# Measuring Norms and Enumerator Effects: Survey Method Matters\*

Pablo Álvarez-Aragón<sup>†</sup> and Hugues Champeaux<sup>‡</sup>

<sup>+</sup>*University of Namur* <sup>‡</sup>*University of Cagliari* 

#### Abstract

While the process of data collection can lead to bias, little empirical evidence investigates the role of the survey method. In this paper, we compare two survey methods: the standard face-to-face interview and an alternative method we call Human-Assisted Self-Administered survey (HASA). In the latter, respondents are guided by an enumerator reading questions, but they answer privately on an electronic device. Taking advantage of an RCT in Benin, we randomize the survey method across respondents. We show that the survey method leads to different results depending on enumerator influence. Identifying this influence, we document that variables that are likely to be influenced by enumerators differ systematically across survey methods. Interestingly, these variables are mainly related to gender norms and women agency. We find that respondents who answer directly on a tablet report less gender-equal values. Investigating the mechanisms, we show that social desirability bias affects responses in face-to-face interviews.

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

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

<sup>\*</sup>P. Álvarez-Aragón is affiliated to the University of Namur. Email: pablo.alvarez@unamur.be. H. Champeaux is affiliated to the University of Cagliari, Italy. Email: hugues.champeaux@unica.it. We are grateful to Jean-Marie Baland, Jules Gazeaud, Karen Macours, Giulia Mancini, William Parienté and all the seminar and conference participants at Namur, Sassari, Cagliari, DIAL, ICDE, LAGV, and CERDI for their valuable comments and suggestions. The authors are particularly grateful to Catherine Guirkinger for her comments, guidance, and financial support. 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. This work was supported by the H2020 Research and Innovation Program, European Research Council (ERC) grant agreement 759294 to C. Guirkinger; and the Fonds National de la Recherche Scientifique (FNRS) / Fonds pour la Recherche en Sciences Humaines (FRESH) [1.F.001.23F]. The study has been registered in the AEA RCT Registry: AEARCTR-0011696.

### 1. Introduction

Over the past 30 years, tremendous efforts have been made to make data widely available worldwide.<sup>1</sup> The growth in the use and collection of survey data has sparked interest in how such surveys are designed and implemented, and in understanding the consequences of different survey methodologies (Beegle et al., 2012; De Weerdt et al., 2016). In particular, this literature has highlighted the tendency to misreport sensitive information in a face-to-face encounter with an enumerator, which is still the most common setting for surveys conducted in most parts of the world (Tourangeau and Yan, 2007).

To reduce the bias of misreporting, various strategies have been implemented and studies have compared the performance of alternative survey methods (e.g., list experiments or self-administered surveys) with the standard face-to-face survey (Lépine et al., 2020; Chuang et al., 2021; Cullen, 2023; Peterman et al., 2024). These proposed alternative methods allow for better control of enumerator influence in the data collection process by limiting interactions or the amount of information respondents provide directly to enumerators. Comparing data collected using different methods reveals whether and how enumerators influence self-reported outcomes. In particular, enumerators influence respondents' answers specially when they are asked about subjective or sensitive items (Di Maio and Fiala, 2020) that are likely to be subject to social desirability bias (Blair et al., 2020). Questions related to the measurement of gender and social norms or women's agency are thus likely to be affected by these influences. However, despite the growing interest in gender issues in recent years, we know little about the biases in the measurement of these outcomes and the mechanisms that drive differences in respondents' responses across survey methods.

In this paper, we compare a new alternative survey design, which we call the Human-Assisted Self-Administered Survey (hereafter HASA), to the standard Face-to-Face Survey (hereafter FFS), which relies on direct questioning and where enumerators are privy to respondents' answers. In this new method, which is expected to reduce the interaction between the respondent and the enumerator, respondents individually self-report their answers to a question asked by an enumerator on a tablet lent to them. We ad-

<sup>&</sup>lt;sup>1</sup>For example, statistics on the extent of poverty in 1991 were based on 22 surveys from 22 countries (Ravallion et al., 1991); by 2004, 454 surveys from 97 countries were being used (Chen and Ravallion, 2004); and by 2010, these numbers had grown to 675 surveys from 115 countries (Chen and Ravallion, 2010).

minister the exact same questionnaire to 847 women in rural Benin, randomizing the survey method. 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)), 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. Then, we study whether responses vary by survey method and the mechanisms driving these differences.

We find that when enumerator influence is high there is a systematic divergence in the responses across survey methods. Interestingly, variables that are highly influenced by the enumerator are related to social and gender norms, and measures of women's agency. Our results document that HASA respondents, who answer in a setting where there is less direct interaction with the enumerator, consistently report more conservative opinions about gender norms, attitudes, and experiences. Moreover, these differences are substantial. For example, we find that women in the HASA 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.

We highlight the role of interviewer effects due to participants' expectations of what interviewers think is socially desirable as the main driver of these differences and propose an innovative way to elicit this bias in surveys. By asking respondents whether they want to answer a sensitive optional question, we show that respondents in HASA are (much) less likely to agree to answer than respondents in FFS (even though HASA surveys are supposed to be more private). We argue that this result is consistent with the existence of a higher degree of social desirability bias in FFS. Because of the greater interaction in FFS, respondents are more likely to not refuse to answer and behave as they think they should. Importantly, we show that our results are not driven by differences in enumerator characteristics, but rather by the survey method design. This finding reveals that enumerators influence respondents' responses beyond the influence of their characteristics (e.g., gender or prior experience) and thus highlights the limitations of standard remedies to mitigate enumerator bias, such as using enumerator fixed effects of their random assignment across treatment arms. When focusing on the influence of alternative methods or enumerator bias, very little work has examined these issues together. In addition, little evidence has focused on the role of both survey method and enumerator effects specifically on the measurement of social and gender norms or women's agency.

