Policy Research Working Paper 10242

Interviewer Design Effects in Household Surveys

Evidence from Sudan

Alvin Etang
Habtamu Fuje
Christopher Root

WORLD BANK GROUP
Poverty and Equity Global Practice
November 2022

Policy Research Working Paper 10242

Abstract

Interviewer design effects occur when data collected by the same interviewer is more similar than data collected by different interviewers. Design effects inflate survey variance and reduce the precision of estimates. Using household survey data collected via computer assisted personal interviewing (CAPI) in Sudan this paper employs a two-level mixed effects regression model to identify interviewer design effects for key variables. The study finds mean interviewer design effect values of 7 with a maximum of 16, implying a significant loss of precision. Recommendations to mitigate interviewer design effects include simplifying questions, sound survey implementation practices, and utilizing multi-way cluster robust standard errors to account for both area and interviewer clustering during data analysis.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at aetangndip@worldbank.org; christopher.n.root@gmail.com; hfuje@worldbank.org.

Interviewer Design Effects in Household Surveys: Evidence from Sudan

Alvin Etang, Habtamu Fuje, and Christopher Root [1]

Keywords: Survey interviewer, Interviewer effects, Measurement error, Mixed effects regression

JEL Classification: C10, C81, C83, C89

[1] Alvin Etang is a Senior Economist in the Poverty and Equity Global Practice of the World Bank. Habtamu Fuje is an Economist with the Poverty and Equity Global Practice of the World Bank. Christopher Root is an Agricultural Economist who works an independent consultant. The authors would like to thank staff at the Sudan Central Bureau of Statistics who worked on the pilot Agricultural Production Survey project. The authors are grateful to Kristen Himelein (Senior Economist, EEAPV) and Gbemisola Oseni (Senior Economist, DECPM) for peer reviewer feedback and suggestions on an earlier draft, and to Eiman Osman (Consultant, EAEPV) for her support with the study. The findings, interpretations, and conclusions of this paper are those of the authors and should not be attributed to the World Bank or its Executive Directors. The authors may be contacted at christopher.n.root@gmail.com; hfuje@worldbank.org; aetangndip@worldbank.org.

1. Introduction

This paper seeks to identify interviewer effects using household data collected in Sudan. The data is from a pilot agricultural survey conducted in 2018 by the Central Bureau of Statistics (CBS) in collaboration with the World Bank in Kassala State. Kassala lies in the eastern part of Sudan bordering Eritrea and is viewed as having high smallholder agricultural potential. It was chosen for the pilot agricultural survey in part for this reason but also because of its diverse agriculture including rainfed, irrigated and mechanized crop production, as well as livestock husbandry. The drier north of the state is pastoral whereas the wetter central and southern parts of the state are primarily sorghum-based agro-pastoral livelihood systems.

The pilot survey is a good opportunity to identify interviewer effects because of the authors' close involvement throughout the survey. This involvement included questionnaire and sample design; enumerator and supervisor training and piloting; post survey debriefing with enumerators and supervisors; and data cleaning and analysis.

Interviewer effects

Interviewer effects are the effects that interviewers can have on survey or census data quality. In the case of surveys, which are the focus here, interviewer effects are the difference between the true population parameter and the survey estimate that is attributable to interviewer errors. Interviewer effects can broadly be divided into three categories: coverage errors, non-response errors and measurement errors. Coverage errors are errors in the sample frame as a result of mistakes in listing or under coverage of difficult to access households. Non-response errors result from interviewer's inability to solicit a response from a sampled respondent. Finally, measurement errors, which are the focus of this paper, occur when an interviewer does not solicit or record the correct response from the respondent (West and Blom, 2017).

There is a long history of the study of interviewer effects on survey data collection (Kish, 1962). In their recent comprehensive review of the literature on interviewer effects, West and Blom (2017) identify six types of interviewer characteristics that have been related to measurement error in the literature. These characteristics are race/ethnicity, age, sociodemographic matching with interviewer's characteristics, survey-specific experience, current survey experience, and gender.

