Measuring Norms and Enumerator Effects: Survey Method Matters

Pablo Álvarez-Aragón, Hugues Champeaux

January 3, 2024

DeFiPP Working Paper 2024-01

DeFiPP
Development Finance & Public Policies
defipp.unamur.be

UNIVERSITÉ DE NAMUR

Measuring Norms and Enumerator Effects: Survey Method Matters

Pablo Álvarez-Aragón Hugues Champeaux [*]

January 3, 2024

Abstract

The reliability of quantitative data is a prerequisite for the study and design of sound public policy. However, the process of data collection and the way in which individuals are interviewed affects the data collected and can lead to bias. While this process directly impacts the quality of the data, little empirical evidence investigates the key role of the survey methods themselves. In this paper, we compare two survey methods: the standard face-to-face interview and an alternative method we call in-group individual survey. In the latter, respondents are guided by an enumerator who reads them the questions, but they answer individually and privately on an electronic device. Taking advantage of an RCT in Benin, we randomize the survey method across respondents while holding the questionnaire constant. We show that the survey method leads to different results depending on the degree of enumerator influence. Identifying this influence by quantifying how much of the variation in the outcome variable is attributable to enumerators, we document that variables that are unlikely to be influenced by enumerators do not differ significantly across survey methods. However, variables that are likely to be affected differ systematically. These variables are mainly related to norms, opinions, and beliefs. In particular, we find that respondents who answer directly on an electronic device report less gender-equal behavior and values. To rule out other mechanisms, we show that social desirability bias is more likely to affect responses in classical face-to-face interviews, where individuals' responses are less confidential.

JEL Classification : C81, C83, C93, J16, O10, O12

Keywords: gender, enumerator effects, survey experiment, social desirability bias, measurement, Benin

[*] P. Álvarez-Aragón is affiliated to the University of Namur, Belgium; H. Champeaux is affiliated to the University of Cagliari, Italy. Emails: pablo.alvarez@unamur.be and hugues.champeaux@unica.it. We are grateful to Jean-Marie Baland, Jules Gazeaud, William Parienté, Karen Macours, and all the participants in the Namur doctoral workshop for their valuable comments and suggestions. The authors are particularly grateful to Catherine Guirkinger for her comments, guidance, and financial support. The data collection of this paper benefited from financial support from the European Research Council under the H2020 Research and Innovation Program/ERC grant agreement 759294, granted to Catherine Guirkinger. We are also grateful to Coline Broka and Anna Jolivet for their useful thoughts and assistance with data collection. Our thoughts are with Martin Sossou, who helped us with so much kindness. Pablo Álvarez-Aragón acknowledges financial support from an FNRS/FRESH doctoral fellowship. The study has been pre-registered in the AEA RCT Registry: AEARCTR-0011696.

1 Introduction

The collection of reliable survey data is essential for the design of economic and social policies. Especially in low-income countries, where public information is often scarce or non-existent, surveys are indispensable tools. The growth in the use and collection of survey data has sparked interest in how such surveys are designed and implemented, as well as in understanding the consequences of different choices of survey methodology, leading to a burgeoning literature highlighting the effects of varying the survey methodology (Beegle et al., 2012; De Weerdt et al., 2016). However, relatively little work has focused on finding ways to avoid errors at the data collection stage, and most of the intellectual effort has been devoted to developing econometric techniques to reduce bias while using the resulting error-prone data (De Weerdt et al., 2020). Nevertheless, attempting to reduce measurement error at the data collection and survey design stages is crucial, especially when the survey deals with sensitive issues (e.g., social, cultural, and gender norms). In particular, some studies have focused on the role of the enumerator in interviews, investigating the influence of enumerators on the respondents' answers (Brunton-Smith et al., 2017; Rodriguez-Segura and Schueler, 2023). This literature has shown that the type of question matters and that bias is more prominent when asking for opinion or sensitive questions (Di Maio and Fiala, 2020; West and Blom, 2017). Another stream of research has compared the performance of alternative survey methods compared to face-to-face interviews, such as phone surveys (Abate et al., 2023) or self-administered data collection procedures, particularly in relation to the disclosure of sensitive issues such as domestic violence (Peterman et al., 2023; Cullen, 2023). When focusing on alternative methods or enumerator effects, very little work has examined these issues together. In addition, little evidence has focused on the role of both survey method and enumerator influence specifically on the measurement of social and gender norms.

In this paper, we attempt to fill this gap in the literature by focusing on the data collection stage and comparing two different survey designs. In particular, we test an alternative survey design to the standard Face-to-Face Survey (hereafter FFS), which we call the In-group Individual Survey (hereafter IIS), in which respondents are grouped together and self-answer questions posed by a common enumerator directly on a tablet provided to them (privately). To this end, we administer the exact same questionnaire to 847 women in rural Benin randomizing the survey method. Specifically, 526 women will answer the questionnaire in groups, while 321 women will answer in standard face-to-face interviews. Based on approaches developed in the literature to identify the extent of enumerator bias (e.g., Himelein, 2016; Laajaj and Macours, 2019; Di Maio and Fiala, 2020; Rodriguez-Segura and Schueler, 2023), we first classify variables according to whether they are likely to be influenced by the enumerator or not. To do this, we quantify the degree of enumerator influence on respondents' answers by looking at the R-squared of a regression of each outcome on enumerator fixed effects. We show that when enumerator influence is high, there is systematic divergence in responses across survey methods. We also show that variables that are highly influenced by the enumerator, and therefore highly affected by the survey method, are associated to behaviors and beliefs related to social norms. Most interestingly, our results document that in-group respondents are more likely to report conservative opinions about gender norms. Moreover, these differences are substantial. For example, when we examine differences on variables that are likely to be influenced by external factors (i.e., those variables that are classified as highly influenced by the enumerator according to our classification procedure), we find that women in the IIS are 36 percentage points less likely to say that they decide how to spend their own money, 38 percentage points more likely to think that a husband's agricultural activities are slowed down by the development of his wife's individual economic activities, or 27 percentage points more likely to think that housework is a woman's job.

Among the possible mechanisms, we emphasize the role of social desirability bias as the main driver of these differences. In particular, we hypothesize that the influence of the enumerator is higher in the FFS and that our effects are driven by respondents' social desirability bias. For example, when asked for information on wealth items or social background, FFS respondents answer in the direction of the most socially desirable options. Furthermore, we show that a sensitive optional question on domestic violence is much more likely to be answered in the FFS than in the IIS, even though in-group surveys are supposed to be more private. This suggests that enumerator influence is much more important in the context of the FFS, consistent with a social desirability bias. These results uncover potential trade-offs. For instance, IIS make easier to isolate respondents from other household members, and likely reduces the influence of the enumerator. On the other hand, we show that attrition rates significantly differ across survey methods because respondents are certainly harder to reach in IIS.

We contribute to several strands of the literature. First, this paper adds to the thin but growing literature on studies investigating the role of the survey method on the quality of the data collected. In a recent review, De Weerdt et al. (2020) document that the effect sizes of varying the survey method are very important.[1] In particular, when studying gender-related outcomes such as domestic violence, formal administrative and household data are known to be lower-bound estimates of their prevalence (Sardinha et al., 2022). To address this concern, alternative survey methods have been explored to reduce misreporting in the collection of sensitive data and information such as crime (Blattman et al., 2016), sexual and reproductive behavior (Lépine et al., 2020; Chuang et al., 2021), and intimate partner violence (IPV) (Bulte and Lensink, 2019; Agüero and Frisancho, 2022). Sim-

[1] Using an experimental framework in Tanzania, De Weerdt et al. (2016) find that the prevalence of hunger ranges from 19% to 59% depending on the consumption module used in the questionnaire. In the same framework, Beegle et al. (2012) and Gazeaud (2020) show that differences in survey methods matter when reporting consumption expenditures and when targeting poor households, respectively.

ilarly, in Nigeria, Cullen (2023) finds that women's reported experience of IPV is 35 percent higher when measured using an indirect list method that anonymizes respondents' responses compared to standard FFS. Using an experimental design in Senegal, and closely related to our setting, Peterman et al. (2023) find that audio computer-assisted self-interview (ACASI) led to a 4 to 7 percentage point increase in IPV compared to standard face-to-face interviews. In this paper, we contribute to the burgeoning literature on the study of alternative survey methods that offer greater confidentiality to respondents. We extend this by testing an alternative survey method - IIS - and by providing new evidence on social norm measures. In particular, we show that the survey method matters when studying gender-related values and norms, and not just self-reported behaviors.