Our study focuses on Benin, which has a high level of gender inequality. According to the Gender Inequality Index (a composite measure of gender inequality in three dimensions: reproductive health, empowerment, and the labor market), Benin ranked 161/167 in 2022 and 142/156 in 2010, always above the sub-Saharan African average.<sup>2</sup> Gender norms are also very conservative. For example, DHS data show that 27% believe that beating a woman is justified if she neglects the children. However, gender norms are changing rapidly and mask important heterogeneity. While the proportion of women who thought a beating was justified if the wife neglected the children was 52% in 2001, it fell to 22% in 2017. In urban areas, this figure was 40% in 2000, while in rural areas it was 60%. Similarly, there is a pronounced education gradient: while the proportion of uneducated women who agree that a beating is justified if the wife neglects the children is 32%, it drops to 3% among women with higher education. These stark differences raise the question of whether respondents from rural and uneducated backgrounds adjust their responses to questions about gender or social norms when faced with a highly educated, urban enumerator who is likely to hold less conservative gender attitudes.

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.<sup>3</sup> 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

<sup>&</sup>lt;sup>2</sup>Similar patterns emerge when we look at the Gender Development Index, which measures gender disparities in outcomes across three basic dimensions of human development: female and male life expectancy at birth, female and male years of schooling, and control over economic resources.

<sup>&</sup>lt;sup>3</sup>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.

(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). For instance, Cullen (2023) finds that women's reported experience of IPV in Nigeria 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. (2024) 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. Using a phone survey experiment in Guatemala, Agüero et al. (2023) study how women's agency varies with the degree of privacy when responding. In their study, they compare women who answer verbally to those who answer using a code (either verbally or via the phone keypad). Interestingly, they find that women who respond with greater privacy report lower levels of agency. In this paper, we contribute to the burgeoning literature on the study of alternative survey methods that offer greater confidentiality to respondents. Compared to most of the literature, which focus on the measurement of specific very sensitive modules, we focus on measuring broader social and gender norms, as well as dimensions of women's agency. In contrast to existing methods, the HASA is more flexible (ACASI is mostly limited to yes/no questions), and highly cost and time efficient (one single enumerator can interview several people at the same time and same place). In particular, we show that the survey method matters when studying gender-related values, norms, and practices, and not just when discussing very sensitive topics such as intimate partner violence or sexual behaviors, in line with Agüero et al. (2023).

Second, we also contribute to those studies that measure the influence of the enumerator and the role of social desirability bias in surveys. 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 interviewers' expectations about social desirability bias (the respondent guesses the interviewer's internalized norms and opinions and adjust his answers accordingly). In this sense, most 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. Moreover, they show that enumerators' characteristics such as gender, whether the enumerator is from an urban area, or past experience matter for respondents' answers. In another recent paper, Rodriguez-Segura and Schueler (2023) show that enumerator effects (e.g., enumerator experience) can be large enough to produce some spurious results in impact evaluations. They also show that enumerator characteristics such as gender, whether the enumerator is from an urban area, or prior experience matter for respondent responses. Moreover, they show that enumerator effects can be particularly important when enumerators are assigned to clusters of respondents (or, more generally, clusters of the unit of observation), because it becomes more difficult for enumerator assignment to be orthogonal to treatment assignment. Finally, using a household survey in Timor-Leste, Himelein (2016) document that enumerator influence is higher for subjective questions than for objective questions, and show that respondent characteristics explain most of the variation. We contribute to the literature on interviewer effects by examining how responses are affected by the survey method itself, holding enumerator characteristics constant. In this paper, we hypothesize that, because the interaction between the respondent and the interviewer is much more intense in FFS than in HASA surveys, the effects on survey responses of assumed interviewers' expectations of what is socially desirable are larger. These effects are not eliminated by standard solutions to reduce enumerator bias (such as including enumerator fixed effects or randomizing enumerator assignment) since they are fundamentally due to the nature of the survey method itself.

Moreover, social desirability bias has been identified as a major source of misreporting (Blair et al., 2020), and scholars usually measure it by exploiting the Marlowe Crowne scale, an index built from a series of items asked during the survey (Dhar et al., 2022; Amaral et al., 2024). However, the use of the Marlowe Crowne scale is subject to several concerns. First, the social desirability scale was developed in the 1960s in the United States, which is far removed from the context of a developing country in the 2020s. Second, recent research questions the validity of the Marlowe Crowne scale, arguing that it may also capture a personality trait or behavior rather than a response bias in a survey (Lanz et al., 2022). By asking a very simple question about the respondent's willingness to answer a sensitive question, we add to this literature by providing another simple and cost-efficient way to detect such bias in developing countries.

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 (that took place at most every two weeks) of group business training before receiving their subsidies.<sup>4</sup>

During the first session of the business training, and prior to any treatment related to gender or social norms, each participant was asked to complete a short survey on an individual tablet. Therefore, in our setting, being in the treatment group means responding using the human-assisted self-administered method, while being in the control group means responding to a standard face-to-face survey. The random assignment to FFS or HASA is the same assignment as to the business training that is part of the larger RCT. However, we do not expect the training program to affect the answers in our survey. The data collection occurred during the first session of the business training, after only a short introduction to the training to meet participants, where the concept of "planning" was explained and respondents got familiar with the tablets. The main remaining concern here would be associated with the potential effects of having expectations associated

<sup>&</sup>lt;sup>4</sup>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.

with assignment to treatment. However, as we argue in Section 3, these expectations are likely to work *against* our main results. This self-administered questionnaire was supervised by an enumerator specially trained for this task.<sup>5</sup> The implementation of the HASA survey started in December 2021 and lasted until January 2022.<sup>6</sup> For the control group, standard FFS were conducted in January 2022 by a professional team of six enumerators.<sup>7</sup> As they had to be trained for two different survey processes, enumerators were not randomly assigned between treatment and control groups.<sup>8</sup> Figure A1 in Appendix A summarizes our experimental design.

