IDB WORKING PAPER SERIES Nº IDB-WP-853

Misreporting in Sensitive Health Behaviors and Its Impact on Treatment Effects:

An Application to Intimate Partner Violence

Jorge Agüero
Verónica Frisancho

Inter-American Development Bank
Department of Research and Chief Economist

December 2017

Misreporting in Sensitive Health Behaviors and Its Impact on Treatment Effects:

An Application to Intimate Partner Violence

Jorge Agüero*
Verónica Frisancho**

* University of Connecticut
** Inter-American Development Bank

December 2017

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Agüero, Jorge.
Misreporting in sensitive health behaviors and its impact on treatment effects: an application to intimate partner violence / Jorge Agüero, Verónica Frisancho.
p. cm. — (IDB Working Paper Series ; 853)
Includes bibliographic references.
1. Intimate partner violence. 2. Women-Violence against. 3. Health behavior. 4. Health surveys. I. Frisancho Robles, Verónica C. II. Inter-American Development Bank. Department of Research and Chief Economist. III. Title. IV. Series.
IDB-WP-853

http://www.iadb.org

Abstract

A growing literature seeks to identify policies that could reduce intimate partner violence. However, in the absence of reliable administrative records, this violence is often measured using self-reported data from health surveys. In this paper, an experiment is conducted comparing data from such surveys against a methodology that provides greater privacy to the respondent. Non-classical measurement error in health surveys is identified as college-educated women, but not the less educated, underreport physical and sexual violence. The paper provides 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-random measurement error, List experiments, Intimate partner violence, Treatment effects

JEL Classification: C83, C21, I12

1 Introduction

Much of the empirical work in economics relies on self-reported data, despite the possible presence of measurement error in survey responses.[1] For different reasons, ranging from random mistakes, limited attention, and lack of recollection to behavioral biases or stigma, respondents could give inaccurate answers that introduce measurement error in the data. 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.

In recent years, measurement error concerns have been increasingly addressed in the literature. An important share of these studies has made use of administrative records to directly measure and characterize misreporting in sensitive topics such as voting (Rosenfeld, Imai, and Shapiro, 2016), mental health conditions (Bharadwaj, Pai, and Suziedelyte, 2015), or personal earnings (Gottschalk and Huynh, 2010). However, in several cases this is not an alternative due to lack of accurate administrative data or self-selection into such reporting.

Using an indirect questioning technique, this paper measures and characterizes misreporting when dealing with a sensitive topic and proposes an alternative to quantify the bias introduced by measurement error in the estimation of treatment effects. In particular, we focus on the measurement of physical and sexual intimate partner violence (IPV), due both to its saliency as a public health issue and the urgency of generating accurate data on its prevalence to guide policy efforts.

Our focus on IPV is also extremely timely as a growing number of studies try to identify the main drivers of this phenomenon (e.g., Angelucci, 2008, Hidrobo and Fernald, 2013, Haushofer and Shapiro, 2013, Bobonis, González-Brenes, and Castro, 2013, Hidrobo, Peterman, and Heise, 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, Heise, Pena, Agurto, and Winkvist, 2001, Kishor, 2005, Aizer, 2010), but little is known about the magnitude and the characteristics of misreporting in this field. The use of inaccurate self-reported data on victimization could introduce important distortions into the estimates of treatment effects in the aforementioned studies. While classical measurement error in the dependent variable only affects the precision of the estimates, the presence of non-classical measurement error makes it impossible to obtain unbiased causal effects of a variable of interest, even under valid exogenous variation in the variable of interest.

To address the issue of misreporting, and in the absence of reliable administrative data, we compare the prevalence rates of physical and sexual IPV estimated by two different survey

[1] See Meyer, Mok, and Sullivan (2015) for a recent discussion of unit nonresponse and item nonresponse problems in household surveys.

methods that provide varying degrees of privacy to the respondent. The first one follows the direct questions as applied by the Demographic and Health Surveys (DHS), a global project that is the main source of IPV data (Klugman, Hanmer, Twigg, Hasan, McCleary-Sills, and Santamaria, 2014) and that has been implemented in 122 surveys covering 61 developing countries.[2] The second method provides further anonymity to the respondent through the use of indirect questions in the form of list experiments (e.g., Blair and Imai, 2012, Glynn, 2013, Karlan and Zinman, 2012). Both methods are applied to a sample of female clients of a microcredit organization operating in several impoverished peri-urban districts in Lima, Peru.

In particular, we randomize two questionnaires at the individual level. The control group receives the nine direct questions that the DHS uses to measure the prevalence of physical and sexual IPV. In addition, the control group receives nine lists of four neutral statements and is asked to provide the number of statements that hold true in each list but not the individual occurrence of each statement. The treatment group does not answer the direct IPV questions. Instead, the treatment answers nine lists of five statements each, where the first four statements 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. As before, respondents only answer *how many* statements are true but not which ones. Randomization guarantees that the average number of neutral statements is equal across treatment arms. Thus, the prevalence rate of a given act of physical or sexual violence is 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. There is no difference for 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 identified when prevalence rates are measured through direct questions.