Second, we also contribute to those studies that measure the influence of the enumerator and the role of social desirability bias in surveys. In the literature, social desirability bias has been identified as a major source of misreporting (Blair et al., 2020). Krumpal (2013) divides interviewer effects into two categories: effects related to the existence of variation in interviewer characteristics (e.g., gender and socioeconomic status), and effects due to assumed interviewer expectations about social desirability bias. In this sense, some papers have explicitly focused on measuring the influence and importance of enumerator characteristics on different outcomes within the same survey. For example, Di Maio and Fiala (2020) find that while the enumerator effect is small for many questions, it can account for over 30 percent of the variation in subjective outcomes such as political preferences. In another recent paper, Rodriguez-Segura and Schueler (2023) show that enumerator effects can be large enough to produce some spurious results in impact evaluations. Moreover, they show that enumerator effects may be particularly important when enumerators are assigned to clusters of respondents (or, more generally, clusters of the unit of observation), because it becomes harder for enumerator assignment to happen orthogonally to treatment assignment. Using a household survey in Timor-Leste, Himelein (2016) documents that enumerator influence is higher for subjective questions than for objective questions. In line with this literature, we show that the degree of enumerator influence varies depending on the type of outcome. Based on this degree of influence, we provide the evidence that the same questions can lead to highly significant differences depending on the survey method. To do so, we rely on respondents' willingness to answer and show that social desirability bias differs across survey methods.

The remainder of the paper is organized as follows. In Section 2 we present the experimental design of our study, the data, and discuss attrition. Then, in Section 3 we present our main results, while in Section 4 we discuss the main mechanisms. Section 5 concludes.

2 Experimental Design and Data

2.1 Experimental design

Our experiment was conducted in southern Benin in collaboration with the Belgian Development Agency (ENABEL). We took advantage of an agricultural intervention designed as a randomized controlled trial (RCT) on 1,009 households applying for an agricultural subsidy. For this intervention, a baseline survey was conducted between March 2020 and June 2020, where all interviews were conducted using a standard face-to-face survey method. After randomization, 673 women were assigned to the intervention, while 336 women formed the control group. The program consisted of a combination of group business training and a subsidy to start and/or expand pineapple production. Treated women were asked to attend 7 sessions of group business training before receiving their subsidies.[2] At the beginning of the first session and prior to the start of the intervention, each participant was asked to complete a short survey on an individual tablet. This self-administered questionnaire was supervised by an enumerator specially trained for this task.[3] The implementation of the individual in-group survey started in December 2021 and lasted until January 2022.[4] For the control group, standard FFS were conducted in January 2022 by a professional team of six enumerators.[5] Therefore, in our setting, being in the treatment group means responding in groups, while being in the control group means responding to a standard face-to-face survey. Because this survey was administered prior to the start of the intervention, responses are not influenced by the training itself. Figure C1 in Appendix C summarizes our experimental design.

In-group Individual Survey - IIS.

Gathered in groups, respondents answered on their tablets without any direct assistance or interaction. However, to avoid handling problems with the tablets, we specifically trained eight enumerators to follow a rigorous protocol for this experiment.Each question was also assigned a color and a number so that the enumerators could check that all respondents were on track. Before reading the question, the enumerator must check the number and color of the question on each tablet. Then, for each question, the enumerator was asked to read several times the question and their answers verbally. As many of the participants were illiterate, we also added pictograms and illustrations to represent the different choices and options (see Appendix G for illustrations). As a particular concern

[2] Groups were designed according to the participants' district location, and people gathered in rooms specifically designated for training. The choice of location was made by the enumerators, and participants' travel costs were covered by ENABEL. The participants were divided into 107 groups to attend the sessions.

[3] For the same group, the enumerator is also responsible for delivering the subsequent business training.

[4] Some make-up sessions were conducted in June 2022 for women who could not attend the regular sessions.

[5] In this case, respondents were interviewed individually in their village or home, and enumerators were instructed to try to avoid the presence of other people (e.g., husbands) close to the respondent during the interview.

in this setting is the interaction between respondents, a minimum distance between participants was requested. We trained enumerators to emphasise that responses must be private, and they were explicitly trained to respect the privacy of responses and to avoid touching respondents' tablets as much as possible, except for technical issues. In this respect, the in-group individual survey can be thought of as a guided, self-administered interview.

2.2 Summary statistics

The surveys we conducted can be divided into three different modules. First, we collected information on the demographic characteristics of the respondents. Second, we collected information on the socioeconomic characteristics of the respondents. Finally, since the original RCT aims to study gender-related outcomes, we also have information on variables related to women's empowerment, intrahousehold decision making, or domestic violence.

Table A1 in Appendix A summarizes our main variables and examines the balance between treatment and control. It shows the means in the control and treatment groups and the control-treatment difference for variables collected during the baseline survey of 1009 women. There are no significant differences between the treatment and control groups, neither in demographic variables nor in socioeconomic and gender-related outcomes, as expected from computerized randomization. In the survey method experiment, we base our analyses on a set of questions related to demographic characteristics and social norms to assess the difference between the two survey methods. All retained outcomes are presented in Appendix B.

2.3 Attrition

A first aspect of interest concerns the different attrition rates between the two survey methods. In the FFS method, enumerators visited respondents in their homes/villages, whereas in the IIS, respondents were asked to join the group somewhere in their district location. It is therefore worth exploring here whether the survey design may influence attrition. To compare attrition by survey method and to examine the characteristics of attritors, we estimate the following equation:

$$A_i = \beta_0 + \beta_1 T_i + X_i' \Phi + X_i' T_i \theta + \epsilon_i \tag{1}$$

Where A_i denotes whether respondent i has participated in the survey method experiment, T_i equals one for respondents in IIS, and X_i' is a vector of individual-level characteristics.

Table D1 in Appendix D shows that the probability of attrition increases by 17 percentage points in IIS. While attrition is 4.4% in face-to-face interviews, it rises to 21.8% in individual in-group surveys. Importantly, only two characteristics differ between attritors in FFS and those in IIS. Attrition is higher in IIS if the respondent does not have a cell phone and if the respondent answers that she can buy furniture with her own money. Moreover, we also show that respondent characteristics remain balanced at baseline once we remove attritors (Table D2 in Appendix D). Therefore, these results suggest that the different survey process is the main driver of attrition in our setting.

This analysis of the influence of the survey method on attrition yields two main results. First, it is clear that the survey method matters significantly for attrition rates, potentially invalidating an experiment, and that there is a trade-off between low attrition rates and the economic cost of contacting participants. Second, we argue that in our study, attrition does not seem to be an irremediable problem, since the characteristics of attritors do not differ systematically across survey methods, and the characteristics of non-attritors respondents are balanced at baseline. Therefore, we assume nonresponse as random throughout the paper.

