Exhaustive or Exhausting? Evidence on Respondent Fatigue in Long Surveys*

Dahyeon Jeong Shilpa Aggarwal Jonathan Robinson Naresh Kumar
Alan Spearot David Sungho Park

January 8, 2022

Abstract

Living standards measurement surveys require sustained attention for several hours. We quantify survey fatigue by randomizing the order of questions in in-person surveys (lasting 2.5 hours on average) fielded in an evaluation of cash transfers in rural Liberia and Malawi. An additional hour of survey time increases the probability that a respondent skips a question by 10-64%. Because skips are more common, the total monetary value of aggregated categories such as assets or expenditures declines as the survey goes on, and this effect is sizeable for some categories: for example, an extra hour of survey time lowers food expenditures by 25%. Evidence from a similar experiment within high-frequency phone surveys shows that the results are not driven by the respondents deliberately choosing to skip questions in order to hasten the end of the survey, suggesting that cognitive burden is the key driver of survey fatigue.

JEL Codes: C83, C93, O12
Keywords: Survey Fatigue, Measurement, Survey Methodology, Design of Experiments

*The author order is randomized. Jeong: World Bank (dahyeonjeong@worldbank.org); Aggarwal: Indian School of Business and J-PAL (shilpa_aggarwal@isb.edu); Kumar: University of California, Santa Cruz (nkumar5@ucsc.edu); Park: University of California, Santa Cruz (davidspark@ucsc.edu); Robinson: University of California, Santa Cruz, BREAD, CEGA, J-PAL and NBER (jmrtwo@ucsc.edu); Spearot: University of California, Santa Cruz (aspearot@ucsc.edu). We thank USAID for funding. We are grateful to Jenny Aker for her collaboration and Sanjana Gupta for outstanding research assistance. For organizing the data collection, we thank Joseph Davis, Arja Dayal, Wilson Dorleleay, Walker Higgins, Andreas Holzinger, Erik Jorgensen, Teresa Martens, Laura McCargo and Camelia Vasilov at IPA Liberia, and Patrick Baxter, Emanuele Clemente, Calvin Mhango, Monica Shandal, Patrick Simbewe, and Asman Suleiman at IPA Malawi. We are extremely grateful to all the enumerators who collected this data in both countries, though there are too many to list individually. We thank seminar participants at UCSC and the IPA-GPRL Methods Conference for helpful comments. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not necessarily represent the views of USAID, the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

1 Introduction

Many of the surveys that are administered in development economics or by multilateral agencies such as the World Bank to measure poverty or as part of evaluations are long and complicated, and require the sustained attention of a respondent for several hours. For any researcher who has observed such a survey, it is clear that some respondents disengage as the survey drags on, because they are exhausted, bored, or because their attention wanders. As a result, response quality during the later part of a long survey may suffer, a phenomenon known as survey fatigue.

While survey fatigue is well-documented in the literature,[1] until recently there has been comparatively little research to rigorously quantify its effects. In this paper, we provide such a quantification by randomizing the order in which modules appear in a long survey, generating exogenous variation in the time-into-survey when a particular question was asked. This random order of questions allows us to compare responses to the *same* question when it is asked sooner in the survey versus when it is asked later, and quantify the divergence in responses. We conduct this experiment within surveys administered at baseline and endline for a randomized evaluation of cash transfers in rural Liberia and Malawi (Aggarwal et al. 2020). These surveys were long, averaging about 2.5 hours, and the experimental randomization induced meaningful variation in the time it took to reach a specific question: the average time to reach a specific question was changed by as much as about 30 minutes as a result of the randomization.

We have two main findings. First, and consistent with other work, we find clear evidence of survey fatigue. We estimate survey fatigue separately for two ways of asking questions. The first is an "open-ended" method which we used for those questions in which there is no top code or pre-listed set of options. For example, for transfers given out, respondents were asked to provide the number of transfers that they gave, and could list as many or as few as they wanted. For such questions, we find that each additional hour of surveying causes a 26-64% decrease in the number of items listed. The second method, or "fixed list" method, is one in which the list of items was precoded. For example, in the food expenditures section, we generated a list of around 35 food items and asked about each of these separately. Survey fatigue might be reduced with this method, if the listing serves as a memory aid for those who need help with recall later in the survey as they begin

[1] For example, survey fatigue has its own entry in the Encyclopedia of Survey Research Methods.

to tire out. In addition, for some categories (such as food expenditure), there are minimal follow-up questions so that listing a value of zero would not reduce survey length substantially. However, we still observe survey fatigue in this method, though much less than in the prior method: for every additional hour, respondents are about 10-19% more likely to report no value for a given question. While survey fatigue appears less prevalent when using the fixed list method, we are unable to definitively attribute this to the question type, since the method is not random – it is also possible that these categories are less subject to survey fatigue.

Second, we quantify the extent to which this skipping reduces the *value* of aggregate categories such as the total value of transfers or expenditures. For any skipped question, the value of that category would be set to zero by default, and so we would expect survey fatigue to lower aggregated values. This effect might be modest if the categories that are skipped tend to be more marginal. However, the effects we find are sizeable: for example, an additional hour of survey time reduces the value of food expenditures by 25%, and has even larger effects (in percentage terms) on smaller categories (such as transfers).

This paper contributes to a recent literature that experimentally evaluates the effect of survey time on survey fatigue. Laajaj and Macours (2021) randomize the order of cognitive, non-cognitive and technical questions in a sample of farmers in Western Kenya but, unlike us, find no effect of survey time on reporting. Two other papers were conducted contemporaneously to this study, and find similar results to ours. Ambler et al. (2021) randomize the order of a household labor supply module, where questions are asked about the labor supply of each household member, but the order in which the household members are listed was randomized. The authors find a 2% reduction in the number of activities reported when a household member is moved back by one position in the household roster. Abay et al. (2021) employ a methodology similar to ours, in which the authors randomize the placement of a dietary diversity module within a phone survey in Ethiopia. Like us, they find large effects: a 15 minute increase in survey time before the module leads to an 8-17% decline in reported dietary diversity.[2] Finally, in a similar but different design in a different context, Backor et al. (2007) conduct a web-based time-use survey in the US in which an extra question is

[2]Another related paper is Kilic and Sohnesen (2019), who find that poverty incidence differs when measured in a short or a long survey in Malawi. However, in their case, since everybody got the same long survey or the same short survey, it is not possible to disentangle the effects of survey length from those of question order, i.e., when your responses are impacted by a question being preceded by another question (see here).

included at a random order, creating variation in how many hours had already been asked about when a particular question appeared in the survey. Similar to these other papers, the authors find that asking about an additional hour lowers the number of activities reported in each subsequent hour by 5 percentage points.

Our experiment also furthers the literature by helping us rule out some of the explanations for *why* survey fatigue occurs. Past research suggests that survey fatigue may be driven by people deliberately choosing to not answer questions in order to expedite the end of the survey, or if people become more likely to inadvertently make mistakes as they become tired. Some researchers have also conjectured that, over time, respondents learn that answering "no" to a question often invokes a skip code that will allow them to skip a number of follow-up questions. This behavior, known as "satisficing," has been documented in survey settings (Krosnick 1991). To examine this, we also randomized the order of modules within phone surveys that we conducted with respondents repeatedly every 2 months over a 6-month period. These surveys took about 30-40 minutes to complete. If satisficing were present, we would expect that respondents would become familiar with the structure of the surveys, including the mechanics of skip patterns over time as they go through multiple rounds of the survey. As an expected result, fewer questions would be answered right from the outset during the later rounds of the phone surveys, and there would be no evidence of experimental survey fatigue *within* a survey round. However, our evidence is not consistent with satisficing: we find evidence of survey fatigue similar to our baseline and endline surveys in the 2-3 rounds of the phone survey. Our results suggest that fatigue is very likely driven by an increase in cognitive burden as the survey progresses.

The rest of this paper proceeds as follows. Section 2 explains the data and experimental design, Section 3 presents results, and Section 4 concludes.

2 Data and Experimental Design

2.1 Setting

We use data from baseline and endline surveys conducted as part of a cash transfer RCT with the NGO GiveDirectly in Liberia and Malawi. In the experiment, the treatment group received cash transfers via mobile money. The average amount of the transfer was $500; however, the amount

and other implementation details were varied experimentally – see our trial registry on the AEA website (Aggarwal et al. 2020) for more details on the design of the underlying experiment.[3]

In each country, the project took place in rural areas, with universal targeting in treatment villages (i.e. all households in treatment villages received transfers). For this reason, the total allocation to a village depends on its size; to ensure liquidity, the NGO decided to only include villages which were small. Operationally, we set a population threshold based on the most recent population census.[4] In Liberia, the study takes place in Bong and Nimba Counties; in Malawi, it takes place in Chiradzulu and Machinga Districts. In each country, the project enrolled 300 villages, with half selected for treatment.

