the review of income and wealth

Review of Income and Wealth
Series 65, Number 1, March 2019
DOI: 10.1111/roiw.12343

SAME QUESTION BUT DIFFERENT ANSWER: EXPERIMENTAL EVIDENCE ON QUESTIONNAIRE DESIGN'S IMPACT ON POVERTY MEASURED BY PROXIES

BY TALIP KILIC* AND THOMAS PAVE SOHNESEN

World Bank

Based on a randomized survey experiment that was implemented in Malawi, the study finds that observationally-equivalent, as well as same, households answer the same questions differently depending on whether they are interviewed with a short questionnaire or its longer counterpart. Statistically significant differences in reporting emerge across all topics and question types. In proxy-based poverty measurement, these reporting differences lead to significantly different predicted poverty rates and Gini coefficients. The difference in poverty predictions ranges from 3 to 7 percentage points, depending on the model specification. A prediction model based only on the proxies that are elicited prior to the variation in questionnaire design yields identical poverty predictions irrespective of the short-versus-long questionnaire treatment. The results are relevant for estimating trends with questionnaires exhibiting inter-temporal variation in design, impact evaluations administering questionnaires of different length and complexity to treatment and control samples, and development programs utilizing proxy-means tests for targeting.

JEL Codes: C83, D12, I32

Keywords: household survey experiment, poverty, proxy-based poverty measurement, questionnaire design, survey-to-survey imputation

1. INTRODUCTION

Does the same question that is asked of the same population yield the same answer in face-to-face interviews when other parts of the questionnaire are altered? If not, what might be correlated with the discrepancies and what would be the resulting implications for proxy-based poverty measurement? The assumption of the same question providing the same answer has been termed the survey-invariance-assumption (Ravallion, 2016), and testing its validity is at the core of this study. While the empirical investigation is conducted in the context of predicting household consumption expenditures, the findings are equally relevant for the estimation of trends based on questionnaires that exhibit variations in design over time, for impact evaluations that administer questionnaires of different

Note: The authors would like to thank Astrid Mathiassen and Bjorn Wold for their inputs into the experiment design; Heather Moylan for research assistance; and Phyllis Ronek for editorial assistance. The earlier versions of the paper benefited from valuable comments from two anonymous reviewers, Alberto Zezza, Astrid Mathiassen, and the participants at (i) the 2015 Center for the Study of African Economies (CSAE) Conference (Oxford, UK), (ii) the 2015 Deutsches Institut fur Entwicklungspolitik International Workshop "Measuring the Social, Economic and Political Effects of Social Protection: How to Overcome the Challenges" (Bonn, Germany), (iii) the 2015 Midwest International Economic Development Conference (Madison, WI), and (iv) the 2015 International Conference of Agricultural Economists (Milan, Italy).

*Correspondence to: Talip Kilic, World Bank, Center for Development Data, Via Labicana 110, 00184, Rome, Italy; (tkilic@worldbank.org); +39-0677710236.

length and complexity to treatment and control samples, and for social assistance and similar programs that rely on proxy-means tests for beneficiary targeting.

Estimating consumption and poverty via proxies is compelling, as consumption measurement is often argued to be complex and costly. The literature on proxy-based poverty measurement highlights the promise of the method in improving the frequency and comparability of poverty estimates at a lower cost. While common applications require primary data collection based on shorter welfare monitoring surveys, secondary survey data, such as those from Demographic Health Surveys or Labor Force Surveys, have also been used to obtain poverty predictions (Christiaensen *et al.*, 2012; Douidich *et al.*, 2013). With increasing pressure placed on national statistical systems to improve the frequency, quality, cost-effectiveness and comparability of poverty statistics, the interest in the method's application is generating continued interest.

Both parametric and non-parametric approaches to estimation of proxy models have been featured in the literature (see Vu and Baulch (2011) for a review). Regardless of the approach, all practical applications of poverty measured by proxies, such as proxy-means tests, would rely on data originating from two non-identical questionnaires: one set of data to establish the underlying model and another set of data with proxies to pair with the model parameters for obtaining predictions. In the case of consumption and poverty, the model is typically established based on data from a multi-purpose household questionnaire that yields a comprehensive welfare aggregate (hereafter referred to as a standard household questionnaire); data on proxies would be solicited through a shorter household questionnaire, often with a shorter field implementation period.[1] Even if questions underlying proxy definitions are worded *identically* across short versus standard household questionnaires, identical questions could yield different answers in questionnaires that exhibit substantial variation in the number or the order of questionnaire modules or questions within modules.

Designing a questionnaire and interview process, without their leading to biases in responses is difficult. Tourangeau *et al.* (2000) posit that question-answering process involves the stages of comprehension, retrieval, judgment, and response production. Theoretically, questionnaire design decisions place different demands on respondents at different stages (Hess *et al.*, 2001).

The concepts validity—does an instrument measure what it is intended to measure?—and reliability—does an instrument measure what it is intended to measure in a consistent fashion—are in some areas of literature, especially the medical literature, often used to evaluate survey instruments. Despite knowing that variations in questionnaire design can lead to biases, the literature on substantial questionnaire variation with identical questions given to comparable samples is thin, and there are no standards or guidelines available to consider when implementing such changes in practice.

Beegle *et al.* (2012), in their comparative assessment of different questionnaire designs and their impacts on measured household consumption in Tanzania,

[1] Although the shorter fieldwork for a short household survey would result in cost savings, the differences in the period of implementation between a short household survey and its standard comparator could affect the values obtained for the seasonality-prone poverty proxies. Our setup ensures that the observed differences between the data obtained from a short vs. standard questionnaire are not due to differences in the period of survey implementation.

find that the recall-based reporting on frequent non-food consumption expenditures is negatively affected by increasing the scope of the food consumption module (whether recall- or diary-based) which is administered prior to the non-food consumption module. Given that the questionnaire wording and structure for the non-food consumption module was identical across the food consumption module variants, the authors suggest respondent burden to be the potential culprit behind their finding. Though not reported in their paper but confirmed in private communication with the authors, Beegle *et al*. (2012) also varied the placement of the labor module of the questionnaire in the same household survey experiment and found that the placement of the labor module prior to the food consumption module had a statistically significant (at the 10 percent level) negative effect on reported food and total consumption.[2]

The evidence on the presence of respondent burden and its effects on data quality is quite heterogeneous. The documented effects are ultimately context- and subject-specific, but there are several (some experimental) studies that document (i) question/module placement effects, whether earlier or later in a questionnaire (Johnson *et al*., 1974; Kraut *et al*., 1975; Herzog and Bachman, 1981; Andrews, 1984) and (ii) motivational underreporting during personal interviews in responses to gateway questions as to avoid follow-up questions (Kreuter *et al*., 2011; Eckman *et al*., 2014). The data collection themes across these studies, however, do not overlap with those featured in our analysis. In addition, variations in responses to same questions might occur due to unintended priming that exhibits variation in short versus standard questionnaires. Although most of the questions in the short questionnaire that is tested in our experiment are “factual” in nature, as opposed to personal or subjective, all questions can be primed by other questions or influenced by the interaction between the respondents and the interviewers.