2.4 Enumerator influence and outcome classification

Data collection through the completion of a questionnaire involves a social interaction between the respondent and his or her environment. Any factor that affects this interaction can potentially affect the quality of the data collected (Di Maio and Fiala, 2020). One example that has been studied in the literature is the effect of the enumerator's behavior and characteristics on the respondent's answers (West and Blom, 2017; Di Maio and Fiala, 2020).Himelein (2016) shows that the enumerators matter more for questions related to sensitive topics or subjective variables. [6] For example, it is reasonable to argue that the influence of the enumerator's observable characteristics is greater in questions related to preferences than in questions related to demographics questions (e.g., age or number of children). Therefore, if the interaction between respondents and their environment varies depending on the survey method (e.g., because of the enumerator, or because of third parties), we should expect differences in the answers.

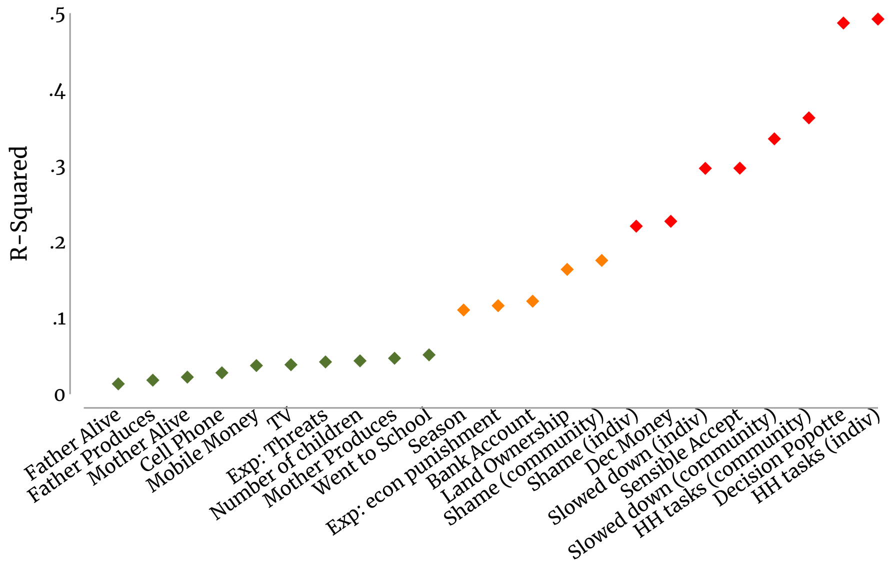
To explore this question, we build on the approach developed in previous papers (e.g., Himelein, 2016; Laajaj and Macours, 2019; Di Maio and Fiala, 2020) and determine whether a variable is likely to be influenced by external factors based on the importance of the enumerator effect. This method consists of examining the explanatory power of enumerators by looking at the R^2 of a regression

[6] However, this may not be the only factor. For example, the presence of other people at the time of the questionnaire can condition the respondent's answers, and this influence can vary greatly depending on the nature of the question (Rasinski et al., 1994, 1999).

of an outcome variable on enumerator fixed effects.[7] Thus, a high R^2 is interpreted as enumerators picking up a large amount of the variation in responses to the question related to that outcome variable, and a low R^2 is interpreted as enumerators having little influence on respondent responses. For example, in the case of Di Maio and Fiala (2020), the R^2 is small for demographic variables such as age, gender, or marital status, but it becomes large when examining political questions, suggesting that responses to political questions may be biased by enumerator characteristics rather than reflecting the true opinions of respondents. This is a straightforward way to identify which variables are likely to contain responses that are influenced by the presence of an enumerator, which may be particularly relevant in the context of face-to-face surveys. Figure 1 shows our results in ascending order. As expected, we find that there is considerable variation in the explanatory power of enumerators. First, the R^2 takes very low values (< 0.1, green color) for variables related to social background or items-owning (e.g. whether the respondent's father/mother is alive or produces pineapples, or the number of children). Second, R^2 takes intermediate values (between 0.1 and 0.2, orange color) for variables such as land ownership, bank account ownership, whether the respondent did not participate in a decision-making process because she was afraid of being punished, or whether the respondent was insulted for making a decision. Finally, R^2 can also take very high values (from 0.2 to 0.5, red color) for variables mostly related to gender norms, intrahousehold decisions (e.g. whether the respondent thinks that housework is a woman's job, whether the respondent thinks that the husband's activities are slowed down by the wife's activities).

[7] All of the retained outcomes and their labels are described in Appendix B.

Figure 1: Enumerator Effect: R^2 of different outcomes



Note. Sample: All women included. Variable Definitions: "*Father (Mother) Alive*" is a dummy variable that equals one if the respondent's father (mother) is alive. "*Father (Mother) Produces*" is a dummy variable that equals one if the respondent's father (mother) has produced pineapples. "*Cell Phone*", "*TV*", "*Bank Account*", "*Land Ownership*", "*Mobile Money*" are dummy variables that take the value of one if the respondent owns a cell phone, a TV, a bank account, land, or a mobile money account, respectively. "*Number of children*" is the total number of children at the time of the survey. "*Went to School*" is an indicator that equals one if the respondent attended primary school. "*Exp: Threats*" is a dummy variable that equals one if the respondent was threatened for making a decision (experienced threats). "*Season*" equals one if the respondent thinks that the last agricultural season was either good or excellent in terms of yields. "*Exp: econ punishment*" equals one if the respondent was punished (economically) for participating in household decisions (experienced punishement). "*Dec money*" is a dummy variable that equals one if the respondent alone decides how to spend her own money. "*Shame (indiv)*" and "*Shame (community)*" are dummy variables that equal one if the respondent thinks (or thinks that the community thinks) that a man feels ashamed when his wife brings home more money than he does. "*Slowed down (indiv)*" and "*Slowed down (community)*" equal one if the respondent thinks (or thinks that the community thinks) that the husband's agricultural activities are slowed down by the development of his wife's individual economic activities. "*HH tasks (indiv)*" and "*HH tasks (community)*" are dummy variables that equal one if the respondent thinks (or thinks that the community thinks) that housework is a woman's job. "*Sensible Accept*" equals one if the respondent is willing to answer a very sensitive question related to domestic violence. Finally, "*Decision Popotte*" equals one if the respondent is the one who decides how to spend the money put in common for food.

Following these results, we expect the choice of survey method to be particularly relevant for those variables with a high R^2, since we believe that they are more likely to be influenced by external factors, such as the presence of the enumerator. In the next Section 3, we use this classification to present our outcomes into different categories depending on whether they have a low, medium, or high R^2 in the regression on enumerator fixed effects.

3 Results

This section presents the main results. As explained above, we use the classification about enumerator influence to investigate the effect of the survey method on the answers. Since we are testing many hypothesis, we always report several checks at the bottom of the tables to correct for multiple hypothesis testing, including Anderson (2008)'s FDR sharpened q-values, List et al. (2016)'s familywise error rate (FWER) p-values, or Bonferroni (1936)'s correction.[8] We show univariate regressions with robust standard errors in parenthesis.[9]

Low influence variables. We first examine the effect of the survey method on those variables that are less likely to be influenced by external factors. Results presented in Table 1 show that variables are not systematically affected by the survey method. However, there are more differences than would be expected by pure chance. In particular, respondents who respond in IIS are 6.4 (4.8) percentage points (pp) more likely to say that her father (mother) was a pineapple producer when she was young, 10.7 pp less likely to report owning a TV, and 7.2 pp less likely to report having attended primary school. Interestingly, these are variables related to social status (education or, in our context, where all respondents are related to the pineapple sector, whether the respondent's parents produced pineapples), or wealth (TV), and therefore answers may be sensitive to social desirability bias. We will explore this issue further in the next section.