In each village, we attempted to enroll 10 households into the survey sample.[5] We chose to target women for the study, though many questions were asked at the household level. Male heads were interviewed only when the female was not present, and would not be reachable within a few days; our sample was ultimately 76% female in Liberia and 94% in Malawi.

Two of the 10 sampled households in each village were further randomly sampled to participate in a monthly panel survey that was conducted over the phone and was designed to measure a pre-defined set of outcomes at a high frequency. While the major focus of these surveys was to measure food security, they also included questions on income, labor supply, transfers, savings, and credit. We designed these surveys such that each household was called every other month, but the 2 households in each village alternated months, such that each village provided a data point every month. The phone surveys took about 30-40 minutes to complete.

Figure A1 shows the timeline of project activities.

2.2 Question order randomization

This experiment takes place within baseline and endline surveys which are similar to World Bank LSMS surveys and take about 2.5 hours to complete on average. The surveys contain 19 self-

[3] In both countries, the size of the transfer was varied between $250, $500, and $750. In addition, in Liberia, cash was disbursed either as a "lump-sum" or via quarterly payments. However, even the lump sum was disbursed in increments of $250 per month, so that cash was paid out over 3 months for the largest transfer.

[4] In Malawi, the upper threshold was 100 household per village according to the 2008 national census. In Liberia, we conducted the experiment in two cohorts; the first cohort included villages that had up to 25 households in the 2008 national census, and the threshold for the second cohort was 125, reflecting the larger village sizes in the study region.

[5] It was not always possible to enroll 10 households per village. The total sample size is 2,715 in Liberia and 2,944 in Malawi

contained sections, including household demographics, agriculture, income, expenditures, savings, assets, labor supply, shocks, and other topics.[6] The beginning of the survey (which included household identifying information, demographics, and agriculture) and the end of the survey (which had a section on intimate partner violence, followed by the collection of household tracking information) were the same across all versions. The remaining sections were grouped into 3 modules, and the order of these 3 modules was randomized, giving us 6 versions of the survey (which we refer to as versions A-F – see Figure A3). The survey software records the amount of time elapsed (since beginning) at each question, allowing us to calculate the specific time at which a question appeared in the survey.

The amount of time it takes to progress through the survey varies depending on a number of factors, including respondent and enumerator characteristics, and the details of a household's circumstance. For example, because our survey had a focus on agriculture, a household which grew multiple crops would be asked a number of questions about each one of them. Table A1 shows information on the average survey duration. The baseline and endline surveys took on average 2.3 and 2.7 hours respectively in Liberia; and 3 and 2.8 hours respectively in Malawi. The standard deviation in survey time is sizeable, ranging from 0.7 to 1.1 hours. Figure A4 shows a CDF of the time until completion of different points of the survey (using survey Version A only) for both countries and for both baseline and endline pooled together (i.e., for 4 country-survey combinations). The figure shows CDFs for various quantiles in the survey time distribution (i.e. relative to completing the question which makes up the p-th percentile of the overall distribution of time to survey completion). The CDFs show that even 10% into the survey, the standard deviation of time is already over 30 minutes and that for all percentiles, there are surveys that take a large amount of time. For example, about 10% of people take over 3 hours to even get halfway through the survey (Panel C).

Finally, although not the main focus of this paper, we also randomized survey order for the final 2-3 rounds of the phone survey. In order to do this, we randomized the location of the Expenditures and Transfers sections to appear at either the very beginning or the very end of the survey, and the order between the two sections, generating 4 possible permutations of section order within the survey (Figure A5). We return to this randomization in the discussion section, when we discuss

[6] See Figure A2 for the list of sections.

possible explanations for survey fatigue.

Table 1 shows the effect of the randomized survey versions on the time until which the first question of each section was administered. The reported means and standard deviations at the bottom of the table are those pertaining to that section for Version A of the survey. As can be seen from this table, the module randomization introduced significant variation in the time-into-survey when a section starts. For example, looking at Column 1, we can see that the Assets section started just after the 80th minute on average for those who got Version A of the survey. However, the full range for when this section started ranges from 77th minute (version B) to 107th minute (Version F) - a difference of 30 minutes. This range of about 30 minutes is consistently observed across all sections.

We use the survey version that was used for each respondent as an instrument for the time-into-survey when a particular set of questions began to be asked of that respondent. The first-stage F-statistics are shown at the bottom of Table 1, and range from 35 to almost 200.

Table 1: Experimental variation in time before which sections were administered

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Time into survey (minutes) at the beginning of following section:						
	Module 1				Module 2	Module 3	
	Assets	Savings	Credit	Transfers	Expenditure	Shocks	Contributions
Version B	-3.30***	-2.84**	-2.79**	-2.63*	6.39***	-12.20***	-8.83***
	(1.28)	(1.36)	(1.36)	(1.37)	(1.37)	(1.47)	(1.50)
Version C	19.21***	17.71***	17.54***	17.67***	-16.83***	-4.74***	-3.24**
	(1.27)	(1.36)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version D	23.96***	22.52***	22.24***	22.35***	-18.00***	-16.52***	-5.45***
	(1.27)	(1.36)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version E	6.61***	6.01***	5.67***	6.04***	5.79***	-25.74***	-15.79***
	(1.27)	(1.35)	(1.36)	(1.37)	(1.36)	(1.46)	(1.50)
Version F	26.06***	24.56***	24.01***	24.07***	-8.38***	-27.57***	-14.49***
	(1.28)	(1.36)	(1.36)	(1.38)	(1.37)	(1.47)	(1.50)
Version A: Mean	80.01	93.47	93.89	95.61	109.39	125.53	134.72
Version A: SD	38.78	40.34	40.26	40.87	44.23	47.39	49.78
F-statistic: joint significance	197.30	151.42	146.55	143.33	127.92	114.25	35.05
Number of respondents	5,591	5,597	5,597	5,597	5,597	5,597	5,592
Observations	10,153	10,226	9,952	10,228	10,227	10,224	10,154

Note: The omitted category is version A. Observations include in-person baseline and endline survey data. Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

2.3 Respondent characteristics and randomization check

Table 2 presents summary statistics for several basic demographic indicators, as well as comparisons across treatment groups. We present these statistics only for those indicators which were asked before the module randomization kicked in as the variables from the later sections would by definition be imbalanced under our central hypothesis for this paper. We show the balance across versions separately for the baseline and endline surveys, but pool them across the 2 countries. For each survey (baseline or endline), we show the mean and standard deviation (for non-binary variables) pertaining to Version A of the survey (chosen arbitrarily), followed by the p-value for the joint test of equality across all 6 versions of the survey. Panel A shows respondent characteristics. Almost 90% of the sample is female, three-quarters are married, and the average age is 41. Average years of education (for the respondent) is only 4.2, and 57% are literate in English (these last 2 variables were measured at baseline only).

Panel B shows household statistics. At baseline, the average household has 4.8 members, and 96% were engaged in farming. About 40% of the sample live in a house with a thatch roof, and 80% live in a house with a mud floor. About 77% own their dwelling and only 2% have electricity. We cannot reject equality across treatments for all of these variables.

Finally, Panel C shows the other experimental treatments. Cash was randomly given out to 50% of villages (and given that we sampled about 10 households per village, it was given, by design, to roughly 50% of the respondents). The phone surveys were administered to 20% of the respondents. As expected, the survey experiment is orthogonal to both of these treatments.

Table 2: Summary Statistics and Randomization Check

	Baseline Survey		Endline Survey	
	Version A (Mean/SD)	p-value: test of equality over 6 versions	Version A (Mean/SD)	p-value: test of equality over 6 versions
Panel A. Respondent Characteristics				
=1 if female	0.87	0.188	0.89	0.308
=1 if currently married or has partner	0.76	0.970	0.74	0.188
Age	40.50	0.661	40.95	0.388
	(15.20)		(14.31)	
Years of education	4.18	0.553		
	(3.75)			
=1 if can read/write in English	0.57	0.667		
	(0.50)			
Panel B. Household Characteristics				
Number of household members	4.77	0.436	4.98	0.744
	(2.11)		(2.16)	
=1 if household engaged in farming past year	0.96	0.786	0.90	0.803
=1 if thatch roof	0.40	0.206	0.24	0.780
=1 if mud/dirt floor	0.80	0.848	0.77	0.392
=1 if owns dwelling	0.77	0.844	0.77	0.840
=1 if has electricity in dwelling	0.02	0.280	0.02	0.523
Panel C. Cross-randomized groups				
Cash Treatment Group	0.53	0.216	0.51	0.914
Phone survey group	0.21	0.640	0.22	0.655
Observations	4,879		5,349	

Note: Column 1 and 3 (Version A) represent control mean with standard deviation in parentheses. Columns 2 and 4 present p-values from the joint test of equality of the means for all the 6 survey versions, A-F.

