0.068)	(0.010)	(0.069)
Sex acts	0.152	0.133	0.019	0.086	0.085	0.000
	(0.125)	(0.032)	(0.130)	(0.070)	(0.013)	(0.072)
Joint test						
χ^2		7.765			4.226	
Prob $> \chi^2$		0.354			0.753	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by marital status are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.5: Difference in estimated prevalence rates of physical and sexual IPV by mother's tongue

	Spanish			Other language		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.470	0.315	0.155	0.134	0.281	-0.148
	(0.060)	(0.021)	(0.064)**	(0.153)	(0.056)	(0.162)
Slap	0.190	0.258	-0.069	0.111	0.317	-0.206
	(0.066)	(0.020)	(0.069)	(0.168)	(0.059)	(0.178)
Punch	0.233	0.216	0.017	-0.078	0.281	-0.360
	(0.070)	(0.019)	(0.073)	(0.168)	(0.056)	(0.177)**
Kick	0.125	0.138	-0.012	0.251	0.203	0.048
	(0.066)	(0.016)	(0.069)	(0.160)	(0.050)	(0.168)
Strangle	-0.002	0.054	-0.056	-0.001	0.063	-0.064
	(0.064)	(0.010)	(0.065)	(0.157)	(0.030)	(0.160)
Knife	0.102	0.056	0.045	-0.246	0.063	-0.308
	(0.067)	(0.010)	(0.068)	(0.131)	(0.030)	(0.134)**
Sex acts	0.124	0.083	0.041	-0.057	0.190	-0.247
	(0.067)	(0.012)	(0.069)	(0.172)	(0.049)	(0.180)
Joint test						
χ^2		13.017			5.033	
Prob $> \chi^2$		0.072			0.656	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by mother tongue are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.6: Difference in estimated prevalence rates of physical and sexual IPV by memory

	Bad memory			Good memory		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.385	0.350	0.035	0.461	0.270	0.190
	(0.080)	(0.028)	(0.084)	(0.070)	(0.027)	(0.075)**
Slap	0.153	0.262	-0.109	0.202	0.267	-0.065
	(0.083)	(0.026)	(0.087)	(0.080)	(0.027)	(0.084)
Punch	0.181	0.248	-0.068	0.201	0.198	0.004
	(0.090)	(0.026)	(0.094)	(0.082)	(0.024)	(0.086)
Kick	0.163	0.155	0.008	0.124	0.135	-0.011
	(0.083)	(0.021)	(0.086)	(0.081)	(0.021)	(0.084)
Strangle	-0.013	0.063	-0.076	0.008	0.047	-0.039
	(0.078)	(0.014)	(0.080)	(0.077)	(0.013)	(0.078)
Knife	-0.015	0.073	-0.088	0.118	0.040	0.078
	(0.082)	(0.015)	(0.084)	(0.077)	(0.012)	(0.078)
Sex acts	0.102	0.116	-0.014	0.097	0.073	0.024
	(0.086)	(0.019)	(0.088)	(0.080)	(0.016)	(0.082)
Joint test						
χ^2		3.325			9.061	
Prob $> \chi^2$		0.853			0.248	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by the memory test are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.7: Difference in estimated prevalence rates of physical and sexual IPV by household head status

	Household head			Not the household head		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.387	0.275	0.112	0.486	0.389	0.097
	(0.069)	(0.023)	(0.073)	(0.082)	(0.037)	(0.089)
Slap	0.125	0.240	-0.114	0.266	0.320	-0.054
	(0.076)	(0.022)	(0.078)	(0.089)	(0.035)	(0.096)
Punch	0.127	0.197	-0.071	0.297	0.282	0.015
	(0.081)	(0.020)	(0.083)	(0.092)	(0.034)	(0.098)
Kick	0.072	0.112	-0.040	0.255	0.218	0.036
	(0.075)	(0.016)	(0.078)	(0.091)	(0.031)	(0.096)
Strangle	-0.005	0.026	-0.031	0.004	0.120	-0.116
	(0.072)	(0.008)	(0.073)	(0.086)	(0.025)	(0.090)
Knife	-0.037	0.034	-0.071	0.205	0.109	0.097
	(0.075)	(0.009)	(0.075)	(0.085)	(0.024)	(0.089)
Sex acts	0.035	0.065	-0.030	0.204	0.160	0.044
	(0.076)	(0.013)	(0.077)	(0.093)	(0.028)	(0.098)
Joint test						
χ^2		8.436			3.046	
Prob $> \chi^2$		0.296			0.881	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by household head status are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.8: Difference in estimated prevalence rates of physical and sexual IPV by employment status