Thus, if independent samples that are drawn from the same population and that are interviewed at the same time, indeed provide different values for the same poverty proxies depending on whether they were subject to a short questionnaire versus its standard counterpart that would be used for establishing the poverty prediction model in a prior period, it is reasonable to expect that the subsequent poverty predictions could be different.

Using a novel experimental setup, this study is the first that provides experimental evidence on this possibility, which is implicitly assumed away in proxy-based poverty measurement exercises if questions underlying proxy definitions are worded identically across short and standard survey instruments.[3] More

[2] The magnitude, statistical significance, and the drivers of these impacts could technically be different than those reported here since their experiment around the placement of the labor module does not represent a shift that is as pronounced as the shift from a standard to a light household survey questionnaire. Newhouse *et al*. (2014) highlight the impact on proxy-based poverty estimation from incomparability of the employment question between standard and short household questionnaires.

[3] Survey mode does not differ between the light and standard household questionnaires used in our experiment. Paper questionnaires were administered by the interviewers in face-to-face interviews. The first data entry was done in the field, and a second data entry with verification was done at the headquarters. There is a rich literature on the comparative effects of survey mode (computer-assisted personal interviewing in face-to-face interviews, telephone interviews, self-administered questionnaires mailed-in or completed online, etc.) that is not covered here. If survey mode differs between the light and standard household questionnaires used in a proxy-based poverty measurement exercise, the variation may affect the proxy measurement and the subsequent poverty predictions.

specifically, the work is based on a randomized household survey experiment that was implemented in Malawi in 2013. The inspiration for the experiment was the discrepancy in the poverty trends based on competing Malawi National Statistical Office (NSO) products during the period of 2004/05–2010/11. Although the direct measurement of household consumption expenditures from the Second and Third Integrated Household Surveys (IHS2 and IHS3) had produced a stagnant headcount poverty trend of 52 percent in 2004/05 and 51 percent in 2010/11, the Welfare Monitoring Survey (WMS)-based poverty predictions that were disseminated between the IHS2 and the IHS3 had implied a steep decline from 50 percent in 2005 to 39 percent in 2009. At conceptualization, the WMS had been designed to provide, among other indicators, poverty predictions on an annual basis in the interim years of the IHS, which is conducted approximately every five years. This objective was fulfilled between the IHS2 and the IHS3 by combining the parameters from a model of household consumption expenditures estimated using the IHS2 with the associated proxies obtained from the 20-page WMS questionnaire that was markedly lighter in inter- and intra-module scope of data collection than the IHS counterpart.[4] The Poverty Predictors module of the WMS was a direct input into the design of the light household survey questionnaire that was at the core of the experiment, as will be explained later.

There are two key findings that emerge from the analysis. First, we find that observationally equivalent households, as well as same households, answer the same questions differently when interviewed with a short questionnaire versus its standard counterpart. The analysis yields statistically significant differences in reporting across all topics and types of questions. The effect is quite pronounced for binary poverty proxies related to consumption of non-food and food consumption items and experience of household shocks. The ordinal categorical variables, particularly those related to subjective welfare and housing, are also impacted by changes in questionnaire design. Second, relying on prediction models based on the national household survey data collected with the standard questionnaire in 2010, we find that the differences in reporting are sufficient to give poverty predictions that are significantly different from each other. The resulting difference in predicted poverty estimates ranges from approximately 3 to 7 percentage points, depending on the model specification. The poverty predictions do not depend on whether the data was collected using the short or the standard questionnaire if we only use the poverty proxies that are elicited prior to the variation in questionnaire design and that include demographic variables from the household roster and location-fixed effects. The findings emphasize the need for further methodological research on module/question placement effects and associated cognitive and behavioral processes. Further, they support the view that light household survey operations designed for proxy-based poverty measurement should judiciously pilot their instruments prior to rollout, in parallel with the questionnaire instruments from which they have evolved.

The paper is organized as follows. Section 2 presents the randomized household survey experiment setup and describes the data. Section 3 shows the

[4] The information on the WMS is available on http://www.nsomalawi.mw/publications/welfare-monitoring-surveys-wms.html.

differences in reporting by survey treatment status and discusses potential reasons for observed patterns. Section 4 evaluates the impact on proxy-based poverty measurement. Section 5 concludes.

2. Data

The methodological experiment on proxy-based poverty measurement (hereafter referred to as "the experiment") was integrated into the Malawi Integrated Household Panel Survey (IHPS) 2013, which was implemented using paper questionnaires and face-to-face interviews. The IHPS attempted to track and resurvey 3,246 households across 204 enumeration areas (EAs) that had been surveyed for the Third Integrated Household Survey (IHS3) 2010/11.[5] The survey was implemented by the National Statistical Office (NSO) had been designed at baseline to be representative at the national, urban/rural, and regional levels, and for the six strata defined by the combinations of region and urban/rural domains. The IHPS targeted all individuals who were part of the IHS3, including those who moved away from the IHS3 dwelling locations between 2010 and 2013. Once a split-off individual was located, the new household that he/she formed or joined since the IHS3 interview was brought into the IHPS sample. The overall IHPS database includes 4,000 households, which could be traced back to 3,104 IHS3 households. Attrition was limited to only 3.8 and 7.4 percent of household and individuals respectively.

The main IHPS fieldwork was carried out during the period of April–October 2013, with residual tracking operations conducted during the period of November–December 2013. The survey was designed with two visits to each household, with approximately three months in between the visits. At baseline, the IHPS EAs had been randomly divided into two halves, known as Sample A and Sample B EAs. Sample A households completed the standard household questionnaire during visit one and only completed an update to the household roster module in visit two. In contrast, Sample B households completed only the household roster module of the standard questionnaire in visit one and completed the full standard questionnaire in visit two. Given the demanding tracking objectives of the survey, the teams managed to implement the two-visit approach for 91.7 percent of the IHPS sample (i.e. 3,667 households). On average, there were 96 days between the two visits.

The standard household questionnaire spanned sixty-six pages and twenty-three modules. Our experiment was administering an additional two-page instrument (included in the Appendix) immediately after the household roster module (i.e. the first module following the cover page) to a subsample of IHPS households during the visit in which the interview would have only necessitated the administration of the household roster module. Toward this end, four households in each IHPS EA, of the households that remained in the original EA between

[5] The IHPS 2013 and the IHS3 were supported by the Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) initiative, whose primary objective is to provide financial and technical support to governments in sub-Saharan Africa in the design and implementation of nationally-representative multi-topic panel household surveys with a strong focus on agriculture. The IHPS 2013 and IHS3 data and documentation are publicly available on www.worldbank.org/lsms.

2010 and 2013, were randomly selected for the experiment and received the additional two-page instrument. Since only households that had remained in the original EA were considered for the experiment, the analysis sample is limited to all households that remained in the original EA between 2010 and 2013 and that were subject to the two-visit approach in 2013. This yields an analysis sample of 2,822 households, of which 765 households were part of the experiment.