### Human-Assisted Self-Administered Survey - HASA.

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 enumerators could check that all respondents were on the same question. Before reading the question, the enumerator had to check the number and color of the question on each tablet. Then, for each question, the enumerator was asked to read the question and its different response options several times. As many of the participants were illiterate, we also added pictograms and illustrations to represent the different choices and options (see Appendix B for illustrations). For convenience due to the subsequent business training and to reduce costs, respondents were gathered in groups. As a particular concern in this setting is the interaction between respondents, a minimum distance between participants was requested. We trained enumerators to emphasize that responses must be private, as well as to respect the privacy of responses and to avoid touching respondents' tablets as much as possible, except for technical issues. In this respect, the human-assisted self-administered survey can be thought of as a guided, self-administered interview. The HASA method was specifically adapted to survey people living in rural areas where many different languages are spo-

<sup>&</sup>lt;sup>5</sup>For the same group, the enumerator is also responsible for delivering the subsequent business training.

<sup>&</sup>lt;sup>6</sup>Some make-up sessions were conducted in June 2022 for women who could not attend the regular sessions.

<sup>&</sup>lt;sup>7</sup>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.

<sup>&</sup>lt;sup>8</sup>To take into account differences across enumerators, we collected information on them before the surveys to control for their characteristics (see Section 4.3).

ken and where people may feel uncomfortable using tablets for the first time, with the goal of both respecting the privacy of responses and ensuring that respondents were able to follow the survey.

### 2.2. Summary statistics

The surveys we conducted can be divided into three different modules. First, we collected information on basic demographic characteristics of the respondents (age, number of children, education or parents' health). Second, we collected information on the socioeconomic characteristics of the respondents (proxies for wealth, source of income). Finally, since the original RCT aims to study gender-related outcomes, we also have information on variables related to women's empowerment, gender norms, intrahousehold decision making, or domestic violence.

Table C1 in Appendix C.1 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 characteristics nor in 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 outcomes and their descriptions are presented in Appendix C.2.

### 2.3. Attrition

A first aspect of interest is the difference in attrition rates between the two survey methods. In the FFS method, enumerators visited respondents in their homes/villages, whereas in the HASA, respondents were asked to join the group somewhere in their district location. It is therefore worth exploring here whether this specific survey design may influence attrition.<sup>9</sup> To compare attrition by survey method and to examine the characteristics of attritors, we estimate the following equation:

<sup>&</sup>lt;sup>9</sup>In our setting, respondents to the HASA surveys were grouped together. This is not a necessary condition for conducting a HASA survey, but it is very cost effective and was very convenient for us, as the HASA respondents were going to attend a business training later on.

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

Where  $A_i$  denotes whether respondent *i* has participated in the survey method experiment,  $T_i$  equals one for respondents in HASA, 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 HASA. While attrition is 4.4% in face-to-face interviews, it rises to 21.8% in HASA surveys. Importantly, only two characteristics differ between FFS and HASA attritors. As expected, attrition is higher in HASA if the respondent does not have a cell phone, because respondents in HASA had to be contacted to tell them when and where the group interviews would take place (as opposed to FFS, where enumerators went to respondents' homes).

Although the characteristics of attritors do not seem to differ systematically across survey methods, we implement three robustness checks to address concerns related to selection bias due to differential attrition. First, we show that respondents' characteristics remain balanced at baseline once we remove attritors (Table 1). Second, we rely on the sensitivity bounds approach developed by Kling and Liebman (2004). We reproduce our main results (Table 4) estimating treatment effects bounds by imputing to the attrited observations the mean outcome of their respective treatment arm  $\pm$  0.10 and  $\pm$  0.25 standard deviations.<sup>10</sup> Finally, we also follow Lee (2009)'s procedure addressing selection bias due to differential attrition by equalizing selection rates across treatment groups. The method trims observations from the group with higher retention, in our case, the control group. The trimming procedure is non-random and targets respondents with the most extreme values in the outcome distributions, either the highest (producing lower bounds) or the lowest (producing upper bounds). The results of lower and upper bounds in Appendix D, Table D2 suggest that attrition is unlikely to drive our findings.

The analysis of the influence of the survey method on attrition yields two results. First, it is clear that the survey method matters significantly for attrition rates, potentially invalidating an experiment. In our setting, there is a trade-off between low attrition rates

<sup>&</sup>lt;sup>10</sup>This latter assumption of  $\pm$  0.25 standard deviations is considered as quite extreme in the literature (Baird et al., 2019; Blattman et al., 2020; Özler et al., 2021; Brune et al., 2022; Gazeaud et al., 2023).

and the economic cost of contacting participants. Although HASA surveys are more cost-efficient than FFS because they are conducted in groups, the fact that meetings take place distant from respondents' villages makes their attendance more uncertain. Finally, 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 still balanced at baseline. Therefore, we assume nonresponse as random throughout the paper.

	(1)			(2)	T-test
	C	ontrol	Tre	atment	Difference
	Ν	Mean	Ν	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)	

# Table 1: Respondent's characteristics after attrition

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.

### 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.<sup>11</sup> For example, it is reasonable to argue that the enumerator's influence on respondents' answers is greater in questions about preferences, norms, or values than in questions about demographic characteristics (e.g., age or number of children). Therefore, if the interaction between respondents and enumerators varies across survey methods (e.g., because some methods require respondents to disclose their answers to the enumerator while others do not), we would expect to see greater differences across survey methods in responses to "subjective" questions, which are more likely to be influenced by enumerators.

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 the enumerator. This method consists of examining the explanatory power of enumerators by looking at the  $R^2$  of a regression of an outcome variable on enumerator fixed effects.<sup>12</sup> 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.<sup>13</sup> 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.<sup>14</sup>

First, the  $R^2$  takes very low values (< 0.1, green color) for variables related to social

<sup>&</sup>lt;sup>11</sup>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).

<sup>&</sup>lt;sup>12</sup>All outcomes and their labels are described in Appendix C.2.

<sup>&</sup>lt;sup>13</sup>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.