	Does not work			Works		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.443	0.351	0.092	0.418	0.296	0.122
	(0.086)	(0.039)	(0.094)	(0.067)	(0.023)	(0.070)*
Slap	0.088	0.272	-0.183	0.211	0.262	-0.051
	(0.098)	(0.036)	(0.104)*	(0.072)	(0.022)	(0.075)
Punch	0.185	0.252	-0.066	0.194	0.213	-0.019
	(0.104)	(0.035)	(0.110)	(0.076)	(0.020)	(0.079)
Kick	0.113	0.185	-0.072	0.153	0.130	0.022
	(0.105)	(0.032)	(0.111)	(0.071)	(0.017)	(0.073)
Strangle	0.066	0.086	-0.020	-0.026	0.044	-0.070
	(0.095)	(0.023)	(0.098)	(0.069)	(0.010)	(0.070)
Knife	-0.001	0.066	-0.067	0.076	0.054	0.022
	(0.096)	(0.020)	(0.099)	(0.071)	(0.011)	(0.072)
Sex acts	0.070	0.113	-0.043	0.110	0.088	0.022
	(0.100)	(0.026)	(0.104)	(0.074)	(0.014)	(0.076)
Joint test						
χ^2		4.712			7.288	
Prob $> \chi^2$		0.695			0.400	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by employment status are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

Table A.9: Difference in estimated prevalence rates of physical and sexual IPV by standing in ADRA

	Young client			Mature client		
Violent act	ρ	p	$(\rho - p)$	ρ	p	$(\rho - p)$
Pull hair	0.404	0.309	0.096	0.517	0.322	0.195
	(0.063)	(0.021)	(0.066)	(0.112)	(0.049)	(0.123)
Slap	0.148	0.279	-0.131	0.318	0.189	0.129
	(0.068)	(0.021)	(0.071)*	(0.126)	(0.041)	(0.133)
Punch	0.197	0.237	-0.039	0.166	0.156	0.010
	(0.073)	(0.020)	(0.076)	(0.127)	(0.038)	(0.132)
Kick	0.098	0.152	-0.053	0.337	0.111	0.226
	(0.068)	(0.017)	(0.071)	(0.124)	(0.033)	(0.129)*
Strangle	-0.038	0.062	-0.099	0.158	0.022	0.136
	(0.066)	(0.011)	(0.067)	(0.113)	(0.016)	(0.114)
Knife	0.039	0.064	-0.025	0.129	0.022	0.107
	(0.067)	(0.011)	(0.069)	(0.125)	(0.016)	(0.126)
Sex acts	0.048	0.090	-0.041	0.326	0.122	0.204
	(0.070)	(0.013)	(0.072)	(0.121)	(0.035)	(0.127)
Joint test						
χ^2		11.178			5.519	
Prob $> \chi^2$		0.131			0.597	

NOTE: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by standing in ADRA are obtained from the model in Equation (2). Significance levels: * 10%; ** 5%; *** 1%.

B Sample Instruments

B.1 Informed Consent

Thanks for agreeing to talk to me. My name is ... I work as a surveyor for the University of Connecticut and the Inter-American Development Bank, who are performing a study about female microentrepreneurs in Peru. I kindly request your participation in this interview. While I read the instructions and questions, please tell me whether there is anything that you do not understand.