Appendix Table X1 provides an overview of the sample. Of 1,428 Sample A households who received the full standard questionnaire in visit one and were revisited in visit two for a household roster update, 393 households received the additional two-page instrument. Similarly, of 1,394 Sample B households, who received only the household roster module in visit one and the full standard questionnaire in visit two, 372 households were administered the additional two-page instrument in visit one. Table X3 in the Appendix presents the sample means for 36 household level attributes computed from the non-experiment modules and the results from the tests of mean differences by whether a household was part of the experiment. No mean differences are statistically significant at the 10 percent level, indicating that the experiment was given to a random sample of households and that any difference in reporting between the two groups can be attributed to the variation in questionnaire design. In addition, there are no differences in item non-response by survey treatment, as the rate of item non-response is only 0.02 percent across all comparable questions in each sample.

Further, 765 experiment households in fact form a subsample of whom the same questions were asked to same households in different questionnaires and at different points in time. The interviews were three months apart with approximately half the sample receiving the standard questionnaire first and vice versa (Appendix Table X1). A separate analysis is done for this subgroup as one can control for both observed and unobserved household characteristics for this group.

In selecting the questions to be included in the two-page instrument for the experiment, inputs were solicited from the Statistics Norway staff that had supported the NSO in producing WMS-based poverty predictions, and the aim was to (i) be able to compute the indicators from the two-page Poverty Predictors module of the WMS questionnaire that was unchanged between 2005 and 2009; (ii) capture the poverty proxies used by past survey-to-survey imputation applications to the Malawi Second Integrated Household Survey (IHS2) 2004/05 data (Houssou and Zeller, 2011); and (iii) include other poverty proxies on food consumption, non-food consumption, and subjective welfare that have been suggested in the literature, but that are not currently used extensively (Christiaensen *et al.*, 2012). The modules that were part of the two-page instrument were selected questions from the following modules in the standard household questionnaire: (i) housing, (ii) food consumption over past one week, (iii) non-food expenditures over past one week and one month, (iv) non-food expenditures over past three months, (v) durable goods, (vi) shocks and coping strategies, and (vii) subjective assessment of well-being. That leaves out 11 other sections in the standard questionnaire that are not included at all in the experiment (Appendix Table X2).

Each question that appears in the standard and experiment questionnaires is identical across these instruments. Both questionnaires included the same household roster module, as the experiment only came after this module. For the housing module, eight questions on ownership of house, quality of roof, toilets, use of cell phone and bed nets were selected out of the 52 questions in the standard questionnaire (Appendix Table X2). The module on subjective well-being was abbreviated in a similar fashion. Regarding the food consumption module, the preprinted list of items was reduced from 124 to 24. In the standard questionnaire, households first answer a yes/no question on the consumption of each food item. Thereafter, additional questions determine the value of each consumed item. In the experiment questionnaire, households only answered the yes/no question for 24 items. The modules on durable goods and non-food consumption with one-month recall and with one-year recall were shortened in a similar fashion, such that the item list was shorter and the follow-up questions seeking to establish values were dropped. In the case of the modules on shocks and non-food consumption with one-week recall, the item list was identical across the standard and experiment questionnaire, but the experiment questionnaire modules only included a yes/no question. (Appendix Table X2). The modules yield a mix of binary, ordered categorical, and continuous poverty proxies, and were administered in the same order in which they appeared in the standard household questionnaire, with the exceptions of the modules on shocks and coping strategies, and subjective assessment of well-being, whose order (was reversed in the two-page instrument for presentation reasons. Since the durable goods module was inadvertently different across survey treatments, the data on the ownership of durable goods was not used in the analysis.[6] A few questions on the ownership of some assets, such as bed nets and cell phones, were included in other questionnaire modules, and the related data are used to analyze reporting differences.

The survey was time-stamped, and Appendix Table X2 presents the median time allocated to the administration of a given module, and the median time the interview had been on-going prior to the administration of each module. The statistics are presented separately for the experiment and standard interviews. The complexity and scope of the standard household questionnaire led to substantially longer interviews, with the experiment taking 23 minutes at the median and 109 minutes for the standard interview. By the time the first poverty proxy question is asked in the standard interview (at the 34th-minute mark at the median), the experiment interview is already conducted in full. The modules on food and non-food consumption, which one seeks to predict in proxy-based poverty measurement, took 25 minutes to implement at the median.

[6] In the IHPS household questionnaire, the ownership of each asset is first established by a yes/no question with the values of 1 and 2 recorded for yes and no answers, respectively. The question on the number of items owned is then asked for assets that are owned. Due to a mistake in the design of the experiment instrument, the yes/no question was dropped, and the question on the number of items owned was included with an instruction for the interviewer to record a value of zero for assets that are not owned. This resulted in an unusual number of experiment households owning two assets in the Visit One data, which led to the discovery of the fact that interviewers were recording a value of 2 in the experiment module for assets that are not owned, similar to the practice followed for the yes/no question in the complex household questionnaire. Although the interviewers were retrained on the correct administration of the experiment module prior to the Visit Two period, we still do not have 100 percent confidence in these data.

3. Analysis: Questionnaire Design's Impact on Reporting

3.1. *Methodology*

To estimate the impact of the tested questionnaire designs on answers given to the identical questions, two types of analyses are undertaken. First, tests are conducted for mean and distributional differences in poverty proxies from the short versus the standard questionnaires. The distributional differences in the poverty proxies by whether or not a household was part of the experiment are assessed through two-sample Kolmogorov-Smirnov tests of equality of distributions. The tests of mean differences are equivalent to bivariate regressions of the following form:

$$y_i = \alpha + \beta e_i \ + \mu_i \tag{1}$$

where *i* stands for household; *y* is a given poverty proxy, which can be a binary, ordinal categorical, or a continuous variable; *e* is the binary variable identifying whether or not a household was part of the experiment; and α and μ identify the constant and the error terms, respectively. The null hypothesis is that that there is no impact from the survey instrument, equivalent to β equal to zero. These regressions are weighted and take into account stratification and clustering as part of the complex survey design. For binary, ordinal categorical, and continuous poverty proxies, Logit, Ordered Logit, and OLS regressions are used, respectively. Given the evidence for the successful randomization of households to standard versus experiment interviews, the results from the bivariate regressions should provide sufficient causal evidence on the questionnaire design's impact on reporting.

Nevertheless, to test the sensitivity of the findings based on Equation 1, multivariate regressions of the following form are estimated:

$$y_i = \alpha + \beta e_i \ + \gamma Z_i + \mu_i \tag{2}$$

where the only difference with respect to Equation 1 is the inclusion of *Z*, a vector of observable household attributes that are computed from the identical, non-experimental modules administered in both standard and experiment interviews prior to the variation in questionnaire design and that also include fixed effects for the months of interview, spanning May through October 2013 and taking April 2013 as the comparison group. The vector *Z* includes the following control variables: (i) household size and sum of household members aged 0–14 and over the age of 65; (ii) age (in years) of head of household; (iii) binary variable identifying female head of households; (iv) binary variables identifying the highest educational attainment among household members, capturing primary, junior secondary, and secondary (and above) educational attainment; (v) binary variables identifying Christian and Muslim head of households; (vi) binary variables identifying Chewa and Tumbuka head of households; (vii) binary variables capturing polygamous, separated, divorced, widowed/widower, never married head of households, (viii) number of months in the last 12 months that head of household has been away; (ix) number of days in the last seven (days that head of household has been away; (x) binary variable identifying rural residence; (xi) binary variables capturing north and south regional location; and (xi) month of interview fixed effects.