Table 1: Survey Method and low influence variables (I)

	The dependent variable is				
	Father Produces	Father Alive	Mother Alive	Cell Phone	Mobile Money
In-Group Survey	0.0639**	0.0416	-0.0249	0.0250	-0.0498
	(0.0264)	(0.0334)	(0.0339)	(0.0575)	(0.0343)
Mean Y for the FFS group	0.143	0.318	0.654	1.570	0.645
Relative effect	44.61	–	–	–	–
R-squared	0.00643	0.00180	0.000634	0.000220	0.00246
N	847	847	847	847	847
Sharpened q-value	0.025	0.137	0.228	0.285	0.097
FWER p-value	0.8103	0.1683	0.8337	0.8967	0.652
Bonferroni adjustment	0.344	1	1	1	1

	The dependent variable is				
	TV	Exp: threats	Number of children	Mother produces	Went to school
In-Group Survey	-0.107***	0.00684	0.251*	0.0478**	-0.0722**
	(0.0339)	(0.0273)	(0.128)	(0.0217)	(0.0352)
Mean Y for the FFS group	0.402	0.178	4.336	0.0872	0.483
% of FFS mean	-26.67	–	5.788	54.75	-14.96
R-squared	0.0121	0.0000740	0.00448	0.00520	0.00499
N	847	847	847	847	847
Sharpened q-value	0.004	0.343	0.049	0.038	0.043
FWER p-value	0.0103	0.7983	0.362	0.256	0.3817
Bonferroni adjustment	0.035	1	1	0.618	0.895

NOTE. Variable Definitions: "*Father Produces*" is a dummy variable that equals one if the respondent's father has produced pineapples. "*Father (Mother) Alive*" is a dummy variable that equals one if the respondent's father (mother) is alive. "*Cell Phone*" and "*Mobile Money*" are dummy variables that take the value of one if the respondent owns a cell phone, or a mobile money account, respectively. "*TV*" is a dummy variable that takes the value of one if the respondent owns a TV. "*Exp: Threats*" is a dummy variable that equals one if the respondent was threatened for making a decision (experienced threats). "*Number of children*" is the total number of children at the time of the survey. "*Mother Produces*" is a dummy variable that equals one if the respondent's mother has produced pineapples. "*Went to School*" is an indicator that equals one if the respondent attended primary school. Robust standard errors in parenthesis. *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

[8] The consideration of multiple hypothesis testing is important in this context. Indeed, under the null hypothesis and independent outcomes, testing one by one leads to a probability of false rejection of 68% when using a critical value of 0.05 $[(1-(1-0.05)^{22})]$, or of 90% when using a critical value of 0.1 $[(1-(1-0.1)^{22})]$.

[9] When available in baseline, we add baseline outcomes as controls. Results in Appendix F show that the point estimates barely move.

Medium influence variables. We then turn to examine the effect of the survey method on variables that are somewhat likely to be influenced by external factors (intermediate R^2 in our classification). For this group of variables, no systematic differences arise.[10]

Table 2: Survey Method and medium influence variables

	(1) Good Season	(2) Exp: econ punishment	(3) Bank account	(4) Land owner	(5) Shame (comm.)
In-Group Survey	0.0309 (0.0334)	0.0327 (0.0334)	-0.0495 (0.0316)	0.0833^{**} (0.0353)	-0.0351 (0.0341)
Mean Y for the FFS group	0.657	0.321	0.293	0.498	0.651
% of FFS mean	–	–	–	16.71	–
R-squared	0.00103	0.00112	0.00298	0.00660	0.00124
N	847	847	847	847	847
Sharpened q-value	0.202	0.196	0.092	0.027	0.191
FWER p-value	0.8197	0.8423	0.632	0.184	0.8747
Bonferroni adjustment	1	1	1	0.405	1

NOTE. Variable Definitions: "*Season*" equals one if the respondent thinks that the last agricultural season was either good or excellent in terms of yields. "*Exp: econ punishment*" equals one if the respondent was punished (economically) for participating in household decisions (experienced punishment). "*Bank Account*", "*Land Ownership*" are dummy variables that take the value of one if the respondent owns a bank account, or land, respectively. "*Shame (Comm.)*" is a dummy variable that equals one if the respondent thinks that the community thinks that a man feels ashamed when his wife brings home more money than he does. Robust standard errors in parenthesis. *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

High influence variables. Finally, we examine the effect of the survey method on questions that are highly influenced by external factors. All of these questions relate to gender norms or are sensitive in nature. For all variables we find striking and significant differences.

Table 3 shows that women who answer in IIS are 15.4 percentage points more likely to say that it is disrespectful to the husband if the wife brings home more money than he does (column 1), 35.9 percentage points less likely to say that they decide how to spend their own money (column 2), or 37.6 percentage points more likely to think that the husband's agricultural activities are slowed down by the development of his wife's individual economic activities (column 3). We also find that women are 22.3 percentage points more likely to think that the community (in this case, the pineapple producers) thinks that that the husband's agricultural activities are slowed down by the development of his wife's individual economic activities (column 4), 26.5 (14.2) percentage points more likely to think (to think that the community thinks) that housework is a woman's job (columns 7 and 5, respectively), or 52.1 percentage points less likely to say that she is the one who decides alone how to spend the money put in common for food (column 6). These effects are quantitatively very important, representing in some cases more than 50% of the control mean, and suggesting that the survey method has a very strong influence on respondents' answers.

[10] The effect on the probability of owning land that is non-significant under some corrections for multiple hypothesis testing.

Table 3: Survey Method and high influence variables

	(1) Shame Ind	(2) Dec Money	(3) Slow down Ind	(4) Slow down Com	(5) Tasks Com	(6) Dec Popotte	(7) Tasks Ind
In-Group Survey	0.154*** (0.0351)	-0.359*** (0.0330)	0.376*** (0.0272)	0.223*** (0.0333)	0.142*** (0.0306)	-0.521*** (0.0265)	0.265*** (0.0309)
Mean Y for the FFS group	0.439	0.579	0.0935	0.280	0.688	0.946	0.607
% of FFS mean	35.04	-61.94	402.5	79.69	20.67	-55.07	43.65
R-squared	0.0224	0.132	0.151	0.0483	0.0275	0.268	0.0940
N	847	847	847	847	847	750	847
Sharp q-val	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Sharp q-val	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Bonferroni	0.0003	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000

NOTE. Variable Definitions: "*Shame (indiv)*" is a dummy variable that equals one if the respondent thinks that a man feels ashamed when his wife brings home more money than he does. "*Dec Money*" is a dummy variable that equals one if the respondent alone decides how to spend her own money. "*Slow down Ind*" equals one if the respondent thinks that the husband's agricultural activities are slowed down by the development of his wife's individual economic activities. "*Slow down community*" equals one if the respondent thinks that the community thinks that the husband's agricultural activities are slowed down by the development of his wife's individual economic activities. "*Tasks (indiv)*" and "*Tasks (community)*" are dummy variables that equal one if the respondent thinks (or thinks that the community thinks) that housework is a woman's job. "*Dec Popotte*" equals one if the respondent is the one who decides how to spend the money put in common for food. Robust standard errors in parenthesis. *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

Overall, these results point to one main conclusion. We find that the survey method matters for questions that are easily influenced by external factors, including the interviewer. The most important and striking differences between face-to-face interviews and individual in-group surveys appear for questions with a high probability of being influenced. Interestingly, these questions mainly relate to issues such as gender norms, domestic violence, intrahousehold decisions, or variables potentially associated with social stigma. In particular, we observe that women responding in groups consistently elicit less gender-equitable responses. These findings highlight the importance of considering the survey method used when interpreting experimental results, especially when questions relate to gender or social norms, and draw attention to the pervasiveness of underreporting.

4 Mechanisms

We find that respondents answering in groups systematically elicit less gender-equal responses. In this section, we discuss some potential mechanisms that may help us better understand these findings.

4.1 Social desirability bias

One possible explanation for our findings relates to the different types of enumerators' interaction and influence across survey methods. In FFS, respondents have to reveal their answers and preferences to the enumerator, while they keep their answers to themselves when responding in groups.