3 Results

3.1 Quantifying survey fatigue

We start by examining the impacts of time-into-survey on the count of items or instances reported in response to the open-ended questions (questions described in Figure A6). To do this, we run the following regression:

$$Y_{icq} = \beta Hours_{icq} + \phi_c + \varepsilon_{icq}, \tag{1}$$

where Y_{icq} refers to the count of items corresponding to question q reported by survey respondent i in country-survey sample c, $Hours_{icq}$ denotes elapsed time into survey (in hours) at which question q is asked to respondent i, instrumented with the randomized module order (Versions A-F) that was fielded to the respondent, ϕ_c represents country-survey fixed effects (e.g., Malawi Baseline),

and ε_{icq} is the error term.

In this analysis, there is no reason to expect heterogeneity in responses based on outcomes – ex ante, we expect similar results for any question category. Therefore, to discipline our analysis, we present results exhaustively for every relevant outcome, and adjust the standard errors to account for a false discovery rate (FDR) using the procedure in Anderson (2008). For each outcome, we present only q-values from this procedure, and statistical significance is ascertained only based on the q-values obtained after FDR-correction.

We present these results in Table 3. We show 5 outcomes: the number of Rotating Savings and Credit Associations (ROSCAs) and Village Savings and Loan Associations (VSLAs) that the respondent reported being part of in the savings section; the reported number of transfers received and given during the past month; and the number of credit purchases during the past month.[7] Four out of 5 of these outcomes are statistically significant at 10% (and 2 are significant at 5%), even with the FDR adjustment. The effect sizes are large: an extra hour reduces the number of items by 26-64%. Because these surveys average 2.5 hours, this implies that the decision to place a question at the beginning rather than the end of the survey can have a large effect.

Table 3: Survey time and the probability of missing responses ("Open-ended" questions)

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Hours into Survey	0.002	-0.058**	-0.074*	-0.209***	-0.095*
	[0.468]	[0.033]	[0.065]	[0.001]	[0.078]
Dependent variable: Mean	0.056	0.205	0.275	0.328	0.366
Hours into Survey: Mean	1.9	1.9	1.9	1.9	1.9
Hours into Survey: SD	1.0	1.0	1.0	1.0	1.0
Number of respondents	5,596	5,597	5,596	5,594	5,597
Observations	10,225	10,224	10,223	10,215	10,228

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. See Table B2 for results by country and Table C2 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

[7] For both transfers and credit purchases, some earlier survey versions included questions recalling for the past 3 months instead. Later for analysis on aggregated values, the monetary values collected from these versions are divided by 3, comparable to the past-month values.

Next, we investigate the impacts of elapsed survey time on choosing an item in questions asked via the fixed-list method (questions described in Figure A7), and run the following regression:

$$Y_{icqj} = \beta Hours_{icq} + \phi_c + \psi_{qj} + \varepsilon_{icqj}, \tag{2}$$

where Y_{icqj} is a binary indicator of whether respondent i in country-survey sample c responded "yes" to having consumed/bought/experienced item j in question q of the survey, $Hours_{icq}$ elapsed time into survey (in hours) at the beginning of question q, instrumented with the randomized module order (Versions A-F), ϕ_c country-survey fixed effects, ψ_{qj} question-item fixed effects, and ε_{isqj} error term. Like before, we adjust the standard errors for multiple testing, and report only the FDR-corrected q-values in our tables.

Table 4 presents this analysis for a set of 9 outcomes: livestock, farm tools, durable goods, savings, loans, food expenditures, non-durables expenditures, household shocks, and public goods contributions. Note that these regressions are at the question level, and so are much better powered than the previous set of outcomes: we find that 4 of 9 outcomes are significant at 5% (and even of those not significant, nearly all are negative signed). Effect sizes are more measured than for the "open-ended" questions, ranging from 10-19% for the statistically significant outcomes. Nevertheless, survey fatigue is clearly evident here as well.[8,9]

[8] Please note, however, that in both Table 3 and Table 4, the effect sizes in percent terms are slightly overestimated due to the fact that the dependent variable means are calculated across all versions and are therefore, depressed due to survey duration effects. Nevertheless, the effects are large enough in an absolute sense to be economically meaningful.

[9] See Appendix B and Appendix C for heterogeneity in these results by country and by survey type (baseline or endline).

Table 4: Survey time and the probability of missing responses ("Fixed list" questions)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Hours into Survey	-0.007**	-0.004	-0.001	-0.002	0.000	-0.025***	-0.038***	-0.025***	-0.002
	[0.037]	[0.144]	[0.406]	[0.365]	[0.468]	[0.001]	[0.001]	[0.001]	[0.378]
Dependent variable: Mean	0.072	0.154	0.176	0.060	0.020	0.203	0.249	0.130	0.050
Hours into Survey: Mean	1.7	1.7	1.8	1.9	1.9	1.9	1.9	2.0	2.0
Hours into Survey: SD	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.2
Number of respondents	5,594	5,594	5,594	5,597	5,597	5,597	5,597	5,597	5,349
Observations	134,831	208,281	212,373	114,045	138,711	366,947	112,497	166,524	48,141

Note: Observations at respondent-question-item level. Each regression is an IV regression, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. See Table B1 for results by country and Table C1 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

3.2 Effect of survey fatigue on aggregated values

The prior section implies that aggregated values of categories such as expenditures or transfers will be attenuated by survey fatigue; in this section, we quantify this attenuation. We run regressions identical to Equation (1), except that the dependent variable is now in dollar amounts, rather than counts; in addition, results are shown for both open-ended and fixed list questions. Results are shown in Table 5. We find that the vast majority (9 of 11) of point estimates are negative, more than half of which (5) are significant at conventional levels despite being corrected for multiple hypothesis testing. In addition, 2 of the coefficients - those for farm tools and public goods - are marginally significant at 15% and 13% respectively. In addition to being statistically significant, the effect sizes are economically meaningful. Focusing on just the statistically significant effects, the coefficient magnitudes range from 25% of the mean (for food expenditure) to 86% (for transfers given).

One surprising result is that our effect sizes for reported monetary values (as shown in Table 5) are in some cases, much larger in percent terms than they are for the counts that were collected via the open-ended and the fixed-list questions in Table 3 and Table 4, respectively. This is especially true for some of the small categories such as transfers given, where an extra hour reduces the value

by $0.59, on a base of just $0.69, or 86%, while the effect of an hour on the count in Table 3 is a reduction of -0.21 transfers on a base of 0.33 (or 64%). But even for a larger category like food, the percent decline in value is 25%, compared to 12% in skipping in Table 4. This is at odds with others in the literature, such as Ambler et al. (2021) and Abay et al. (2021), who find that respondents are likely to forget the more marginal categories as they progress through the survey. While we can only conjecture as to what may cause this, our results are consistent with recent work such as Brzozowski et al. (2017), who show that recall errors in surveys tend to not be mean zero, but are in fact, negatively correlated with true behavior - i.e., when respondents make mistakes, they tend to overstate the low values and understate the high values.

Table 5: Survey Fatigue and Reported Total Monetary Value of Aggregated Categories

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Hours into Survey	-13.47	-1.32	4.68	-1.47	0.73	-4.12***	-2.52***	-0.10	-0.51**	-0.59***	-0.65***
	[0.334]	[0.148]	[0.365]	[0.365]	[0.345]	[0.001]	[0.001]	[0.121]	[0.015]	[0.001]	[0.002]
Dependent variable: Mean	95.78	10.48	58.11	15.52	6.40	16.22	7.93	0.14	0.95	0.69	0.81
Hours into Survey: Mean	1.7	1.6	1.8	1.9	1.9	1.9	2.0	2.1	1.9	1.9	1.9
Hours into Survey: SD	1.0	1.3	1.0	1.0	1.0	1.3	1.3	1.6	1.0	1.0	1.0
Number of respondents	5,594	5,349	5,594	5,597	5,597	5,597	5,597	5,349	5,597	5,597	5,597
Observations	10,189	5,349	10,189	10,226	9,952	10,227	10,227	5,349	10,228	10,228	10,228

Note: All values in USD. Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. For transfers and credit purchases, some earlier survey versions included questions recalling for the past 3 months instead of past month. The monetary values collected from these versions are divided by 3, making them comparable to the past-month values. See Table B3 for results by country and Table C3 for results by survey type (baseline/endline). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

3.3 Are effects driven by satisficing?

In this subsection, we investigate whether the practice known as "satisficing" is likely an explanation behind the observed pattern of results. Satisficing is a term used to describe the phenomenon where respondents may be answering questions in such a way that helps them avoid follow-ups, and therefore, reduce survey length. In this case, satisficing would entail responding "no" to questions, or not listing additional values of items such as transfers, in order to avoid follow-up questions on

those items. In order to check for this, we utilize our phone surveys. As mentioned in Section 2.1, we randomly selected 20% of our sample to participate in phone surveys, which began shortly after the baseline survey. Respondents were called once every 2 months for about 16-26 months (or 8-13 rounds).