Further, all bivariate, distributional, and multivariate analyses are conducted for two different analysis samples: namely A1, which has a total of 2,822 households that were either subject to the standard interview (2,057) or the experiment interview (765); and A2, which is inclusive of only the subset of the (765) households that were part of the standard as well as the experiment interview at two different points in time that are three months apart. (See the sample composition in Appendix Table X1.) The preferred impact estimates originate from the A1 analysis sample that is the foremost and intended product of the experimental design. Recalling that the timing of the standard versus the experiment interviews were randomized for the 765 households that make up the A2 analysis sample and assuming that the variables in vector Z account for potential seasonality in reporting patterns, the results based on the A2 analysis sample are included as a robustness check. Using this sample would account for any remaining observable and unobservable heterogeneity that may jointly predict household survey treatment status and the outcome variables of interest.

3.2. *Reporting differences in standard versus experiment questionnaires*

There are significant discrepancies in how households answer the same questions in different questionnaires. Table 1 reports the number of poverty proxies that are associated with statistically significant different reporting at least at the 10 percent level by standard versus interview status. As expected, the results based on Equation 1 are similar to those based on Equation 2, confirming that the randomization has worked. There are significant differences in a minimum 32 of the 83 variables, equivalent to significant differences in approximately 40 percent of the variables. While the exact numbers vary slightly between the bivariate and multivariate approaches to estimation and between the analyses samples of A1 and A2, they are overwhelmingly consistent and robust across columns 1, 3, 4, and 6. The distributional differences are also present in 21 poverty proxies in the analysis sample A1 and 12 poverty proxies in the analysis sample A2, as reported in columns 2 and 5. There are reporting differences for all types of questions (binary, ordinal categorical, and continuous), and for all topics (food, non-food, shocks, housing, subjective questions, and durable assets). Though not identified explicitly, all variables that exhibit statistically significant differences at the mean and across the distribution based on the analysis sample A2 also exhibit the same sets of statistically significant differences based on the analysis sample A1.

Among housing variables, a statistically significant difference in the reporting is observed only for the toilet type, and not regarding other housing attributes (the roof and floor quality and the number of rooms in dwelling). Experiment households report higher values for the categorical question on the toilet type, which is associated with toilet facilities of a worse quality. The roof and floor type were assessed by the interviewers without asking the respondents. Hence, although there might be a difference depending on whether a question is filled in by asking the household or by the interviewer, this cannot be argued persuasively,

TABLE 1

Module-Specific Poverty Proxies with Statistically Significant Differences in Reporting by Standard/Experiment Interview Status

		# of Poverty Proxies with Statistically Significant Differences in Reporting					
		Analysis Sample: A1			Analysis Sample: A2		
	Total # of Poverty Proxies	(1) Bivariate Regression (Equation 1)	(2) Two-Sample Kolmogorov-Smirnov Equality of Distributions Test	(3) Multivariate Regression (Equation 2)	(4) Bivariate Regression (Equation 1)	(5) Two-Sample Kolmogorov-Smirnov Equality of Distributions Test	(6) Multivariate Regression (Equation 2)
Binary							
Food	22	8	6	7	7	5	8
Non-Food	25	17	11	14	14	7	13
Shocks	23	6	2	6	7	0	6
Ordered Categorical							
Housing	4	1	1	1	0	0	1
Subjective Welfare	4	2	0	2	2	0	2
Durable Assets	4	3	1	2	2	0	2
Continuous							
Cell Phone Expenditures	1	1	0	0	1	0	0
TOTAL	83	38	21	32	33	12	32

Note: The minimum statistical significance level used is 10 percent. Binary poverty proxy related differences in reporting in columns 1 and 3, and in columns 4 and 6 are based on bivariate Logit and multivariate Logit) regressions, respectively. Ordered categorical poverty proxy related differences in reporting in columns 1 and 3, and in columns 4 and 6 are based on bivariate Ordered Logit and multivariate Ordered Logit regressions, respectively. Continuous poverty proxy related differences in reporting in columns 1 and 3, and in columns 4 and 6 are based on bivariate OLS and multivariate OLS regressions, respectively. The regressions control for variables included in vector Z of Equation 2, as specified in Section 3.1. All regressions are weighted, take into account stratification as part of the complex survey design, and have standard errors clustered at the enumeration area level.

as there are only two questions whose answers are recorded in accordance with the interviewer observations.

The four questions on subjective well-being include three questions asking households to place themselves, their friends, and their neighbors on a six-point scale going from poor to rich, and a question asking households if they find their consumption less than, equal to, or more than adequate. In all four questions, experiment households report lower values for the ordinal categorical questions, which is conceptually associated with a worse welfare status. Two of these are significant. We cannot identify with 100 percent certainty why the average predicted poverty rates, as reviewed later in the paper, and the average levels of subjective poverty indicators move in the opposite direction because of the experiment questionnaire treatment. Survey treatment effects could be question-dependent, given differential cognitive burdens or respondents' perceptions of different questions—which we do not clearly understand with the data at hand. The subjective well-being questions are more multi-faceted, and arguably more demanding compared to the questions on food and non-food consumption. This differential complexity, and the difficulty of unpacking the movements in the opposition directions, are signaled in at least three ways. First, only 2 out of 4 subjective wellbeing questions have means that are statistically significantly different between the standard and the experiment samples. Second, the experiment questionnaire treatment in fact does not lead to statistically significant distributional changes in any of the subjective well-being indicators, as reported in Table 1. Lastly, as reported later in the paper, irrespective of the survey treatment, the standard statistical reliability measures for the subjective well-being questions, in addition to those on shocks, are lowest among all categories of questions.

Regarding the four ordinal categorical questions on durable assets (i.e. number of bed nets in the household, number of phones in the household, sets of clothing for the head of household, and the quality of bed sheets for the head of household), a statistically significant difference is recovered only for one (Table 1), which, though not explicitly stated in the table, is the quality of bed sheets for the head of household. Here, the experiment households also report smaller numbers, which is associated with better quality of bed sheets.

For the remainder of the analysis in Sections 3.2 and 3.3, the focus is on the analysis of binary poverty proxies, as explained below, based on Logit regressions. Unless otherwise specified, the presented regression results are the marginal effects and the standard errors associated with the binary variable that identifies whether a household was subject to the experiment or not (i.e. the variable e in Equations 1 and 2). The estimations are weighted, take into account stratification as part of the complex survey design, and have the standard errors clustered at the household level given the pooling of the data at that level. We focus on the binary questions as they constitute the overwhelming majority of the questions, and the observed biases in reporting seem more systematic in the binary questions than the ordinal categorical questions. In the interest of brevity, some results are shown for sample A1 only, since the findings based on the analysis of sample A2 are near-identical. Full results are available upon request. Hence, most of the analysis is based on a sample composed of 2,822 households, of which 765 were part of the experiment.