This can be particularly important in the case of sensitive or opinion questions. For example, it has been shown that domestic violence outcomes from household surveys (usually conducted in a face-to-face manner) are lower-bound estimates of the true values (Sardinha et al., 2022; Peterman et al., 2023). Similarly, Rasinski et al. (1994) show that women are less likely to tell the truth when the questions are administered by an interviewer than when they are self-administered, due to the reduced privacy of face-to-face interviews and the perceived risk of embarrassment from the interviewer's reaction. In our setting, the differences in enumerator interaction between survey methods may lead to the existence of differential enumerator effects. For example, respondents may perceive enumerators as advocates of gender-equal responses, as our enumerators are highly educated (all attended university), mostly from urban areas (63%), and more supportive of gender-equal attitudes.[11] In FFS, respondents may be more likely to report more gender-equal beliefs because enumerators have the most direct influence on their responses. In other words, FFS could lead to greater response bias for questions about opinions and beliefs.

For some items and social background questions, we can also find an effect of the survey methods on the responses. In Table 1, we show that respondents who respond to individual in-group surveys are about 10 pp less likely to report owning a TV and 7 pp less likely to report having attended primary school.[12] In the Benin context, TV ownership is clearly perceived as a signal of wealth. In our sample, 40.2% of respondents in the FFS method report owning a TV, while 29.5% in the IIS.[13] For education, reporting having attended school is also a component of social background, and being educated may be perceived as socially desirable. In the FFS group, 48.4% of respondents declared having attended school, compared to 41.1% in the IIS.[14] These results suggest that different methods lead to changes, and that these changes are large enough to be highly unlikely to be the result of random measurement error.[15]

To shed more light on this mechanism related to enumerator influence and social desirability bias, we introduced an optional question where we asked the respondent whether she agrees to be asked a very sensitive question about her experience of domestic violence (without knowing the exact question).[16] Results are presented in Table 4. We find that respondents in IIS are 43 percentage

[11] This information comes from self-administered and anonymous surveys of enumerators. All enumerators agree that women should try to develop their own business out of their households, and 38% do not think it is better for a woman (as opposed to a man) to do the housework and childcare, compared to 22% in our sample of women. In total, there were 14 different enumerators, 8 for IIS and 6 for FFS. The assignment of enumerators was based on the geographical distribution of respondents, and could not be randomized.

[12] In our baseline survey, as previously shown in Table A1 in Appendix, there is no significant difference on these characteristics (the survey method was standard face-to-face interviews for both groups).

[13] For comparison, in the 2018 Benin Demographic and Health Survey (BDHS), 28.7% of households own a TV.

[14] For comparison, 44.85% of women declared having attended school in the 2018 BDHS.

[15] For education, since all of our enumerators attended university, the differences we find here cannot be attributed to enumerator bias in the usual sense - that is, differences in enumerator characteristics. Instead, we hypothesize that respondents changed their answers when faced an educated enumerator because of their interviewer's presumed expectations.

[16] See the variables "agree to answer" and "domestic violence" in the Appendix B.

points (46% reduction compared to the control group) less likely to agree to answer the sensitive question. At first glance, these results may seem surprising, as it is assumed that individual group interviews are more private than face-to-face interviews and we might therefore expect higher response rates. However, by contrast, this result shows that in face-to-face interviews, respondents are less likely to refuse to answer a sensitive question. This result is consistent with the existence of social desirability bias, where respondents act in accordance with their perception of the enumerator's expectations. Because of the higher interaction in FFS, respondents are more likely to not refuse to answer and to behave as they think they should. This may be due to pressure from the enumerator who is waiting for an answer and/or because the respondent feels bad about refusing to answer because she perceives the interview as the enumerator's job.[17] Interestingly, when we compare the self-selected samples on domestic violence, respondents in the IIS are 13 percentage points more likely to report domestic violence. This result suggests that the in-group method is well perceived as respecting privacy when asking sensitive questions and that IIS respondents feel comfortable answering that they have experienced domestic violence, ruling out the possibility that IIS leads to more intrusive and less private surveys.

Table 4: Two-step question on domestic violence

	(1)	(2)
	Agree to answer	Domestic violence
In-Group Survey	-0.433***	0.129***
	(0.0255)	(0.0285)
Mean Y for the FFS group	0.941	0.0695
% of FFS mean	-46.05	185.5
R-squared	0.200	0.0366
N	847	569

NOTE. Outcome definitions: "Agree to answer" is a dummy variable that equals one if the respondent answers "yes" to the following question: *"Would you be willing to answer a question about domestic violence in the home?"*. "Domestic violence" is a dummy variable that equals one if the respondent answers "yes" to the following question: *"In the past year, has anyone hit you at home to the extent that it has prevented you from working?"*. Robust standard errors are reported in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Taken together, these findings suggest that the influence of the enumerator may be important not only for highly sensitive questions such as reporting domestic violence, but also for general questions related to gender norms or social background and some wealth-related items. This could hinder our understanding of existing levels and trends in gender (in)equality if we rely solely on face-to-face interviews.

[17] We do not think that these results can be explained by free-riding behavior in the IIS. For example, it could be that people agree to answer more when they are face-to-face because they are curious about the question, but free-ride in groups because they will hear the question if at least one person in the group agrees to answer (since the enumerator reads each question aloud). However, since our groups are not very large on average, there is a significant probability that no one will answer, and all members of the group decide at the same time whether or not to answer without interacting with each other.

4.2 Patterns of Answers

Another potential driver of our estimates may be related to specific response patterns associated with tablet handling issues for IIS respondents. Unlike face-to-face interviews, where the enumerator reads the question and enters the respondent's answer into a tablet, IIS respondents have to record their answers themselves. Although their enumerators were specially trained to reduce concerns about handling the tablets, we introduced some pictograms to represent the different choices and to help respondents (see Figure G2 in Appendix G).[18]While these illustrations help respondents, they can also distort their responses. In particular, in the case of Likert scales, respondents may be led to select the extreme responses ("No, not at all" and "Yes, absolutely") because of their bigger size relative to the intermediate statements ("Rather no" and "Rather yes"). Because they are more salient, extreme responses may be more likely to be selected in the individual in-group surveys. However, we have redefined all Likert scaled variables with dummies, meaning that we consider an extreme or moderate statement to be equivalent, and therefore our results cannot be affected by such concerns.

Another potential channel that could explain our findings may be related to the effort and concentration required to complete the survey. In a group setting, respondents may be distracted by others and less involved in the survey process than in face-to-face interviews, or may not understand the functioning of tablets. To control for such respondent behavior, we test for say-yes bias, which is the probability of mechanically answering "yes" to a binary choice, which is a proxy of fatigue. Table G1 in Appendix G that these mechanisms cannot explain our results.

5 Conclusion

In recent decades, the development of data analysis and the central role of quantitative work in the implementation of public policies has been observed. The most common method is the standard face-to-face survey, requiring a direct exchange of information between the respondent and the enumerator. However, in face-to-face surveys, enumerators can influence respondents by shaping how they report or explain their answers (e.g., in Laajaj and Macours, 2019; Di Maio and Fiala, 2020; Rodriguez-Segura and Schueler, 2023). Despite all the recent efforts devoted to data collection, little evidence documents the impact of the survey method itself on data quality.

In this paper, we attempt to fill this gap by taking advantage of an RCT in Benin where two different survey methods were randomly assigned to women involved in the pineapple production sector.

[18] In Section 2.1 we explained in detail the rigorous protocol followed by the enumerators to deal with these issues.