We argue that our results have ample applications in settings where the dependent variable suffers from non-random measurement error (Bound, Brown, and Mathiowetz, 2001, Butler, Burkhauser, Mitchell, and Pincus, 1987) and where administrative records are not a data source alternative. As a general result, we review the implications of systematic misreporting on the estimation of causal effects. A common strategy to deal with endogeneity biases has been to rely on the exogenous variation introduced in the variable of interest through a randomized controlled trial (RCT) or a quasi-experimental approach using instrumental variables (IV). We show that RCTs and (valid) IVs still yield biased treatment effects in the presence of non-classical measurement

[2] The World Health Organization has conducted similar surveys about IPV but applied them in a smaller set of countries (10 in total). For simplicity, we refer to both surveys as *DHS-type*.

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.

Our experimental approach provides researchers with a simple and inexpensive strategy to test for measurement error, classical or not, in contexts where administrative records are not available and fieldwork is being conducted. By providing full anonymity to the respondent, we minimize the costs of being exposed as a victim and obtain a benchmark measure that can be used to gauge the characteristics of the reporting error for a given sensitive outcome. Furthermore, our approach allows researchers to correct their treatment effect estimates. These contributions are particularly valuable for the case of IPV, since previous efforts have neither attempted to quantify the severity and patterns of underreporting in such sensitive behavior nor explored the implications of misreporting for the estimation of treatment effects.

The paper is divided in five sections including this introduction. Section 2 reviews the literature on misreporting when sensitive information is collected. Section 3 reviews the design of the indirect method we relied upon, describes the data and the sample, provides details on the estimation strategy, and presents the results. Section 4 discusses the implications of our results on the estimation of the causal effects of risk factors on IPV, presents simulation results that quantify the magnitude of the bias introduced under different scenarios, and provides practical guidelines to deal with measurement error bias in the estimation of treatment effects. The last section concludes.

2 Misreporting in Sensitive Survey Questions

There is an extensive literature showing that measurement error in survey data on certain topics is not random but rather correlates with an array of characteristics. For instance, income and asset data are prone to systematic measurement error due to distrust, lack of recall, strategic reporting, stigma, or a desire to reduce the interview time, among other reasons. Poorer households, for example, could underreport their income if they perceive the data collected are going to be used to distribute social welfare benefits. At the same time, they tend to have more sources of income at diverse frequency rates, which could lead them to misreport due to lack of perfect recall. Indeed, Meyer, Mok, and Sullivan (2008) shows that welfare benefits may be misreported. Also, by relying on tax records, Gottschalk and Huynh (2010) shows that there is substantial measurement error in earnings and that this error is correlated with (true) earnings and positively correlated across time.

Reporting of health outcomes also suffers from such bias (Bound, Brown, and Mathiowetz, 2001). For example, Butler, Burkhauser, Mitchell, and Pincus (1987) show evidence of non-classical error in the measurement of arthritis while Johnston, Propper, and Shields (2009) finds a similar pattern in hypertension self-reporting. O'Neill (2012) identifies a negative correlation between self-reported and anthropometric measures of body mass index. More recently, Bharad-

waj, Pai, and Suziedelyte (2015) relies on administrative records and finds that underreporting in mental health medication is correlated with age, gender, and ethnicity.

Non-classical measurement error in the dependent variable makes it impossible to obtain unbiased causal effects of a particular characteristic or attribute, especially if the latter is correlated with misreporting behavior. For example, a well-known puzzle in the development economics literature is that of an inverse plot size-productivity relationship. Two recent studies (Gourlay, Kilic, and Lobell, 2017, Desiere and Jolliffe, 2018) show that whenever self-reported measures of yields are replaced by more accurate measures, the relationship between plot size and productivity vanishes.

In the case of risky behaviors such as crime or violence, the identification of causal relationships is particularly crucial since these findings tend to guide costly policy efforts and targeting strategies. Even when exogenous variation in the hypothesized risk factor is introduced, misleading conclusions may emerge if the dependent variable is systematically misreported.

The Case of Intimate Partner Violence

There is a growing consensus on the best practices in how to ask questions about IPV. They have been compiled and proposed by the WHO Organization et al. (1997) and further discussed by Ellsberg and Heise (1999). For example, participants should be provided several opportunities to respond about issues related to IPV. Generic and subjective questions such as "Have you ever experienced domestic violence?" must be avoided and, instead, several specific acts of violence should be inquired about.

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. For instance, Ellsberg, Heise, Pena, Agurto, and Winkvist (2001) argues that when greater privacy measures are enforced, higher rates for IPV are reported, relative to the DHS methodology. While this evidence is suggestive, the authors cannot isolate the fact that the compared surveys were conducted in different years and without an experimental design.

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-identifying 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 cannot provide a benchmark for the measurement of prevalence rates since the nature of IPV implies that administrative records from the police or health establishments capture a non-random sample of the true cases. 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. Using surveys from 24 countries in the DHS program, Palermo, Bleck, and Peterman (2014) shows that only 7 percent of women who experienced such violence made a formal report that would be captured in administrative data (e.g., police, medical, or social services). Moreover, the authors also show that reporting depends on women's socioeconomic characteristics such as age, marital status, education, and urban location. Since women who make an active effort to report are also the ones who face lower exposure costs, selection into reporting is most likely present in administrative records.

Our paper relies on list experiments to measure and characterize the reporting error in the prevalence of physical and sexual lifetime experience of IPV as committed by the woman's last partner.[3] List experiments provide full anonymity to respondents, which minimizes the costs of being exposed as a victim and/or exposing the aggressor. Thus, we provide a significant contribution to the literature on violence against women by establishing a benchmark and characterizing misreporting.[4]

[3] 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, Coffman, and Keith (2013) to measure the size of LGBT population and anti-gay sentiment, Imai, Park, and Greene (2014) to examine vote-selling, and Rosenfeld, Imai, and Shapiro (2016) to study anti-abortion support.