TABLE 2

Differences in Reporting in Pooled Binary Poverty Proxies by Standard/Experiment Interview Status

	A1			A2		
Analysis Sample	(1)	(2)	(3)	(4)	(5)	(6)
Control Variables	NO	YES	YES	NO	YES	YES
Interviewer Fixed Effects	NO	NO	YES	NO	NO	YES
Experiment	0.027***	0.023***	0.022***	0.027***	0.023***	0.022***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Observations	197,513	197,443	197,443	107,080	107,010	107,010

Note: ***/**/* indicate statistical significance at the 1/5/10 percent level. Experiment is equal to 1 if the household was subject to the experiment questionnaire treatment, and 0 otherwise. The estimations are based on the pooled binary poverty proxy data at the household level. Logit regressions are used and marginal effects are reported. The regressions control for variables included in vector Z of Equation 2, as specified in Section 3.1. All regressions are weighted, take into account stratification as part of the complex survey design, and have the standard errors clustered at the household level given the pooling of the binary poverty proxy data at that level.

For proxy-based consumption and poverty measurement, it matters greatly if the differences in reporting are systematic. Systematic differences in the reporting will lead to systematic bias in the proxy-based poverty measures, while unsystematic reporting differences between the standard and experiment questionnaires would not. Although not shown in Table 1 explicitly, of the 27 binary outcomes that exhibit statistically significant differences in column 3, 24 of them have a higher mean in the experiment sample.

To investigate this pattern further, Equation 1 and the variants of Equation 2 are estimated using the pooled data on all binary poverty proxies, which constitute the majority of the poverty proxies considered for Table 1 and which seem to be associated with more systematic bias in reporting in comparison to non-dichotomous poverty proxies. The results are reported in Table 2. Columns 1, 2, and 3 show, respectively, the findings from the estimations of Equation 1, Equation 2, and Equation 2 augmented with interviewer-fixed effects, all using sample comparison A1. The regressions yield coefficients that have near-identical magnitudes and that are statistically significant at the 1 percent level. Columns 4,5, and 6 show almost identical results using only same households in sample A2. The matching of enumerators to households and questionnaires was not managed, leaving potential room for introduction of some kind of selection bias. However, all but one of the 52 enumerators administered both standard and experiment questionnaires. Further, 66 percent of households that were interviewed twice were interviewed by the same enumerator. The results in Columns 3 and 6 are conditional on enumerator-fixed effects, which lead to minimal variation in the coefficient of interest, indicating that observed reporting results are not driven by interviewer effects. The core results reported in column 2 indicate that the experiment questionnaire treatment, on average, translates into 2.3 percentage point increase in the probability of a positive answer for the binary poverty proxies. At the mean of 25.7 percentage points for the standard sample, this effect is equivalent to 8.9 percent higher reporting in the experiment sample.

TABLE 3

Heterogeneity Of Experiment Questionnaire Treatment Impact On Pooled Binary Poverty Proxies Across Analysis Samples & Questionnaire Modules

	All		Food		Non-Food		Shocks	
Analysis Sample	A1	A2	A1	A2	A1	A2	A1	A2
Overall	0.023***	0.023***	0.026***	0.031***	0.029***	0.032***	0.014**	0.006
	(0.005)	(0.004)	(0.008)	(0.007)	(0.006)	(0.006)	(0.006)	(0.004)
Observations	197,443	107,010	62,060	33,636	70,504	38,210	64,879	35,164
1st Half of Fieldwork	0.023***		0.022*		0.035***		0.010	
	(0.008)		(0.013)		(0.010)		(0.010)	
Observations	98,466		30,952		35,156		32,358	
2nd Half of Fieldwork	0.022***		0.025**		0.023**		0.019**	
	(0.007)		(0.012)		(0.009)		(0.008)	
Observations	98,977		31,108		35,348		32,521	

Note: ***/**/* indicate statistical significance at the 1/5/10 percent level. The reported coefficients and standard errors are those associated with the binary variable identifying whether a household was subject to the experiment questionnaire treatment. The estimations are based on Logit regressions, using the pooled data at the household level for all 70 binary poverty proxies from food, non-food and shocks modules. The regressions control for variables included in vector Z of Equation 2, as specified in Section 3.1. The regressions are weighted, take into account stratification as part of the complex survey design, and have the standard errors clustered at the household level given the pooling of the binary poverty proxies data at that level. The results are robust to including interviewer fixed effects.

3.3. *Heterogeneity in reporting differences*

The systematic higher reporting associated with the binary poverty proxies in the experiment sample is likely to result in systematically different estimation of consumption and poverty. However, is the impact equal for all modules and comparison groups, or is it concentrated in a few? Does it exhibit temporal variation throughout the fieldwork period? Answering these questions might provide insights into the mechanisms driving the reporting differences. The results reported in Table 3 are from Logit regressions with specifications identical to Equation 2. These regressions are based on alternative pooled binary poverty proxy datasets that are split in accordance with the survey module (food consumption, non-food consumption versus shocks) and by whether the data was collected in the first versus second half of the fieldwork.

First, the results are generally not very sensitive to using the data from either the first or the second half of the fieldwork. Second, the positive experiment questionnaire treatment effect on the binary poverty proxies is present in all survey modules, but notably larger among the proxies related to food and non-food consumption. Evaluating the coefficients in the context of the mean from the corresponding module in the standard sample, it is noted that on the whole, the experiment questionnaire treatment corresponds to a higher reporting in the amount of 7.1 percent for food consumption, 12.4 percent for non-food consumption, and 7.9 percent for experience of shocks. Results for same households only (sample A2) are very similar and a little higher for food and non-food. The results vary more for shocks than for the other sections. Here the impact is concentrated in the second half of the fieldwork, and the result is not significant for sample A2.

On the other hand, the impact for non-food is larger in the first half of the fieldwork. In sum, there do not seem to be systematic patterns between the first or second round of fieldwork, which could otherwise have indicated a learning effect.

Of interest is also whether the experiment questionnaire treatment effect varies with household attributes. If it does, the predictions based on poverty proxies that are not immune to the experiment questionnaire treatment are likely to result in a different shape of the consumption distribution, as opposed to a mere level effect. To shed light on this possibility, a variant of Equation 2 is estimated:

$$y_i = \alpha + \beta e_i \ + \gamma Z_i + \theta I_i + \mu_i \tag{3}$$

where the only difference with respect to Equation 2 is the inclusion of I, a vector of interaction terms between e and selected household attributes in the vector Z. The household attributes that are interacted with e include (i) household size, (ii) sum of household members aged 0–14 and over the age of 65, (iii) age (in years) of head of household, (iv) a binary variable identifying female head of households, (v) binary variables identifying the highest educational attainment among household members, and (vi) a binary variable identifying rural residence. These variables were chosen among the list of many possible variables as they are commonly highly correlated with consumption and poverty. The aim of this analysis is to assess if the questionnaire impact varies with household characteristics indicating different impact at different points of the consumption distribution. Since the experiment was not designed to tease out the mechanisms underlying potential differences in reporting, our analysis does not intend to assess who underreports at large. Table 4 reports the results from the estimations of Equation 3 and specifically the marginal effects associated with the interaction terms included in the vector I.

