Measuring Violence Against Women with Experimental Methods[*]

Jorge M. Agüero[†] Veronica Frisancho[‡]

August 2020

Abstract

The prevalence of intimate partner violence is a central indicator of the Sustainable Development Goals for women's agency. However, measuring this indicator largely relies on self-reports that could suffer from severe misreporting if women face high costs of revealing their victim status. We study the degree of misreporting in surveys that have been identified as the best source of data, such as the widely used Demographic and Health Surveys (DHS). Focusing on a sample of women in impoverished urban areas of Lima, Peru, we conduct an experiment that replicates direct measures from these surveys and compares them against list experiments, a method that provides greater privacy to respondents. We find no significant differences across direct and indirect methods in any of the seven acts of physical and sexual violence considered. This result largely persists when testing across sixteen different subgroups and accounting for multiple hypothesis testing.

Keywords: Women's agency, intimate partner violence, measurement, list experiments, direct elicitation
JEL-codes: I12, C83, C21.

[*] We would like to thank the staff at ADRA for allowing us to collect data during their regular group meetings. We also thank the valuable comments received from Janet Currie, Kosuke Imai, and discussants and participants at the 2017 Conference on Risk and Risky Behavior, the 2017 Congress of the Peruvian Economic Association, NEUDC 2017, ASSA 2018, Southern Economics Association meetings 2018, and several seminar presentations. This research was approved by the University of Connecticut's Institutional Review Board, Protocol #H15-164: "Measuring Violence Against Women with Experimental Methods." RCT registration ID: AEARCTR-0000517. A previous version of this paper was circulated under the title "Misreporting in Sensitive Health Behaviors and its Impact on Treatment Effects: An Application to Intimate Partner Violence."

[†] University of Connecticut, Department of Economics and El Instituto. 365 Fairfield Way, Storrs, CT 06269-1063. E-mail: jorge.aguero@uconn.edu.

[‡] Inter-American Development Bank, Research Department. 1300 New York Ave. NW, Washington, DC 20577. E-mail: vfrisancho@iadb.org.

1 Introduction

Women's voice and agency plays an important role in economic development. Empowering women leads to changes in decision making that directly foster development [Duflo, 2012]. Under the view that development requires the expansion of human capabilities, violence against women is one of the central inhibitors of economic and social development [Panda and Agarwal, 2005]. The elimination of violence against women is thus not only a major public health issue [Krug et al., 2002; Bott et al., 2012] but also a key objective of the Sustainable Development Goals (SDG) [United Nations, 2015].

High-quality data on intimate partner violence (IPV) is crucial to identify the most vulnerable sub-groups, monitor the evolution of violence over time, and facilitate the design of effective policies to prevent and reduce IPV. Unlike other SDG, for which administrative records are the primary source of data,[1] information collected by the services used by violence victims is biased due to selection into reporting and lack of trust to report truthfully [United Nations, 2014]. This is partly explained by limited access to police stations and health care centers, distrust of the authorities at male-dominated institutions, and lower levels of law enforcement in developing countries. Indeed, a study covering 24 developing countries shows that only seven percent of victims from violence against women made a report that can be captured by administrative records [Palermo et al., 2014].

Most of the data used to track SDG on women's agency rely on self-reports from victims. The United Nations' guidelines for producing statistics on violence against women identify population-based surveys as "the best source of data" for estimating the prevalence of these acts of violence [United Nations, 2014, p. 2]. Over time, these specialized surveys have achieved great progress in the development of instruments and protocols. Yet the private

[1] For example, the SDG rely on administrative records for targets and indicators related, but not limited to, social protection, production, trade and utilization of major food crops and livestock, tuberculosis incidence, education and domestic financial institutions. For more details about the primary sources for each target and indicator see *E-Handbook on Sustainable Development Goals Indicators* available at https://unstats.un.org/wiki/display/SDGeHandbook/Home (last accessed on May 30, 2019). Also, see Calvo et al. [2019] for a discussion of data collection for governance-related issues in the SDG.

nature of IPV imposes costs to truthful reporting that even the most rigorous safety and privacy protocols for surveys using *direct* questions could not completely overcome. Women who survive IPV face emotional costs as well as fear of retaliation, loss of economic support, and stigma when reporting their survival status as perpetrators tend to be their intimate partners.[2]

Our paper is the first to measure the level of misreporting present in population-based surveys that rely on direct questions on IPV. We conduct a randomized experiment where we compare the prevalence rates of IPV under two different instruments that vary in the degree of privacy they provide to the respondent. While the control group receives a questionnaire that follows the UN best-practices, the treatment group answers an alternative instrument that guarantees full anonymity in the report of IPV. The provision of greater privacy when collecting data on IPV could improve the levels of truthful reporting, especially for women with higher costs of exposure.

To maximize the external validity of our study, the questionnaire provided to the control group was based on the Demographic and Health Surveys (DHS). The DHS are large nationally-representative surveys applied in over 90 developing countries. Since 2000, the DHS includes a module on IPV that follows the UN recommendations. This module has been applied in over 40 countries and provides the main data source to monitor progress on the IPV indicators of SDG-5 [United Nations, 2014]. This makes our comparison to the alternative instrument widely relevant. The DHS module on IPV is an adaptation of the Conflict Tactic Scale,[3] which asks about specific and objective acts of physical and sexual violence. The Conflict Tactic Scale reduces bias arising from subjective perceptions of violence and provides participants with several opportunities to expose victimization [Ellsberg and Heise, 1999; Kishor and Johnson, 2004; Bender, 2017].

[2] Indeed, the United States Bureau of Justice Statistics [2017] reports that the most prevalent reasons not to report domestic violence victimization to the police in the US are personal privacy (32 percent), protecting the offender (21 percent), considering that the crime was minor (20 percent), and fear of retaliation (19 percent).

[3] Originally developed by Straus [1979] and later modified and expanded by the World Health Organization (WHO, 1997).

To indirectly record IPV prevalence rates, we rely on *list experiments*. This methodology increases privacy as the respondent's status is never revealed neither to the enumerator nor to the researchers [Glynn, 2013; Blair and Imai, 2012]. In the standard list experiment, assignment to the treatment is randomized to guarantee that the control group can serve as a counterfactual for the former. The control group is presented with a list of neutral statements and asked to disclose the *number* of those that are true. Since respondents never reveal which statements are true, privacy levels are greatly increased. In the treatment group, a sensitive statement is added to the list and, once more, respondents only report how many statements are true. The prevalence rate of the sensitive issue is estimated as the difference in the average response to the lists across arms. In our application, we create a list of statements for each of the seven IPV questions in the DHS on physical and sexual violence and ask women in the control group to answer the direct DHS questions and the neutral lists of statements. Women in the treatment arm never receive the direct questions and only answer the extended lists.

We focus on a sample of female microcredit clients from impoverished urban districts in Lima, Peru. This focus is not necessarily a limitation in a region like Latin America and the Caribbean, where 66% of the microcredit client base is female [Convergences, 2018]. Particularly in Peru, where over five million clients are served by the microfinance sector, our study population becomes extremely relevant.[4] Given this widespread access to microcredit in the country, our sample is able to cover women with diverse backgrounds and socioeconomic status with potentially varying costs of reporting, allowing us to explore important heterogeneous effects of the provision of greater levels of privacy.

Our main results is that, on average, there are no significant differences in reporting across direct and indirect methods in any of the seven acts of physical and sexual violence. This finding is robust to the use of multiple hypothesis testing and joint tests of statistical

[4] For the past ten years, Peru has been at the top of the ranking of the Global Microscope, an annual report on the Environment for Financial Inclusion created by The Economist Intelligence Unit. See http://www.eiu.com/landing/Global-Microscope for details. Last accessed on May 30, 2019.

significance. Since we closely followed the UN guidelines, this result validates the adequacy of the best available practices when using direct survey methods to collect data on IPV. In fact, our survey is close to a "standalone" survey on violence because it did not include all the other modules observed in a typical DHS (e.g., maternal and child health, birth history, etc.). This focus provides an additional strength to the design since our power calculations were specifically tailored to the IPV questions as opposed to other variables as in the DHS. This allowed us to obtain a large enough sample to accurately measure even the low prevalence violent events included in the conflict tactics scale.[5]

We also test for differences in reporting across 16 subgroups. In all but one case, IPV reports using direct questions perform as well as with indirect methods. Thus, for the vast majority of subgroups studied, the use of UN's best practices to elicit IPV prevalence is supported. It is worth noting that, even after controlling for multiple hypotheses testing, we find that women with completed tertiary education report *higher* rates of violence under the list experiments than under the direct methods, while there is no significant difference among the least educated women. This evidence could be consistent with a negative correlation between years of education and traditional views of gender roles [e.g., Angelucci, 2008; Marcus and Harper, 2015]. While this finding cannot be extrapolated, it highlights that even when the best-practices to collect IPV data are implemented, differential misreporting patterns could still emerge in certain contexts and sub-populations.

Our study contributes to the scarce body of work measuring misreporting in the case of IPV. It significantly improves upon recent work by Joseph et al. [2017] and Peterman et al. [2017] on IPV using list experiments in developing countries. As explained in more detail below (see section 2), both studies ignore UN recommendations and use a general statement about violence as the sensitive statement in the list experiments. Moreover, neither study includes a direct question equivalent to the sensitive item in the control questionnaire. In-

[5] As discussed in section 3.2, we implemented all the recommended protocols in the control and treatment groups. For example, female enumerators were trained specifically to deal with IPV-related issues, we had an emergency questionnaire to be used when interrupted during the IPV section, and support information for victims and surveyors was provided when required.

stead, we follow the best recommended practices and compare our indirect method to the DHS direct questions to understand misreporting patterns.

Furthermore, our large sample size overcomes a common problem found in the applications of list experiments in other contexts [e.g., Karlan and Zinman, 2012; De Cao and Lutz, 2018], where the same subjects are exposed to direct *and* indirect questioning. This strategy, the double list experiment, eliminates the anonymity that our experiment is able to provide by assigning subjects in the sample to either a treatment or a control group.