CAPI and PAPI

The prevalence of computer assisted personal interviewing (CAPI) should theoretically reduce interviewer induced measurement error through automated skip logic as well as real time response validation. Caeyers *et al.* (2012) conducted a randomized field experiment to compare CAPI with pen and paper interviewing (PAPI). The authors indeed found fewer data errors in the CAPI-collected data than PAPI data as well as lower mean and variance of consumption estimates.

However, there are still sources of interviewer related measurement errors that are not likely to be attenuated by CAPI. These include an interviewer's ability to understand and explain more complex questions, the skill with which they engage respondents to maintain their enthusiasm and truthfulness throughout an interview. Furthermore, an interviewer's background

characteristics such as gender or ethnicity may make a respondent more or less forthright in their responses (Haber, 2018).

Intraclass Correlation Coefficients

Classical statistical models depend on the assumption that observations are independent and identically distributed (IID). Non-random interviewer measurement error violates this assumption through errors clustered by interviewer. The result is loss of precision through higher variances. The clustering of responses by interviewer is analogous to the geographic clustering that occurs through a multistage cluster sample and similarly leads to larger standard errors. This homogeneity of responses resulting from clustering can be expressed through the intraclass correlation coefficient (ICC) (Brunton-Smith *et al.*, 2016).

The ICC expresses the share of the total sample variance that can be explained by the within-cluster variance. In the case of interviewer measurement error, this is the share of the total sample variance that can be explained by interviewer-based measurement error that is not randomly distributed across interviewers. The ICC ranges from 0 to 1, with a value of zero indicating zero correlation in responses by interviewer and a value of one indicating that all response variance is due to interviewers. Intuitively, an ICC of zero implies no loss of sample efficiency (i.e. no increase in standard errors) resulting from clustering and therefore no adjustment needs to the sample to account for clustering. On the other hand, an ICC value of one would indicate that each cluster must be treated as a single observation, obviously implying a much larger sample requirement.

Design effects

The ICC affects variance and, therefore, sample size requirements through the design effect, which is commonly used to adjust for clustered sample design. The design effect is the ratio of the actual sample variance to the presumed variance under simple random sampling. The design effect formula is shown below:

$$DF = 1 + \rho(c - 1)$$

where *DF* is the design effect, *c* is the size of the cluster and ρ the ICC. In the case of interviewer effects, *c* is the number of respondents per interviewer. To correct for clustering of responses and the consequent increase in variance, the design effect is multiplied by the sample size requirement obtained through simple random sampling.

In their seminal work on interviewer design effects, O'Muircheartaigh and Campanelli (1998) used an interpenetrated design, randomly assigning interviewers to respondents in a household survey in the United Kingdom (UK). They calculated interviewer ICCs for 820 variables and found median interviewer design effects of 1.8 with a maximum of 5, implying an 80 percent and 400 percent increase in the sample size required to maintain the same variance. Brunton-Smith et al. (2016) found evidence of interviewer design effects in UK Household Longitudinal Survey data ranging between 2.5 and 3.3. Schnell and Kreuter (2005) analyzed face-to-face and mail-in

crime surveys conducted in Germany and found that interviewers were accountable for 77 percent of the total clustered variance with interview areas only accounting for 23 percent.

Overview of research

These findings, which are typically not accounted for in survey designs, mean that surveys are likely underpowered. This paper sets out to quantify interviewer design effects in the context of a developing country household survey conducted with CAPI. To do this, we attempt to isolate the portion of the variance attributable to interviewers from the portion attributable to interview areas. We also control for interviewer – respondent interaction effects that may be associated with the gender of interviewer and respondents. We make recommendations for survey design and data analysis to mitigate the effect of interviewer design effects.

The paper proceeds as follows. Section 2 describes the data, including sampling and the variables selected for this study. Section 3 describes our analytical approach, while section 4 presents and discusses the results. Finally, section 5 presents a brief conclusion and recommendations.