We find that larger households and those residing in urban areas are, on average, more likely to answer yes to questions on both food and non-food consumption when interviewed with the experiment questionnaire (columns 2 and 3). As the number of dependents decline and the household is subject to the experiment questionnaire treatment, the likelihood of reporting positive non-food consumption also increases. On the other hand, the experiment questionnaire treatment effect on the reporting of shocks does not seem to vary by the selected household attributes (column 4). The household attributes that are underlining the statistically significant interaction effects are commonly associated with richer households. To investigate this further, Appendix Table X4 and Figure X1 show how the treatment effect is different in different parts of the consumption distribution. There is indication of a u-shaped effect with poorer and especially richer household having a larger treatment effect, the latter being consistent with Table 4.[7] Hence, the variation in survey design impact is likely to influence both poverty (level of predicted consumption) and inequality (the shape of the consumption distribution). Section 4 test if the impact leads to significant different poverty and inequality measures.

[7] We used the 2010 official consumption aggregates to define household consumption quintiles (we preferred the 2010 data to the 2013 consumption data as it is less related to the 2013 poverty proxies of interest) and subsequently estimated the experiment questionnaire impact on pooled binary poverty proxies in each consumption quintile, following the format of Table 2.

TABLE 4

Heterogeneity of Experiment Questionnaire Treatment Impact on Binary Variables by Analysis Sample

Analysis Sample	A1	A2
Female head of household	−0.005	0.003
	(0.008)	(0.009)
Female head of household*Experiment	−0.011	−0.015*
	(0.011)	(0.008)
Head Age (Years)	−0.001***	−0.001***
	(0.000)	(0.000)
Head Age (Years)*Experiment	0.001**	0.000**
	(0.000)	(0.000)
Highest HH Education: No Education†	−0.063***	−0.060***
	(0.008)	(0.010)
Highest HH Education: No Education†*Experiment	0.020**	0.006
	(0.010)	(0.008)
Highest HH Education: Primary†	−0.025***	−0.035***
	(0.008)	(0.010)
Highest HH Education: Primary†*Experiment	−0.000	0.001
	(0.012)	(0.011)
Household Size	0.006***	0.008***
	(0.002)	(0.003)
Household Size*Experiment	0.010***	0.007***
	(0.003)	(0.002)
Number of dependents in household	−0.006**	−0.007*
	(0.002)	(0.004)
Number of dependents in household*Experiment	−0.009*	−0.007*
	(0.005)	(0.004)
Rural	−0.027***	−0.040***
	(0.007)	(0.008)
Rural*Experiment	−0.022**	−0.011
	(0.010)	(0.009)
Observations	**197,443**	**107,010**

Note: ***/**/* indicate statistical significance at the 1/5/10 percent level. † indicates a binary variable. The standard errors are in parentheses. The reported coefficients are marginal effects associated with the interactions between the selected household attributes and the binary variable identifying whether a household was subject to the experiment questionnaire treatment. The estimations are based on Logit regressions, using the pooled data at the household level for all 70 binary poverty proxies from food, non-food and shocks modules. The regressions control for variables included in vector Z of Equation 2, as specified in Section 3.1. The regressions are weighted, take into account stratification as part of the complex survey design, and have the standard errors clustered at the household level given the pooling of the binary poverty proxies data at that level. The results are robust to including interviewer fixed effects.

3.4. *Which questionnaire is providing "correct" data?*

Although essential, it is unfortunately an inquiry that we cannot definitively answer. While some questions may present opportunities for validation, such as those on cell phone expenditures and the presence of bed nets and sheets, responses to other questions, such as those on food and non-food consumption, are difficult to externally validate, free of potential biases introduced by the variation in the measurement approach itself. One option is to gauge the reliability of data through Cronbach's Alpha coefficients estimated for different survey modules and across different analysis samples, as presented in Table 5.

There is a tendency for lower alphas in the standard questionnaire compared to the experiment sample for sections where the regression analysis also yields a

TABLE 5

Cronbach's Alpha Estimations by Module and Sample

Module	Experiment Sample	Standard Sample A2	Standard Sample A1
Food	0.80	0.74	0.72
Non food	0.81	0.68	0.71
All Consumption	0.88	0.82	0.81
Shocks	0.61	0.62	0.64
Housing	0.82	0.83	0.83
Subjective Welfare	0.65	0.64	0.60
All Questions	0.89	0.88	0.86
Simple Mean Across All Modules	0.76	0.72	0.72

larger impact, notably in modules food and non-food consumption. This would indicate that the standard questionnaire might measure consumption less reliably than the experiment sample. However, this is of limited use, as the experiment sample did not measure actual consumption, but only whether each item was consumed or not. Hence, it is not a better alternative per se than the standard questionnaire. The remaining sections have very similar alphas across different samples and do not, therefore, present further guidance on whether any of them might be preferred for reliability purposes.

3.5. *Potential reasons for reporting differences*

One can imagine several reasons why reported answers differ significantly between a short and a long questionnaire. Unfortunately, data is not sufficient to provide concrete results on this, but the following reflections are among potential explanations. The length of the questionnaire could be one reason. As both interviewer and interviewee become more tired, the volume and quality of reporting might decline such that the reporting differences could be larger for later modules than earlier modules. (See Appendix Table X2 and the large variation in time elapsed before each section between the two questionnaire designs.) The findings reported in Table 3, however, do not support such a hypothesis as the magnitude of the experiment questionnaire treatment effect on questions appearing in the later modules, such as shocks, is not larger than those observed for questions appearing in earlier modules, such as food and non-food consumption. This implies that interview length alone cannot explain the discrepancies.

Further, sensitivity to change in the overall questionnaire design might vary by "question type". The binary variables are subject to the largest survey treatment effects, and the experiment versions of their respective modules were also the ones where the change in the immediate context of the question was the largest. For instance, in the standard questionnaire, the food consumption module has additional questions to establish value of consumption, which are not included in the experiment questionnaire. Similar adjustments were made to the modules on non-food consumption and shocks in the context of the experiment. Hence, if standard questionnaire respondents realized the higher likelihood for follow-up questions conditional on answering yes to the screening question and intentionally underreported with respect to their counterparts subject to the experiment, this could potentially explain the findings. If such learning took place, it likely took place through visual inspection of the paper questionnaire, and not through learning from the repeated interviews. We argue this since, as reported in Table 3, the results are not sensitive to using the data from either the first half or the second half of the fieldwork (i.e. including in the sample experiment households that had received the standard questionnaire three months prior).

Another mechanism at work could be that interviewers may have exerted different levels of effort while administering different questionnaires. One could speculate that with a shorter list of items that are not coupled with follow-up questions, interviewers may have been more dedicated. Since survey treatment effects in Table 2 did not change after including interviewer-fixed effects in the regressions, such variation in effort would have to be similar for all interviewers. This variation in effort would also be a source of bias in a typical proxy-based poverty estimation (though not in the experiment) that relies on a different set of interviewers at two different points in time for different questionnaire instruments.

Finally, priming could have, inadvertently, taken place. Effect of question priming is well-illustrated by Strack *et al.* (1988). They first ask students (i) “How happy are you with your life in general?”; then, secondly (ii) “How many dates did you have last month?” The simple correlation between these two questions was insignificant −0.01. However, after reversing the order of the questions, the correlation between the two questions increased to 0.66 for another set of students. The change in reporting is argued to come about as the students link the number of dates they have been on with general life happiness. In the experiment, entire sections of the standard questionnaire were left out, and hence the order of questions was very different.