[4] Alternative methods include qualitative approaches as in Blattman, Jamison, Koroknay-Palicz, Rodrigues, and Sheridan (2016). The authors combine surveying with ethnographic techniques to uncover misreporting. Their approach does not provide additional anonymity to the respondents. It is also quite expensive since it requires the survey team to stay for longer periods in the field and its success depends heavily on the surveyors' ability to make the respondent feel safe and comfortable to truthfully report or reveal her answers or behavior. Surveyors' training becomes crucial, which only adds to the cost of the fieldwork, making it hard to scale up. There are other indirect questioning techniques such as endorsement experiments or randomized response techniques, which are often used in the political science literature but that are not appropriate to measure IPV prevalence. The former is not adequate since it is designed to measure attitudes rather than behavior. Randomized response methods do measure behavior but generate high non-response

Similar to Karlan and Zinman (2012), we recruit a large enough sample to only ask the control group about their previous violence experience using face-to-face DHS-type survey questions. This allows us to ensure full protection to the treatment group, who only answers the list experiment questions that include the sensitive statement.

Two recent studies are closely related to our paper: Joseph, Usman Javaid, Andres, Chellaraj, Solotaroff, and Rajan (2017) and Peterman, Palermo, Handa, and Seidenfeld (2017). They both rely on list experiments to measure prevalence rates of physical violence only. Their contribution is valuable but they have several limitations. First, Joseph, Usman Javaid, Andres, Chellaraj, Solotaroff, and Rajan (2017) measures prevalence rates at the household level, which implies that the respondent is not necessarily a woman. 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 is the perpetrator himself. Second, their sensitive statement is quite general (*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 WHO guidelines for the measurement of violence, which require asking about several and specific violent events. The same holds for Peterman, Palermo, Handa, and Seidenfeld (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, Usman Javaid, Andres, Chellaraj, Solotaroff, and Rajan (2017) nor Peterman, Palermo, Handa, and Seidenfeld (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 indirect questions to a control group that differs from the treatment group, and iv) comparing the prevalence rates obtained from the indirect method to the ones that come from the DHS direct method.

3 Measuring Reporting Bias in Violence Against Women

3.1 List Experiments: Design

List experiments have been traditionally used to gather opinions and/or record 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

rates since the burden to conduct the randomization is imposed on the respondent. This method can be hard to grasp, even among highly educated respondents.

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. Below we show how the comparison between the average number of true statements across groups yields the prevalence rate of the sensitive statement while providing full anonymity to the respondent.

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. However, we observe the number of responses 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 assumption of no design effects[5], that is, 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:

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

In other words, 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. The prevalence rate of the sensitive statement can thus be 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 intimate partner physical and sexual violence during a woman's lifetime as committed by her last partner. In particular, the sensitive statements used in the lists reproduce the ones asked directly in the DHS when trying to directly measure physical and sexual IPV prevalence.

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 from answering honestly.

[5] We later conduct two tests that validate this assumption for our particular exercise.

[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 this data include neutral lists of four statements.

On the other hand, a list that avoids the problems stated above will most likely introduce greater variability into the responses, which could then increase the variance of the estimator. Glynn (2013) provides some guidance in the development of lists so as to maximize the level of protection while sacrificing little variance. He shows that introducing negative correlation between the responses to the neutral items in the list limits the variability of the responses while minimizing the likelihood of ceiling effects. 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. Since the data collected under this method does not allow the researcher to link prevalence rates to respondents' other characteristics, the analysis of correlations between the experience of violence and other variables is limited. However, with sufficiently large samples sizes one can measure prevalence rates by sub-samples as we do here (see Section 3.4).

3.2 Sample Description and Data

The population of interest for our study is composed of adult women in Lima who receive microloans from the Adventist Development and Relief Agency (ADRA), an international non-governmental organization (NGO) running a village banking program in Peru's peri-urban and rural areas.[7] ADRA's clients in Lima are microentrepreneurs from the most impoverished districts such as San Juan de Lurigancho, Villa Maria del Triunfo, Villa El Salvador, Ventanilla, Huaycan, and Los Olivos.

From the total pool of 1,873 clients in 112 village banks in ADRA's microcredit 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 1,776 clients. We draw six 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 1,690 clients in 106 banks. Finally, we work with all banks with monthly meetings scheduled during July 2015, which restricts the population of interest to 1,562 women in 98 village banks. We targeted this restricted universe and were able to interview 1,223 women between July 1 and August 25, 2015.

Randomization of the treatment was done at the individual level and was conducted by the surveyor. The questionnaire was implemented via tablet computers. 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 are left with a sample of 1,078 valid surveys.

[7] This research was approved by the University of Connecticut's Institutional Review Board, Protocol #H15-164: "Measuring Violence Against Women with Experimental Methods."

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]

Table A.1 in Appendix A confirms that the randomization was successful. There is only a small significant difference in the share of women that are household heads across treatment arms (at the 5 percent level). All our estimates include a full set of controls, including a binary variable that indicates if the woman is the household head.

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.