2. Data description

The data used in this study is from the Agricultural Production Survey (APS), a pilot survey conducted by the Central Bureau of Statistics of Sudan in collaboration with the World Bank in March-April 2018. The survey was conducted in Kassala State of Sudan. The APS questionnaire was designed largely based on the World Bank's Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA). The questionnaire included four modules: (1) the household module, (2) the agricultural planting module, (3) the agricultural harvest module, and (4) the livestock module. For this study, we only use the household module because this was the only module completed for all households and therefore provides the largest sample size. The data was collected with tablets using the Survey Solutions CAPI program.

Sampling

The sample was drawn using a two-stage cluster design. First, the simple random sample size was calculated with the standard parameters of a 5 percent error margin, a 95 percent confidence level and a key parameter expected value of 0.5.[2] Next the area cluster design effect was calculated to account for interview area clustering using formula 1 above with cluster size *c* equal to 10 households and ICC *p* equal to 0.15.[3] The result was a design effect of 2.35, which was multiplied by the SRS sample size (of 384) to yield a total sample size of 903 households.[4]

In the first stage of sampling, Popular Administrative Units (PAUs) were randomly selected prior to the survey. Then each day within each cluster (PAU), the survey team carried out a listing

[2] The sample size is determined using this formula: $n_{srs} = \frac{t^2 \times p(1-p)}{m^2}$, where n_{srs} is the sample size with SRS sampling, *t* is the confidence level, p is the estimated value for the key parameter and m is the margin of error. Note that 0.5 is the most conservative value for *p* in that it results in the largest sample size.

[3] This ICC was estimated based on values used in other agricultural studies in other countries.

[4] Three replacement clusters were also included in the survey. These were mistakenly included in data collection resulting in a larger than necessary sample size of 93 clusters with 930 respondents.

exercise and then randomly selected households from the lists using a tablet-based random number generator. Only households engaged in crop agriculture and/or raising livestock were eligible for listing.

The survey was conducted by 20 interviewers and 5 supervisors, organized into five teams, over a month-long period. Interviewers were not randomly assigned to respondents. Interviewers and supervisors took part in a five-day training and a one-day pilot prior to data collection.

Selection of Variables

For calculating interviewer ICCs, we use the ten numerical variables from the household module with over 800 responses.[5] All variables selected are analyzed as recorded by interviewers and are not based on post0interview processing. This avoids potential attenuation or amplification of enumerator effects. For example, rather than calculating hours worked per year, we analyze separately months worked in past 12 months, days worked in those months, and hours worked in those days – as was asked and recorded during the interview. Similarly, weekly expenditures on food items and monthly expenditures on non-food items were directly recorded for each item and our analysis is based on these separate variables and not on post interview multiplication of price and quantity. We use the two food and two non-food item expenditure results with more than 800 responses.

We classify our ten dependent variables into three categories based on their difficulty as described in Table 1. We expect that more difficult to record variables will elicit more interviewer-based measurement error and have higher associated ICCs. However, the evidence that would allow for a clear classification of questions on the basis of their propensity for inducing interviewer effects is ambiguous (Schnell and Kreuter, 2005).

Table 1: Dependent variables

Classification	Variable
Easier questions that involve NO recall or computation beyond counting	Number of people in HH Age of HH head
More difficult question requiring recall OR computation beyond counting	Months out of past 12 worked in primary occupation by HH head
Most difficult questions requiring recall AND computation beyond counting	Days per month worked in primary occupation by HH head Hours per day worked in primary occupation by HH head Meals per day for HH members over 5 HH expenditure on sorghum (including sorghum flour) in last 7 days (SDG) HH expenditure on sugar or high-sugar products in last 7 days (SDG) HH expenditure on matches in last 30 days (SDG) HH expenditure on laundry detergent in last 30 days (SDG)

[5] The exception is volume of food items consumed due to food weight units which was deemed unreliable.

Interviewer characteristics

The mean values for interviewer characteristics are provided in Table 2. These characteristics are used as covariates in the analytical approach, which is described in the section below.