After deciding to implement the survey order randomization into the longer in-person surveys, we later decided to also randomize the order in the phone surveys. Importantly, the randomization began around the 8th round of the survey in Liberia and the 11th in Malawi, so respondents already had lots of experience with the questionnaire.[10] If satisficing is an explanation, we would therefore expect survey fatigue to be minimal in this experiment (since people would be equally able to skip questions wherever they appeared in the survey).[11] The randomization was very similar to the longer surveys, though less involved: specifically, as shown in Figure A5, we varied the location of the expenditure and transfers sections within the survey.

Table 6: Impacts of survey time on open-ended and fixed list questions, phone surveys

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Number of distinct items reported for the following:					=1 if item is selected (not skipped):			
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases	Savings	Loans	Food expend	Non-durables
Hours into Survey	0.050	0.308	0.091	-0.246	-0.346*	0.048	-0.014	-0.103***	-0.069*
	[0.315]	[0.108]	[0.308]	[0.105]	[0.091]	[0.108]	[0.185]	[0.001]	[0.091]
Dependent variable: Mean	0.088	0.372	0.205	0.190	0.283	0.140	0.031	0.216	0.351
Hours into Survey: Mean	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Hours into Survey: SD	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Number of respondents	780	780	780	780	780	780	780	780	780
Observations	1,762	1,762	1,762	1,762	1,762	18,678	24,654	63,059	20,083

Note: For columns 1-4, observations at respondent-question-item level, and regressions include country-sample fixed effects and question-item level fixed effects. For columns 5-9, observations at respondent level, and regressions include country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). See **??** for results by country. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

In Table 6, we analyze the effect of hours-into-survey on responses in the phone survey. Columns 1-5 analyze responses to open-ended questions, and Columns 6-9 show outcomes for questions that

[10] See Figure A1 for the specific survey rounds when order randomization was implemented.

[11] Another implication of survey fatigue is that the total survey time, and thus the value of categories, should decline over time as respondents learn the skip codes. However, we have no way of testing this since the number of rounds is colinear with time trends.

follow the fixed list pattern. To study these, we run the same regressions as in Equation (1) and Equation (2) respectively, except that the outcomes are now drawn from the phone survey. Similarly, in Table 7, we show the impacts on the value of aggregated categories, a replication of the analysis that we show in Table 5.

Contrary to the predictions of a satisficing hypothesis, we find evidence of negative effects of survey duration on both, the counts as well as the value of objects/outcomes reported by the respondents. Taken together, these results suggest that the operative channel for survey fatigue effects is the cognitive burden imposed by long surveys, and not deliberate gaming by the respondents.

Table 7: Effect of survey time of total value of aggregated categories, phone surveys

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Total value of reported items for the following:						
	Savings	Loans	Food Expend	Non-durables	Transfers received	Transfers given	Credit purchases
Hours into Survey	8.56	-2.45	-9.63**	-2.62	2.68*	-0.73	-2.38*
	[0.185]	[0.333]	[0.039]	[0.185]	[0.091]	[0.153]	[0.091]
Dependent variable: Mean	10.19	8.63	13.37	7.49	1.55	0.59	1.28
Hours into Survey: Mean	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Hours into Survey: SD	0.1	0.1	0.1	0.1	0.1	0.2	0.2
Number of respondents	780	780	780	780	780	780	780
Observations	1,762	1,762	1,762	1,762	1,762	1,762	1,762

Note: Observations at respondent level, and regressions include country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). See **??** for results by country. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

4 Conclusion

In this paper, we randomize the order of questions asked as part of the baseline and endline surveys of a cash transfer experiment to provide evidence on the impact of survey duration on the quality of responses elicited during the survey. Our results point to strong fatigue effects, on the order of a 10-64% reduction in the count of reported items, which leads to even bigger effects on the reported monetary values of categories that aggregate over these items.

An important implication of these results is that the effects of any program might be attenuated if effects are measured later in the survey. For example, if the effect of survey fatigue is to pro-

portionally reduce the number of items mentioned, then treatment-control differences will become smaller (in absolute value, though not in percentages) if measured later in the survey. This effect could be magnified if there exist non-linearities, for example if there is some threshold level of cognitive load that the treatment group is more likely to encounter because they have more to report. In our case, we can examine if the effect of the cash transfer differs when outcomes are measured later in the survey. However, we find no compelling evidence of this effect in this data, perhaps because power is limited because this analysis can only be conducted on the endline and because the cash treatment requires clustering at the village level (results are shown in tables Table A3 and Table A4, in which we regress outcomes on cash, time into the survey, and its interaction).[12] We leave a further evaluation of this to future work.

Is there a way for these findings to inform survey design? Survey fatigue is not a recent discovery, and practitioners suggest a variety of remedies to address this concern, most of which boil down to fielding shorter surveys, or splitting surveys into multiple shorter versions. For example, Aggarwal et al. (2021) is an example of a multi-day baseline survey. Other strategies involve sacrificing detail in order to survey fatigue, for example by splitting the survey into shorter versions, administering only one of the versions to each respondent, and imputing responses to the unasked questions (Herzog and Bachman 1981; Raghunathan and Grizzle 1995). Another strategy is to replace ordinal questions with binary ones (Dolnicar et al. 2011). However, each of these remedies comes with its own set of problems, either in terms of detail and measurement error, or in cost.

While we have no easy fixes to recommend, an obvious remedial step would be to place the most important questions (for example, those about the primary outcomes in an RCT), as early as possible in the survey. Relatedly, it may also be a good survey practice for enumerators to suggest taking a short break before they start asking important questions that are placed later in the survey. This may be an important consideration especially for interventions in which the primary outcome is sensitive (for example, intimate partner violence, which was placed at the end of these surveys for exactly this reason).[13] Researchers often choose to place such sensitive questions later in the survey to allow respondents some time to become familiar with the enumerator and with

[12] In the regressions, we demean the hours variable, but find the interaction term is negative only for 5 of 9 outcomes, and none are significant (Table A4).

[13] See Park et al. (2021) and Park and Kumar (2021) for related work on the pitfalls of measuring IPV in this and a related sample in Liberia.

the survey, but this paper suggests that this consideration should be balanced against the risk of survey fatigue.

A final implication from this paper is that, for those working with secondary data collected via long surveys, such as the LSMS or the DHS surveys, it may be useful to recognize that cross-country comparisons or even within country comparisons across survey waves may be complicated because of varying survey duration. It may be important to design panel surveys such that outcomes are measured at similar points in the survey over waves.

References

Abay, Kibrom A., Guush Berhane, John Hoddinott, and Kibrom Tafere Hirfrfot (2021). "Assessing response fatigue in phone surveys: Experimental evidence on dietary diversity in Ethiopia". World Bank Policy Research Working Paper No. 9636.

Aggarwal, Shilpa, Jenny Aker, Dahyeon Jeong, Naresh Kumar, David S. Park, Jonathan Robinson, and Alan Spearot (2020). "The Effect of Cash Transfers and Market Access on Households in Rural Liberia and Malawi". AEA RCT Registry.

Aggarwal, Shilpa, Rebecca Dizon-Ross, and Ariel Zucker (2021). "Incentivizing Behavior Change: The Role of Time Preferences". National Bureau of Economic Research Working Paper No. 27079.

Ambler, Kate, Sylvan Herskowitz, and Mywish K. Maredia (2021). "Are we done yet? Response fatigue and rural livelihoods". *Journal of Development Economics* 153: 102736.

Anderson, Michael L (2008). "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association* 103 (484): 1481–1495.

Backor, Kristen, Saar Golde, and Norman Nie (2007). "Estimating Survey Fatigue in Time Use Study". SIQSS Working Paper.

Brzozowski, Matthew, Thomas F. Crossley, and Joachim K. Winter (2017). "A comparison of recall and diary food expenditure data". *Food Policy* 72: 53–61.

Dolnicar, Sara, Bettina Grün, and Friedrich Leisch (2011). "Quick, Simple and Reliable: Forced Binary Survey Questions". *International Journal of Market Research* 53 (2): 231–252.

Herzog, A. Regula and Jerald G. Bachman (1981). "Effects of Questionnaire Length on Response Quality". *The Public Opinion Quarterly* 45 (4): 549–559.