This study is also related to previous work using list experiments to measure prevalence rates of risky or socially sensitive attitudes or behaviors. Recent applications of list experiments include, for example, Karlan and Zinman [2012] to measure loan proceeds from microfinance loans, Jamison et al. [2013] to collect information on sexual behavior, McKenzie and Siegel [2013] to elicit illegal migration rates, Coffman et al. [2013] to measure the size of LGBT population and anti-gay sentiment, Imai et al. [2014] to examine vote-selling, and Rosenfeld et al. [2016] to study anti-abortion support. We argue that the particular nature of IPV imposes costs of self-exposing as a victim which may be quite different from those faced by a criminal or drug abuser.

Finally, our results have ample applications on the topic of measurement error for other SDG indicators. Our approach is especially useful where administrative records are not a reliable data source to measure sensitive behavior as observed in other settings [Bound et al., 2001; Butler et al., 1987].

The remainder of the paper is structured as follows. Section 2 presents some background information on misreporting IPV in survey questions while Section 3 provides details on the design of our study and the instruments used to implement direct and indirects questions on IPV. Section 4 presents and discusses the results and Section 5 concludes.

2 Misreporting intimate partner violence in surveys

Our paper tests for possible underreporting biases in the measurement of IPV as collected by population-based surveys. These surveys have been identified as the best source of data to monitor progress of the Sustainable Development Goal related to women empowerment [United Nations, 2014].

We argue that two features of IPV generate large potential for error in the measurement of prevalence rates even in well-designed specialized surveys. First, this violence is perpetrated by people known to the victim: her current or ex-partner. Second, it tends to be invisible as much of it happens behind closed doors and in the privacy of the home.

The nature of IPV may thus lead to measurement biases as it introduces very large costs to self-identify as an IPV victim. For instance, there is an emotional cost due to her attachment to the offender and the potential sanctions (social or legal) that he may face. If her status as a victim is revealed, a woman may also face the potential loss of her partner's economic support and the risk of retaliation through an escalation of violence against her or her children. Women may also fear stigmatization, either from intrinsic or extrinsic sources [Overstreet and Quinn, 2013]. However, the costs of being exposed are likely to be heterogeneous. This implies that privacy concerns may differentially prevent women from truthfully reporting their previous experience of violence, making misreporting more prevalent for specific subgroups.

Despite increasing levels of support for victims to speak up [Klugman et al., 2014], it is still difficult to disclose these violent events and expose their aggressors. As much as security and privacy protocols can be improved upon and emphasized in the fieldwork, direct methods still demand that a woman identifies herself as an IPV victim to a surveyor, potentially exposing her and her aggressor. A few non-experimental studies suggest that higher privacy may lead to more reporting. For example, Ellsberg et al. [2001] compared the prevalence of IPV in Nicaragua using two surveys conducted in different years and find that when safety and privacy protocols are enforced more strictly, higher prevalence rates of IPV are reported.

Additionally, the WHO multi-country study asked about past abuse (sexual abuse as a child) twice so as to compare responses under UN guidelines to those obtained by secretly marking a happy or sad face in a sheet that was not observed by the surveyors. The latter method reported a higher incidence of abuse [World Health Organization, 2005].

A few studies have tried to reduce exposure costs by relying on indirect methods that provide full anonymity to respondents and foster truthful reporting. Among these methods, list experiments are the most frequently used.[6] Three recent examples closely related to our paper are Joseph et al. [2017], Peterman et al. [2017], and Bulte and Lensink [2019]. Even though their contribution is valuable, they face several limitations. First, Joseph et al. [2017] measures prevalence rates at the household level, asking anyone who opens the door about the violence experienced by all women in the household. Thus, it may be the case that the respondent does not know about the IPV experience of each women in the household or that he himself is the perpetrator. Second, the sensitive statement in the lists is quite general and it is based on the single-question generic approach (*Has at least one woman member of your household faced physical aggression from her husband anytime during her life?*), greatly departing from the well-established multiple-question UN guidelines for the measurement of violence. The same holds for Peterman et al. [2017], who target women as respondents but use 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?*). Bulte and Lensink [2019] also rely on a unique question on IPV which even varies

[6] Alternative methods include qualitative approaches as in Blattman et al. [2016]. The authors combine surveying with ethnographic techniques to uncover misreporting. The method requires that the surveyor team stays for longer periods of time in the field, increasing the costs of data collection. Moreover, since it does not provide additional anonymity to the respondents, the success of the method depends heavily on the surveyors' ability to make the respondent feel safe and comfortable to truthfully report her answers or behavior. Surveyors training then becomes crucial, adding to the cost of the fieldwork and limiting the scaling-up of the technique. Kataoka et al. [2010] relies on a small sample (N=382) of pregnant women in Tokyo who were interviewed multiple times at a prenatal clinic. The authors compare face-to-face interviews against a self-administered questionnaire that provides more privacy either dealing with potential external stigma costs of reporting IPV to the nurse or hiding her responde from her partner. Other indirect questioning techniques such as endorsement experiments or randomized response techniques are often used in the political science literature and recent papers have adapted them to measure health outcomes such as abortions [Lara et al., 2006].

across methods (*I am regularly hit by my spouse* in the list experiment and *How often did your husband push, slap, beat, or hit you during the last 6 months?* in the direct question). Moreover, the authors include in the same questionnaire the list experiment questions (either the control or the treatment format) followed by the direct IPV question. This repetition of IPV questions under the two methods for the treatment group could potentially bias the responses to the direct question. In sum, none of these studies is able to measure misreporting relative to the *best available* IPV direct reporting method and they do not include a direct question equivalent to the sensitive item in the control questionnaire while adhering to the WHO best practices and protocols.

Our design overcomes all these limitations by (i) focusing on women as respondents, (ii) following the UN guidelines for direct questions as well as their privacy and safety protocols throughout the application of the questionnaire, (iii) asking the direct questions only to the control group, and (iv) comparing the prevalence rates obtained from the indirect method to the ones that come from the DHS direct method.[7] Indeed, we are the first in the field of IPV to experimentally measure the bias in survey-based data.

An additional advantage of our paper is that we are able to analyze the level of misreporting across sub-groups defined by individual characteristics. This is particularly relevant given the evidence in reporting biases documented in other sensitive variables and behaviors. For instance, Gottschalk and Huynh [2010] show that there is substantial measurement error in earnings and that this error is correlated with (true) earnings and positively correlated across time. Dillon et al. [2019] show that self-reported measurement bias in land size leads to overreporting for small plots and underreporting for large plots. In the health literature, Butler et al. [1987] show evidence of non-classical error in the measurement of arthritis while Johnston et al. [2009] finds a similar pattern in hypertension self-reporting. O'Neill [2012] identifies a negative correlation between self-reported and anthropometric measures

[7] Recently, De Cao and Lutz [2018] used a list experiment to measure the reporting bias relative to direct questions about attitudes towards female genital cutting in the Afar province of Ethiopia. However, they share some of the limitations previously mentioned for the studies on IPV such as the loss of full anonymity that comes from asking the direct questions to both the treatment and control groups.

of body mass index. More recently, Bharadwaj et al. [2015] relies on administrative records and finds that underreporting in mental health medication is correlated with age, gender, and ethnicity.[8]

Heterogeneous differences in misreporting may distort targeting strategies and estimated treatment effects of programs intended to reduce IPV prevalence. Many studies have tried to identify the main drivers of this type of violence [e.g., Angelucci, 2008; Hidrobo and Fernald, 2013; Haushofer and Shapiro, 2013; Bobonis et al., 2013; Hidrobo et al., 2016; Erten and Keskin, 2018]. When an outcome variable, such as IPV, suffers from classical measurement error, the precision of the treatment effects estimates (i.e., the standard errors) is affected. However, when the error is not classical but rather correlated with a risk-factor (e.g., misreporting varies by income, education attainment, etc.), then it is impossible to obtain unbiased causal estimates of the variable of interest (see B). Indeed, Gillen et al. [2019] replicate three classic and influential studies on behavioral economics in the United States (but unrelated to IPV) and find that measurement error in control and causal variables explains 30-40% of variance in choices. Accounting for non-classical measurement error substantially affects the findings of the studies, showing that the bias introduced is not negligible. Thus, a key goal of our study is to investigate whether there is misreporting across different subgroups in our sample.

3 The Experiment

3.1 Experimental Design

Our study relies on a survey experiment to measure the impact of an increase in privacy on IPV reporting. Greater privacy is created by reducing respondents' expectations that her report will be shared with others. The treatment varies the survey module on physical and

[8] A recent article also shows that self-reports on bargaining power coming from men and women within the same household systematically differ; indeed, disagreement varies across assets and activities [Ambler et al., 2019].

sexual IPV, which allows us to compare prevalence rates across direct and indirect methods. While the control group reveals exposure to physical and sexual IPV through direct DHS questions, the treatment group does so only indirectly, through list experiments.

List experiments are often used to gather opinions and record behaviors related to sensitive issues that are prone to underreporting. The basic design of a list experiment features a control group, which is given a list of S neutral statements, and a treatment group, who receives $S+1$ statements including the same list received by the control and an added statement referring to a sensitive issue. Both groups are asked to report the *number* of statements that hold true, without indicating which ones are in fact true. Random assignment of the treatment implies that the average number of neutral statements that hold true is equal across groups. Thus, the control group serves as a counterfactual for the treatment group and the prevalence rate of the sensitive statement can be estimated by the difference in the average number of true statements reported across groups.[9]

This paper applies this methodology to measure the bias in prevalence rates of physical and sexual IPV under direct reporting. To mimic the DHS questionnaire, we ask if the respondent ever experienced IPV as perpetrated by her *last partner* in both the direct questions and the randomized lists. 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 2 and 3). To ease transition into more sensitive questions on physical and sexual IPV, all respondents were given direct questions on emotional violence in module 4. In the physical and sexual violence module, the questionnaires differed depending on the treatment assignment. In the control group, we administered direct questions on physical and sexual IPV (module 5A) followed by a module with the lists of neutral statements (5B). The treatment group skipped the direct questions and was only provided with the list experiment questions with the added sensitive statement.

Our study focuses on female microentrepreneurs in impoverished peri-urban areas in a

[9] Note that full anonymity precludes gathering individual-level data on prevalence for each violent act.

middle income country. The experimental sample is drawn from a *village* banking program run by the Adventist Development and Relief Agency (ADRA) in Lima, Peru's capital. We impose several constraints on the total pool of 1873 clients (112 banks) in Lima. First, we focus on borrowers aged between 18 and 60. Second, we reserve 6 banks at random for piloting purposes. Finally, we target all clients in banks with monthly meetings scheduled during July 2015. Our final sample includes 1119 women in 98 banks who were interviewed between July 1st and August 25th, 2015.