Table 2: Interviewer characteristics[6]

Interviewer characteristic	Value
Female (%)	40
Age (mean years)	40.2
Lives or has lived in the survey state (Kassala) (%)	50
Attended university or higher (%)	77.8
Works with CBS (%)	82.4
Conducted 10 or more CAPI surveys previously (%)	61.1
Self-rated skill with tablets or smartphones before survey (1 – 10) (mean)	9.3
"On a scale of 1 to 10, how much do you generally trust other people?" (mean)	6.8
"On a scale of 1 to 10, how much do you think other people generally trust you?" (mean)	7.6

3. Analytical approach

This research seeks to identify interviewer effects and their causes through four different specifications of a model that measures interviewer ICC. Following Brunton-Smith *et al*. (2016), we use two-level mixed effects linear regression models with respondents nested in interviewers. We deploy four different versions of this model. The first specification simply measures the interviewer ICC for each variable. Specifications II and III aim to control for covariates including interviewer area clusters and the gender of interviewers and respondents. A fourth specification includes interviewer characteristics in an attempt to identify their contribution to interviewer-related measurement error. Note that for each specification, we are only interested in the ICCs that are obtained post-estimation and not the coefficients themselves.

The first specification simply identifies the ICC without controlling for possible covariates:

$$\text{Specification I:} \quad y_{ij} = \beta + u_j + \epsilon_{ij}$$

where *y* is each of the ten dependent variables in table 1; *i* is the households interviewed by interviewer *j*, where j runs from 1 through 20; β is the is the intercept, u_j is the level 2 (interviewers) random intercept; and ϵ_{ij} is the level 1 error term. For this model, the ICC is calculated using the formula below.

$$\rho = ICC = \frac{\sigma_u^2}{\frac{\pi^2}{3} + \sigma_u^2}$$

[6] Except for female where all 20 interviewers responded, all responses are based on n = 18 except age and worked with CBS, which only had 17 respondents.

where σ_u^2 is the variance of u_j and $\frac{\pi^2}{3}$ the variance of the error term ϵ_{ij}, which is assumed have a logistic distribution.

An analogous specification is used for estimating interview area cluster ICCs, with a_j interviewer clusters random intercept replacing the u_j interviewers random intercept.

$$Area\ clusters\text{:}\ y_{ij} = \beta + a_j + \epsilon_{ij}$$

Next, we control for the fact that interviewers were not randomly assigned to households by including interviewer-survey area fixed effects.

$$\text{Specification II:}\ y_{ij} = \beta_0 + u_j + X_{ej} + \epsilon_{ij}$$

where X_{ej} is the area cluster fixed effects, with e indicating clusters 1 through 93.

Next, we add the gender of the interviewer, the respondent, and an interaction term. This aims to account for differences in interviewer errors between men and women (West and Blom, 2017) as well as the potential advantages of matching interviewer and respondent gender including trust for sensitive questions (Lupu and Michelitch, 2018).

$$\text{Specification III:}\ y_{ij} = \beta_0 + u_j + X_{ej} + I_j + R_i + I_j * R_i + \epsilon_{ij}$$

where I_j is the interviewer $j's$ gender and R_i is respondent $i's$ gender.

In the final model, we add interviewer characteristics (table 1) individually to Specification III above in an attempt to identify the extent to which they account for interviewer effects as measured by ICCs. Specification IV is only used for the three dependent variables associated with the highest ICCs. Note that because of missing data, results for Specification IV are only suggestive and are therefore only presented in the annexure.

$$\text{Specification IV:}\ y_{ij} = \beta_0 + u_j + X_{ej} + C_j + \epsilon_{ij}$$

where C_j is the vector of interviewer $j's$ characteristics shown in table 1. This specification is repeated for each C_j interviewer characteristic. As with all of the four specifications, the outcome of interest here is the associated ICC.

4. Results

Descriptive

The number of interviews as well as means and standard errors for each dependent variable are shown in Table 3. Annex 1 presents these results disaggregated by interviewer. In comparison with the other variables, all four expenditure variables have relatively high standard deviations compared with their means. This is consistent with the intuition that wealth inequality is likely to correlate with high expenditure standard deviations. Nevertheless, since ICC measures the relative share of the variance attributable to clustering, we should not expect to see higher ICCs for variables with higher standard deviations.