<sup>&</sup>lt;sup>14</sup>Appendix E replicates the same figure using only the sample of respondents answering in a face-toface format. We show that the ordering of variables by  $R^2$  magnitude is exactly the same.

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).

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 or cell phone, a TV, a bank account, and, 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: con punishment" equals one if the respondent was punished (economically) for participating in household decisions (experienced punishment). "Dec money" is a dummy variable that equals one if the respondent altended community thinks) that a man feels ashamed when his wife brings home more money than he does. "Slawed down (indiv)" and "Slawed down (community)" equal one if the respondent thinks (or thinks that the community thinks) that a man feels ashamed of his wife's individual economic activities. "HH tasks (indiv)" and "HH tasks (community)" are dummy variables that equals one if the respondent thinks (or thinks that the community that equals one if the respondent thinks (or thinks that the community that equals one if the respondent thinks (or thinks that the community thinks) that equal one if the respondent thinks (or thinks that the community thinks) that equal one if the respondent thinks (or thinks that the community thinks) that equal one if the respondent thinks (or thinks that the community thinks) that the quevelopment of his wife's individual economic activities. "HH tasks (ind

Our main contribution is twofold. First, we show that the patterns showed by Di Maio and Fiala (2020) replicate when looking at gender and social norms rather than at political questions, as both are "subjective" topics that may be even sensitive in some contexts. Second, we use this simple statistical procedure to hypothesize when survey methods may particularly influence responses. In particular, 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.<sup>15</sup> We show univariate regressions with robust standard errors in parenthesis.

**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 2 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 HASA 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.<sup>16</sup>

<sup>&</sup>lt;sup>15</sup>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  $[(1 - (1 - 0.1)^{22})]$ .

<sup>&</sup>lt;sup>16</sup>Although the effect on the number of children is slightly significant, compared to our other dependent variables, the magnitude of the effect is very low (less than 6%).

	The dependent variable is							
	Father Produces	Father Alive	Mother Alive	Cell Phone	Mobile Money			
HASA survey	0.0639** (0.0264)	0.0416 (0.0334)	-0.0249 (0.0339)	0.0250 (0.0575)	-0.0498 (0.0343)			
Mean Y FFS group	0.143	0.318	0.654	1.570	0.645			
% of FFS mean	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			
HASA survey	-0.107***	0.00684	0.251*	0.0478**	-0.0722**			
2	(0.0339)	(0.0273)	(0.128)	(0.0217)	(0.0352)			
Mean Y 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			

### Table 2: Survey Method and low influence variables

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 owns 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 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.

**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.<sup>17</sup>

<sup>&</sup>lt;sup>17</sup>The effect on the probability of owning land is non-significant under some corrections for multiple hypothesis testing.

	(1) Good Season	(2) Exp: econ punishment	(3) Bank account	(4) Land owner	(5) Shame (comm.)
HASA survey	0.0309 (0.0334)	0.0327 (0.0334)	-0.0495 (0.0316)	0.0833** (0.0353)	-0.0351 (0.0341)
Mean Y 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

Table 3: Survey Method and medium influence variables

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, dimensions of women's agency, or are sensitive in nature. For all variables we find striking and significant differences.

Table 4 shows that women who answer in HASA are 15.4 percentage points more likely to say that a man feels ashamed when his 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 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 suggest that the survey method has a very strong influence

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Shame Ind	Dec Money	Slow down Ind	Slow down Com	Tasks Com	Dec Popotte	Tasks Ind
HASA survey	0.154***	-0.359***	0.376***	0.223***	0.142***	-0.521***	0.265***
	(0.0351)	(0.0330)	(0.0272)	(0.0333)	(0.0306)	(0.0265)	(0.0309)
Mean Y 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
Ν	847	847	847	847	847	750	847
Sharp q-val	0.001	0.001	0.001	0.001	0.001	0.001	0.001
FWER p-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

Table $\Lambda$ : Survey	Method	and high	influence	variables
Tuble 4. Dui vey	method	and mon	minucitee	variabieb

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 are slowed down by the development of his wife's individual 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 human-assisted self-administered surveys appear for questions with a high probability of being influenced. Interestingly, these questions mainly relate to gender-related outcomes such as gender norms, intra-household decisions, or variables potentially associated with social stigma. In particular, we observe that women responding to HASA surveys 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. Interestingly, these findings align with the work of Agüero et al. (2023), who show that women who respond in a more private setting in a phone survey report lower levels of agency.

A potential concern in this setting relates to the validity of randomization. In fact,

<sup>&</sup>lt;sup>18</sup>One potential concern relates to the fact that HASA surveys were conducted in groups, where women engaged simultaneously with the same enumerator, while FFS were conducted at the individual level. Our results do not change when we cluster at the training group level or when reporting bootstrapped standard errors. Since we only have training groups for respondents in HASA surveys, we created similar sized counterfactual groups for respondents in FFS based on their place of residence, which was the main criterion used to construct these groups for the HASA respondents.

respondents in HASA were asked to attend a business training (including a session on gender) as part of a larger RCT, and attendance at the training was required to receive an agricultural subsidy. This setting raises two issues. On the one hand, it is possible that the treatment and control groups differ due to differences in unobservable characteristics due to differential attrition. On the other hand, differential motivation as a result of being selected for the training group could affect our results even though the training had not yet started. However, we think this situation is unlikely for two reasons. First, we do not observe systematic differences between the observable characteristics of those who received the training and those in the control group (see Table 1), even for variables related to gender norms or proxies for women's empowerment. Second, our main results in Table 4 go in the opposite direction from what one would expect if respondents who attend the training are more motivated or if their responses are partially driven by experimenter effects. In fact, even though the training focuses on improving gender equality, women in HASA surveys consistently give less gender-equal responses, which is more consistent with an increase in privacy due to the survey method than with a conscious effort to give responses that please the experimenter. Furthermore, if our results are driven by selection into the training, there is no reason why less gender-equitable women should be more likely than others to take a training focused on women's business development.