Kilic, Talip and Thomas Pave Sohnesen (2019). "Same Question But Different Answer: Experimental Evidence on Questionnaire Design's Impact on Poverty Measured by Proxies". *Review of Income and Wealth* 65 (1): 144–165.

Krosnick, Jon A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". *Applied Cognitive Psychology* 5 (3): 213–236.

Laajaj, Rachid and Karen Macours (2021). "Measuring skills in developing countries". *Journal of Human Resources* 56 (4): 1254–1295.

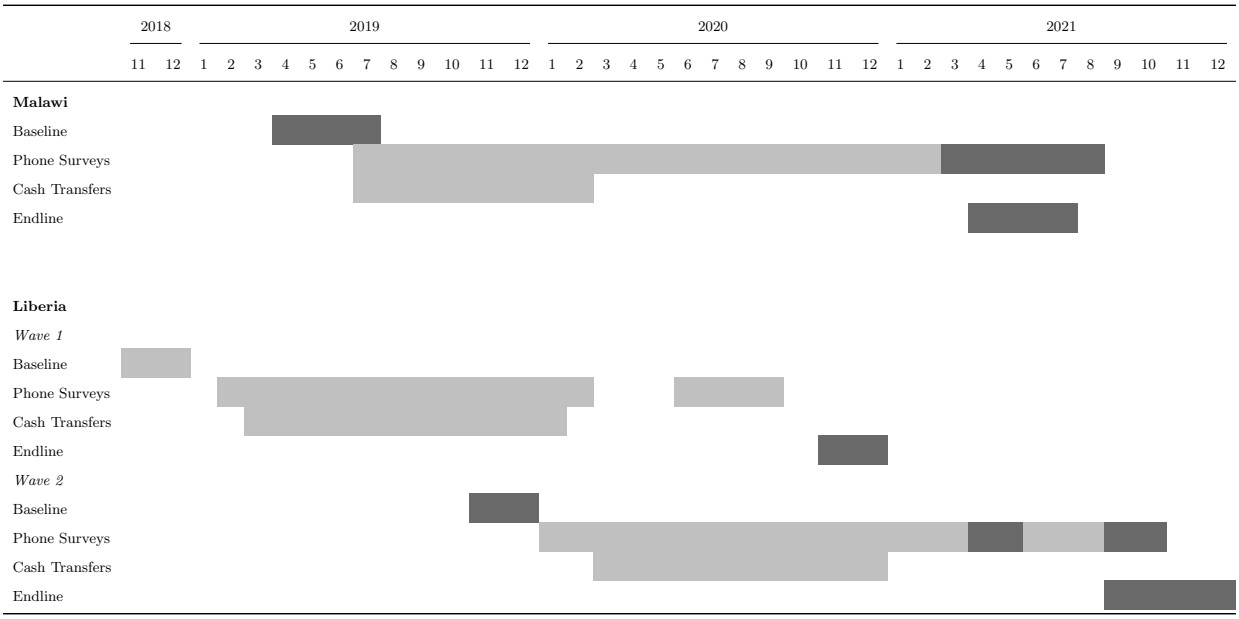
Park, David Sungho, Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson, and Alan Spearot (2021). "Private but Misunderstood? Evidence on Measuring Intimate Partner Violence via Self-Interviewing in Rural Liberia and Malawi". National Bureau of Economic Research Working Paper No. 29584.

Park, David Sungho and Naresh Kumar (2021). "Reducing Intimate Partner Violence: Evidence from a Multifaceted Female Empowerment Program in Urban Liberia". Unpublished.

Raghunathan, Trivellore E. and James E. Grizzle (1995). "A Split Questionnaire Survey Design". *Journal of the American Statistical Association* 90 (429): 54–63.

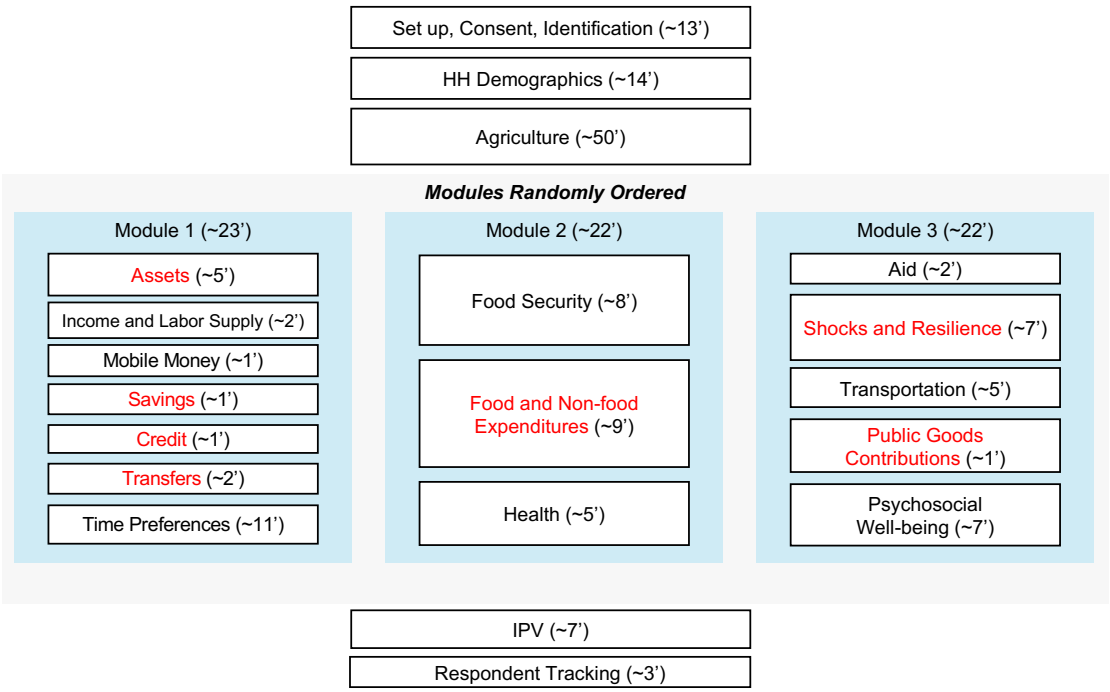
Appendix A: Main Appendix Figures and Tables

Figure A1: Timeline of Survey Activities



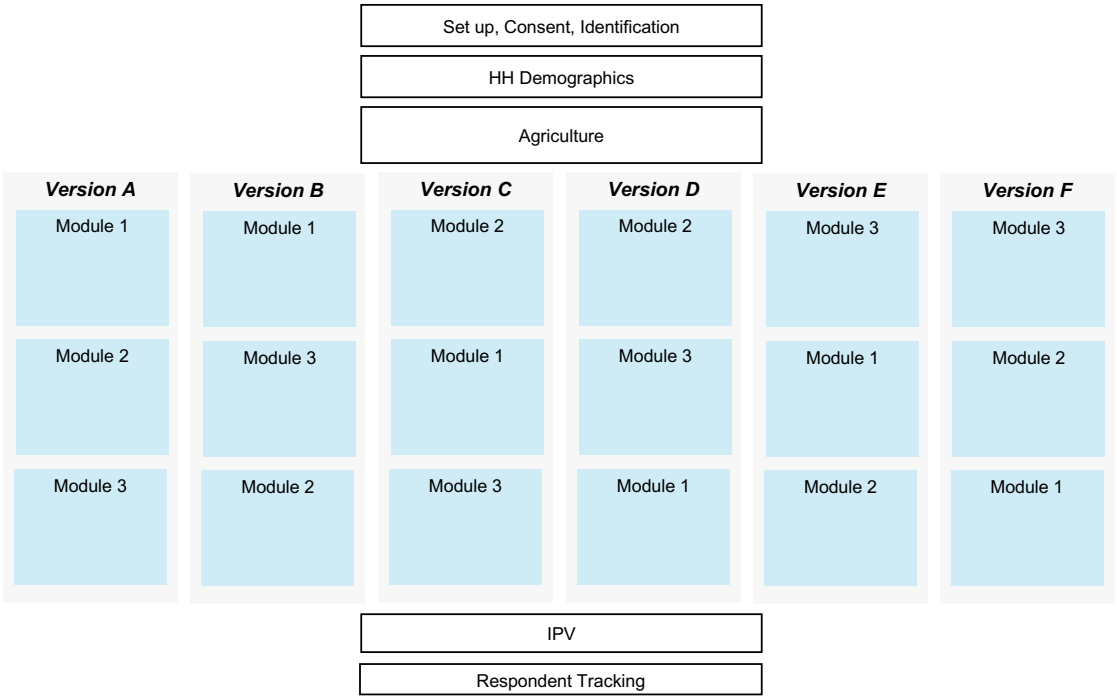
Note: Darker grey blocks indicate the survey rounds where module order randomization was conducted and thus data for which are included for analysis in this paper.