Females participating in ADRA's microcredit programs are different than the average woman in Lima. As shown in Appendix Table A.1, ADRA's clients are older, more likely to be married, less educated, and poorer than the women interviewed in Lima in the 2015 Peruvian DHS.[10] ADRA's clients also experience higher levels of physical and sexual violence. Our study does not try to characterize women in Lima, but rather focus on a sub-group of women who tend to be self-employed and who have been the focus of attention of several development interventions such as microcredit programs.

Data collected using DHS-type direct questions reveals very high prevalence rates of ever experiencing violent acts inflicted by the woman's last partner. As shown in Table 2, about 72 percent of the women in the control group have experienced some type of violence, either emotional, physical, or sexual. The prevalence rate of physical and/or sexual violence is also high at 46 percent.

Randomization of the treatment was conducted in the field, at the individual level, right after the surveyor obtained the informed consent of the respondent. Table 3 confirms that the randomization was successful. After cleaning the data, we are left with a sample of 992 valid surveys. According to our power calculations, this sample was large enough to detect an effect as small as 3 percentage points.[11]

[10] The DHS restricts its sample to women aged 15 to 49 as it focuses on fertility. Microcredit programs, including ADRA's, do not follow such age restriction and neither does our sample.

[11] The baseline violence prevalence rates in the area studied were obtained from the 2015 Peruvian DHS survey. We focused on one of the least frequently reported acts of violence: being forced to perform sex acts she does not approve of. Initial prevalence rate is set at 0.027 with a standard deviation of 0.161. As discussed in our pre-analysis plan, our sample size follows a randomization conducted at the individual level,

3.2 Instruments: Development and Application

Collecting data on violence is quite challenging and demands implementing strict ethical and safety protocols and well-designed instruments. The direct questions on physical and sexual IPV were obtained from the Peruvian DHS. The DHS follows a multiple-question approach that provides participants with several opportunities to respond about physical and sexual violence experience [Kishor and Johnson, 2004]. This question format focuses on specific and objective acts of violence and it is less likely to be affected by different perceptions about what constitutes violence (see Ellsberg and Heise [1999] and Bender [2017] for an extensive discussion).

The list experiments module includes one list of statements for each victimization act covered in the direct questions. While the control group receives lists that consist of four neutral statements, each list provided to the treatment group ends with an added sensitive statement that corresponds to a given act of violence. We consider seven physical and sexual violence acts as inflicted by their actual or past partners: having her hair pulled; being slapped or having her arm twisted; being punched or something that may have hurt her; being kicked or dragged; being strangled o burnt; being threatened with a knife, gun, or other weapon; and being forced to perform sex acts she does not approve of.

To guarantee respondents' safety and well-being, we implement all the protocols recommended by the UN when collecting data on IPV, including the application of the survey in a private space and the use of an emergency questionnaire whenever someone interrupts the interview during the violence module. Moreover, to ensure that the protocols were adequately implemented, we place special attention to the selection and training of the surveyors. We selected a team of female surveyors with previous experience on gender issues and gender-based violence. All candidates attended a three-day training workshop and only the top performers in the practice sessions were recruited. The workshop itself included a sensitization session provided by a local civil rights organization, *Centro de la Mujer Peru-*

with a minimum detectable effect of 0.03, a significance level of 10 percent and power of 0.8.

ana Flora Tristán, which works on gender issues and women's empowerment. All surveyors were trained in the application of both treatment and control questionnaires, which only differed in the way in which we asked about IPV, either direct or indirect elicitation. The questionnaire application, irrespective of the result of the randomization, was undertaken under the same guidelines and protocols recommended by the UN when asking about IPV. Since respondents may be less familiar with indirect questioning techniques such as the list experiments, we expanded the protocols to include the use of visual aids in module 5B. Surveyors were instructed to provide each respondent with a printed copy of the list experiment questions that corresponded to her randomization outcome. 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.

For the list experiments to effectively protect respondents' privacy while providing a good estimator of the prevalence rate, the selection and grouping of neutral statements is crucial. The design of each list has to take into account the trade-off between protecting the respondent and reducing the variability of the responses. On one hand, we want to avoid *ceiling effects* or *floor effects*: if a large share of the population is likely to respond that all or none of the neutral statements are true, the respondent is no longer protected. On the other hand, a list that avoids ceiling and floor effects will tend to introduce greater variability in the responses, which could increase the variance of the estimator. Glynn [2013] provides some guidance for the development of lists and 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 or floor effects.

To develop the questionnaires, we piloted 41 neutral statements with a sample of 31 individuals and measured the prevalence rates of each statement. These prevalence rates allowed us to measure the adequacy of the statements in our setting and helped us decide how to group them depending on their correlation patterns. Based on the correlation of responses across pairs of statements in the pilot data, we developed an algorithm that induced 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. Table A.2 in the Appendix shows the prevalence rates of the 26 statements we kept for the list experiments, after removing those with very low prevalence rates. Two statements used in the final instrument were not tested in the pilot. Table A.3 reports the correlation of prevalence rates in each set of statements grouped together. For each of the seven 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 percent confidence level.[12]

One may argue that the inclusion of the direct questions on physical and sexual IPV in the control group could bias the responses in the rest of the survey, including answers to the lists of neutral statements. If that were the case, then answers to all other non-sensitive questions in module 6 would have been affected. Table 4 tests for differences in the answers and non-response rates to the last module on client's satisfaction with ADRA, which was applied to control and treatment groups. In only one case the answers across groups differ, but the magnitude is quite small and significance only holds at the 10 percent level (see panel A). Item non-response rates are also similar across arms (see panel B). This evidence rules out the possibility that the treatment assignment biased the report of the prevalence rates of neutral statements.

[12] An alternative approach to test for the internal validity of list experiments has been proposed by Chuang et al. [2019]. They compare the responses when the list with S statements includes only neutral statements against a list with less neutral statements. The latter list is intended to make the sensitive issue (sexual behavior in their study) less salient and see how this affects the report. The authors found a higher prevalence rate when the sensitive statement was accompanied with other sensitive statements. While they do not compare these two arms against a direct question as in our case, their findings suggest that our experiment, carried out with true neutral lists, may represent a lower bound. Also, it is not clear whether additional biases, if any, are created by the lists that include other sensitive statements. In any case, we argue that the emotional violence module in our questionnaire served as a smooth transition into the physical and sexual IPV questions, reducing the saliency of the sensitive statements in module 5B for the treatment group.

3.3 Estimation

Let T_i denote the binary treatment assignment to the list experiment and let L_i be the number of statements that hold true for individual i. The difference-in-means estimator ρ approximates the prevalence rate of the sensitive statement:

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

where additional controls can also be added to the regression model. Let the prevalence rates under the direct questions be denoted as p. The level of misreporting between the list experiment and the direct questions is measured by $(\rho - p)$. 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 when DHS-type questions are implemented.

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

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

where x_i is a binary variable indicating that individual i has characteristic x. Additional controls besides x_i can potentially be added to the regression model in (2). The term $(\rho' + \zeta)$ captures the prevalence rate as measured by experimental methods among individuals with $x_i = 1$ while ρ' will measure the prevalence rate for those with $x_i = 0$.[13] These prevalence rates can be compared to their counterpart measures obtained through direct reporting, conditional on x_i.

Since the data produced by list experiments preclude us from obtaining individual-level prevalence rates, we cannot deal with the issue of multiple outcomes by creating aggregate IPV indexes. Thus, we correct for the potential issue of simultaneous inference within

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

each sub-group as defined by x_i using multiple hypothesis testing. We rely on the false discovery rate (FDR) adjusted q-values as introduced by Benjamini and Hochberg [1995]. We implement this correction for the family of seven IPV outcomes for the overall analysis and for each sub-group used in the heterogeneity analysis. The FDR estimates the expected proportion of tests that are false positives among all significant tests. For example, an FDR adjusted q-value of 0.1 indicates that ten percent of the *significant tests* are false positives. In comparison, an unadjusted p-value of 0.1 implies that ten percent of *all tests* (significant or not) are false positives. Finally, all regressions include controls for the three variables that are not balanced between treatment and control samples, that is education level, memory test score at the beginning of the survey, and household head status (see Table 3). Additionally, we follow Bruhn and McKenzie [2009] and control for additional variables that may have a strong link with the outcome of interest such as marital status and working status.

4 Results

Table 5 presents the main findings of our paper. The first and second columns show the prevalence rates using indirect (ρ) and direct reporting methods (p) for physical and sexual IPV, respectively. The last column measures the gap ($\rho - p$) for each act of IPV while the last two rows report the results from a joint test of significance of these gaps for all acts of violence analyzed. On average, IPV rates obtained with direct questions do not differ when compared to experimental methods that provide more privacy to the respondent. For six out of seven acts of physical violence, the prevalence rates obtained through randomized lists do not significantly differ from those measured using direct DHS-type questions. In fact, when correcting for multiple hypotheses testing, using the FDR q-values, none of the seven IPV acts differ across reporting methods. This result is further reinforced with the joint test, which fails to reject the null hypothesis that the seven gaps are jointly zero. This allows us to rule out an overall reporting bias.[14]

This result validates the quality of IPV data collected under UN guidelines and supports the reliability of population-based surveys. However, it does not rule out misreporting among different groups. More vulnerable groups with higher costs of being exposed could be less likely to truthfully report violence under direct methods that provide limited privacy in the field.

Estimates of the prevalence rates and gaps by different individual characteristics are obtained from estimating equation (2).[15] In particular, we focus on eight characteristics that may be correlated with respondents' costs of being exposed as a victim. These include age, marital status, education level, mother tongue (as a proxy for ethnicity), memory test scores, household head status, employment status, and tenure at ADRA. For each of the sixteen sub-samples (two groups for each of the eight characteristics), Table 6 reports the joint significance tests that prevalence rates in all measured acts of physical and sexual IPV are jointly zero. The key finding here is that for all but one of the 16 sub-samples, we fail to find IPV reporting differences between direct and indirect methods. This is further confirmed in Figures A.2-A.8, which depict the gaps in prevalence rates for each violent act and sub-sample. Overwhelmingly, the evidence indicates that the best practices to elicit IPV prevalence rates do not lead to systematic misreporting patterns.