Table 3: Summary of dependent variables

Variable	n	Mean	Standard deviation
Number of people in HH	930	6	2
Age of HH head	927	48	13
Months out of past 12 worked in primary occupation by HH head	923	9	3
Days per month worked in primary occupation by HH head	923	24	7
Hours per day worked in primary occupation by HH head	923	8	3
Meals per day for HH members over 5	930	3	0
HH expenditure on sorghum (incl. sorghum flour) in last 7 days (SDG)	837	106	114
HH expenditure on sugar or high-sugar products in last 7 days (SDG)	882	91	63
HH expenditure on matches in last 30 days (SDG)	881	12	9
HH expenditure on laundry detergent in last 30 days (SDG)	913	133	110

Intraclass Correlation Coefficient

Table 4 shows the interviewer ICCs for specifications I to III, described above. Additionally, Table 4 shows ICCs for the area clusters. These can be directly compared to the ICCs for interviewers from Specification I as they are estimated with an analogous model.

Specification I without any covariates results in ICCs that range from 0.015 for number of people in the household to 0.359 for hours per day worked in their primary occupation by the head of household. The mean ICC using Specification I is 0.161. This implies that data collected by the same interviewer is on average 16.1 percent more likely to have the same values as data collected by different interviewers, with a range of 1.5 to 35.9 percent.

As expected, the two variables classified as easier do indeed have the lowest average ICCs, followed by the variable classified as more difficult and then the mean of the seven variables classified as most difficult (see table 1 for classification). This relationship holds across all three specifications.

On average, the ICC for specification I is 0.161 compared to 0.099 for the ICC based on area clustering. The implication is that interviewers induced more correlation in responses than did geography. Put another way, data collected by the same interviewers is more similar, due to measurement error, than data collected within the same interview areas. Disaggregating mean ICCs by question difficulty reveals a modest difference in mean area cluster ICCs between the easier and most difficult questions for area clusters (0.068 versus 0.097). For interviewer ICCs, this difference is much more pronounced at 0.027 for easier questions, compared to more difficult 0.207 questions. This finding is consistent with intuition: question difficulty should interact with interviewer ability, not geography.

The inclusion of area cluster fixed effects in specification II reduces interviewer effects (ICCs) in all cases. Across the ten variables, the average ICC is 0.137 with Specification II, compared to 0.161 for Specification I. This indicates that there is some correlation between interviewers and interview areas which is not unexpected since interviewers were not randomly assigned to clusters.

Finally, we add variables for the gender of interviewer and respondent along with an interaction term (Specification III). This results in very small decreases in the ICC across most of the variables and reduces the mean ICC from 0.137 for Specification II to 0.128 for Specification III.

Table 4: ICCs for interviewer area clusters and Models 1 -3

		Interviewer clusters		
	Area clusters	**I**	**II**	**III**
Number of people in HH	.047	.015	.013	.013
	(.021)	(.011)	(.011)	(.011)
Age of HH head	.088	.039	.021	.022
	(.025)	(.018)	(.013)	(.013)
Months out of past 12 worked in primary occupation by HH head	.174	.101	.032	.022
	(.033)	(.035)	(.016)	(.013)
Days per month worked in primary occupation by HH head	.057	.12	.117	.117
	(.022)	(.039)	(.039)	(.039)
Hours per day worked in primary occupation by HH head	.072	.359	.335	.316
	(.024)	(.076)	(.074)	(.072)
Meals per day for HH members over 5	.1	.235	.227	.223
	(.026)	(.062)	(.06)	(.06)
HH expenditure on sorghum (including sorghum flour) in last 7 days (SDG)	.118	.087	.072	.072
	(.029)	(.032)	(.028)	(.028)
HH expenditure on sugar or high-sugar products in last 7 days (SDG)	.145	.27	.212	.209
	(.031)	(.067)	(.058)	(.058)
HH expenditure on matches in last 30 days (SDG)	.071	.15	.115	.081
	(.024)	(.046)	(.039)	(.03)
HH expenditure on laundry detergent in last 30 days (SDG)	.118	.23	.221	.203
	(.028)	(.061)	(.059)	(.056)
Total mean ICC[7]	.099	.161	.137	.128
Easier questions, average ICC	.068	.027	.017	.018
More difficult question, average ICC	.174	.101	.032	.022
Most difficult questions, average ICC	.097	.207	.186	.174