Figure A2: Sections in In-person Surveys



Note: Approximate duration for each section (in minutes) are reported in parentheses. In red are the sections for which survey questions are relevant for analysis in this paper.

Figure A3: Randomized Order of Modules in In-person Surveys

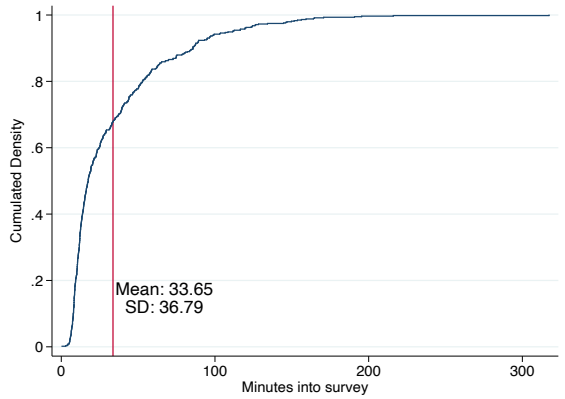


Note: A respondent is randomly provided with one among Versions A-F. The order in which the sections not included in the modules is not randomized. For every version among A-F, survey set-up, demographics, and agriculture come at the beginning, while IPV and respondent tracking are at the end.

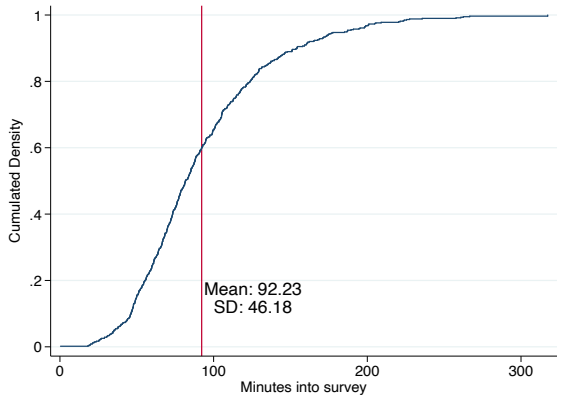
Figure A4: Distribution of Survey Time

Distribution of time to reach the question where on average the survey is:

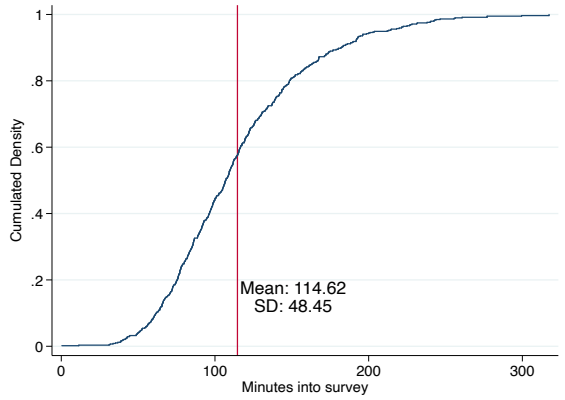
(a) 10% completed



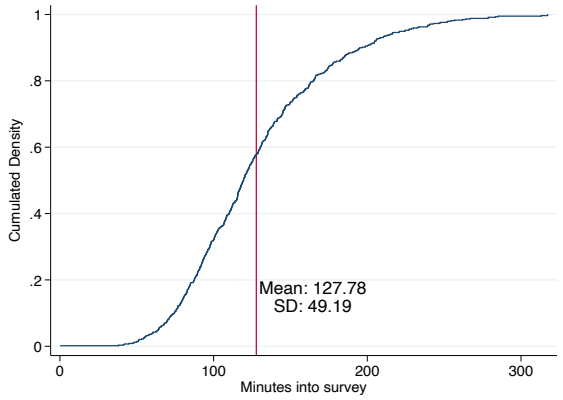
(b) 25% completed



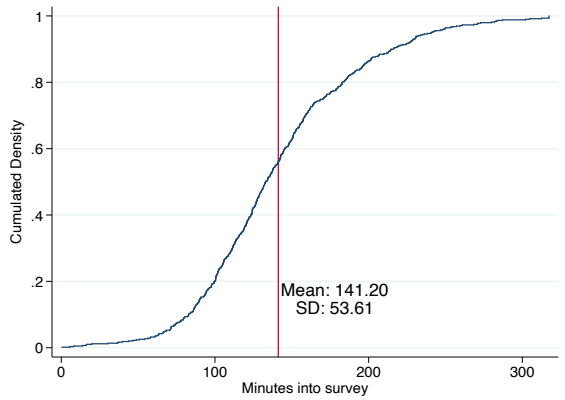
(c) Median



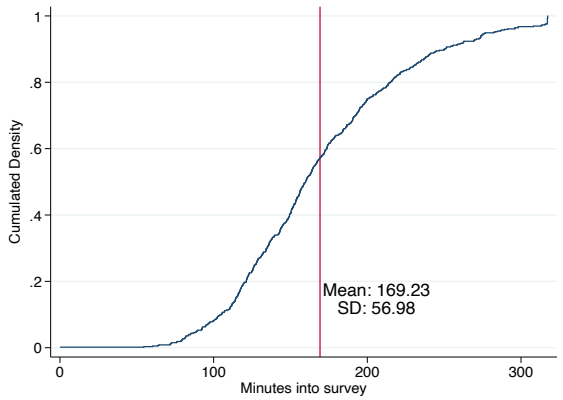
(d) 75th percentile



(e) 90th percentile

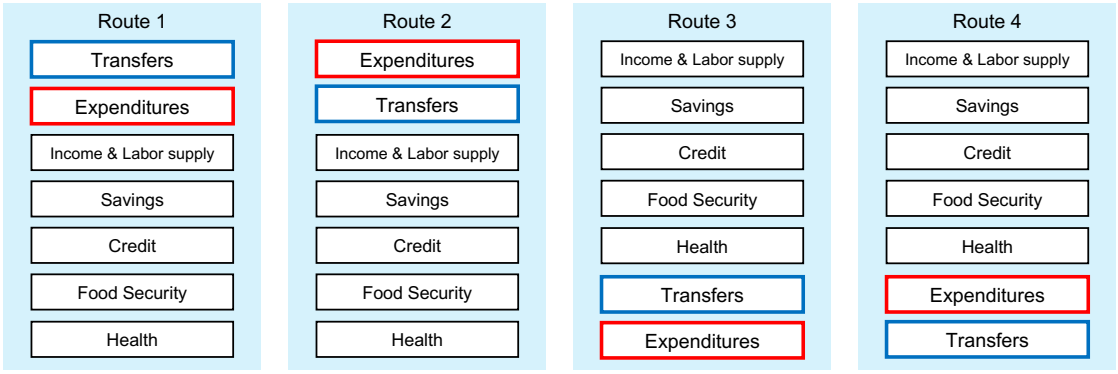


(f) Total survey length



Note: Based on Version A only.

Figure A5: Randomized Order of Modules in Phone Surveys



Note: A respondent is randomly provided with one among Routes 1-4.