### 4. Mechanisms

We find that respondents answering to HASA surveys systematically elicit less genderequal 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 is the different degree of interaction between respondents and enumerators in the different survey methods. In FFS, respondents have to reveal their answers and preferences to the enumerator, while in HASA they keep their answers private. In the former, respondents are more likely to adjust their answers according to the perceptions they have of what interviewers consider socially desirable. Although related, this differs from conventional social desirability bias, in which respondents tailor their answers to conform to dominant cultural norms and present themselves in a favorable light. In our case, social desirability refers more to interviewers' perceived expectations of social desirability bias. Since the enumerators' values and norms may differ from the dominant social values in Benin (due to their social background or higher education), respondents infer the enumerators' internalized norms and expectations and adjust their answers accordingly.

This bias can be particularly important in the case of sensitive or subjective questions (Tourangeau and Yan, 2007). 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; Cullen, 2023; Peterman et al., 2024). 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 selfadministered, 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 also lead to the existence of differential enumerator effects on declared norms. 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.<sup>19</sup> In FFS, respondents may be more likely to report more genderequal beliefs because enumerators interact closely with respondents and therefore have a greater influence on their responses.

For some questions on wealth and social background, we also observe an effect related to the survey method. In Table 2, we show that respondents who participate in HASA surveys are about 10 pp less likely to report owning a TV and 7 pp less likely to report having attended primary school.<sup>20</sup> 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

<sup>&</sup>lt;sup>19</sup>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 HASA and 6 for FFS. The assignment of enumerators was based on the geographical distribution of respondents, and could not be randomized.

<sup>&</sup>lt;sup>20</sup>In our baseline survey, as already shown in table C1 in online appendix C.1, there is no significant difference on these characteristics (the survey method was standard face-to-face interviews for both groups).

report owning a TV, while 29.5% in the HASA.<sup>21</sup> 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 reported having attended school, compared to 41.1% in the HASA.<sup>22 23</sup>

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).<sup>24</sup> Results are presented in Table 5. We find that respondents in HASA are 43 percentage 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 human-assisted self-administered surveys 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.<sup>25</sup> Interestingly, HASA respondents are 13 percentage points more likely to report domestic violence, conditional on their willingness to answer. This finding suggests that the HASA method is well perceived

<sup>&</sup>lt;sup>21</sup>For comparison, in the 2018 Benin Demographic and Health Survey (BDHS), 28.7% of households own a TV.

<sup>&</sup>lt;sup>22</sup>For comparison, 44.85% of women reported having attended school in the 2018 BDHS.

<sup>&</sup>lt;sup>23</sup>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 with an educated enumerator because of their interviewer's presumed expectations.

<sup>&</sup>lt;sup>24</sup>See the variables "agree to answer" and "domestic violence" in the Appendix C.2.

<sup>&</sup>lt;sup>25</sup>We do not think that these results can be explained by free-riding behavior in the HASA. 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 HASA surveys 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.

as respecting privacy when asking sensitive questions and that HASA respondents feel comfortable answering that they have experienced domestic violence, ruling out the possibility that HASA leads to more intrusive and less private surveys.<sup>26</sup>

	(1) Agree to answer	(2) Domestic violence
HASA 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
Ν	847	569

Table 5: Two-step question on domestic violence

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.

### 4.2. Patterns of Answers

Another potential driver of our estimates may be related to specific response patterns associated with tablet handling issues for HASA respondents. Unlike face-to-face interviews, where the enumerator reads the question and enters the respondent's answer into a tablet, HASA 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

<sup>&</sup>lt;sup>26</sup>Note that identification is limited when we look at the prevalence of domestic violence conditional on agreeing to answer, as there may be selection bias. However, if we assume that the prevalence of domestic violence is higher among those who refuse to answer than among those who agree to answer, these results are a lower bound.

(see Figure B2 in Appendix B).<sup>27</sup> 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 human-assisted self-administered 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.

Similarly, 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.

Finally, although we tried to make the survey friendly to respondents, the possibility of misunderstanding questions may still be a concern, especially in HASA where the enumerator is less able to clarify and resolve doubts. In Appendix G, Tables G<sub>2</sub> and G<sub>3</sub>, we show that our main results do not differ by respondents' education level at baseline, or by whether they own or not a mobile phone. If misunderstandings and problems with tablets were widespread, we would expect to see larger differences between survey methods for uneducated women or respondents without a cell-phone. However, we find that the differences in responses are, if anything, slightly larger when respondents are more educated, suggesting that our results are unlikely to be driven by misunderstanding of the questions or difficulties with tablet use.

### 4.3. Enumerators' characteristics

We argue that because the degree of interaction between respondents and enumerators differs considerably between survey methods (being much higher in face-to-face surveys), responses systematically diverge due to significant social desirability bias in FFS.

<sup>&</sup>lt;sup>27</sup>In Section 2.1 we explained in detail the rigorous protocol followed by the enumerators to deal with these issues.

Another concern is that our results may be driven by systematic differences in enumerator characteristics across treatment groups. Although we cannot include enumerator fixed effects because there is no variation in treatment status within enumerators, we can show how our results vary when controlling for a comprehensive set of enumerator characteristics. Table 6 reproduces our main results when controlling for whether the enumerator's place of birth was rural or urban, the enumerator's age and gender, months of experience as an enumerator, and a dummy variable that equals one if the enumerator believes that women rather than men should do household chores. Our results remain largely unchanged when we control for the characteristics of the enumerators. These findings suggest that our results are unlikely to be driven by differences in enumerator characteristics. Moreover, these results suggest that standard solutions to eliminate or reduce enumerator bias (such as including enumerator fixed effects or randomizing enumerator assignment) do not address the bias highlighted in this paper, which is fundamentally due to the nature of the survey method itself. These attempts reduce concerns related to the influence of enumerator characteristics, but are unable to address bias due to survey method-induced variation in responses.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Shame Ind	Dec Money	Slow down Ind	Slow down Com	Tasks Com	Popotte	Tasks Ind
HASA survey	0.139***	-0.407***	0.311***	0.170***	0.134***	-0.606***	0.264***
	(0.0504)	(0.0471)	(0.0357)	(0.0442)	(0.0380)	(0.0333)	(0.0398)
Enum controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Y FFS group	0.439	0.579	0.0935	0.280	0.688	0.946	0.607
% of FFS mean	31.74	-70.27	333.0	60.74	19.47	-64.07	43.51
R-squared	0.0899	0.140	0.226	0.121	0.150	0.415	0.189
Ν	847	847	847	847	847	750	847