Design effects

We use these ICCs to calculate interviewer design effects using an average interviewer workload of 46.5 interviews. To calculate interviewer design effects, we use the ICC results from Specification II. Using the design effect formula from Section 1, we find a mean design effect of 7.2. In other words, clustering of responses by interviewers results in survey data with average variances that are 7.2 times higher that they would be without interviewer effects. Design effects range from 1.6 for number of people in the household to 16.2 for hours per day worked in primary occupation by the household head. By comparison, there were 10 interviews per area cluster which together with the mean ICC of 0.099 results in a mean design effect of 1.9.

The design effects found here are high compared to O'Muircheartaigh and Campanelli (1998) who found a mean interviewer design effect of 1.8 with a high of 5. However, this was based on a smaller survey cluster size of around 24 respondents per interviewer and the highest ICC

[7] Calculated as the arithmetic mean of the ten ICCs.

reported was 0.171 or about half of the high ICC estimated in this research. Brunton-Smith *et al.* (2016) found interviewer design effects between 2.5 and 3.3 using a similar number of interviews per interviewer as the survey we use here.

Interviewer characteristics

Annex 2 shows the results using Specification IV. This specification adds to Specification II interviewer characteristics for the three dependent variables with the highest ICCs in Table 3 above. These results are presented only in the annexes because they are based on a limited sample since data from all interviewers was not available. Even if data for all 20 interviewers were available, this would not be a large enough sample from which to draw statistically robust conclusions. These results therefore should be interpreted with caution and be considered only as suggestive rather than empirical. The inclusion of different interviewer characteristics reduces ICCs by between 6 and 14 percent with a mean of 10 percent when compared to ICCs calculated using the same sample and model but without interviewer characteristic covariates. The most significant interviewer characteristics are having lived in Kassala State (14.1 percent lower ICCs), having a university degree or higher (13.4 percent lower) and experience working for CBS before (13.1 percent lower).

5. Conclusions and recommendations

Researchers have long been aware of the need to consider geographic clustering of responses in multi-stage cluster designs. Estimated design effects are used to calculate the sample size increase needed to account for increased variance due to within-cluster homogeneity of responses. No equivalent practice, however, is used to account for potential non-random distribution in measurement errors that results in homogeneity in data collected by the same interviewer. Using data from a pilot survey in Sudan, we find that clustering of responses by interviewers can in fact lead to larger design effects than area clustering. This finding is consistent with the findings of Schnell and Kreuter in Germany (2005).

This paper provides evidence that simpler questions lead to smaller interviewer design effects. This finding reinforces the intuition that time invested in designing, testing and refining questions to make them as simple as possible is likely to yield data quality improvements.

In general, good survey implementation practices can contribute to lower interviewer design effects. These include recruiting qualified interviewers; budgeting adequate time for questionnaire design, review and programming for tablets; conducting a high-quality training using a close to finalized questionnaire; budgeting adequate time for piloting and finalizing the questionnaire; and managing survey implementation well, including early identification of reoccurring interviewer-level issues that might affect data quality.

Another recommendation emerging from these findings is that interviewer identifiers should always be included in all survey data. Interviewer identifications in data allow for multi-way cluster robust standard errors by both area cluster and interviewer (Cameron *et al.*, 2011). Given the evidence of interviewer design effects in survey data collected in developed countries and

now in developing countries, researchers working with survey data should strongly consider implementing multi-way cluster robust standard errors.