We only find differential reporting across DHS-type and list experiments methods among women with tertiary education. Table 7 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 and that these differences survive the multiple hypotheses test-

[14] In the control group, the non-response rate for the IPV module with the direct questions is 5.4 percent. 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 percent in the treatment group and close to null in the control group.

[15] Appendix Table A.4 shows power calculations and the negative predicted value for ten sub-groups using data from DHS. Both estimates indicate the we have enough power for at least seven of these sub-groups. Potential misreporting in the DHS itself further reduces the possibility that our study is underpowered. Since measurement error in IPV is most likely to emerge due to *underreporting*, power calculations using DHS "biased" baseline values would impose more demanding requirements on the sample size needed to identify an effect.

ing correction and the joint test. Less-educated women do not exhibit any significant differences by reporting method. The effect among educated women is not capturing a better understanding of the list experiment questions since there are no significant differences by characteristics that may proxy better performance in this module (see Figures A.4 for differences by mother's tongue and A.5 for those based on a memory test implemented within the survey). Also, the IPV rates under full anonymity are large enough that reverse the education gradient in violence. Panel (a) in Figure 1 shows that direct reporting produces a negative correlation between education level and prevalence rates: less educated women report more violence than those with more years of schooling. However, the provision of greater privacy under the list experiment reverses the direction of this relationship as shown in Panel (b). Once the costs of being exposed as a victim are minimized through greater provision of privacy, women with complete tertiary education exhibit *higher* prevalence rates of physical and sexual IPV, irrespective of the violent act considered.[16]

The switch in the sign of the correlation between years of education and IPV when moving from direct to indirect methods is also present in very recent evidence from Africa.[17] This reversal, present across different contexts, suggests that the protective effect of education identified by previous studies [Ackerson et al., 2008; Eswaran and Malhotra, 2011; Yuan and Hesketh, 2019] may be in part driven by misreporting issues when using direct methods. Instead, our result on the positive relationship between education and IPV, under indirect methods, provides supportive evidence for instrumental theories of violence as discussed in Angelucci [2008]; Tankard et al. [2019]; Erten and Keskin [2018].

The difference in reporting among highly educated women cannot necessarily be extended

[16] Putting forward a full explanation for why schooling leads to systematic misreporting in our sample goes beyond the scope of this paper. Nevertheless, we speculate that the source of non-random measurement error among more educated women is a higher stigma cost due to more equal views on gender roles. Education provides access to new knowledge, ideas, and lifestyles that expose individuals to gender roles different from the ones accepted in their communities or localities [Marcus and Harper, 2015]. Since contact with more gender-equal social norms increases the costs imposed by stigma we rationalize the greater propensity to misreport among the most educated in our sample through this channel. See Lindbeck et al. [1999] for an example of how social norms and stigma are related in the case of welfare recipients.

[17] Cullen [2020] compares direct and indirect methods in Nigeria and Rwanda and finds that the IPV-education relation also depends on how IPV is measured.

to the female population in Lima or to other contexts. As discussed above, female clients of microcredit organizations are not a random sample but rather represent a particular sub-group with lower socioeconomic status and potentially higher levels of empowerment.[18] Nevertheless, the results by education level portrayed here are still useful to highlight the possibility of non-random measurement error, potentially context-specific, when measuring IPV.

5 Conclusion

This is the first study to measure and characterize the possible bias in direct reporting of violence against women in an experimental setting. We compare lifetime prevalence rates obtained from the most common source of IPV data in developing countries, the violence module of the Demographic and Health Surveys (DHS), to estimates using indirect questioning techniques that provide full anonymity and minimize exposure costs.

The results show that direct questions are a good source of data to measure prevalence of physical and sexual IPV. It is important to note that our questionnaire is closer to a "standalone" survey on violence because we did not include all the other modules embedded in a typical DHS (e.g., maternal and child health, birth history, etc.). Overall, our study gives evidence-based support for the UN recommendation of using standalone surveys relative to multiple-objective surveys such as the DHS.

Furthermore, we also reject the existence of significant differences in reporting across a

[18] Clearly, microcredit access may not only mimic an income transfer, but it could potentially have empowerment effects that can also play a role in the prevalence of IPV. Indeed, we acknowledge that our sample, females participating in ADRA's microcredit programs, are different than the average woman in Lima (see Table A.1). Although we are not aware of studies that directly look at the compound effect of microcredit on IPV prevalence, several studies provide evidence on the relationship between woman's income and IPV. For instance, IPV may decrease as woman's income goes up, improving the value of her outside option and/or raising her participation constraint in the partnership [Tauchen et al., 1991; Farmer and Tiefenthaler, 1997; Bobonis et al., 2013; Eswaran and Malhotra, 2011; Haushofer and Shapiro, 2013]. However, other papers show that this effect may be reversed whenever the man feels threatened by the woman's increase in bargaining power or if he decides to use violence as a means to extract resources from her [Angelucci, 2008; Tankard et al., 2019]. In either case, while we recognize that access to micro-credit can affect the *level* of IPV prevalence in our sample, it certainly does not play a role in our findings since both women in the treatment and control groups in our experiment are members of microcredit programs.

large set of subgroups, reinforcing the finding that direct questioning techniques are quite accurate to monitor the evolution of IPV prevalence rates. However, differences by education are present even after accounting for multiple hypothesis testing: college-educated women underreport physical and sexual violence while there is no bias among the less educated. This highlights the possibility of context-specific heterogenous effects that should be further explored to avoid misdiagnosis and targeting issues.

References

Ackerson, L. K., Kawachi, I., Barbeau, E. M. and Subramanian, S. [2008], ‘Effects of individual and proximate educational context on intimate partner violence: A population-based study of women in india’, *American Journal of Public Health* **98**(3), 507–514.

Ambler, K., Doss, C., Kieran, C. and Passarelli, S. [2019], ‘He says, she says: Spousal disagreement in survey measures of bargaining power’, *Economic Development and Cultural Change* p. forthcoming.

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

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

Benjamini, Y. and Hochberg, Y. [1995], ‘Controlling the false discovery rate: a practical and powerful approach to multiple testing’, *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

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

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

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K. and Sheridan, M. [2016], ‘Measuring the measurement error: A method to qualitatively validate survey data’, *Journal of Development Economics* **120**, 99 – 112.

Bobonis, G. J., González-Brenes, M. and Castro, R. [2013], ‘Public Transfers and Domestic Violence: The Roles of Private Information and Spousal Control’, *American Economic Journal: Economic Policy* pp. 179–205.

Bott, S., Guedes, A., Goodwin, M. and Mendoza, J. A. [2012], *Violence Against Women in Latin America and the Caribbean: A comparative Analysis of Population-Based Data from 12 Countries*, Washington, DC: Pan American Health Organization.

Bound, J., Brown, C., Duncan, G. J. and Rodgers, W. L. [1994], ‘Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data’, *Journal of Labor Economics* **12**(3), 345–368.

Bound, J., Brown, C. and Mathiowetz, N. [2001], ‘Measurement Error in Survey Data’, *Handbook of econometrics* **5**, 3705–3843.

Breiding, M. J., Black, M. C. and Ryan, G. W. [2008], ‘Prevalence and Risk Factors of Intimate Partner Violence in Eighteen US States/territories, 2005’, *American Journal of Preventive Medicine* **34**(2), 112–118.

Bruhn, M. and McKenzie, D. [2009], 'In pursuit of balance: Randomization in practice in development field experiments', *American Economic Journal: Applied Economics* **1**(4), 200–232.

Bulte, E. and Lensink, R. [2019], 'Women's empowerment and domestic abuse: Experimental evidence from vietnam', *European Economic Review* **115**(June), 172–191.

Bureau of Justice Statistics [2017], Police Response to Domestic Violence, 2006-2015, Technical report, U.S. Department of Justice, Office of Justice Programs. Special Report.

Butler, J. S., Burkhauser, R. V., Mitchell, J. M. and Pincus, T. P. [1987], 'Measurement Error in Self-Reported Health Variables', *The Review of Economics and Statistics* **69**(4), 644–650.

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

Calvo, T., Razafindrakoto, M. and Roubaud, F. [2019], 'Fear of the state in governance surveys? empirical evidence from african countries', *World Development* **123**, forthcoming.

Capaldi, D. M., Knoble, N. B., Shortt, J. W. and Kim, H. K. [2012], 'A Systematic Review of Risk Factors for Intimate Partner Violence', *Partner Abuse* **3**(2), 231–280.

Chuang, E., Dupas, P., Huillery, E. and Seban, J. [2019], Sex, Lies, and Measurement: Do Indirect Response Survey Methods Work?, Technical report, Working paper.

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

Convergences [2018], Microfinance barometer 2018, Technical report, Convergences and Zero Exclusion Carbon Poverty.

Cullen, C. [2020], Method matters: Underreporting of intimate partner violence in nigeria and rwanda, Technical report, Policy Research Working Paper No. 9274, The World Bank.

De Cao, E. and Lutz, C. [2018], 'Sensitive survey questions: Measuring attitudes regarding female genital cutting through a list experiment', *Oxford Bulletin of Economics and Statistics* **80**(5), 871–892.

De Koker, P., Mathews, C., Zuch, M., Bastien, S. and Mason-Jones, A. J. [2014], 'A Systematic Review of Interventions for Preventing Adolescent Intimate Partner Violence', *Journal of Adolescent Health* **54**(1), 3–13.

Dillon, A., Gourlay, S., McGee, K. and Oseni, G. [2019], 'Land measurement bias and its empirical implications: Evidence from a validation exercise', *Economic Development and Cultural Change* **67**(3), 595–624.

Duflo, E. [2012], 'Women empowerment and economic development', *Journal of Economic Literature* **50**(4), 1051–1079.

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

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

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

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

Farmer, A. and Tiefenthaler, J. [1997], 'An economic analysis of domestic violence', *Review of social Economy* **55**(3), 337–358.

Fulu, E., Jewkes, R., Roselli, T. and Garcia-Moreno, C. [2013], 'Prevalence of and Factors Associated with Male Perpetration of Intimate Partner Violence: Findings from the UN Multi-country Cross-sectional Study on Men and Violence in Asia and the Pacific', *The Lancet Global Health* **1**(4), e187–e207.

Gillen, B., Snowberg, E. and Yariv, L. [2019], 'Experimenting with measurement error: Techniques with applications to the caltech cohort study', *Journal of Political Economy* **127**(4), forthcoming.