Figure A6: Example of "Open-Ended" Question Order

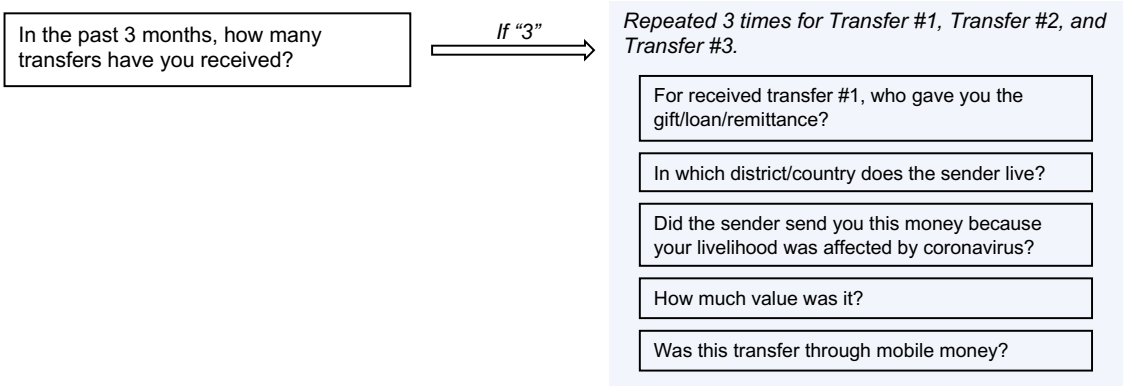


Figure A7: Example of "Fixed List" Question

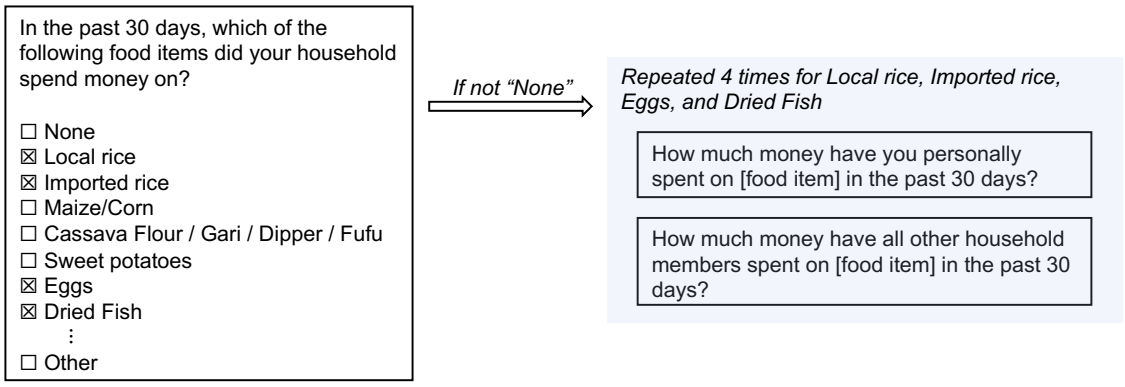


Table A1: Average Duration by Survey Versions (in hours)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Survey Version						Overall
	A	B	C	D	E	F	
Panel A: Liberia							
Baseline	2.28	2.27	2.24	2.31	2.29	2.24	2.27
	(0.69)	(0.65)	(0.69)	(0.75)	(0.67)	(0.70)	(0.69)
Endline	2.73	2.64	2.74	2.68	2.72	2.77	2.71
	(1.04)	(1.05)	(1.12)	(1.02)	(1.09)	(1.16)	(1.08)
Panel B: Malawi							
Baseline	3.15	3.03	3.06	3.03	3.01	3.04	3.05
	(1.02)	(0.89)	(0.93)	(0.92)	(0.91)	(0.90)	(0.93)
Endline	2.75	2.81	2.80	2.76	2.75	2.78	2.77
	(0.80)	(0.82)	(0.81)	(0.79)	(0.82)	(0.82)	(0.81)

Note: Standard deviations in parentheses.

Table A2: Experimental variation in time before sections were administered (phone surveys)

	(1)	(2)	(3)	(4)
	Time into survey (minutes) at the beginning of following section:			
	Savings	Credit	Transfers	Expenditure
Version B	-0.17	-0.08	8.66***	-1.45***
	(0.32)	(0.34)	(0.34)	(0.29)
Version C	-9.14***	-9.03***	10.48***	9.95***
	(0.31)	(0.33)	(0.34)	(0.28)
Version D	-9.66***	-9.59***	17.52***	8.53***
	(0.31)	(0.32)	(0.33)	(0.28)
Version A: Mean	15.47	16.53	3.21	4.81
Version A: SD	6.89	6.98	3.10	3.70
F-statistic: joint significance	585.88	523.70	941.79	837.01
Number of respondents	780	780	779	780
Observations	1,762	1,762	1,759	1,760

Note: Observations include only phone survey data. Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively.

Table A3: The effect of survey time on the measurement of the effect of cash

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	=1 if tem is selected (not skipped)									Number of distinct items reported for the following				
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Time into Survey (hr)	0.01	-0.00	-0.01	-0.01	-0.01	-0.02	-0.08	0.00	-0.01	-0.02	-0.13	0.03	-0.15	-0.22
	[0.591]	[0.766]	[0.591]	[0.256]	[0.256]	[0.591]	[0.222]	[1.000]	[0.594]	[0.594]	[0.228]	[0.705]	[0.228]	[0.228]
Cash × Time into Survey (hr)	-0.02	-0.01	-0.02	0.00	0.01	0.01	0.08	-0.02	0.00	-0.02	0.21	-0.29	-0.05	0.16
	[0.304]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[0.276]	[1.000]	[1.000]	[1.000]	[0.276]	[0.135]	[1.000]	[1.000]
Cash	0.01***	0.01*	0.02***	0.01***	0.00	0.01*	0.02***	-0.01*	-0.00	0.01*	0.04*	0.02	0.05**	-0.02
	[0.001]	[0.072]	[0.001]	[0.001]	[0.266]	[0.087]	[0.003]	[0.072]	[0.376]	[0.087]	[0.098]	[0.152]	[0.021]	[0.178]
Control Mean	0.06	0.13	0.15	0.05	0.02	0.18	0.20	0.08	0.04	0.03	0.24	0.16	0.15	0.34
Hours into Survey: Mean	1.7	1.7	1.8	1.9	1.9	1.9	2.0	2.2	2.2	0.0	-0.0	-0.0	-0.0	0.0
Hours into Survey: SD	0.9	0.9	0.9	1.0	0.9	0.9	0.9	1.0	1.0	0.9	0.9	0.9	0.9	0.9
Observations	54,714	80,419	82,023	44,761	51,489	141,028	43,582	63,392	35,658	3,961	3,962	3,962	3,958	3,962

Note: Regressions include baseline measurement of outcome, fixed effects for cash treatment randomization strata, and country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values (calculated from p-values based on standard errors clustered at village level) in brackets.

Table A4: The effect of survey time on the measurement of the effect of cash

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Expenditure		Assets					Transfers	
	Food	Nondurables	Livestock	Farm tools	Durables	Savings	Loans	Given	Received
Time into Survey (hr)	-0.55	-0.39	6.94	-2.47	-7.09	-7.71	-3.98	0.14	-1.94
	[1.000]	[1.000]	[1.000]	[0.999]	[1.000]	[1.000]	[0.999]	[1.000]	[1.000]
Cash × Time into Survey (hr)	-0.21	1.12	-45.35	0.30	36.20	-5.41	7.99	-2.70	-2.15
	[1.000]	[1.000]	[1.000]	[1.000]	[0.962]	[1.000]	[0.286]	[0.286]	[1.000]
Cash	0.19	0.27	26.00**	1.47***	21.02***	4.56***	-0.19	0.26	1.33**
	[0.197]	[0.227]	[0.034]	[0.004]	[0.001]	[0.007]	[0.545]	[0.197]	[0.019]
Control Mean	3.08	6.30	90.00	9.75	56.32	8.68	6.94	1.66	6.85
Control SD	4.90	9.37	367.73	10.66	138.21	55.59	19.14	6.58	14.53
Hours into Survey: Mean	1.9	2.0	1.7	1.7	1.8	1.9	1.9	1.9	1.9
Hours into Survey: SD	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.9	0.9
Observations	3,962	3,962	3,962	3,962	3,962	3,962	3,687	3,962	3,962

Note: Regressions include baseline measurement of outcome, fixed effects for cash treatment randomization strata, and country-sample fixed effects. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values (calculated from p-values based on standard errors clustered at village level) in brackets.

Appendix B: Heterogeneity by Country

Table B1: Heterogeneity by Country in Fixed List Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Panel A. Liberia									
Hours into Survey	-0.012**	0.002	0.001	-0.002	0.001	-0.033***	-0.053***	-0.011	0.006
	[0.026]	[0.654]	[0.713]	[0.571]	[0.522]	[0.001]	[0.001]	[0.256]	[0.522]
Dependent variable: Mean	0.097	0.196	0.171	0.048	0.009	0.189	0.234	0.065	0.075
Hours into Survey: Mean	1.5	1.6	1.6	1.7	1.7	1.7	1.7	1.8	1.8
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.5
Number of respondents	2,653	2,653	2,653	2,653	2,653	2,653	2,653	2,653	2,566
Observations	49,511	94,521	90,020	54,012	62,397	166,537	49,511	72,016	23,094
Panel A. Malawi									
Hours into Survey	-0.002	-0.011**	-0.001	-0.002	-0.001	-0.016***	-0.025**	-0.028***	-0.014*
	[0.535]	[0.015]	[0.551]	[0.551]	[0.551]	[0.001]	[0.015]	[0.001]	[0.099]
Dependent variable: Mean	0.057	0.119	0.179	0.071	0.028	0.214	0.261	0.180	0.028
Hours into Survey: Mean	1.9	1.9	1.9	2.0	2.0	2.1	2.1	2.2	2.2
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9
Number of respondents	2,941	2,941	2,941	2,944	2,944	2,944	2,944	2,944	2,783
Observations	85,320	113,760	122,353	60,033	76,314	200,410	62,986	94,508	25,047

Note: Observations at respondent-question-item level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table B2: Heterogeneity by Country in Open Ended Questions Question