We took special care in the design of the questionnaires. As described earlier, we piloted the non-sensitive statements in a small sample of ADRA's clients who were not part of the experimental sample. We came up with a list of 41 statements and asked 31 individuals to provide a yes/no answer in order to measure the prevalence rates of each statement. The questions were framed around the subject's lifetime experience to be in line with the sensitive items on violence intended to measure prevalence rates in a woman's lifetime.

The prevalence rates of the non-sensitive statements were useful in two ways. On one hand, they measured the adequacy of the statements for 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 (Glynn, 2013).[9] Table A.2 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.[10] Table A.3 in Appendix A reports the correlation of prevalence rates in each set of statements grouped together.

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 control groups.[11] This reduces potential biases that may be introduced when asking the same respondent both the

[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.

[9] 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.

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

[11] See sample instruments in Appendix C.

direct and indirect questions as done in Karlan and Zinman (2012) and Joseph, Usman Javaid, Andres, Chellaraj, Solotaroff, and Rajan (2017).

The structure of the questionnaire for the treatment and control groups is shown in Table 1. 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, also asked after the emotional violence module.

Table 1. 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 2 we test for differences in the answers and non-response rates to the last module across treatment arms (module 6). The eight questions in this module refer to client's satisfaction with ADRA. In only one case did answers across treatment and control groups differ significantly, and then only at the 10% level. Non-response rates are also similar and in only one out of the eight cases is the treatment group 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 randomizing the order of the direct and indirect questions on IPV in the control questionnaire.

Table 2. Difference in Responses and Non-Response Rates 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 no-response rates			
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: * significant at 10%; ** significant at 5%; *** significant at 1%.

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 B 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 in the topics of gender and gender biased violence. They all attended a three-day training workshop and we selected the top performers in the practice sessions. The workshop itself included a sensitization session provided by a local NGO, *Centro de la Mujer Peruana Flora Tristán*, which works on gender issues and women's empowerment.

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 C.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 3 reports the prevalence rates of ever experiencing different violent acts as collected by DHS-type direct questions. Prevalence of emotional violence against women was collected for the entire sample while only the control group answered the direct questions related to physical and sexual IPV. We included nine different acts of physical and sexual violence as inflicted by their actual or past partner: having her hair pulled; being pushed, shaken, or having something thrown at her; being slapped or having her arm twisted; being punched or hit with something that may have hurt her; being kicked or dragged; being strangled or burned; being threatened with a knife, gun, or other weapon; being forced to have sex; and being forced to perform sex acts she does not approve of. Based on these direct questions, we crafted nine corresponding sensitive statements to be added to the lists of neutral statements provided to the control.

Prevalence rates as measured by the direct questions are shockingly high in our sample. Almost 80% 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 are about 64% and 62% for any type of physical or sexual violent act, respectively.

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

Table 3. 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	0.62
Pull hair	560	0.31
Push	559	0.46
Slap	559	0.26
Punch	559	0.22
Kick	558	0.15
Strangle	560	0.06
Knife	560	0.06
Forced sex	559	0.23
Unapproved sex practices	558	0.09
IPV	560	0.78

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

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

3.4 Results

Our main goal is to test if there are statistically significant differences in the report of violence across direct and experimental data collection methods. A positive gap between ρ and p would suggest the presence of underreporting in DHS-type surveys.

Table 4 presents the prevalence rates using indirect and direct reporting methods for physical and sexual IPV. The last column corresponds to a test of the difference between ρ and p for each act of IPV. The last two rows of the table report the results from a joint test of significance of the gap between ρ and p for the seven acts of violence analyzed.

On average, the results in Table 4 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 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 4 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 next explore such potential outcomes relying on equation (2).

Keep in mind 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 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, civil status, and the number of children of the victim, we designed the survey instrument to be able to test for differences across 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

[12] These are the multivariate regression estimators obtained under linearity in x_i and $(T_i \cdot x_i)$ as proposed in Blair and Imai (2012).

[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.

Table 4. Difference in Estimated Prevalence Rates of Physical and Sexual IPV

Violent act	List experiments (ρ)	Direct reporting (p)	$(\rho - p)$
Pull hair	0.418	0.311	0.107*
	(0.060)	(0.020)	
Slap	0.170	0.265	-0.094
	(0.065)	(0.019)	
Punch	0.174	0.224	-0.049
	(0.070)	(0.018)	
Kick	0.126	0.145	-0.019
	(0.067)	(0.015)	
Strangle	-0.022	0.055	-0.077
	(0.065)	(0.010)	
Knife	0.046	0.057	-0.011
	(0.065)	(0.010)	
Sex acts	0.052	0.095	-0.043
	(0.068)	(0.012)	
Joint test			
χ^2		8.12	
Prob $> \chi^2$		0.322	

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

truthfully since they do not fear the loss of economic support from their partner. But they may also be more likely to underreport if the burden of stigmatization is greater among them.

We find evidence of misreporting among the most educated women in the sample. Table 5 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, which is not observed in the group of less educated women.