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

Gottschalk, P. and Huynh, M. [2010], 'Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error', *The Review of Economics and Statistics* **92**(2), 302–315.

Haushofer, J. and Shapiro, J. [2013], 'Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya'.

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

Hidrobo, M., Peterman, A. and Heise, L. [2016], 'The Effect of Cash, Vouchers, and Food Transfers on Intimate Partner Violence: Evidence from a Randomized Experiment in Northern Ecuador', *American Economic Journal: Applied Economics* **8**(3), 284–303.

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

Jamison, J. C., Karlan, D. and Raffler, P. [2013], ‘Mixed-method evaluation of a passive mhealth sexual information texting service in uganda’, *Information Technologies & International Development* **9**(3), 1–28.

Jewkes, R., Levin, J. and Penn-Kekana, L. [2002], ‘Risk Factors for Domestic Violence: Findings from a South African Cross-sectional Study’, *Social Science & Medicine* **55**(9), 1603–1617.

Johnston, D. W., Propper, C. and Shields, M. A. [2009], ‘Comparing Subjective and Objective Measures of Health: Evidence from Hypertension for the Income/health Gradient’, *Journal of health economics* **28**(3), 540–552.

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

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

Kataoka, Y., Yaju, Y., Eto, H. and Horiuchi, S. [2010], ‘Self-administered questionnaire versus interview as a screening method for intimate partner violence in the prenatal setting in japan: A randomised controlled trial’, *BMC Pregnancy and Childbirth* **10**(84), 1–7.

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

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

Koenig, M. A., Ahmed, S., Hossain, M. B. and Mozumder, A. K. A. [2003], ‘Women’s Status and Domestic Violence in Rural Bangladesh: Individual-and Community-level Effects’, *Demography* **40**(2), 269–288.

Krug, E. G., Mercy, J. A., Dahlberg, L. L. and Zwi, A. B. [2002], ‘The World Report on Violence and Health’, *The Lancet* **360**(9339), 1083–1088.

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

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

Marcus, R. and Harper, C. [2015], How do gender norms change?, Technical report, Overseas Development Institute, London.

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

O'Neill, D. [2012], The Consequences of Measurement Error when Estimating the Impact of BMI on Labour Market Outcomes, Technical report, IZA Discussion Paper No. 7008.

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

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

Palermo, T., Bleck, J. and Peterman, A. [2014], 'Tip of the Iceberg: Reporting and Gender-based Violence in Developing Countries', *American Journal of Epidemiology* **179**(5), 602–612.

Panda, P. and Agarwal, B. [2005], 'Marital violence, human development and women's property status in india', *World development* **33**(5), 823–850.

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

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

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

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

Tankard, M., Levy Paluck, E. and Prentice, D. [2019], 'The effect of a savings intervention on women's intimate partner violence victimization: heterogeneous findings from a randomized controlled trial in colombia', *BMC Women's Health* **19**(1), 17.

Tauchen, H., Dryden Witte, A. and Long, S. [1991], 'Domestic violence: A nonrandom affair', *International Economic Review* **32**(2), 491–511.

United Nations [2014], *Guidelines for Producing Statistics on Violence Against Women: Statistical Surveys*, United Nations. Statistical Office.

United Nations [2015], 'Transforming our world: The 2030 agenda for sustainable development', *Resolution adopted by the General Assembly* .

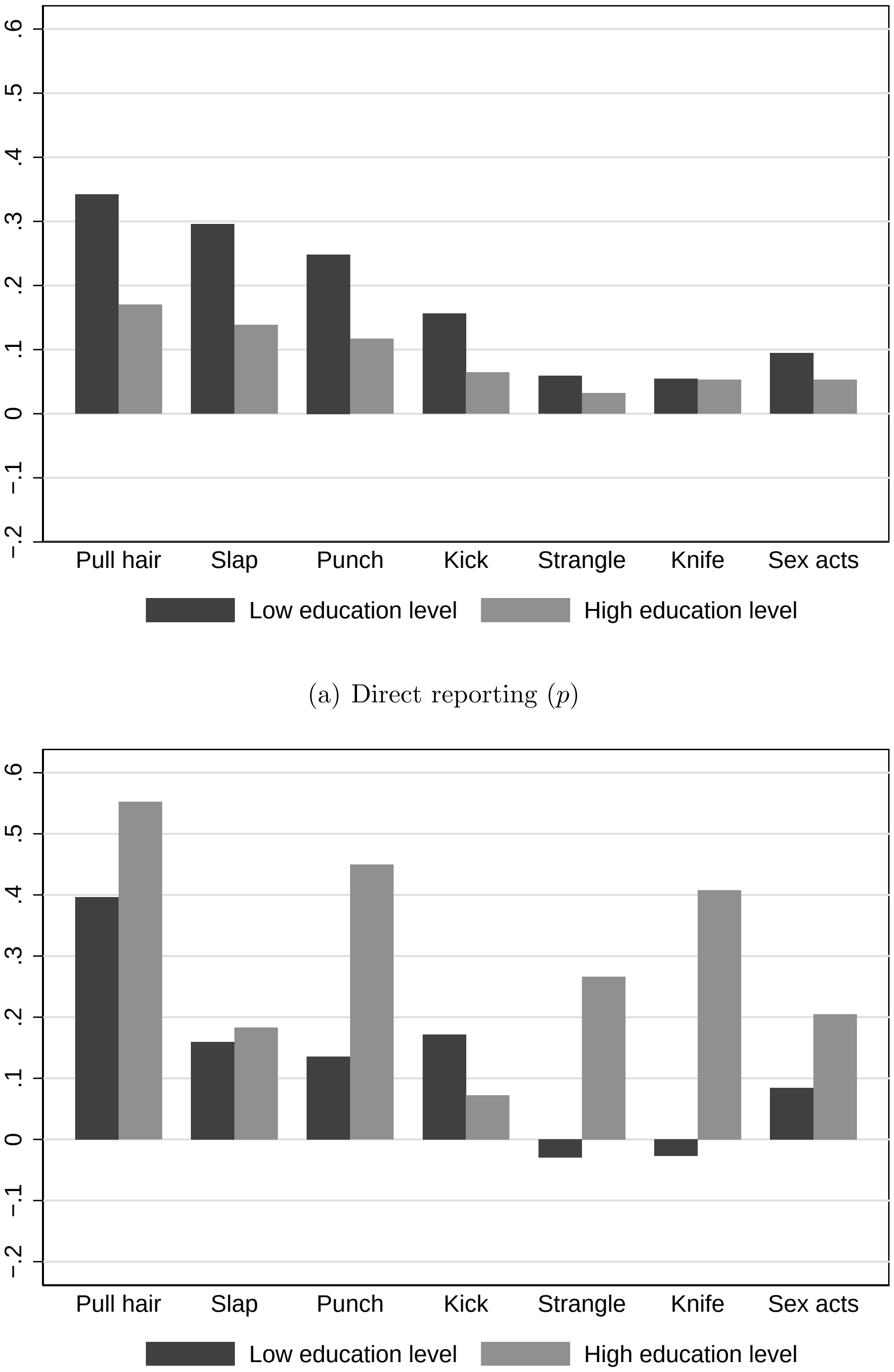
World Health Organization [2005], ‘Who multi-country study on women’s health and domestic violence against women: initial results on prevalence, health outcomes and women’s responses’.

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

Yuan, W. and Hesketh, T. [2019], ‘Intimate partner violence against women and its association with depression in three regions of china: a cross-sectional study’, *The Lancet* **394**(S5).

Figures and Tables

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



(a) Direct reporting (p)

(b) Indirect reporting (ρ)

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

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	

Table 2: Prevalence rates of IPV under DHS-type questions

	N. Observations	Prevalence Rate
Emotional IPV	992	0.65
Humiliate	990	0.38
Insult	989	0.35
Called Lazy	990	0.27
Threatens to harm	990	0.15
Threatens to Leave	990	0.33
Physical and sexual IPV	518	0.46
Pull hair	518	0.31
Slap	517	0.27
Punch	517	0.22
Kick	516	0.14
Strangle	518	0.05
Knife	518	0.05
Unapproved Sex practices	516	0.09
Any IPV	518	0.72

Note: Direct questions about emotional violence were asked to both treatment and control groups. Direct questions about physical and sexual violence applied only to the control group.

Table 3: Summary Statistics and Balance Check

	Control	(T-C)	N
Demographic Characteristics			
Age	42.178	0.759	992
	(10.423)	[0.644]	
Married	0.801	0.001	992
	(0.400)	[0.025]	
Literate	0.967	0.008	992
	(0.178)	[0.011]	
Spanish is not mother tongue	0.106	0.014	992
	(0.308)	[0.020]	
Household head	0.305	0.071	992
	(0.461)	[0.030]**	
Works	0.732	0.009	992
	(0.444)	[0.028]	
Less than complete primary	0.093	0.000	992
	(0.290)	[0.018]	
Primary education	0.263	-0.028	992
	(0.440)	[0.027]	
Secondary education	0.463	-0.014	992
	(0.499)	[0.032]	
Higher education	0.181	0.042	992
	(0.386)	[0.026]*	
Number of children	2.833	0.013	989
	(1.585)	[0.093]	
Number of children under 12 under her care	0.908	0.042	977
	(1.091)	[0.068]	
Memory test: % words remembered right after	0.855	0.046	992
	(0.352)	[0.021]**	
Memory test: % words remembered at the end	0.498	0.046	992
	(0.500)	[0.032]	
Always lived in current locality	0.635	-0.021	992
	(0.482)	[0.031]	
Financial Situation			
Average loan size in past 4 cycles	1535.462	-29.836	945
	(1178.896)	[74.047]	
Average savings balance in past 4 cycles	785.622	18.831	945
	(876.089)	[61.594]	
High loan size, savings balance, and tenure	0.158	0.008	992
	(0.365)	[0.023]	
Emotional IPV			
Humiliates her in public	0.377	0.001	990
	(0.485)	[0.031]	
Calls her ignorant or idiot	0.359	-0.024	989
	(0.480)	[0.030]	
Calls her lazy, useless, or sleepy	0.267	-0.003	990

Continued on next page

	Control	(T-C)	N
	(0.443)	[0.028]	
Threatened to harm her or someone close to her	0.157	-0.009	990
	(0.364)	[0.023]	
Threatened to leave, take children, or cut off financial support	0.335	-0.012	990
	(0.473)	[0.030]	
Survey Application			
Interruption by men	0.042	0.002	992
	(0.202)	[0.013]	
Interruption by partner	0.008	-0.006	992
	(0.088)	[0.004]	
Presence partner	0.019	-0.007	992
	(0.138)	[0.008]	

NOTE: Differences between control and treatment group are obtained from regressing each variable on the treatment dummy. Standard errors in parenthesis. Significance levels (* 10%; ** 5%; *** 1%).