4. Analysis: Questionnaire Design’s Impact on Poverty Measured by Proxies

Numerous methods to proxy poverty via proxies already exist. We focus on methods that rely on a consumption regression to deduct proxy weights (i.e. beta coefficients), as exemplified by

$$y_i = \beta_j x_{ij} + \varepsilon_i \qquad (4)$$

where y_i is log household consumption expenditures per capita (hereafter referred to as consumption), x_{ij} the vector of proxy variables, and β_j the coefficients (weights) of interest. Examples of such methods include Elbers *et al.* (2003), Tarozzi (2007), and Mathiassen (2013).

To predict consumption, and in extension thereof poverty and inequality, the prediction methods developed in Elbers *et al.* (2003) are used. This prediction method has the advantage of also producing standard errors of poverty and inequality estimates, and implementation is tractable with the publicly-available PovMap software. To ensure that the results are not model-driven and to gauge the sensitivity of poverty and inequality predictions to differences in questionnaire design, consumption is predicted with four different models of varying poverty proxy scope. In all cases, the model is estimated using the IHS3 data and the IHS3 subsample interviewed during the months of April-October (i.e. the implementation period for the IHPS), and consumption predictions are obtained using the IHPS data. To compute predicted poverty rates, the official IHS3 poverty line of 37,002 Malawi Kwacha per person per year is used.

The first prediction model is the original poverty prediction model used in Malawi based on the WMS. The model is updated with coefficients from the IHS3 data. In the other three models, variables were selected in PovMap by *stepwise*; a statistical method used to avoid selection by researchers. Although the accuracy of the models is not the main interest in comparing predictions based on observationally equivalent samples that are subject to different survey treatment, the complete set of results from the prediction models is provided in the Appendix Tables X5 through X8. The list of possible poverty proxies included in each of the four prediction models are as follows:

1. Experiment only: only variables derived from the experiment modules that are administered following the household roster;
2. Experiment and non-experiment: variables derived from the experiment modules as well as demographic, education and locational variables computed from the modules administered prior to the experiment modules;
3. WMS-linked poverty proxies as specified in NSO (2005);[8] and
4. Non-experiment only: demographic, education, and locational variables computed from the modules administered prior to the experiment modules.

Table 6 presents the differences in the predicted headcount poverty rates and Gini coefficients across different models and sample comparisons. The predicted poverty rates and Gini coefficients across scenarios are available upon request. Overall, the variation in questionnaire design is sufficient to generate significant different estimates of both poverty and inequality. Using models 1 through 3, the predicted poverty rate based on the experiment sample is 3 to 7 percentage points *lower* than the predicted poverty rate based on the standard sample (column 1). In all three cases, the predicted poverty rate based on the experiment sample is outside the estimated 95 percent confidence interval for the predicted poverty rate based on the standard sample. Similar movements are observed in the predicted Gini coefficients, which are 3 to 4 percentage points *higher* in the experiment sample (column 1). Lower predicted poverty and higher predicted inequality originating from models 1 through 3 are consistent with the heterogeneity of short questionnaire impact highlighted during the discussion of Table 4, specifically the fact that the household attributes underlining the statistically significant interaction effects in Table 4 are those that are commonly associated with richer

[8] Three variables based on actual expenditures for cooking oil, sugar, and soap are not included due to the need to rely on consumer price index series to adjust them over time. In private communication with Astrid Mathiassen, we were able to confirm that the exclusion of these variables from the WMS model does not affect the poverty predictions based on the annual WMS data from 2005 to 2008. We also exclude three binary variables on ownership of bed, iron, and refrigerator due to the aforementioned issues in the data collection on durable asset ownership as part of the experiment.

TABLE 6

Differences in Poverty Headcount and Gini Coefficients Resulting from Changes in Questionnaire Design

Model/Column	(1)	(2)
	Differences in Headcount Poverty Rate Predictions	
1. Experiment Only	**0.05**	0.01
2. Experiment & Non-Experiment	**0.07**	0.00
3. WMS Model	**0.03**	0.00
4. Non-Experiment Only	0.01	0.00
	Differences in Gini Coefficient Predictions	
1. Experiment Only	**−0.03**	−0.01
2. Experiment & Non-Experiment	**−0.04**	−0.01
3. WMS Model	**−0.03**	−0.01
4. Non-Experiment Only	−0.01	−0.01

Note: Column 1 presents the difference between the prediction from standard interviews and the prediction from experiment interviews. Column 2 presents the difference between the prediction from standard interviews of non-experiment households and the prediction from standard interviews of experiment households. Bold indicates scenarios in which the experiment sample based prediction is outside of the 95 percent confidence interval for the prediction based on the comparator sample (standard interviews for columns 1 and standard interviews of non-experiment households in column 2).

households. Not reported but available upon request are the differences obtained by using the analysis sample A2, which are near-identical to those reported in column 1.

On the other hand, working with model 4 (i.e. only with the poverty proxies that are solicited prior to the variation in questionnaire design), there is only 1 percentage point difference in the predicted poverty rate and Gini coefficient between the experiment and standard samples, and the difference is no longer statistically significant. Moreover, looking at column 2, none of the differences between the predictions from the standard interviews of the non-experiment households and the predictions from the standard interviews of the experiment households are statistically significant. Hence, there is strong evidence that the variation in the predicted poverty and inequality statistics is driven by the variation in questionnaire design underlying the poverty proxy definitions.

5. Conclusion

Our key finding is that observationally-equivalent as well as same households answer the same questions differently when interviewed with a short questionnaire vs. the longer counterpart. We find statistically significant differences in reporting across all topics and types of questions, particularly those related to consumption of non-food and food items, experience of household shocks, subjective welfare, and housing. Relying on prediction models based on the national household survey data collected with the standard questionnaire in 2010, we find that the differences in reporting are sufficient to give predicted poverty rates and Gini coefficients that are significantly different from each other. While the difference in predicted poverty estimates ranges from approximately 3 to 7 percentage points depending on the model specification, restricting the proxies to those that are determined prior to the variation in questionnaire design predicts the same poverty rates in both samples.

Although the poverty proxy comparisons are made across different samples without the luxury of the truth, this point matters less in this case precisely because of the focus on proxy-based poverty measurement. The analyst, who would employ the method in the interim years of a household consumption survey, also does not know the truth, and would work under the assumption that the available short household survey data would be consistent with the data that would have been collected through the same complex household survey that had generated the poverty prediction model. The short household survey instrument tested in the experiment is one variant out of many that would have been deemed, prior to implementation, sensible and feasible by the research community focused on proxy-based poverty measurement. Abstracting away from possible interview-mode effects, the findings should also be of interest to those thinking of using new technologies, such as mobile phones, for collecting consumption or poverty proxy data through succinct interviews.

Furthermore, two broader points relate to direct consumption measurement in household surveys. First, in the case of Malawi, we have shown that the standard questionnaire modules on food and non-food consumption that we seek to proxy take less than 25 minutes to administer as a package at the median. Thus, with respect to a household survey for proxy-based poverty measurement, collecting consumption data, in and of itself, may not be as complex and costly as commonly perceived. Here, "perceived" is the operative word as the cost savings in implementing household surveys with a poverty focus net of consumption data is not rigorously documented due to lack of/or weaknesses in comparative budgetary and survey process data.