Finally, there are two theoretically simple ways to reduce the design effects caused by the clustering of responses by interviewers. One is to simply expand the sample size using an expected interviewer design effect while keeping the number of respondents per interviewer constant. This approach is what is used to account for area cluster design effects. Another approach is to reduce the number of respondents per interviewer. In this analysis, all things being equal, a doubling of the number of interviewers from 20 to 40 would result in a reduction in the average design effect from 7.2 to 4. In both cases however there are limits to how many interviewers can be trained and managed without effects on data quality. Additionally, there may be limits to the number of high-quality interviewers available to work on a given survey. Nevertheless, survey implementors should increase the number of interviewers to the maximum number that can all be trained and managed well.

This paper provides some tentative evidence that interviewers with college degrees or higher, those with local knowledge of the survey area and those with experience conducting surveys, produce data with lower interviewer design effects. Future research could address interviewer characteristics associated with reduced interviewer design effects more robustly. Since the cost of interviewing interviewers during a survey is minimal, interviews of interviewers should become a common practice in survey implementation. Such practice would allow for meta studies to identify the relative importance of different interviewer characteristics for data quality.

References

Brunton-Smith, Ian, P. Sturgis & G. Leckie. 2016. Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society – Series A, Statistics in Society,* 18(2): 551-568.

Caeyers, B., Chalmers, N. & De Weerdt, J. 2012. Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment. *Journal of Development Economics,* 98: 19-33.

Cameron, A.C., J. Gelbach and D. Miller. 2011. Robust Inference with Multiway Clustering. *Journal of Business & Economic Statistics,* 29(2): 238-249.

Haber, Noah. P.J. Robyn, S. Hamadou, G. Yama, H Hien, D. Louvouezo & G. Fink. 2018. Surveyor Gender Modifies Average Survey Responses: Evidence from Household Surveys in Four Sub-Saharan African Countries. *arXiv,* 1810.01981.

Kish, Leslie. 1962. Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association,* 57: 92–115.

Lupu, Noam & K. Michelitch. 2018. Advances in Survey Methods for the Developing World. *Annual Review of Political Science,* 21: 195-214.

O'Muicheartaigh, Colm & P. Campanelli. 1998. The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161(1): 63-77.

Schnell, Rainer & F. Kreuter. 2005. Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics*, 21(5): 389-410.

West, Brady T. & A.G. Blom. 2017. Explaining Interviewer Effects: A Research Synthesis. *Journal of Statistics and Methodology*, 5: 175-211.

Annex 1: Dependent variable means (standard deviations in parenthesis)