Table 4: Difference in Responses and Item Non-Responses to the Last Module Across Treatment Arms

	Control	(T-C)	N
Panel A. Differences in answers			
Likely to assume role in VB committee	0.501	0.027	987
	(0.500)	(0.032)	
Likely to recommend ADRA to others	0.956	-0.025	990
	(0.206)	(0.015)*	
Likely to stay in VB	0.794	-0.022	982
	(0.405)	(0.026)	
Satisfied with family talks	0.830	0.020	990
	(0.376)	(0.023)	
Satisfied with loans	0.869	-0.004	990
	(0.338)	(0.022)	
Satisfied with sports events	0.588	-0.028	990
	(0.493)	(0.031)	
Satisfied with training	0.811	0.012	991
	(0.392)	(0.025)	
Panel B. Differences in item non-response			
Likely to assume role in VB committee	0.006	-0.002	992
	(0.076)	(0.004)	
Likely to recommend ADRA to others	0.002	0.000	992
	(0.044)	(0.003)	
Likely to stay in VB	0.008	0.005	992
	(0.088)	(0.006)	
Satisfied with family talks	0.002	0.000	992
	(0.044)	(0.003)	
Satisfied with loans	0.000	0.004	992
	(0.000)	(0.003)	
Satisfied with sports events	0.002	0.000	992
	(0.044)	(0.003)	
Satisfied with training	0.000	0.002	992
	(0.000)	(0.002)	

NOTE: Differences between control and treatment group are obtained from regressing each variable on the treatment dummy. Standard errors in parenthesis. Significance levels (* 10%; ** 5%; *** 1%).

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

Violent act	List experiments (ρ)	Direct reporting (p)	($\rho - p$)
Pull hair	0.427	0.311	0.117
	(0.061)	(0.071)	(0.064)*
Slap	0.164	0.267	-0.103
	(0.065)	(0.074)	(0.068)
Punch	0.198	0.224	-0.026
	(0.071)	(0.077)	(0.073)
Kick	0.152	0.140	0.012
	(0.067)	(0.074)	(0.069)
Strangle	0.029	0.054	-0.025
	(0.065)	(0.071)	(0.066)
Knife	0.060	0.054	0.006
	(0.066)	(0.074)	(0.067)
Sex acts	0.108	0.087	0.021
	(0.069)	(0.077)	(0.071)
Joint test			
$\chi 2$		8.406	
Prob > $\chi 2$		0.298	

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. Stars denote significance levels (* 10%; ** 5%; *** 1%) based on unadjusted p-values. Daggers denote significance levels († 10%; †† 5%; ††† 1%) based on FDR q-values.

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

Characteristics	$\chi 2$	$Prob > \chi 2$
Age		
<50	8.117	0.322
50+	6.318	0.503
Marital status		
Single	6.171	0.520
Married	5.304	0.623
Education level		
Less than tertiary	7.932	0.339
Completed tertiary	17.896	0.012
Mother tongue		
Spanish	11.883	0.104
Other language	5.930	0.548
Memory test		
Low score	2.891	0.895
High score	10.096	0.183
Household head		
Not the head	9.791	0.201
Head	3.288	0.857
Employment		
Does not work	3.573	0.827
Works	8.311	0.306
Standing in ADRA		
Young client	10.283	0.173
Mature client	6.302	0.505

Note: The null hypothesis reported for each subgroup is that the biases $(\rho - p)$ for the seven acts of physical and sexual violence are jointly zero. See Tables 5 and A.4-A.10 for details about each regression model.

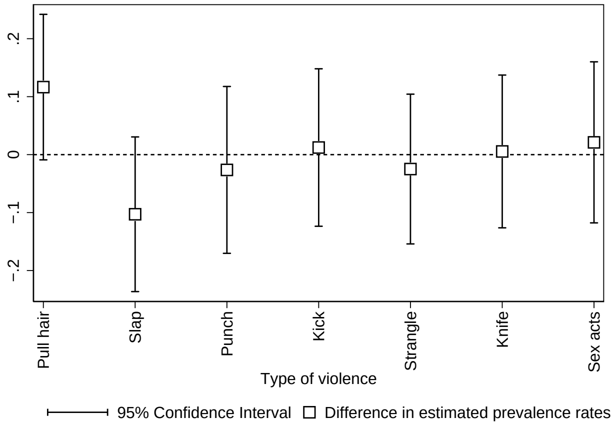
Table 7: Difference in estimated prevalence rates of physical and sexual IPV by education

	Less than tertiary education			Tertiary education		
Violent act	ρ	p	$\rho - p$	ρ	p	$\rho - p$
Pull hair	0.397	0.342	0.055	0.552	0.170	0.382
	(0.069)	(0.023)	(0.072)	(0.129)	(0.039)	(0.134)***
Slap	0.159	0.296	-0.136	0.183	0.138	0.045
	(0.075)	(0.022)	(0.078)*	(0.129)	(0.036)	(0.132)
Punch	0.136	0.248	-0.113	0.450	0.117	0.332
	(0.081)	(0.021)	(0.083)	(0.148)	(0.033)	(0.150)**
Kick	0.172	0.156	0.016	0.072	0.065	0.008
	(0.076)	(0.018)	(0.079)	(0.139)	(0.025)	(0.144)
Strangle	-0.030	0.059	-0.089	0.266	0.032	0.234
	(0.074)	(0.011)	(0.075)	(0.132)	(0.018)	(0.132)*
Knife	-0.027	0.054	-0.081	0.408	0.053	0.354
	(0.075)	(0.011)	(0.076)	(0.138)	(0.023)	(0.141)**
Sex acts	0.085	0.095	-0.010	0.205	0.053	0.151
	(0.078)	(0.014)	(0.080)	(0.147)	(0.023)	(0.148)
Joint test						
$\chi 2$		8.406			17.896	
$Prob > \chi 2$		0.298			0.012	

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. Differential effects by education level are obtained from the model in Equation (2). Stars denote significance levels (* 10%; ** 5%; *** 1%) based on unadjusted p-values. Daggers denote significance levels († 10%; †† 5%; ††† 1%) based on FDR q-values.

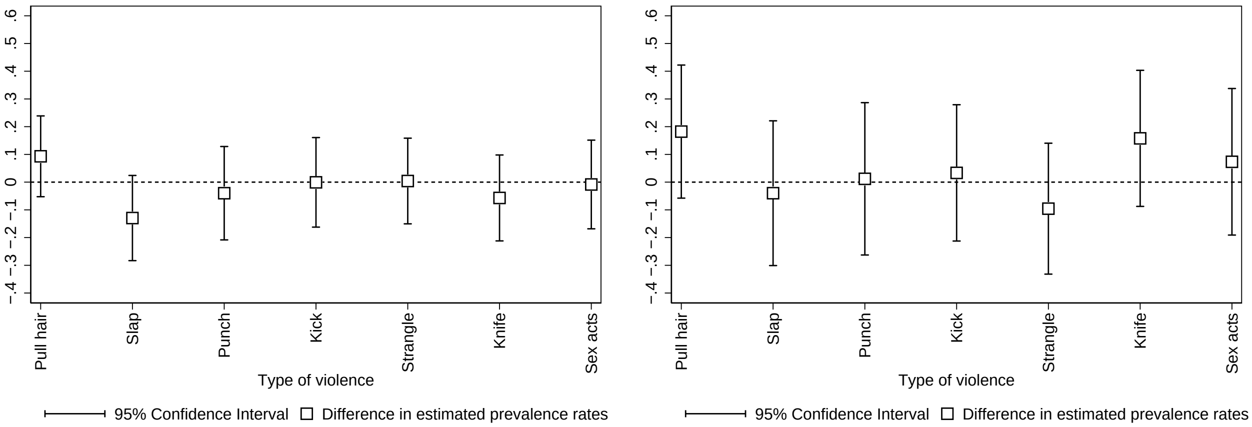
A Additional Figures and Tables (Online Appendix: Not for publication)

Figure A.1: Difference in estimated prevalence rates of physical and sexual IPV



Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant.

Figure A.2: Difference in estimated prevalence rates of physical and sexual IPV by age

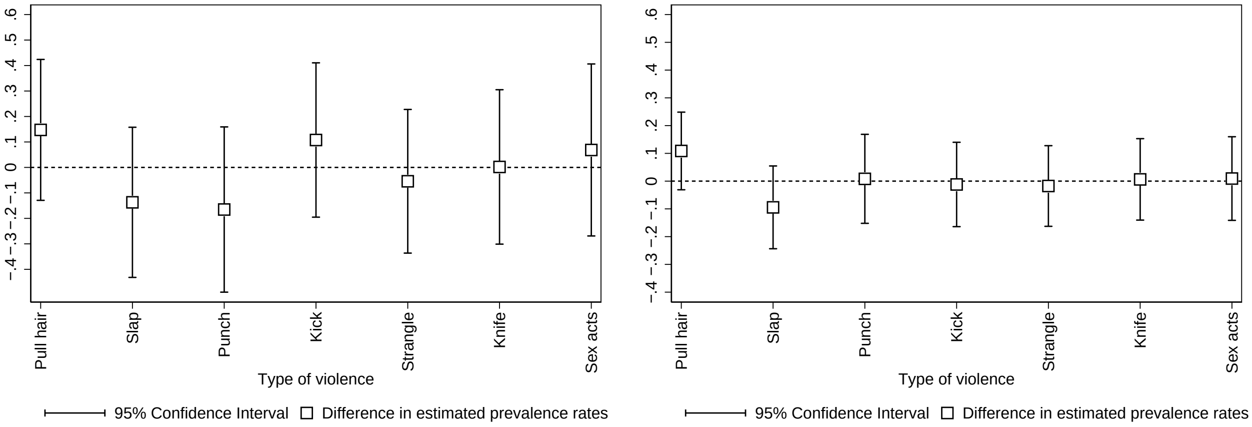


(a) Under 50

(b) 50 or more

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.3: Difference in estimated prevalence rates of physical and sexual IPV by marital status