We compare the standard face-to-face survey to an alternative method we call in-group individual survey, in which each respondent answers privately in a tablet during a group session. By examining the variation in responses, we show that there are systematically significant differences in outcomes over which enumerators have influence. Interestingly, these outcomes measure beliefs about social norms, such as gender norms. Indeed, in-group respondents reported less equal gender values than face-to-face respondents. Relying on respondents' willingness to answer an optional question on domestic violence, we state that face-to-face surveys are most likely to social desirability bias driving estimates.

Our findings support the idea that survey methods matter, especially when measuring values and beliefs and when enumerators may influence respondents. This suggests that scholars and institutions should consider the importance of the survey process when collecting data, but also when using data that has already been collected. In this sense, measures of social norms should be used and interpreted with caution, especially when used to assess the impact of public policies.

References

Abate, G. T., De Brauw, A., Hirvonen, K., and Wolle, A. (2023). Measuring consumption over the phone: Evidence from a survey experiment in urban ethiopia. *Journal of Development Economics*, 161:103026. 2

Agüero, J. M. and Frisancho, V. (2022). Measuring violence against women with experimental methods. *Economic Development and Cultural Change*, 70(4):1565–1590. 3

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495. 9

Beegle, K., De Weerdt, J., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, 98(1):3–18. 2, 3

Blair, G., Coppock, A., and Moor, M. (2020). When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments. *American Political Science Review*, 114(4):1297–1315. 4

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. 3

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62. 10

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(2):551–568. 2

Bulte, E. and Lensink, R. (2019). Women's empowerment and domestic abuse: Experimental evidence from vietnam. *European economic review*, 115:172–191. 3

Chuang, E., Dupas, P., Huillery, E., and Seban, J. (2021). Sex, lies, and measurement: Consistency tests for indirect response survey methods. *Journal of Development Economics*, 148:102582. 3

Cullen, C. (2023). Method matters: The underreporting of intimate partner violence. *World Bank Economic Review*, 37(1):49–73. 2, 4

De Weerdt, J., Beegle, K., Friedman, J., and Gibson, J. (2016). The challenge of measuring hunger through survey. *Economic Development and Cultural Change*, 64(4):727–758. 2, 3

De Weerdt, J., Gibson, J., and Beegle, K. (2020). What can we learn from experimenting with survey methods? *Annual Review of Resource Economics*, 12(1):431–447. 2, 3

Di Maio, M. and Fiala, N. (2020). Be wary of those who ask: A randomized experiment on the size and determinants of the enumerator effect. *World Bank Economic Review*, 34(3):654–669. 2, 4, 7, 8, 15

Gazeaud, J. (2020). Proxy means testing vulnerability to measurement errors? *The Journal of Development Studies*, 56(11):2113–2133. 3

Himelein, K. (2016). Interviewer Effects in Subjective Survey Questions: Evidence from Timor-Leste. *International Journal of Public Opinion Research*, 28(4):511–33. 2, 4, 7

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual Quant*, 47:2025–2047. 4

Laajaj, R. and Macours, K. (2019). Measuring skills in developing countries. *Journal of Human Resources*. 2, 7, 15

Lépine, A., Treibich, C., and d'Exelle, B. (2020). Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours. *Social Science & Medicine*, 266:113326. 3

List, J., Shaikh, A., and Xu, Y. (2016). Multiple Hypothesis Testing in Experimental Economics. Artefactual Field Experiments 00402, The Field Experiments Website. 9

Peterman, A., Dione, M., Le Port, A., Briaux, J., Lamesse, F., and Hidrobo, M. (2023). Disclosure of violence against women and girls in senegal. *IFPRI Discussion Paper 2195. Washington, DC: International Food Policy Research Institute (IFPRI)*. 2, 4, 13

Rasinski, K., Baldwin, A., Willis, G., and Jobe, J. (1994). Risk and loss perceptions associated with survey reporting of sensitive topics. *National Opinion Research Center (NORC), Chicago*, page 497–502. 7, 13

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W., and Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology*, 13(5):465–484. 7

Rodriguez-Segura, D. and Schueler, B. E. (2023). Assessors influence results: Evidence on enumerator effects and educational impact evaluations. *Journal of Development Economics*, 163:103057. 2, 4, 15

Sardinha, L., Maheu-Giroux, M., Stöckl, H., Meyer, S. R., and García-Moreno, C. (2022). Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *The Lancet*, 399(10327):803–813. 3, 13

West, B. T. and Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of survey statistics and methodology*, 5(2):175–211. 2, 7

Appendix A Summary statistics and balance

Table A1: Baseline characteristics

	(1)		(2)		T-test
	Control		Treatment		Difference
	N	Mean/SE	N	Mean/SE	(1)-(2)
Number children	336	4.244	673	4.141	0.103
		(0.126)		(0.090)	
Age	334	37.802	668	37.163	0.639
		(0.568)		(0.382)	
Father alive	336	0.360	672	0.385	-0.025
		(0.026)		(0.019)	
Mother alive	330	0.700	667	0.715	-0.015
		(0.025)		(0.017)	
Bank account	336	0.167	673	0.132	0.034
		(0.020)		(0.013)	
Phone	336	0.667	673	0.678	-0.011
		(0.026)		(0.018)	
Attended school	336	0.399	673	0.407	-0.008
		(0.027)		(0.019)	
Owns TV	336	0.375	673	0.348	0.027
		(0.026)		(0.018)	
Father produces	334	0.171	665	0.159	0.011
		(0.021)		(0.014)	
Pineapple not women	336	0.170	673	0.192	-0.022
		(0.021)		(0.015)	
Pineapple respect	336	0.836	673	0.856	-0.020
		(0.020)		(0.014)	
Buy: furniture	336	0.747	673	0.750	-0.003
		(0.024)		(0.017)	
Buy: motorbike	336	0.741	673	0.713	0.028
		(0.024)		(0.017)	

NOTE. Sample: 1009 female respondents from baseline survey. The value displayed for t-tests are the differences in the means across the groups. "Pineapple not for woman" equals one if the respondent says that producing pineapple is not a women's activity. "Pineapple respect" equals one if the respondent says that producing pineapple might increase respect towards women. "Empowerment: Buy furniture" equals one if the respondent says that she can buy furniture with her own money if she wants. "Empowerment: Buy motorbike" equals one if the respondent says that she can buy a motorbike with her own money if she wants. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Appendix B Codebook and variables