Second, the differences in the propensity to consume food and non-food consumption items suggest that consumption in the standard sample might have been different from consumption in the experiment sample. While we do not have evidence on the relative accuracy of reporting from the experiment and standard samples, underreporting of consumption is usually assumed to be the main problem in the literature. (See, for instance, Beegle *et al.*, 2012.) In our case, consumption in the standard sample would appear to be underreported. Counterexamples of systematic overreporting might exist, though we are unaware of any from general populations in developing countries. Interestingly, the poverty proxy approach leads to lower estimated poverty in the experiment sample, while the reported subjective poverty is higher in the experiment sample, relative to the standard sample. Hence, the survey instrument impacts both poverty indicators, but in opposite directions.

If there is misreporting or underreporting in y_i in equation (4) so that $y^{Standard}$ and $y^{Experiment}$ are systematically different from each other, and the same is observed for at least some proxies (x), then β will be biased as well. Based on the results in Table 3 and Table 4, it would seem reasonable to assume that the misreporting in x and y are correlated and have means different from zero. With measurement errors on both sides of the regression, there are no boundaries on size or direction of bias in β (Bound *et al.*, 2001). Although direct measurement of consumption in household surveys is often considered as the best approximation for true consumption, we can only note that the propensity for reporting consumption is sensitive to questionnaire design, and that consumption regressions from such surveys could be biased due to misreporting.

In future methodological experiments, comparable questionnaire modules could be assigned different orders for different random subsets of the samples that receive experiment versus standard questionnaires, holding the content of the modules, the order of questions in each module, and the interview mode constant. This would, in turn, provide an opportunity to assess whether the reporting differences hold uniformly irrespective of module placement. Similar exercises could be carried out to assess the effect of the order of key questions, holding the content of the modules, the order of modules, and the interview mode constant in alternative questionnaire instruments. These efforts could be complemented by the applications of pretesting techniques, such as cognitive interviews and behavior coding, that could help illuminate cognitive and behavioral processes that play out in answering the same questions as part of different questionnaires (Presser *et al.*, 2004). Moving forward, household survey operations designed for proxy-based poverty measurement should, prior to full rollout, consider piloting their instruments in parallel with the questionnaire instruments from which they have evolved. This methodological exercise could be designed as a randomized household survey experiment to test whether the data for poverty predictors differ depending on whether they were solicited in an experiment versus a standard questionnaire.

References

Andrews, F. M., "Construct Validity and Error Components of Survey Measures: A Structural Modelling Approach," *Public Opinion Quarterly*, 48, 409–42, 1984.

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson, "Methods of Household Consumption Measurement Through Surveys: Experimental Results from Tanzania," *Journal of Development Economics*, 98, 3–18, 2012.

Bound, J., C. Brown, and N. Mathiowetz, "Measurement Error in Survey Data," In J. Heckman and E. Leamer (ed.), *Handbook of econometrics*, Elsevier, Amsterdam, 5, 3707–3843, 2001.

Christiaensen, L., P. Lanjouw, J. Luoto, and D. Stifel, "Small Area Estimation-Based Prediction Methods to Track Poverty: Validation and Applications," *Journal of Economic Inequality*, 10, 267–97, 2012.

Douidich, M., A. Ezzrari, R. Van der Weide, and P. Verme, "Estimating Quarterly Poverty Rates Using Labor Force Surveys: A Primer," World Policy Research Working Paper No. 6466, 2013.

Eckman, S., F. Kreuter, A. Kirchner, A. Jaeckle, R. Tourangeau, and S. Presser, "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys," *Public Opinion Quarterly*, 78, 721–33, 2014.

Elbers, C., J. O. Lanjouw, and P. Lanjouw, "Micro-Level Estimation of Poverty and Inequality," *Econometrica*, 71, 355–64, 2003.

Herzog, R. A. and G. J. Bachman, "Effects of Questionnaire Length on Response Quality," *Public Opinion Quarterly*, 45, 549–59, 1981.

Hess, J., J. Moore, J. Pascale, J. Rothgeb, and C. Keeley, "The Effects of Person-Level Versus Household-Level Questionnaire Design on Survey Estimates and Data Quality," *Public Opinion Quarterly*, 65, 574–84, 2001.

Houssou, N. and M. Zeller, "To Target or Not to Target? The Costs, Benefits, & Impacts of Indicator-Based Targeting," *Food Policy*, 36, 627–37, 2011.

Johnson, W. R., N. A. Sieveking, and E. S. Clanton III, "Effects of Alternative Positioning of Open-Ended Questions in Multiple-Choice Questionnaires," *Journal of Applied Psychology*, 59, 776–8, 1974.

Kraut, A. I., A. D. Wolfson, and A. Rothenberg, "Some Effects of Position on Opinion Survey Items," *Journal of Applied Psychology*, 60, 774–6, 1975.

Kreuter, F., S. McCulloch, S. Presser, and R. Tourangeau, “The Effects of Asking Filter Questions in Interleafed Versus Grouped Format,” *Sociological Methods & Research*, 40, 88–104, 2011.

Mathiassen, A., “Testing Prediction Performance of Poverty Models: Empirical Evidence from Uganda,” *Review of Income and Wealth*, 59, 91–112, 2013.

National Statistical Office (NSO), “Welfare Monitoring Survey (WMS) Report,” 2005. Retrieved from https://goo.gl/xUzfU7.

Newhouse, D., S. Shivakumaran, S. Takamatsu, and N. Yoshida, “How Survey-to-Survey Imputation Can Fail,” World Bank Policy Research Working Paper Series No. 6961, 2014.

Presser, S., M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. “Methods of Testing and Evaluating Survey Questions,” *Public Opinion Quarterly*, 68, 109–30, 2004.

Ravallion, M., *The Economics of Poverty: History, Measurement, and Policy.* Oxford University Press, New York, 2016.

Strack, F., L. Martin, and N. Schwarz, “Priming and Communication: Social Determinants of Information use in Judgments of Life Satisfaction,” *European Journal of Social Psychology*, 18, 429, 1988.

Tarozzi, A., “Calculating Comparable Statistics from Incomparable Surveys, with an Application to Poverty in India,” *Journal of Business & Economic Statistics*, 25, 314–36, 2007.

Tourangeau, R., L. J. Rips, and K. Rasinski. *The Psychology of Survey Response.* Cambridge University Press, New York, 2000.

Vu, L. and B. Baulch, “Assessing Alternative Poverty Proxy Methods in Rural Vietnam,” *Oxford Development Studies*, 39, 339–67, 2011.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher’s web-site:

APPENDIX

Table X1: Sample Split by Visit and Interview Type

Table X2: Module & Interview Durations by Standard/Experiment Interview Status

Table X3: Sample Means by Household Experiment

Table X4: Heterogeneity of Experiment Questionnaire Treatment Impact on Pooled Binary Poverty Proxies by (2010) Household Consumption Quintile

Figure X1: Experiment Questionnaire Treatment Effects by (2010) Consumption Quintile

Table X5: Prediction Model 1- Experiment Only

Table X6: Prediction Model 2 - Experiment & Non-Experiment

Table X7: Prediction Model 3 - WMS Model

Table X8: Prediction Model 4 - Non-Experiment Only

Experiment Questionnaire Modules