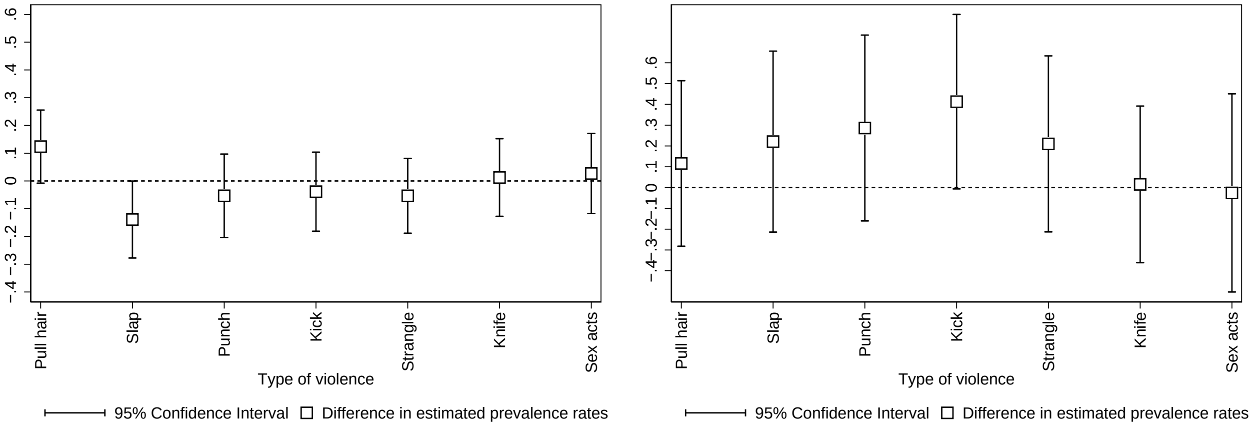


(a) Single

(b) Married

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.4: Difference in estimated prevalence rates of physical and sexual IPV by mother tongue

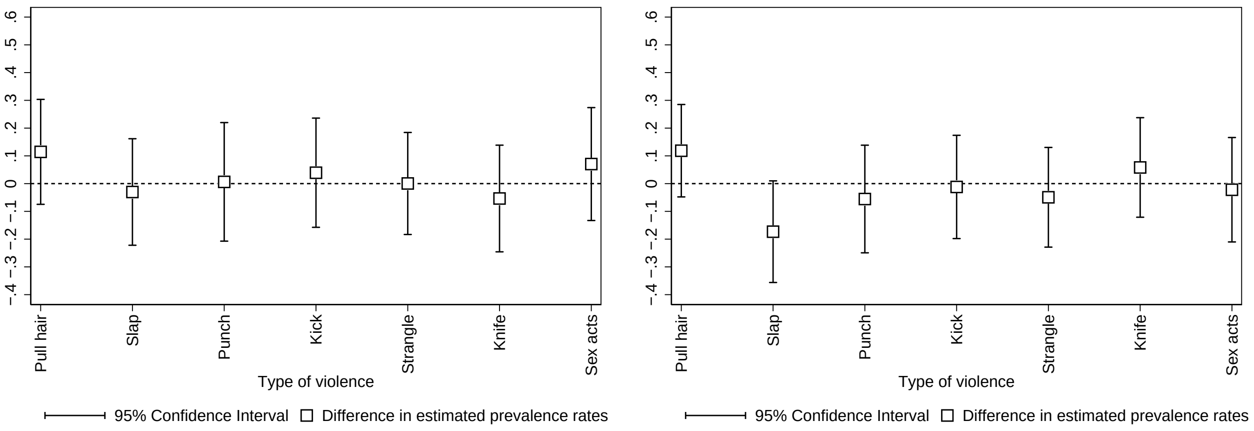


(a) Spanish

(b) Other language

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.5: Difference in estimated prevalence rates of physical and sexual IPV by memory test score

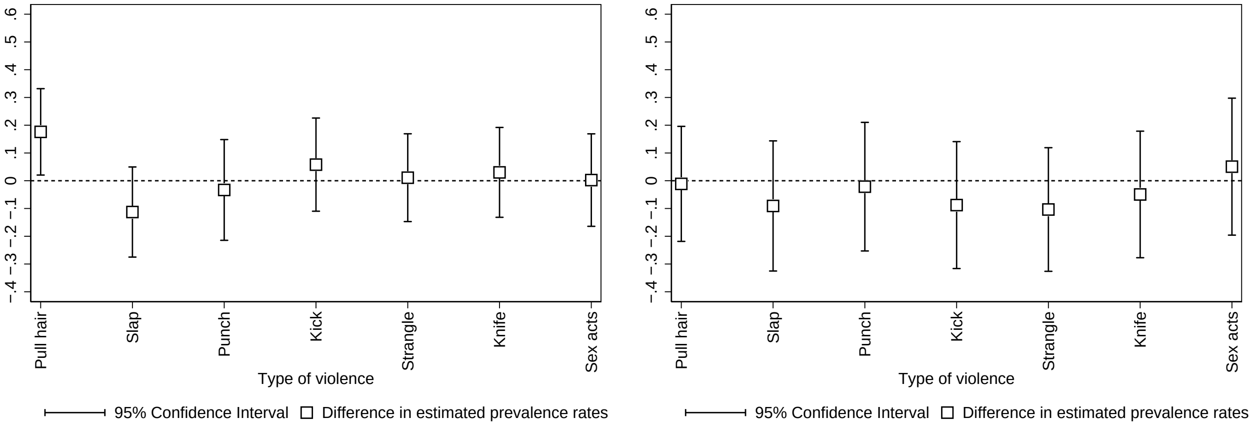


(a) Low score

(b) High score

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.6: Difference in estimated prevalence rates of physical and sexual IPV by household head status

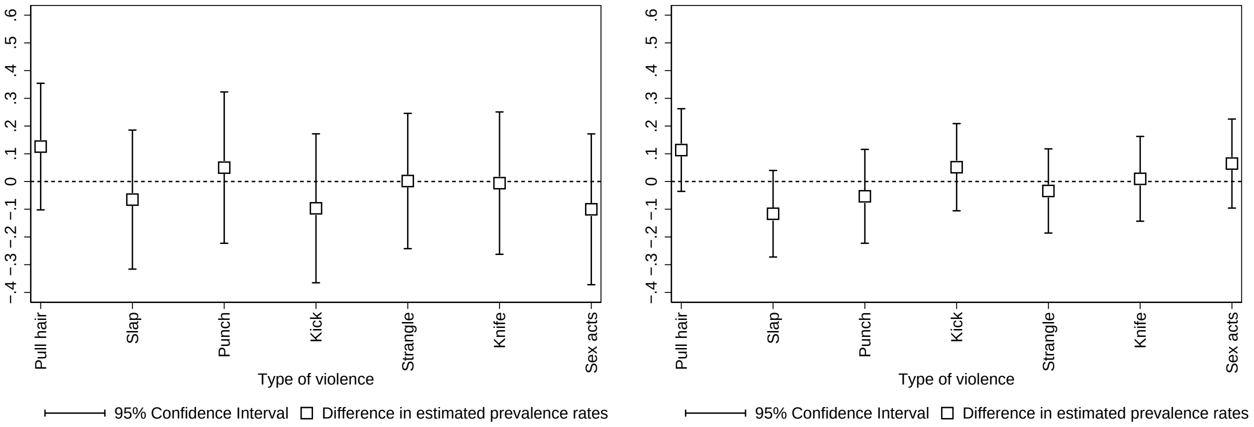


(a) Not the head

(b) Head

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.7: Difference in estimated prevalence rates of physical and sexual IPV by employment status

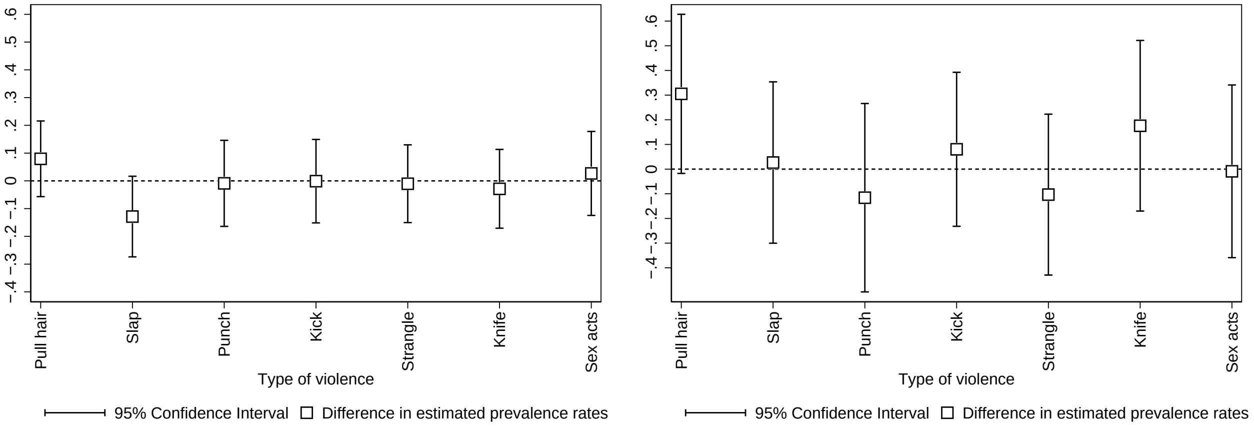


(a) Does not work

(b) Works

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, and marital status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Figure A.8: Difference in estimated prevalence rates of physical and sexual IPV by standing in ADRA



(a) Mature client

(b) Young client

Note: Robust standard errors in parenthesis. Estimates of ρ are obtained from a regression of the indirect answer on the treatment dummy, with the following additional controls: education level, memory test score at the beginning of the survey, household head status, marital status, and working status. Estimates of p are obtained from a regression of the direct answer on a constant. FDR q-values reported no significance for any type of violence, nor any of the two groups compared in the figure.