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

Table 5. 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	0.398	0.340	0.058		0.510	0.173	0.336	**
Slap	0.160	0.293	-0.133	*	0.219	0.133	0.086	
Punch	0.126	0.247	-0.121		0.393	0.112	0.281	*
Kick	0.144	0.163	-0.019		0.043	0.062	-0.019	
Strangle	-0.086	0.061	-0.146	**	0.267	0.031	0.236	*
Knife	-0.034	0.058	-0.093		0.410	0.051	0.359	***
Sex acts	0.040	0.104	-0.065		0.105	0.051	0.054	
Joint test								
χ^2		10.62				22.02		
Prob $> \chi^2$		0.156				0.003		

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

rates reverses once indirect methods are used. The gap in prevalence rates across education levels turns positive for all but one act of violence under indirect methods, revealing a positive correlation between education and experience 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.

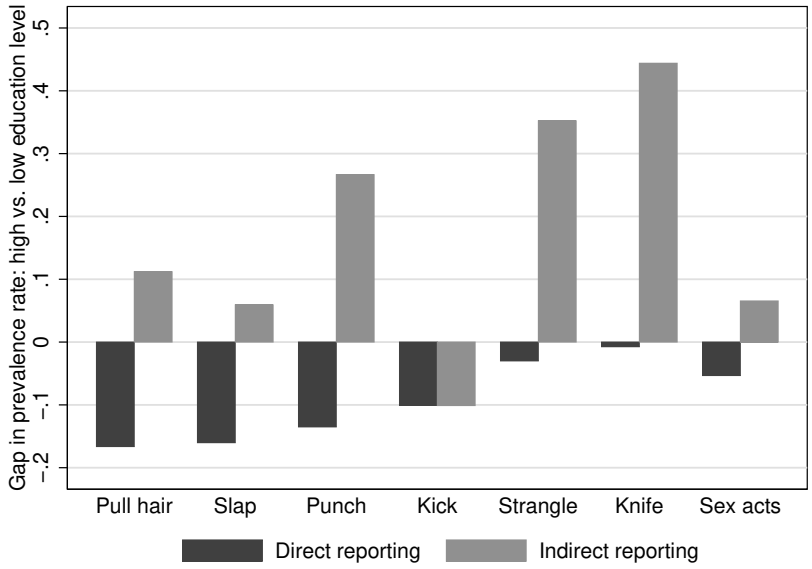
Surprisingly, no other measure of empowerment is correlated with significant biases in the report of violence at the 95% confidence level. Table 6 reports the joint significance tests that the bias in the seven acts of physical and sexual IPV is different from zero by sub-samples. While some modest differences emerge at the 10% in the sub-samples of single women and those with worse or new standing in ADRA (i.e., lower loan size, lower savings balance, and lower tenure), these do not seem to follow a clear pattern as in the case of education. Table A.5 in Appendix A shows that even though the biases are jointly and significantly different from zero, no specific bias among single women is statistically significant. Moreover, the differences identified by standing in ADRA seem to favor *overreporting* under direct methods among clients with worse standing, but this pattern is only barely significantly different from zero for two acts of violence (see Table A.10 in Appendix A).

Table 6. Joint Significance Test of $(\rho - p)$: Heterogeneous Effects

Characteristic	χ^2	Prob $> \chi^2$
Age		
<50	4.124	0.765
50+	8.219	0.314
Civil status		
Single	13.436	0.062
Married	4.318	0.742
Education level		
Less than tertiary	10.617	0.156
Completed tertiary	22.018	0.003
Mother tongue		
Spanish	10.934	0.142
Other language	7.306	0.398
Memory test		
Low score	3.993	0.781
High score	6.598	0.472
Household head		
Not the head	8.781	0.269
Head	4.729	0.693
Employment		
Does not work	6.218	0.515
Works	6.481	0.485
Standing in ADRA		
Young client	13.30	0.065
Mature client	6.64	0.467

NOTE: Joint test that the seven biases are different from zero. See Table 5 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.

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.

We argue that the effect among more educated women is not capturing a better understanding 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.6, for language, and A.7 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, Hanmer, Twigg, Hasan, McCleary-Sills, and Santamaria, 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' 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 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).

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, Nyberg, and Weibull (1999) for an example of how social norms and stigma are related in the case of welfare recipients.

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

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.[15]

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[16] 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:

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

[15] Bound, Brown, Duncan, and Rodgers (1994) provide a general framework where x_i is a vector instead of a scalar.
[16] See Calvi, Lewbel, and Tommasi (2017) for an example where x is endogenous and measured with error but where y is observed without error.

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 on 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.[17] 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, Peterman, and Heise (2016), Haushofer and Shapiro (2013), Angelucci (2008), and Bobonis, González-Brenes, and Castro (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.[18] Other studies have tried to look at the impact of social norm interventions under an experimental design (see Pronyk, Hargreaves, Kim, Morison, Phetla, Watts, Busza, and Porter (2006) and World Health Organization (2009)). Another common strategy for dealing with endogeneity problems is the use IV techniques as in Erten and Pinar (forthcoming), 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, and better access to labor market opportunities, to name a few of the factors that may affect the reporting of IPV.