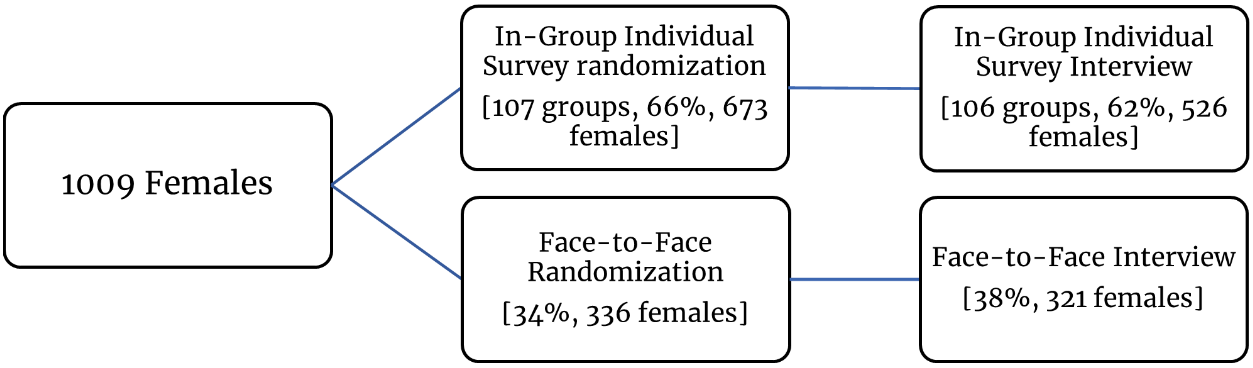
Table B1: Codebook and variables

Name	Label	Type
Father Alive	Is your father alive?	Binary
Mother Alive	Is your mother alive?	Binary
Father Produces	When you were younger, was your father a pineapple grower?	Binary
Mother produces	when you were younger, was your mother a pineapple grower?	Binary
Number of children	How many children have you had in your life (including those who have died)?	Numeric
Went to school	Did you attend primary school?	Binary
TV	Do you have a TV at home?	Binary
Cell phone	Do you have a cell phone ?	Binary
Mobile Money	Have you ever used mobile money on your own phone?	Binary
Good Season	What do you think of last season in terms of yields?	Likert Scale
Bank account	Do you have one or more bank accounts?	Binary
Land owner	Are you owner of an agricultural field?	Binary
Exp: Econ punishment	Have you already been punished economically for taking part in household decisions?	Binary
Exp: Threats	Have you already been physically threatened because for taking part in household decisions?	Binary
Decision Money	Who usually makes decisions about the money I make?	Decision Making
Decision Popotte	Who usually makes decisions about the Popotte money ? (household's kitty)	Decision Making
Shame (Ind.)	In my opinion, a man is ashamed when his wife brings in more money than him.	Likert Scale
Shame (Comm.)	In other pineapple-producing households, a man is ashamed when his wife brings in more money than him.	Likert Scale
Slowed down (Ind.)	In my opinion, the husband's agricultural activities are slowed down by those of the wife.	Likert Scale
Slowed down (Comm.)	In other pineapple-producing households, the husband's agricultural activities are slowed down by those of the wife.	Likert Scale
HH Tasks (Ind)	In my opinion, it is better for a family if a woman has the main responsibility for cooking and other household chores, rather than a man.	Likert Scale
HH Tasks (Comm.)	In other pineapple-producing households, it is better for a family if a woman has the main responsibility for cooking and other household chores, rather than a man.	Likert Scale
Agree to answer	Would you be willing to answer a question about domestic violence in the home?	Binary
Domestic violence	In the past year, has anyone hit you at home to the extent that it has prevented you from working?	Binary

NOTE. The "Binary" type corresponds to a binary choice between "Yes" or "No". The "Likert Scale" refers to a gradation: "Not at all"; "Rather no"; "Rather yes"; "Yes, absolutely". The "Decision making" type presents four choices: "Me"; "My husband"; "Both"; "Another person".

Appendix C Experimental design

Figure C1: Experimental Design


1009 Females
In-Group Individual Survey randomization
[107 groups, 66%, 673 females]
Face-to-Face Randomization
[34%, 336 females]
In-Group Individual Survey Interview
[106 groups, 62%, 526 females]
Face-to-Face Interview
[38%, 321 females]

Appendix D Attrition

Table D1: Survey method and attrition

	(1) P(Attrition)	(2) P(Attrition)
In-group Survey (T)	0.174***	0.400***
	(0.0195)	(0.116)
Number of children		0.00934
		(0.00864)
Number of children x T		-0.00616
		(0.0120)
Age		0.000878
		(0.00146)
Age x T		-0.00287
		(0.00244)
Father Alive		0.00285
		(0.0245)
Father Alive x T		-0.00608
		(0.0443)
Bank account		-0.0262
		(0.0207)
Bank account x T		0.0544
		(0.0554)
Mobile Phone		-0.0264
		(0.0286)
Mobile Phone x T		-0.0964**
		(0.0472)
Attended School		0.00231
		(0.0201)
Attended School x T		0.0318
		(0.0413)
Father produces pineapple		-0.0117
		(0.0282)
Father produces pineapple x T		-0.0726
		(0.0490)
Pinneaple not for women		-0.0123
		(0.0302)
Pinneaple not for women x T		-0.0448
		(0.0489)
Pineapple increases respect		-0.0626
		(0.0425)
Pineapple increases respect x T		0.0318
		(0.0640)
Buy furniture		0.0468**
		(0.0182)
Buy furniture x T		-0.0499
		(0.0508)
Buy motorbike		0.0141
		(0.0205)
Buy motorbike x T		-0.0188
		(0.0488)
Mean Y for the FFS group	0.0446	–
% of FFS mean	389.3	–
R-squared	0.0498	0.0813
N	1009	991

NOTE. Robust standard errors are reported in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

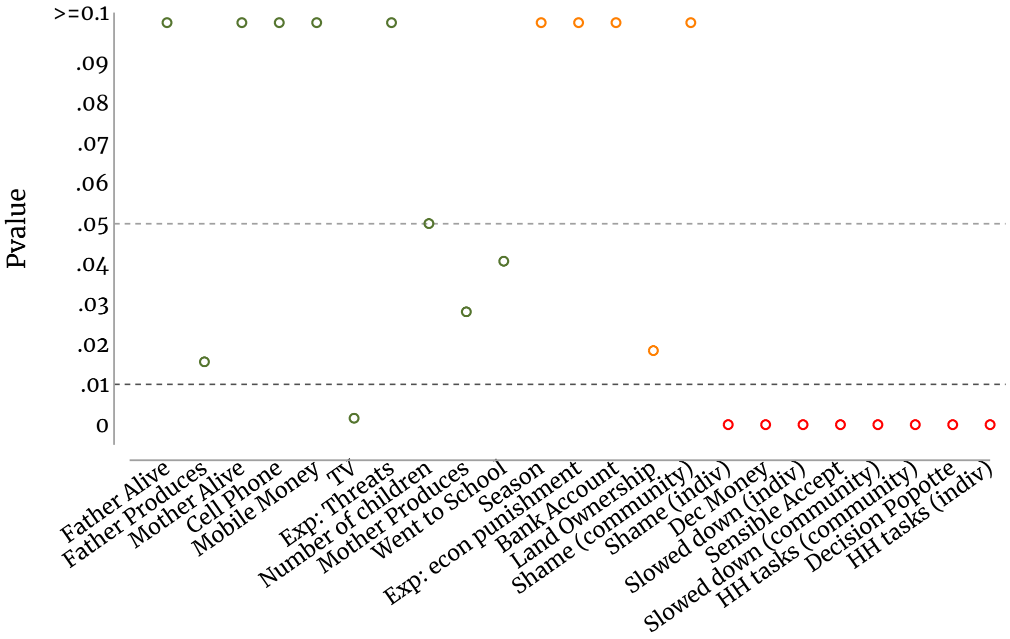
Table D2: Respondent's characteristics after attrition

	(1)		(2)		T-test
	Control		Treatment		Difference
	N	Mean	N	Mean	(1)-(2)
Number children	321	4.190	526	4.167	0.023
		(0.124)		(0.099)	
Age	319	37.655	522	37.479	0.176
		(0.583)		(0.435)	
Father alive	321	0.361	526	0.380	-0.019
		(0.027)		(0.021)	
Mother alive	315	0.705	522	0.716	-0.012
		(0.026)		(0.020)	
Bank account	321	0.171	526	0.133	0.038
		(0.021)		(0.015)	
Phone	321	0.673	526	0.711	-0.038
		(0.026)		(0.020)	
Attended school	321	0.405	526	0.395	0.010
		(0.027)		(0.021)	
Owns TV	321	0.374	526	0.354	0.020
		(0.027)		(0.021)	
Father produces	319	0.172	518	0.174	-0.001
		(0.021)		(0.017)	
Pineapple not women	321	0.171	526	0.203	-0.032
		(0.021)		(0.018)	
Pineapple respect	321	0.844	526	0.861	-0.017
		(0.020)		(0.015)	
Buy: furniture	321	0.738	526	0.753	-0.015
		(0.025)		(0.019)	
Buy: motorbike	321	0.735	526	0.715	0.020
		(0.025)		(0.020)	

NOTE. Sample: respondents from baseline survey that also appear in the endline survey after attrition. The value displayed for t-tests are the differences in the means across the groups. "Pineapple not for woman" equals one if the respondent says that producing pineapple is not a women's activity. "Pineapple respect" equals one if the respondent says that producing pineapple might increase respect towards women. "Empowerment: Buy furniture" equals one if the respondent says that she can buy furniture with her own money if she wants. "Empowerment: Buy motorbike" equals one if the respondent says that she can buy a motorbike with her own money if she wants. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Appendix E Summary: Results

Figure E1: P-values of in-group survey for different outcomes



Note: This figure reports the p-values of answering in groups for different outcomes. The colours represent whether the outcome belongs to the category of low influence (green), medium influence (orange), or high influence (red). Robust standard errors are used and outcome variables at baseline are always included as controls when available.