Table 6: Survey Method and high influence variables, controlling for enumerator characteristics

Note. Outcome Definitions: "*Shame (indiv)*" is a dummy variable that equals one if the respondent thinks that a man feels ashmed 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 are slowed down by the development of his wife's individual economic activities are slowed down by the development of his wife's individual economic 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. "Enum controls" include: urban/rural place of birth, age, gender, years of experience as enumerator, and a dummy variable that equals one if the enumerator thinks that women rather than men should take care of household chores. Robust standard errors in parenthesis. \*\*\* for p < 0.01, \*\* for p < 0.05, \* for p < 0.1.

### 5. Conclusion

Despite the growing interest in measuring gender-related outcomes in recent decades, particularly those related to gender norms and various dimensions of women's agency, the ways in which survey methods affect survey responses are not well understood. Moreover, although we know the importance of enumerator effects on survey quality, we still have a poor understanding of how enumerator effects vary across survey methods and the mechanisms that drive these differences.

In this paper, we shed some light on these questions by taking advantage of an RCT in Benin that randomly assigned female respondents to two different survey methods. We compare the standard Face-to-Face Survey (FFS) with an alternative method we call Human-Assisted Self-Administered Survey (HASA), in which respondents answer privately on a tablet lent to them, guided by an enumerator who reads the questions and clarifies doubts. We proceed in two steps. First, we classify variables according to whether they are likely to be influenced by the enumerator or not by looking at the R-squared of a regression of each outcome on enumerator fixed effects. We then examine whether responses vary by survey method and the mechanisms driving these differences. We show that while survey method has a limited effect on outcomes over which enumerators have little influence, there are systematically significant and quantitatively very large differences in outcomes over which enumerators have control, mostly related to gender norms and women's agency. In particular, we find that HASA respondents are less likely than FFS influence to report gender-equitable answers.

Relying on respondents' willingness to answer an optional question on domestic violence (acceptance rates are much higher in FFS), we provide suggestive evidence that these results are likely driven by participants' perceptions of what interviewers consider socially desirable. In FFS, where the interaction between enumerator and respondent is much higher than in HASA, respondents infer the interviewer's internalized norms and expectations and adjust their responses accordingly. Our results are consistent with other work showing that women who respond in a more private setting report lower levels of agency (Agüero et al., 2023).

Our findings have additional implications. For example, these results suggest that standard practices to reduce enumerator bias associated with enumerator characteristics

(e.g., randomizing enumerator assignment or including enumerator fixed effects in the regression equation) may not be sufficient because the differences in responses we find are due to differences in survey design. Researchers and institutions should consider the importance of survey design when collecting data, but also when using data that has already been collected.

While we lack a "true benchmark" against which to compare the responses of the different methods, our results suggest that increasing respondent confidentiality can significantly alter population snapshots. Indeed, descriptive statistics or public policy evaluation can be significantly affected by enumerator influence (both enumerator characteristics and perceptions of what interviewers consider socially desirable) when survey design involves direct questioning, raising concerns about how to interpret existing measures of gender-related outcomes. Our HASA method ensures greater confidentiality and enforces stricter privacy for respondents compared to standard direct questioning. This method also has several other advantages. First, it minimizes enumerator influence while maintaining a high level of respondent understanding, as participants can ask enumerators for clarification if needed. As compared to other self-administered surveys with no external support, our method is more flexible in contexts characterized by high levels of dialect variation or illiteracy, or for contexts where respondents may be unfamiliar or uncomfortable with touchscreen devices. Moreover, it is especially appropriate for individuals who require greater engagement and are prone to distraction, such as adolescents and young adults. Because respondents are guided by the pace of the enumerator, they cannot rush through the questionnaire. By following the protocol outlined in section 2, HASA can also be administered in group settings, thereby reducing the cost of data collection, which is an important consideration in contexts where research funding is limited. If adapting the survey design is not feasible due to practical constraints, we recommend that researchers discuss the potential bias due to enumerator influence using the methodology and classification we have adapted from Di Maio and Fiala (2020), where outcomes can be identified as more or less likely to be influenced by enumerators.

Given the great heterogeneity of contexts in developing countries, an interesting avenue for future research might be examining the determinants of enumerator influence across environments and settings rather than across survey methods. Such an approach would yield valuable insights into the reliability of measurement and encourage researchers to use mixed-method survey designs, particularly when addressing norms or subjective issues in specific contexts, and not only sensitive experiences or opinions.