Table A.1: Comparison between ADRA clients and women in Lima

Characteristics	Lima	ADRA
Physical and sexual IPV	0.31	0.46
Age	33.1	42.2
Married	0.57	0.80
Literate	0.98	0.97
Spanish is not mother tongue	0.00	0.11
Household head	0.14	0.31
Works	0.70	0.73
Less than complete primary	0.05	0.09
Primary	0.14	0.26
Secondary	0.33	0.46
Higher education	0.48	0.18
Number of children	1.56	2.83
Always lived here	0.60	0.64

Note: Sample for Lima comes from the 2015 Peruvian Demographic and Health Survey. This survey interviews women aged 15-49, while the ADRA sample focused on women aged 18-60. For a better comparison, we restrict the DHS sample to those above 18. ADRA's reported statistics refer to the control group as they answer the direct questions on physical and sexual violence.

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
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
had to truncate your studies to care for your family?	0.742	0.445
pursued a technical degree?	0.387	0.495
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
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
had problems with your partner because of money issues?	0.839	0.374

NOTES: Prevalence rates reported correspond to the pilot sample (N=31).

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

List 1: Pull hair	1a	1b	1c	1d
1a. Purchased a TV with HD	1.00			
1b. Been out for a walk with your children	-0.29	1.00		
1c. Helped your children with their homework	0.12	-0.03	1.00	
1d. Bought expensive clothes	0.33	0.10	-0.34	1.00

List 2: Slap	2a	2b	2c	2d
2a. Pursued a technical degree	1.00			
2b. Business' sales are insufficient to cover household expenses	-0.29	1.00		
2c. Traveled with your friends	-0.12	-0.16	1.00	
2d. Been to the movies with your family	0.34	-0.29	-0.35	1.00

List 3: Punch	3a	3b	3c	3d
3a. Witnessed robberies in your neighborhood	1.00			
3b. Been robbed on the street	-0.29	1.00		
3c. Had insurance from ESSALUD, the armed forces or the police	0.25	-0.02	1.00	
3d. Have been depressed				

List 4: Kick	4a	4b	4c	4d
4a. Felt insecure in your neighborhood	1.00			
4b. Had multiple businesses at the same time	-0.37	1.00		
4c. Reared farm animals for consumption	-0.07	0.22	1.00	
4d. Used the subway as a means of transportation	-0.06	-0.07	-0.37	1.00

List 5: Strangle	5a	5b	5c	5d
5a. Run out of money to cover the household's monthly expenses	1.00			
5b. Traveled with your children	-0.28	1.00		
5c. Been part of a Christian church	-0.23	-0.10	1.00	
5d. Had to truncate your studies to care for your family	-0.05	0.14	-0.31	1.00

List 6: Knife	6a	6b	6c	6d
6a. Paid rent for the place where you live	1.00			
6b. Participated in other microfinance programs	-0.54	1.00		
6c. Bought any high-end clothes	0.15	0.03	1.00	
6d. Participated in a committee or association in your neighborhood	0.09	-0.13	-0.28	1.00

List 7: Sex acts	7a	7b	7c	7d
7a. Made improvements to your dwelling	1.00			
7b. Had problems with your partner because of money issues	-0.24	1.00		
7c. Have received a loan from MiBanco				
7d. Suffered serious medical condition that required medical assistance	-0.04	-0.11		1.00

Note: Questions 3 and 7 include correlations for only three statements because one of them in each list did not come from the set of statements tested in the pilot (see Table A.2).

Table A.4: Power Calculations by Sub-Sample

	Sample size	Power	NPV
	(1)	(2)	(3)
Single	197	0.717	0.761
Married	795	0.703	0.752
Less than tertiary	792	0.910	0.909
Completed tertiary [a/]	200	0.274	0.554
Spanish	880	0.863	0.868
Other language	112	0.562	0.673
Not the head	656	0.714	0.759
Head	336	0.757	0.787
Does not work	262	0.238	0.542
Works	730	0.903	0.903

[a/] We can detect a statistically significance difference between direct and indirect methods in this sub sample.

Note: The first column reports the actual sample sizes by sub-groups. Columns 2 reports the power implied by our sample size for each sub-group under the initial values from the 2015 Peruvian DHS, a significance level set at 10 percent and a proportional MDE, which varies in each sub-sample following the ratio of the MDE of 0.03 and the overall baseline prevalence rate used for the global sample. Column 3 uses these power calculations and reports the negative predictive value (NPV), the probability of a true no-difference in the prevalence rates across direct and indirect methods given that a non-significant effect was found.

B Non-classical measurement error in the outcome (Online Appendix: Not for publication)

Our results show that, on average, there is no evidence of misreporting of physical and sexual IPV experience. However, we found evidence of non-classical measurement error with the provision of anonymity through list experiments. In particular, more educated women underreport when using DHS-type direct questions.

In this appendix we show that such finding has important implications on the empirical literature that tries to identify the main drivers and triggers of intimate partner violence.

B.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 parametric econometric model. Suppose that a researcher wants to estimate β:

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

For example, 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 distributed $N(0, 1)$. For simplicity, Equation (B.1) 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.[19]

Now consider the case when y_i is measured with some noise. A 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[20] and define it as follows:

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

That is, the risk factor x is correlated with ϵ_i whenever $\gamma \neq 0$, introducing endogeneity in the estimation of β. In our paper, we found that education could play the role of x, but in this appendix we consider the general case where x is any risk factor (e.g., education, income, etc).

We now model the measurement error as a mix between a classical and a non-classical component:

$$\omega_i = \phi x_i + \nu_i \tag{B.2}$$

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

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

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

B.2 Causal estimation under endogeneity and measurement error biases

Consider the case where x_i is endogenous ($\gamma \neq 0$) 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:

$$\begin{aligned} \hat{\beta}_{\text{OLS}} &= \beta + \frac{\text{cov}(\epsilon_i, x_i)}{\text{var}(x_i)} + \frac{\text{cov}(\omega_i, x_i)}{\text{var}(x_i)} \\ &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi \qquad \text{(B.3)} \end{aligned}$$

In equation (B.3), the second term captures the endogeneity bias because $\gamma \neq 0$ but the third element (ϕ) corresponds to the non-classical measurement error bias.

B.3 Implications for current evidence

Many papers in the literature have tried to estimate Equation (B.1) via ordinary least squares using only cross-sectional variation to identify the impact of risk factors on violence against women.[21] More recent papers have tried to reduce or eliminate the endogeneity bias relying on exogenous variations introduced by RCTs. For example, Hidrobo and Fernald [2013], Hidrobo et al. [2016], Haushofer and Shapiro [2013], Angelucci [2008], and Bobonis et al. [2013], among others, have explored the role of income on IPV using the random allocation of conditional cash transfers (CCTs) to women in developing countries.[22] Other studies have tried to look at the impact of social norm interventions under an experimental design (see Pronyk et al. [2006] and World Health Organization [2009]). Another common strategy to deal with endogeneity problems is the use IV techniques as in Erten and Keskin [2018], where the authors rely on a school reform in Turkey as an instrument to evaluate the impact of women's education on the prevalence of violence.

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

Thus, non-classical measurement error imposes a limit to the gains that randomization

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

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

or IV provide to obtain less biased estimates of treatment effects. Since ϕ in Equation (B.3) does not go away under these identification strategies, 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 has the opposite sign of the correlation between x_i and ω_i.[23]

From Equation (B.2), notice that ϕ is the slope of the relation between the risk factor of interest (x_i) and the measurement error in the dependent variable (ω_i). By conducting an experiment similar to ours, researchers can directly estimate $E[\omega_i|x_i = x]$ and obtain ϕ by correlating it with different values x. This will allow them to compute the bias in their estimates of β. We thus argue that the lists experiments used in our study provide an inexpensive way to directly measure ϕ and correct biased estimates from RCTs or IV methods. Based on our study's budget and sample size, the cost per women to conduct our experiment was close to US$8. For projects already conducting fieldwork, as those implementing a RCT, the marginal cost of adding the questions required to conduct list experiments is even smaller.

B.4 Non-linear measurement error

In the previous section, we consider the possibility of a linear source of non-classical measurement error as in Blattman et al. [2016]. We extend this case to non-linear *and* non-classical measurement error as the one we identify in our sample. We redefine the measurement error introduced in Equation (B.2) as follows:

$$\omega_i = \pi_i(\phi x_i + \nu_i) + (1 - \pi_i)(\nu_i) \tag{B.4}$$

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:

$$\begin{aligned} \beta_{OLS} &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi\frac{\text{cov}(x_i, \pi_i x_i)}{\text{var}(x_i)} \\ &= \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi E(\pi_i) \end{aligned} \tag{B.5}$$

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.[24]

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

[24] Proof of Equation (B.5): Under the presence of non-linear and non-classical measurement error, the OLS estimator of β becomes:

$$\beta_{OLS} = \beta + \gamma\frac{\text{var}(\epsilon_i)}{\text{var}(x_i)} + \phi\frac{\text{cov}(x_i, \pi_i x_i)}{\text{var}(x_i)} \tag{B.6}$$

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 during data collection efforts. This will allow them to measure the magnitude of the bias in the estimated treatment effects introduced by non-classical measurement error based on the risk factor of interest.

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

Let

$$\text{cov}(x_i, \pi_i x_i) = E(\pi_i x_i^2) - E(x_i)E(\pi_i x_i) \tag{B.7}$$

where

$$\begin{aligned} E(\pi_i x_i^2) &= E(\pi_i x_i^2|\pi_i = 1)P[\pi_i = 1] + E(x_i \pi_i x_i|\pi_i = 0)P[\pi_i = 0] \\ &= E(x_i^2)P[\pi_i = 1] \end{aligned} \tag{B.8}$$

and

$$\begin{aligned} E(\pi_i x_i) &= E(\pi_i x_i|\pi_i = 1)P[\pi_i = 1] + E(\pi_i x_i|\pi_i = 0)P[\pi_i = 0] \\ &= E(x_i)P[\pi_i = 1] \end{aligned} \tag{B.9}$$

Plugging B.8 and B.9 into B.7 yields:

$$\begin{aligned} \text{cov}(x_i, \pi_i x_i) &= E(x_i^2)P[\pi_i = 1] - E(x_i)E(x_i)P[\pi_i = 1] \\ &= P[\pi_i = 1][E(x_i^2) - E^2(x_i)] \\ &= P[\pi_i = 1]\text{var}(x_i) \\ &= E(\pi_i)\text{var}(x_i) \end{aligned} \tag{B.10}$$

If we replace B.10 into B.6, we obtain the last line in B.5:

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

.