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Panel A. Liberia					
Hours into Survey	-0.001	-0.036	-0.037	-0.157**	-0.180**
	[0.884]	[0.145]	[0.522]	[0.026]	[0.026]
Dependent variable: Mean	0.106	0.063	0.297	0.381	0.349
Hours into Survey: Mean	1.7	1.7	1.8	1.8	1.8
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2
Number of respondents	2,652	2,653	2,652	2,650	2,653
Observations	4,500	4,500	4,498	4,494	4,501
Panel A. Malawi					
Hours into Survey	0.004	-0.077	-0.122*	-0.240***	-0.016
	[0.551]	[0.121]	[0.072]	[0.001]	[0.551]
Dependent variable: Mean	0.017	0.316	0.258	0.285	0.380
Hours into Survey: Mean	2.0	2.0	2.1	2.1	2.1
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8
Number of respondents	2,944	2,944	2,944	2,944	2,944
Observations	5,725	5,724	5,725	5,721	5,727

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table B3: Heterogeneity by country on total monetary values of aggregated categories

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Panel A. Liberia											
Hours into Survey	-22.28	0.14	-3.04	-0.47	1.06	-6.05***	-4.10**	-0.15	-0.59	-0.87**	-1.25***
	[0.557]	[0.867]	[0.713]	[0.867]	[0.522]	[0.007]	[0.013]	[0.247]	[0.145]	[0.019]	[0.007]
Dependent variable: Mean	155.11	11.12	53.59	27.39	4.62	21.42	10.59	0.28	1.50	1.23	1.39
Hours into Survey: Mean	1.5	1.5	1.6	1.7	1.8	1.7	1.8	1.9	1.8	1.8	1.8
Hours into Survey: SD	1.3	1.6	1.3	1.2	1.2	1.8	1.8	2.2	1.2	1.2	1.2
Number of respondents	2,653	2,566	2,653	2,653	2,653	2,653	2,653	2,566	2,653	2,653	2,653
Observations	4,501	2,566	4,501	4,501	4,501	4,501	4,501	2,566	4,501	4,501	4,501
Panel A. Malawi											
Hours into Survey	-4.48	-3.86*	15.58	-1.51	0.22	-2.41**	-1.08	-0.01	-0.39**	-0.26***	-0.02
	[0.551]	[0.094]	[0.147]	[0.202]	[0.551]	[0.046]	[0.111]	[0.551]	[0.024]	[0.003]	[0.551]
Dependent variable: Mean	48.83	9.89	61.68	6.18	7.87	12.13	5.83	0.02	0.52	0.26	0.36
Hours into Survey: Mean	1.9	1.8	1.9	2.0	2.0	2.1	2.1	2.3	2.1	2.1	2.1
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8	0.8
Number of respondents	2,941	2,783	2,941	2,944	2,944	2,944	2,944	2,783	2,944	2,944	2,944
Observations	5,688	2,783	5,688	5,725	5,451	5,726	5,726	2,783	5,727	5,727	5,727

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Appendix C: Heterogeneity by Survey type

Table C1: Heterogeneity by Survey (Baseline or Endline) in Fixed List Questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	=1 if item is selected (not skipped):								
	Livestock	Farm tools	Durable	Savings	Loans	Food expend	Non-durables	Shocks	Public goods
Panel A. Baseline surveys									
Hours into Survey	-0.008*	-0.003	0.008*	0.005	0.002	-0.032***	-0.044***	-0.026***	
	[0.052]	[0.190]	[0.065]	[0.190]	[0.190]	[0.001]	[0.001]	[0.001]	
Dependent variable: Mean	0.076	0.166	0.191	0.069	0.022	0.227	0.289	0.194	
Hours into Survey: Mean	1.9	1.8	1.9	2.0	2.0	1.9	2.0	2.0	
Hours into Survey: SD	0.7	0.7	0.7	0.7	0.8	0.8	0.7	0.8	
Number of respondents	4,840	4,840	4,840	4,877	4,877	4,878	4,878	4,875	
Observations	64,860	98,735	102,610	52,640	67,977	174,600	53,658	80,940	
Panel A. Endline surveys									
Hours into Survey	-0.005	-0.005	-0.012*	-0.009	-0.002	-0.017**	-0.032**	-0.017	-0.002
	[0.313]	[0.316]	[0.077]	[0.148]	[0.394]	[0.012]	[0.032]	[0.196]	[0.469]
Dependent variable: Mean	0.068	0.144	0.162	0.053	0.018	0.180	0.212	0.070	0.050
Hours into Survey: Mean	1.6	1.6	1.7	1.8	1.8	1.8	1.9	2.0	2.0
Hours into Survey: SD	1.1	1.2	1.1	1.2	1.2	1.2	1.2	1.2	1.2
Number of respondents	5,349	5,349	5,349	5,349	5,073	5,349	5,349	5,349	5,349
Observations	69,971	109,546	109,763	61,405	70,734	192,347	58,839	85,584	48,141

Note: Observations at respondent-question-item level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects and question-item level fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table C2: Heterogeneity by Survey (Baseline or Endline) in Open-Ended Questions

	(1)	(2)	(3)	(4)	(5)
	Number of distinct items reported for the following:				
	ROSCAs	VSLAs	Transfers received	Transfers given	Credit purchases
Panel A. Baseline surveys					
Hours into Survey	0.020	-0.076**	-0.085	-0.267***	-0.043
	[0.183]	[0.035]	[0.106]	[0.001]	[0.275]
Dependent variable: Mean	0.067	0.204	0.382	0.494	0.414
Hours into Survey: Mean	2.0	2.0	2.0	2.0	2.0
Hours into Survey: SD	0.8	0.8	0.8	0.8	0.8
Number of respondents	4,877	4,875	4,874	4,870	4,879
Observations	4,877	4,875	4,874	4,870	4,879
Panel A. Endline surveys					
Hours into Survey	-0.010	-0.029	-0.061	-0.139**	-0.159*
	[0.440]	[0.384]	[0.216]	[0.032]	[0.077]
Dependent variable: Mean	0.046	0.205	0.178	0.176	0.323
Hours into Survey: Mean	1.8	1.8	1.8	1.9	1.9
Hours into Survey: SD	1.2	1.2	1.2	1.2	1.2
Number of respondents	5,348	5,349	5,349	5,345	5,349
Observations	5,348	5,349	5,349	5,345	5,349

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.

Table C3: Heterogeneity by survey type (baseline or endline) on total monetary values of aggregated categorie

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Total value of reported items for the following:										
	Livestock	Farm Tools	Durables	Savings	Loans	Food Expend	Non-durables	Public goods	Transfers received	Transfers given	Credit purchases
Panel A. Baseline surveys											
Hours into Survey	-15.61*		6.40	-4.45	1.78	-4.65***	-2.50***		-0.36*	-0.44**	-0.27*
	[0.065]		[0.190]	[0.183]	[0.106]	[0.001]	[0.002]		[0.056]	[0.018]	[0.060]
Dependent variable: Mean	51.24		46.06	18.16	6.60	16.93	8.31		0.83	0.71	0.54
Hours into Survey: Mean	1.8		1.8	2.0	2.0	1.9	2.0		2.0	2.0	2.0
Hours into Survey: SD	0.7		0.7	0.8	0.8	0.8	0.7		0.8	0.8	0.8
Number of respondents	4,840		4,840	4,877	4,878	4,878	4,878		4,879	4,879	4,879
Observations	4,840		4,840	4,877	4,878	4,878	4,878		4,879	4,879	4,879
Panel A. Endline surveys											
Hours into Survey	-11.42	-1.32	3.29	0.33	-0.53	-3.06	-2.06	-0.10	-0.59	-0.69*	-1.06**
	[0.480]	[0.219]	[0.502]	[0.576]	[0.480]	[0.104]	[0.127]	[0.186]	[0.167]	[0.062]	[0.041]
Dependent variable: Mean	136.08	10.48	69.01	13.11	6.21	15.57	7.57	0.14	1.05	0.67	1.06
Hours into Survey: Mean	1.6	1.6	1.7	1.8	1.8	1.9	1.9	2.1	1.8	1.9	1.9
Hours into Survey: SD	1.3	1.3	1.3	1.2	1.2	1.7	1.7	1.6	1.2	1.2	1.2
Number of respondents	5,349	5,349	5,349	5,349	5,074	5,349	5,349	5,349	5,349	5,349	5,349
Observations	5,349	5,349	5,349	5,349	5,074	5,349	5,349	5,349	5,349	5,349	5,349

Note: Observations at respondent level. Reported are TOT estimates, where elapsed time into survey (in hours) is instrumented with the randomized module order (Versions A-F). Regressions include country-sample fixed effects. ***, **, and * represent significance at 1%, 5%, and 10%, respectively, based on the false discovery rate (FDR) sharpened q-values in brackets.