### References

- Agüero, J. M. and Frisancho, V. (2022). Measuring violence against women with experimental methods. *Economic Development and Cultural Change*, 70(4):1565–1590.
- Agüero, J. M., Perego, V. M., and Romero, J. (2023). Privacy and measurement error in phone surveys: The case of women's agency. *unpublished working paper*.
- Amaral, S., Dinarte-Diaz, L., Dominguez, P., and Perez-Vincent, S. M. (2024). Helping families help themselves: The (un) intended impacts of a digital parenting program. *Journal of Development Economics*, 166:103181.
- 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.
- Baird, S., McIntosh, C., and Özler, B. (2019). When the money runs out: Do cash transfers have sustained effects on human capital accumulation? *Journal of Development Economics*, 140:169–185.
- 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.
- 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.
- Blattman, C., Fiala, N., and Martinez, S. (2020). The long-term impacts of grants on poverty: Nine-year evidence from uganda's youth opportunities program. *American Economic Review: Insights*, 2(3):287–304.
- 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.
- 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.
- Brune, L., Karlan, D., Kurdi, S., and Udry, C. (2022). Social protection amidst social upheaval: Examining the impact of a multi-faceted program for ultra-poor households in yemen. *Journal of Development Economics*, 155:102780.
- Bulte, E. and Lensink, R. (2019). Women's empowerment and domestic abuse: Experimental evidence from vietnam. *European economic review*, 115:172–191.
- Chen, S. and Ravallion, M. (2004). How have the world's poorest fared since the early 1980s? *The World Bank Research Observer*, 19(2):141–169.
- Chen, S. and Ravallion, M. (2010). The developing world is poorer than we thought, but no less successful in the fight against poverty. *The Quarterly Journal of Economics*, 125(4):1577–1625.
- 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.

- Cullen, C. (2023). Method matters: The underreporting of intimate partner violence. *World Bank Economic Review*, 37(1):49–73.
- 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.
- 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.
- Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American economic review*, 112(3):899–927.
- 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.
- Gazeaud, J. (2020). Proxy means testing vulnerability to measurement errors? *The Journal* of *Development Studies*, 56(11):2113–2133.
- Gazeaud, J., Khan, N., Mvukiyehe, E., and Sterck, O. (2023). With or without him? experimental evidence on cash grants and gender-sensitive trainings in tunisia. *Journal of Development Economics*, 165:103169.
- Himelein, K. (2016). Interviewer Effects in Subjective Survey Questions: Evidence from Timor-Leste. *International Journal of Public Opinion Research*, 28(4):511–33.
- Kling, J. R. and Liebman, J. B. (2004). Experimental analysis of neighborhood effects on youth. *Working Paper 483, Industrial Relations Section, Princeton University*.
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual Quant*, 47:2025–2047.
- Laajaj, R. and Macours, K. (2019). Measuring skills in developing countries. *Journal of Human Resources*.
- Lanz, L., Thielmann, I., and Gerpott, F. H. (2022). Are social desirability scales desirable? a meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, 90(2):203–221.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.
- 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.
- List, J., Shaikh, A., and Xu, Y. (2016). Multiple Hypothesis Testing in Experimental Economics. Artefactual Field Experiments 00402, The Field Experiments Website.
- Peterman, A., Dione, M., Le Port, A., Briaux, J., Lamesse, F., and Hidrobo, M. (2024). Disclosure of violence against women and girls in senegal. *The World Bank Economic Review*, page lhae039.
- 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.

- 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.
- Ravallion, M., Datt, G., and van de Walle, D. (1991). Quantifying absolute poverty in the developing world. *Review of Income and Wealth*, 37(4):345–361.
- 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.
- 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.
- Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, 133(5):859.
- West, B. T. and Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of survey statistics and methodology*, 5(2):175–211.
- Özler, B., Çiğdem Çelik, Cunningham, S., Cuevas, P. F., and Parisotto, L. (2021). Children on the move: Progressive redistribution of humanitarian cash transfers among refugees. *Journal of Development Economics*, 153:102733.

# **Online Appendix**

# Measuring Norms and Enumerator Effects: Survey Method Matters

April 24, 2025

Pablo Álvarez-AragónHugues ChampeauxUniversity of NamurUniversity of Cagliari

# A. Experimental design



Figure A1: Experimental Design

### **B.** Patterns in the answers



Figure B1: Example of available answers (I)

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

# Figure B2: Example of available answers (II)



*Note:* This figure shows an example of how respondents in HASA 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".

### C. Summary statistics and variables

# C.1. Balance

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

Table C1: Baseline characteristics

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.

# C.2. Codebook and variables

Name	Label	Туре
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

# Table C2: Codebook and variables

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".

# D. Attrition

(1) (2)	
P(Attrition) P(Attri	tion)
HASA survey (T) 0.174*** 0.400*	***
(0.0195) (0.11	6)
Number of children 0.009	34
(0.008	64)
Number of children x T -0.006	16
(0.012	20)
Age 0.0008	878
(0.001	46)
Age x T -0.002	87
(0.002	, 44)
Father Alive 0.002	85
(0.02/	-) (5)
Father Alive x T -0.006	n97 108
	(2)
Bank account	137 52
	)2 ))
Reply account v.T.	)
Dank account x 1 0.054	4
(0.055	54)
Mobile Phone -0.020	54
(0.028	36)
Mobile Phone x T -0.096.	4**
(0.047	72)
Attended School 0.002	31
(0.020	<b>)</b> 1)
Attended School x T 0.031	8
(0.041	3)
Father produces pineapple -0.01	17
(0.028	32)
Father produces pineapple x T -0.072	26
(0.049	90)
Pinneaple not for women -0.012	23
(0.030	)2)
Pinneaple not for women x T -0.04.	18
(0.048	39)
Pineapple increases respect -0.062	26
(0.042	25)
Pineapple increases respect x T 0.031	8
(0.06/	)
Buy furniture	r~/ }**
0.040	, 32)
Buy furniture x T	) 20
	99 \\$\
Buy matarbile	
	+1 )
(0.020	<i>151</i>
buy motorbike x 1 -0.010	50
(0.048	oð)
Mean Y for the FFS group 0.0446 –	
% of FFS mean 389.3 -	
R-squared 0.0408 0.081	3
N 1009 991	2

Table D1: Survey method and attrition

Note. Robust standard errors are reported in parenthesis.