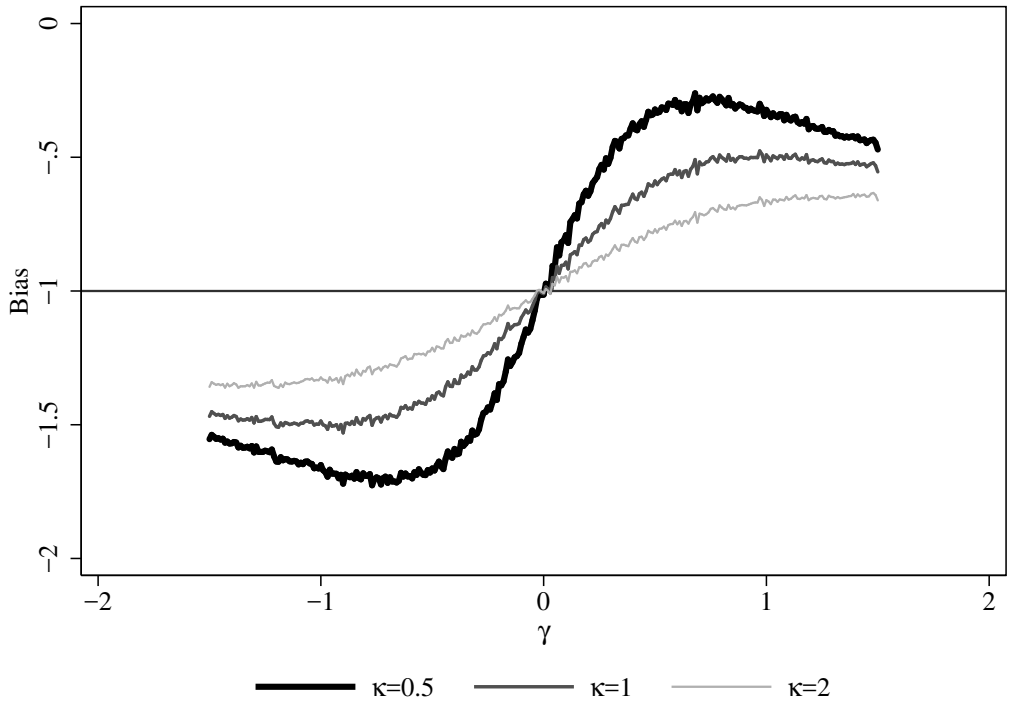
Thus, non-classical measurement error imposes a limit to the gains that randomization or IV provide to obtain less biased estimates of treatment effects. Since ϕ in (5) does not go away under these methodologies, estimates of β could be still far off from the true value. In fact, OLS may yield *less* 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. For instance, if education creates a stigma so that more educated women underreport violence ($cov(\omega_i, x_i) < 0$), as shown in our list experiments, but education is positively correlated with unobserved ability, as expected in human capital models (e.g., Card (2001)), the two biases partially cancel out.

[17] See Jewkes, Levin, and Penn-Kekana (2002), Koenig, Ahmed, Hossain, and Mozumder (2003), Breiding, Black, and Ryan (2008), Fulu, Jewkes, Roselli, and Garcia-Moreno (2013), where demographic and socioeconomic variables are considered among a long list of possible risk factors. See also Capaldi, Knoble, Shortt, and Kim (2012) for a recent review.

[18] See also De Koker, Mathews, Zuch, Bastien, and Mason-Jones (2014) for a review of RCT papers in the United States.

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 ($\gamma\frac{\mathbf{var}(\epsilon_i)}{\mathbf{var}(x_i)} + \phi$) 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 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 ω_i and obtain ϕ by correlating it with x_i. 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, Jamison, Koroknay-Palicz, Rodrigues, and Sheridan (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.[19] As an example, we provide an estimate of the bias remaining when estimating treatment effects of college education on IPV using RCT or IV methods. Using the findings from Table 5 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 lifetime: 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,

[19] See Appendix D for derivation of equation 7.

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 reporting.

We are the first to measure misreporting of IPV when using direct questions, the current best-practice and widely used in health survey worldwide. We find that, on average, there are no significant differences in direct versus indirect reporting. 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 for 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 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 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, Jamison, Koroknay-Palicz, Rodrigues, and Sheridan, 2016) that require intensive qualitative approaches. This opens up the possibility of replicating 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.

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.

BHARADWAJ, P., M. M. PAI, AND A. SUZIEDELYTE (2015): "Mental Health Stigma," Discussion paper, National Bureau of Economic Research.

BLAIR, G., AND K. IMAI (2012): "Statistical Analysis of List Experiments," *Political Analysis*, 20(1), 47–77.

BLATTMAN, C., J. JAMISON, T. KOROKNAY-PALICZ, K. RODRIGUES, AND M. SHERIDAN (2016): "Measuring the Measurement Error: A Method to Qualitatively Validate Survey Data," *Journal of Development Economics*, 120, 99 – 112.

BOBONIS, G. J., M. GONZÁLEZ-BRENES, AND R. CASTRO (2013): "Public Transfers and Domestic Violence: The Roles of Private Information and Spousal Control," *American Economic Journal: Economic Policy*, pp. 179–205.

BOUND, J., C. BROWN, G. J. DUNCAN, AND W. L. RODGERS (1994): "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12(3), 345–368.

BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): "Measurement Error in Survey Data," *Handbook of econometrics*, 5, 3705–3843.

BREIDING, M. J., M. C. BLACK, AND G. W. RYAN (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., R. V. BURKHAUSER, J. M. MITCHELL, AND T. P. PINCUS (1987): "Measurement Error in Self-Reported Health Variables," *The Review of Economics and Statistics*, 69(4), 644–650.

CALVI, R., A. LEWBEL, AND D. TOMMASI (2017): "LATE with Mismeasured or Misspecified Treatment: An Application to Women's Empowerment in India," .