Appendix F Robustness to baseline controls

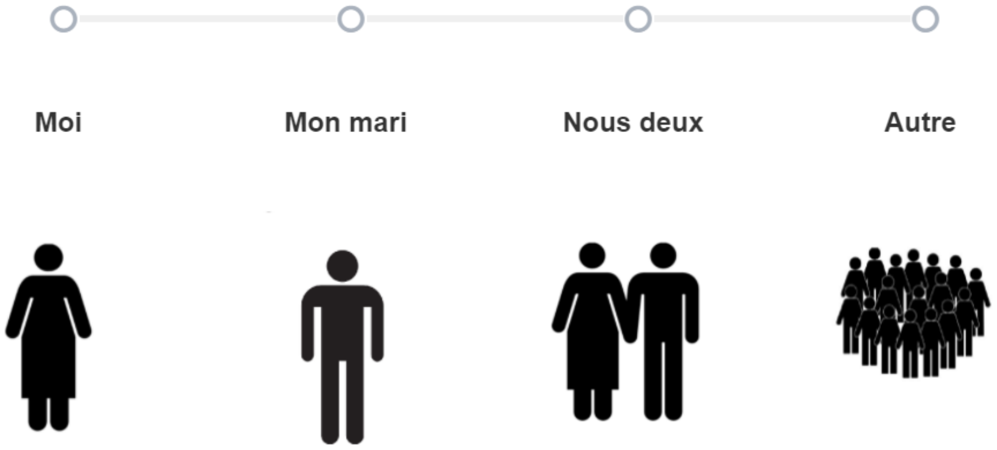
Table F1: Variables with baseline controls

	(1) F prod	(2) F alive	(3) M alive	(4) Phone	(5) MoMo	(6) TV	(7) Children	(8) M prod	(9) School	(10) Bank
In-Group Survey	0.0658***	0.0271	-0.0402*	0.0454	-0.0331	-0.0988***	0.263***	0.0527**	-0.0658**	-0.0333
	(0.0253)	(0.0197)	(0.0211)	(0.0545)	(0.0342)	(0.0306)	(0.0912)	(0.0205)	(0.0258)	(0.0294)
Mean Y FFS group	0.144	0.318	0.657	1.570	0.684	0.402	4.336	0.0878	0.483	0.293
Relative effect	45.63		-6.115			-24.59	6.066	60.02	-13.63	
R-squared	0.103	0.615	0.582	0.0908	0.159	0.190	0.431	0.0981	0.441	0.119
N	837	847	837	847	697	847	847	843	847	847

NOTE. Variable Definitions: "*F (M) Produces*" is a dummy variable that equals one if the respondent's father (mother) has produced pineapples. "*F (M) Alive*" is a dummy variable that equals one if the respondent's father (mother) is alive. "*Phone*" and "*MoMo*" are dummy variables that take the value of one if the respondent owns a cell phone, or a mobile money account, respectively. "*TV*" is a dummy variable that takes the value of one if the respondent owns a TV. "*Children*" is the total number of children at the time of the survey. "*School*" is an indicator that equals one if the respondent attended primary school. "*Bank*" is a dummy variable that takes the value of one if the respondent owns a bank account. Robust standard errors in parenthesis. *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

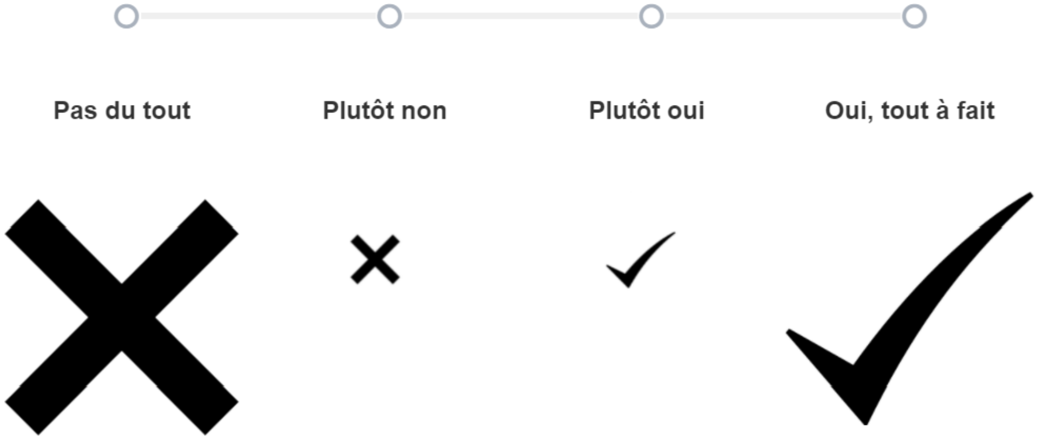
Appendix G Patterns in the answers

Figure G1: Example of available answers (I)



Note: This figure shows an example of how respondents in IIS had to answer the questions about decision making at home.

Figure G2: Example of available answers (II)



Note: This figure shows an example of how respondents in IIS had to answer the questions. It presents a Likert scale of four items: "Pas du tout" can be translated as "No, not at all", "Plutôt non" can be translated as "Rather no", "Plutôt oui" can be translated as "Rather yes", and "Oui, tout à fait" can be translated as "Yes, absolutely".

Probability to select an extreme answer and say-yes bias: Table G1 reports estimates of the probability of selecting an extreme answer in general (column 1), and the probability of selecting positive (negative) extreme responses in column 2 (3). We show that while the probability of choosing an extreme response is 15 percentage points higher for IIS respondents, they are also less likely to choose an extreme negative statement. These estimates suggest that the IIS method does not encourage the selection of extreme answers, even when respondents are unfamiliar with electronic devices (in the absence of treatment effects, both extreme positive and extreme negative answers should have a similar coefficient).

Finally, column 4 shows that there is no say-yes bias between the two survey methods. In other words, in-group respondents answered "yes" at a similar rate as face-to-face respondents, and there is no sign of systematic survey fatigue. This result allows us to rule out that our results are due to differences in fatigue and understanding how tablets work.

Table G1: Patterns of answers and survey method

	(1) P(Extreme Answer)	(2) P(Extreme Yes)	(3) P(Extreme No)	(4) Say-Yes Bias
In-Group Survey	0.148^{***}	0.204^{***}	-0.0555^{***}	-0.000212
	(0.0174)	(0.0163)	(0.0148)	(0.0115)
Mean Y	0.660	0.374	0.286	0.460
R-squared	0.0760	0.146	0.0148	0.00284
N	847	847	847	847

NOTE. To compute the probability of extreme answers (Columns 1-3), we rely on the answers to the 7 Likert-scaled variables used in our analyses. Here, an 'Extreme Answer' refers to a choice at the extremity of the Likert scale : "Not at all" or "Yes, absolutely". In Column (4), we compute the probability to answer "Yes" for all the binary-type questions asked in the survey. All used variables and labels are shown in Table B1. Robust standard errors are reported in parenthesis. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.