You have been selected to participate in this interview as a client of ADRA Peru. We are interviewing ADRA's clients in the districts of Metropolitan Lima to collect information about the current situation of their families. I would like to ask you some questions about you, your household, and the decisions that are taken in your family. The interview lasts for about 15 minutes and your participation is voluntary. I will write down your answers as we move forward. I ask you to please answer honestly and with the best of your knowledge. If you agree to participate, you can decide to end the interview at any moment or to skip any question that you would prefer not to answer.

Your answers are anonymous and their confidentiality will be protected at every moment. Your name will not be used in any report and your answers will not be delivered to any person that does not belong to the research team. The analysis of the data that we collect will allow us to understand the current situation of women and their families in the area. If you have any question, you can ask me or other person in our research team. If you have any problem or inquiry about this study after the interview, please feel free to contact: Jorge Agüero or Veronica Frisancho, the researchers in charge of the project.

B.2 IPV and List Experiment Modules

B.2.1 Control Questionnaire: Direct Questions on Physical and Sexual IPV

1	Has your partner ever pulled your hair?
2	Has your partner ever pushed, shaken, or thrown something at you?
3	Has your partner ever slapped you or twisted your arm?
4	Has your partner ever punched you or hit you with something dangerous?
5	Has your partner ever kicked or dragged you?
6	Has your partner ever tried to strangle or burn you?
7	Has your partner ever threatened you with a knife, gun, or other weapon?
8	Has your partner ever forced you to have sexual intercourse when you did not want to?
9	Has your partner ever forced you to practice sexual practices that you do not approve of?

B.2.2 Control Questionnaire: List Experiment Questions

I am going to read to you a list of statements. Could you please tell me how many of them are true? Do not tell me which ones are true, only how many of them are true.

Have you ever...?

1. (a) Purchased a TV with HD
   (b) Been out for a walk with your children
   (c) Helped your children with their homework
   (d) Bought expensive clothes
2. (a) Played any games in your cellphone
   (b) Purchased new clothes for your children on important dates (e.g. Christmas, birthdays, others)
   (c) Traveled with your family on holidays
   (d) Seen *Al fondo hay sitio*[22]
3. (a) Pursued a technical degree
   (b) Experienced that your business' sales are insufficient to cover your household expenses
   (c) Traveled with friends
   (d) Been to the movies with your family
4. (a) Witnessed robberies in your neighborhood
   (b) Been robbed on the street
   (c) Had insurance from ESSALUD, the armed forces, or the police
   (d) Been depressed
5. (a) Felt insecure in your neighborhood
   (b) Had multiple businesses at the same time
   (c) Reared farm animals for consumption
   (d) Used the subway as a means of transportation
6. (a) Run out of money to cover the household's monthly expenses
   (b) Traveled with your children
   (c) Been part of a Christian church
   (d) Had to truncate your studies to care for your family
7. (a) Paid rent for the place where you live
   (b) Participated in other microfinance programs
   (c) Bought high-end clothes
   (d) Participated in a committee or association in your neighborhood
8. (a) Lost your cell phone

[22] *Al fondo hay sitio* is a very popular soap opera than ran for several years in Peru.

(b) Read *El Comercio*[23]

(c) Seen any soap opera

(d) Visited the Lima's cathedral

9. (a) Made improvements to your dwelling

(b) Had problems with your partner because of money issues

(c) Received a loan from *Mi Banco*

(d) Suffered from a serious medical condition that has required medical assistance

B.2.3 Treatment Questionnaire: List Experiment Questions

I am going to read to you a list of statements. Could you please tell me how many of them are true? Do not tell me which ones are true, only how many of them are true.

Have you ever...?

1. (a) Purchased a TV with HD

(b) Been out for a walk with your children

(c) Helped your children with their homework

(d) Bought expensive clothes

(e) Had your hair pulled by your partner?

2. (a) Played any games in your cellphone

(b) Purchased new clothes for your children on important dates (e.g. Christmas, birthdays, others)

(c) Traveled with your family on holidays

(d) Seen *Al fondo hay sitio*[24]

(e) Been pushed, shaken, or thrown something at you by your partner?