CAPALDI, D. M., N. B. KNOBLE, J. W. SHORTT, AND H. K. KIM (2012): "A Systematic Review of Risk Factors for Intimate Partner Violence," *Partner Abuse*, 3(2), 231–280.

CARD, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5), 1127–1160.

COFFMAN, K., L. COFFMAN, AND M. KEITH (2013): "The Size of the LGBT Population and the Magnitude of Anti-Gay Sentiment are Substantially Underestimated," Discussion paper, NBER Working Paper No. 19508.

DE KOKER, P., C. MATHEWS, M. ZUCH, S. BASTIEN, AND A. J. MASON-JONES (2014): "A Systematic Review of Interventions for Preventing Adolescent Intimate Partner Violence," *Journal of Adolescent Health*, 54(1), 3–13.

DEKESEREDY, W. S., AND M. D. SCHWARTZ (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*.

DESIERE, S., AND D. JOLLIFFE (2018): "Land Productivity and Plot Size: Is Measurement Error Driving the Inverse Relationship?," *Journal of Development Economics*, 130, 84–99.

ELLSBERG, M., AND L. HEISE (1999): "Putting womens safety first: ethical and safety recommendations for research on domestic violence against women," *Geneva, Switzerland: World Health Organization*.

ELLSBERG, M., L. HEISE, R. PENA, S. AGURTO, AND A. WINKVIST (2001): "Researching Ddomestic Violence Against Women: Methodological and Ethical Considerations," *Studies in Family Planning*, 32(1), 1–16.

ERTEN, B., AND K. PINAR (forthcoming): "For Better or Worse? Education and Prevalence of Domestic Violence in Turkey," *American Economic Journal: Applied Economics*.

ESWARAN, M., AND N. MALHOTRA (2011): "Domestic Violence and Women's Autonomy in Developing Countries: Theory and Evidence," *Canadian Journal of Economics*, 44(4), 1222–1263.

FARMER, A., AND J. TIEFENTHALER (1996): "Domestic Violence: The Value of Services as Signals," *American Economic Review*, 86(2), 274–279.

FULU, E., R. JEWKES, T. ROSELLI, AND C. GARCIA-MORENO (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 M. HUYNH (2010): "Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error," *The Review of Economics and Statistics*, 92(2), 302–315.

GOURLAY, S., T. KILIC, AND D. LOBELL (2017): "Could the Debate Be Over? Errors in Farmer-Reported Production and Their Implications for Inverse Scale - Productivity Relationship in Uganda," Discussion paper, Working paper.

HAUSHOFER, J., AND J. SHAPIRO (2013): "Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya," .

HIDROBO, M., AND L. FERNALD (2013): "Cash Transfers and Domestic Violence," *Journal of Health Economics*, 32(1), 304–319.

HIDROBO, M., A. PETERMAN, AND L. HEISE (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., B. PARK, AND K. GREENE (2014): "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models," *Political Analysis*, 23, 180–196.

JEWKES, R., J. LEVIN, AND L. PENN-KEKANA (2002): "Risk Factors for Domestic Violence: Findings from a South African Cross-sectional Study," *Social Science & Medicine*, 55(9), 1603–1617.

JOHNSTON, D. W., C. PROPPER, AND M. A. SHIELDS (2009): "Comparing Subjective and Objective Measures of Health: Evidence from Hypertension for the Income/health Gradient," *Journal of health economics*, 28(3), 540–552.

JOSEPH, G., S. USMAN JAVAID, L. A. ANDRES, G. CHELLARAJ, J. L. SOLOTAROFF, AND S. I. RAJAN (2017): "Underreporting of Gender-Based Violence in Kerala, India: An Application of the List Randomization Method," Discussion paper, Policy Research Working Paper N. 8044, World Bank.

KARLAN, D., AND J. ZINMAN (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.

KLUGMAN, J., L. HANMER, S. TWIGG, T. HASAN, J. MCCLEARY-SILLS, AND J. SANTAMARIA (2014): *Voice and Agency: Empowering Women and Girls for Shared Prosperity*. Washington, DC: World Bank Group.

KOENIG, M. A., S. AHMED, M. B. HOSSAIN, AND A. K. A. MOZUMDER (2003): "Women's Status and Domestic Violence in Rural Bangladesh: Individual-and Community-level Effects," *Demography*, 40(2), 269–288.

LINDBECK, A., S. NYBERG, AND J. W. WEIBULL (1999): "Social norms and economic incentives in the welfare state," *The Quarterly Journal of Economics*, 114(1), 1–35.

MACMILLAN, R., AND R. GARTNER (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 M. SIEGEL (2013): "Eliciting Illegal Migration Rates through List Randomization," Discussion paper, Policy Research Working Paper N. 6426, World Bank.

MEYER, B., W. MOK, AND J. SULLIVAN (2008): "The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences," Discussion paper, National Bureau of Eco-

nomic Research, Working paper NB08-12.

MEYER, B. D., W. K. MOK, AND J. X. SULLIVAN (2015): "Household surveys in crisis," *The Journal of Economic Perspectives*, 29(4), 199–226.

O'NEILL, D. (2012): "The Consequences of Measurement Error when Estimating the Impact of BMI on Labour Market Outcomes," Discussion paper, IZA Discussion Paper No. 7008.

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 D. QUINN (2013): "The Intimate Partner Violence Stigmatization Model and Barriers to Help-Seeking," *Basic Appl Soc Psych.*, 35(1), 109–122.