	# in household	Age of HH head	Months per year worked by HH head	Days per month worked by HH head	Hours per day worked by HH head	Meals per day for HH members over 5	HH expenditure on sorghum in last 7 days (SDG)	HH expenditure on sugar in last 7 days (SDG)	HH expenditure on matches in last 30 days (SDG)	HH expenditure on laundry detergent in last 30 days
Total	*6 (2)*	*48 (13)*	*9 (3)*	*24 (7)*	*8 (3)*	*3 (0)*	*106 (114)*	*91 (65)*	*12 (9)*	*133 (110)*
Interviewer 1	5 (3)	43(12)	9 (3)	25 (6)	6 (2)	3 (0)	123 (165)	70 (40)	8 (3)	90 (78)
Interviewer 2	6 (3)	54 (12)	9 (3)	26 (5)	8 (1)	3 (0)	145 (159)	38 (28)	9 (5)	66 (30)
Interviewer 3	7 (2)	48 (11)	11 (2)	30 (2)	15 (3)	3 (0)	115 (97)	95 (71)	12 (4)	236 (199)
Interviewer 4	6 (2)	45 (14)	10 (3)	24 (6)	7 (2)	3 (0)	108 (73)	179 (102)	13 (7)	190 (102)
Interviewer 5	5 (2)	52 (13)	9 (4)	20 (8)	8 (3)	2 (0)	121 (95)	102 (53)	19 (10)	159 (104)
Interviewer 6	6 (2)	46 (11)	10 (2)	23 (8)	7 (3)	3 (0)	90 (86)	92 (48)	12 (5)	108 (119)
Interviewer 7	6 (2)	51 (12)	8 (3)	24 (6)	7 (3)	3 (1)	56 (85)	108 (45)	9 (4)	140 (83)
Interviewer 8	6 (3)	52 (14)	8 (4)	24 (6)	9 (1)	3 (0)	45 (111)	27 (13)	5 (3)	65 (18)
Interviewer 9	6 (2)	42 (12)	8 (4)	23 (5)	8 (2)	3 (0)	87 (71)	62 (34)	12 (4)	106 (54)
Interviewer 10	7 (2)	49 (14)	10 (3)	26 (5)	7 (2)	3 (0)	91 (79)	104 (53)	19 (6)	244 (128)
Interviewer 11	6 (2)	48 (13)	11 (3)	28 (4)	8 (3)	3 (0)	73 (78)	77 (33)	12 (8)	85 (47)
Interviewer 12	6 (2)	48 (12)	9 (4)	28 (4)	12 (4)	3 (1)	95 (67)	69 (37)	13 (8)	111 (61)
Interviewer 13	7 (2)	47 (10)	10 (3)	24 (5)	11 (4)	3 (0)	218 (243)	101 (55)	12 (6)	124 (58)
Interviewer 14	6 (3)	45 (14)	7 (4)	23 (10)	7 (2)	3 (0)	83 (78)	130 (93)	12 (8)	170 (130)
Interviewer 15	6 (2)	51 (16)	8 (2)	25 (2)	8 (1)	2 (0)	76 (70)	63 (24)	15 (4)	113 (45)
Interviewer 16	5 (2)	48 (13)	10 (3)	19 (11)	7 (3)	3 (1)	156 (103)	120 (79)	20 (29)	182 (105)
Interviewer 17	6 (2)	49 (14)	10 (3)	24 (7)	7 (4)	3 (0)	103 (76)	106 (65)	13 (6)	142 (72)
Interviewer 18	6 (2)	51 (15)	9 (3)	21 (10)	7 (4)	2 (1)	138 (94)	53 (32)	9 (3)	198 (95)
Interviewer 19	6 (2)	51 (14)	10 (3)	26 (7)	7 (2)	3 (0)	103 (70)	102 (53)	8 (3)	58 (49)
Interviewer 20	6 (3)	47 (16)	10 (3)	24 (5)	7 (2)	3 (0)	98 (68)	136 (69)	13 (6)	83 (134)

Annex 2: ICCs with interviewer covariates (Specification IV)

	Hours per day worked in primary occupation by HH head	Meals per day for people over 5	Amount spent on laundry detergent in past 30 days	Mean difference with respective comparison group (%)
Comparison 1 (n = 739, 746, 734)	***.196***	***.229***	***.194***	
	(.061)	***(.068)***	***(.061)***	
Age (years)	.163	.22	.164	12.1
	(.054)	(.066)	(.055)	
Comparison 2 (n = 785, 792, 778)	***.182***	***.219***	***.181***	
	(.057)	***(.064)***	***(.057)***	
Lives or has lived in Kassala (%)	.138	.188	.174	14.1
	(.047)	(.06)	(.055)	
Attended university or higher (%)	.172	.166	.162	13.4
	(.055)	(.053)	(.053)	
Conducted 10 or more CAPI surveys previously	.172	.204	.145	10.7
	(.055)	(.061)	(.049)	
Self-rated skill with tablets or smartphones before survey (1 – 10)	.15	.211	.18	7.3
	(.05)	(.062)	(.057)	
"On a scale of 1 to 10, how much do you generally trust other people?"	.182	.184	.177	6.1
	(.057)	(.057)	(.056)	
"On a scale of 1 to 10, how much do you think other people generally trust you?"	.182	.184	.177	6.1
	(.057)	(.057)	(.056)	
Comparison 3 (n = 738, 745, 731)	***.178***	***.184***	***.152***	
	(.058)	***(.059)***	***(.052)***	
Work with CBS (%)	.163	.161	.124	13.1
	(.054)	(.054)	(.045)	