3. (a) Pursued a technical degree

(b) Experienced that your business' sales are insufficient to cover your household expenses

(c) Traveled with friends

(d) Been to the movies with your family

(e) Been slapped or had your arm twisted by your partner?

4. (a) Witnessed robberies in your neighborhood

(b) Been robbed on the street

[23] *El Comercio* is one of the most read newspapers in the country, particularly in Lima.

[24] *Al fondo hay sitio* is a very popular soap opera than ran for several years in Peru.

(c) Had insurance from ESSALUD, the armed forces, or the police

(d) Been depressed

(e) Been punched or hit with something dangerous by your partner

5. (a) Felt insecure in your neighborhood

(b) Had multiple businesses at the same time

(c) Reared farm animals for consumption

(d) Used the subway as a means of transportation

(e) Been kicked or dragged by your partner

6. (a) Run out of money to cover the household's monthly expenses

(b) Traveled with your children

(c) Been part of a Christian church

(d) Had to truncate your studies to care for your family

(e) Had your partner trying to strangle or burn you

7. (a) Paid rent for the place where you live

(b) Participated in other microfinance programs

(c) Bought high-end clothes

(d) Participated in a committee or association in your neighborhood

(e) Been threatened with a knife, gun, or other weapon by your partner

8. (a) Lost your cell phone

(b) Read *El Comercio*[25]

(c) Seen any soap opera

(d) Visited the Lima's cathedral

(e) Been forced to have sexual intercourse when you did not want to by your partner

9. (a) Made improvements to your dwelling

(b) Had problems with your partner because of money issues

(c) Received a loan from *Mi Banco*

(d) Suffered from a serious medical condition that has required medical assistance

(e) Been forced to practice sexual practices that you do not approve of by your partner

[25] *El Comercio* is one of the most read newspapers in the country, particularly in Lima.

C 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 these 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 C.1 reports some statistics on the prevalence rates of the set of non-sensitive statements that go along with each sensitive statement, relying on data from the pilot. For each sensitive sentence, the first column reports the mean prevalence of the four accompanying neutral statements, while the second and third report the standard deviation and the 75th percentile of the prevalence of the same neutral statements. 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 neutral 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 C.1: Distribution of Prevalence Rates of Neutral Statements Accompanying each Sensitive Statement

	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.

D β_{OLS} under the presence of non-linear and non-classical measurement error

Under the presence of non-linear and non-classical measurement error, the OLS estimator of β becomes:

$$\beta_{OLS} = \beta + \gamma \frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi \frac{\text{cov}(x_i, \pi_i x_i)}{\text{var}(x_i)} \tag{D.1}$$

Let

$$\text{cov}(x_i, \pi_i x_i) = E(\pi_i x_i^2) - E(x_i)E(\pi_i x_i) \tag{D.2}$$

where

$$\begin{aligned} E(\pi_i x_i^2) &= E(\pi_i x_i^2 | \pi_i = 1)P[\pi_i = 1] + E(x_i \pi_i x_i | \pi_i = 0)P[\pi_i = 0] \\ &= E(x_i^2)P[\pi_i = 1] \end{aligned} \tag{D.3}$$

and

$$\begin{aligned} E(\pi_i x_i) &= E(\pi_i x_i | \pi_i = 1)P[\pi_i = 1] + E(\pi_i x_i | \pi_i = 0)P[\pi_i = 0] \\ &= E(x_i)P[\pi_i = 1] \end{aligned} \tag{D.4}$$

Plugging D.3 and D.4 into D.2 yields:

$$\begin{aligned} \text{cov}(x_i, \pi_i x_i) &= E(x_i^2)P[\pi_i = 1] - E(x_i)E(x_i)P[\pi_i = 1] \\ &= P[\pi_i = 1][E(x_i^2) - E^2(x_i)] \\ &= P[\pi_i = 1]\text{var}(x_i) \\ &= E(\pi_i)\text{var}(x_i) \end{aligned} \tag{D.5}$$

If we replace D.5 into D.1, we obtain the last line in 7:

$$\beta_{OLS} = \beta + \gamma \frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi E(\pi_i)$$