		KL bounds +/- 0.10SD		KL bounds +/- 0.25SD		Lee bounds	
	Main estimates (1)	Lower bound (2)	Upper bound (3)	Lower bound (4)	Upper bound (5)	Lower bound (6)	Upper bound (7)
Chama Ind	0.154***	0.141***	0.167***	0.122***	0.186***	0.056	0.279***
Shame mu	(0.0351)	(0.0314)	(0.0314)	(0.0314)	(0.0314)	(0.0420)	(0.0432)
D M	-0.359***	-0.370***	-0.348***	-0.387***	-0.331***	-0.488***	-0.265***
Dec Money	(0.0330)	(0.0299)	(0.0299)	(0.0300)	(0.0300)	(0.0417)	(0.0401)
<b>CI I T I</b>	0.376***	0.364***	0.388***	0.346***	0.407***	0.355***	0.470***
Slow down ind	(0.0272)	(0.0231)	(0.0231)	(0.0231)	(0.0231)	(0.0296)	(0.0218)
Clarus darum Cam	0.223***	0.210***	0.236***	0.191***	0.256***	0.161***	0.383***
Slow down Com	(0.0333)	(0.0294)	(0.0294)	(0.0295)	(0.0295)	(0.0384)	(0.0429)
Teeles Com	0.142***	0.132***	0.153***	0.117***	0.168***	-0.011	0.212***
lasks Com	(0.0306)	(0.0278)	(0.0278)	(0.0279)	(0.0279)	(0.0407)	(0.0367)
Dee Deerette	-0.521***	-0.534***	-0.508***	-0.553***	-0.489***	-0.575***	-0.512***
Dec Popotte	(0.0265)	(0.0217)	(0.0217)	(0.0218)	(0.0218)	(0.0228)	(0.0279)
Te alta Ira d	0.265***	0.256***	0.275***	0.241***	0.289***	0.130***	0.352***
Iasks Ind	(0.0309)	(0.0284)	(0.0284)	(0.0285)	(0.0285)	(0.0403)	(0.0380)

# Table D2: Treatment effect bounds

Note. This table reports the results from three methods to test the sensitivity of our results to attrition. Column 1 reports the estimates from Table 4 for reference. Each row is an outcome variable. Columns 2-5 replace outcome values for the attrited observations with semi-extreme values, following Kling and Liebman (2004). Missing treatment (HASA) and control (FFS) observations are set to have a 0.2 or 0.5 SD difference in their outcomes: we impute mean +0.25 SD for missing treatment and mean -0.25 SD for missing controls, and vice versa (KL bounds). Columns 6 and 7 implement the trimming procedure of Lee (2009) (18% of observations in the control group are trimmed to close the attrition gap). 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 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 thinks (or thinks that the community thinks) that housework is a woman's job. "*Dec Popotte*" equals 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 thinks that 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.

### E. Enumerator influence and outcome classification: FFS



Figure E1: Enumerator effect:  $R^2$  of different outcomes

Note: Sample: Surveys conducted in Face-to-Face format. 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.

# F. Summary: Results



### Figure F1: P-values of HASA survey for different outcomes

*Note:* This figure reports the p-values of answering in HASA survey 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.

#### G. Extreme Answers and Say-yes bias

**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 HASA respondents, they are also less likely to choose an extreme negative statement. These estimates suggest that the HASA 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, HASA 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.

	(1) P(Extreme Answer)	(2) P(Extreme Yes)	(3) P(Extreme No)	(4) Say-Yes Bias
HASA survey	0.148*** (0.0174)	0.204 <sup>***</sup> (0.0163)	-0.0555 <sup>***</sup> (0.0148)	-0.000212 (0.0115)
Mean Y	0.660	0.374	0.286	0.460
R-squared	0.0760	0.146	0.0148	0.00284
Ν	847	847	847	847

Table G1: Patterns of answers and survey method

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 C2. Robust standard errors are reported in parenthesis. \*\*\*, \*\*, and \* indicate significance at the 1, 5, and 10 percent critical level.

Tables G2 and G3 show that our main results do not differ by respondents' education level or cell phone ownership at baseline. If misunderstandings and problems with tablets were widespread, we would expect to see larger differences between survey methods for uneducated women or respondents without a cell phone. However, we do not observe systematic differences in the main patterns, which reduces concerns that our results are driven by difficulties with tablet management.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Shame Ind	Dec Money	Slow down Ind	Slow down Com	Tasks Com	Popotte	Tasks Ind
HASA survey	0.161***	-0.312***	0.364***	0.174***	0.129***	-0.536***	0.261***
	(0.0453)	(0.0432)	(0.0367)	(0.0443)	(0.0393)	(0.0314)	(0.0402)
HASA x Went to school	-0.0185	-0.117*	0.0294	$0.121^{*}$	0.0339	0.0338	0.0115
	(0.0716)	(0.0667)	(0.0542)	(0.0664)	(0.0627)	(0.0555)	(0.0630)
Mean Y FFS group	0.439	0.579	0.0935	0.280	0.688	0.946	0.607
R-squared	0.0251	0.136	0.154	0.0552	0.0281	0.272	0.0942
N	847	847	847	847	847	750	847

Table G2: Survey Method and high influence variables

Note. "Went to school" equals one if the respondent reports that she attended school at baseline. 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 are slowed down by the development of his wife's individual economic activities are slowed down by the development of his wife's individual economic activities are slowed down by the development of his wife's individual economic 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 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.

### Table G<sub>3</sub>: 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
HASA survey	0.133*	-0.265***	0.332***	0.141*	0.182***	-0.511***	0.343***
	(0.0774)	(0.0714)	(0.0622)	(0.0760)	(0.0674)	(0.0562)	(0.0679)
HASA x cell-phone	0.0262	-0.116	0.0557	0.104	-0.0498	-0.0139	-0.0979
	(0.0869)	(0.0804)	(0.0692)	(0.0845)	(0.0757)	(0.0637)	(0.0763)
Mean Y FFS	0.439	0.579	0.0935	0.280	0.688	0.946	0.607
R-squared	0.0225	0.142	0.152	0.0507	0.0282	0.268	0.0962
Ν	847	847	847	847	847	750	847

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 are slowed down by the development of his wife's individual economic 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 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.