PALERMO, T., J. BLECK, AND A. PETERMAN (2014): "Tip of the Iceberg: Reporting and Gender-based Violence in Developing Countries," *American Journal of Epidemiology*, 179(5), 602–612.

PETERMAN, A., T. PALERMO, S. HANDA, AND D. SEIDENFELD (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., J. HARGREAVES, J. KIM, L. MORISON, G. PHETLA, C. WATTS, J. BUSZA, AND J. PORTER (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., K. IMAI, AND J. N. SHAPIRO (2016): "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions," *American Journal of Political Science*, 60(3), 783–802.

WORLD HEALTH ORGANIZATION (2009): "Changing Cultural and Social Norms that Support Violence," Discussion paper, Series of briefings on violence prevention: the evidence.

A Additional Figures and Tables

Table A.1. Summary Statistics and Balance Check

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

Continued on next page

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

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2. Prevalence Rates of Non-Sensitive Statements in the Pilot

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

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 ran for several years in Peru.

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

Table A.3. Correlation of Prevalence Rates among Non-Sensitive Statements

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

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

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

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

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

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

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

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

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

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

Table A.4. Difference in Estimated Prevalence Rates of Physical and Sexual IPV by Age

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

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.5. Difference in Estimated Prevalence Rates of Physical and Sexual IPV by civil status

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

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.6. Difference in Estimated Prevalence Rates of Physical and Sexual IPV by Mother Tongue

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

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.7. 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	0.477	0.350	0.127		0.362	0.270	0.092
Slap	0.253	0.262	-0.009		0.091	0.267	-0.176 *
Punch	0.247	0.248	-0.001		0.105	0.198	-0.093
Kick	0.165	0.155	0.011		0.088	0.135	-0.047
Strangle	0.006	0.063	-0.057		-0.049	0.047	-0.096
Knife	0.061	0.073	-0.013		0.032	0.040	-0.008
Sex acts	0.137	0.116	0.021		-0.029	0.073	-0.102
Joint test							
χ^2		3.99				6.60	
Prob $> \chi^2$		0.781				0.472	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.8. 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	0.455	0.275	0.180	**	0.348	0.389	-0.040
Slap	0.174	0.240	-0.066		0.163	0.320	-0.157
Punch	0.136	0.197	-0.061		0.246	0.282	-0.035
Kick	0.131	0.112	0.019		0.117	0.218	-0.102
Strangle	-0.012	0.026	-0.038		-0.041	0.120	-0.161
Knife	0.057	0.034	0.023		0.025	0.109	-0.083
Sex acts	0.024	0.065	-0.041		0.103	0.160	-0.057
Joint test							
χ^2		8.78				4.73	
Prob $> \chi^2$		0.269				0.693	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.9. Difference in Estimated Prevalence Rates of Physical and Sexual IPV by Employment

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

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

Table A.10. 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	0.408	0.309	0.099		0.466	0.322	0.144
Slap	0.147	0.279	-0.133	*	0.282	0.189	0.093
Punch	0.194	0.237	-0.042		0.081	0.156	-0.075
Kick	0.114	0.152	-0.038		0.180	0.111	0.069
Strangle	-0.061	0.062	-0.122	*	0.158	0.022	0.135
Knife	0.091	0.064	0.027		-0.162	0.022	-0.184
Sex acts	0.023	0.090	-0.067		0.186	0.122	0.064
Joint test							
χ^2		13.30				6.64	
Prob $> \chi^2$		0.065				0.467	

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%. OLS estimates. Mature clients are those with loan size and savings balance above the 75th percentile and a tenure greater than two loan cycles. Estimates of ρ are obtained from a regression that includes surveyor fixed effects as well as controls such as: household head dummy, age, civil status, education level, number of children, Spanish is the woman's mother tongue, working woman, literacy, tenure in ADRA, high average loan size, high savings balance, and an indicator of good memory. 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 (2).

B Ceiling Effects

Although using a very small sample (31 observations), the pilot data allow 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 B.1 reports some statistics on the prevalence rates of the sets of non-sensitive statements with data from the pilot. The first column reports the mean prevalence of the 4 statements, while the second and third report the standard deviation and the 75th percentile of this 4 prevalence rates. The non-sensitive statements grouped with the sensitive ones on pushing and forced sex have very high average prevalence rates and low variance. Moreover, the 75th percentile of prevalence rates for these sets of 4 statements is very high, which shows that many statements in these groups have high prevalence rates. In fact, one of the statements grouped with forced sex has a prevalence rate of 1 ("ever watched a soap opera").

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

Table B.1. Prevalence of 4 Non-Sensitive Statements by Question

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

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

C Sample Instruments

C.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 Verónica Frisancho, the researchers in charge of the project.

C.2 IPV and List Experiment Modules

C.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?

C.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*[20]

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
   (b) Read *El Comercio*[21]
   (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

C.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

[20] *Al fondo hay sitio* is a very popular soap opera than ran for several years in Peru.
[21] *El Comercio* is one of the most read newspapers in the country, particularly in Lima.

(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*[22]
(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
(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*[23]

[22] *Al fondo hay sitio* is a very popular soap opera than ran for several years in Peru.
[23] *El Comercio* is one of the most read newspapers in the country, particularly in Lima.

(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

D β_{OLS} in the Presence of Non-Linear and Non-Classical Measurement Error

In 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:

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