

Call Me Maybe: Experimental Evidence on Using Mobile Phones to Survey Microenterprises

Robert
Garlick

Duke
University

Kate Orkin

Oxford
University

Simon
Quinn

Oxford
University

July 27, 2016

ERID Working Paper Number 224

This paper can be downloaded without charge from the Social
Science Research Network Electronic Paper Collection:

<http://ssrn.com/abstract=2816064>

**Economic Research Initiatives at Duke
WORKING PAPERS SERIES**



Call Me Maybe: Experimental Evidence on Using Mobile Phones to Survey Microenterprises*

Robert Garlick[†], Kate Orkin[‡], Simon Quinn[§]

July 27, 2016

[LATEST VERSION](#) · [ONLINE APPENDIX](#)
[PRE-ANALYSIS PLAN](#) · [QUESTIONNAIRES](#)

Abstract

High-frequency data is useful to measure volatility, reduce recall bias, and measure dynamic treatment effects. We conduct the first experimental evaluation of high-frequency phone surveys in a developing country or with microenterprises. We randomly assign microenterprise owners to monthly in-person, weekly in-person, or weekly phone interviews. We find high-frequency phone surveys are useful and accurate. Phone and in-person surveys yield similar measurements, with few large or significant differences in reported outcome means or distributions. Neither interview frequency nor medium affects reported outcomes in a common in-person endline. Phone surveys reduce costs without increasing permanent attrition from the panel.

*We are grateful for helpful comments from Nicola Branson, Markus Eberhardt, Nathan Fiala, Simon Franklin, Markus Goldstein, David Lam, Murray Leibbrandt, Owen Ozier, Gareth Roberts, Volker Schoer, and Duncan Thomas; seminar audiences at the Development Economics Network Berlin, Duke University, the University of Cape Town, the University of Oxford, and the University of the Witwatersrand; and conference participants at ABCDE, CSAE, and NEUDC. We thank Bongani Khumalo, Thembela Manyathi, Mbuso Moyo, Mohammed Motala, Egines Mudzingwa, and the staff at the Community Agency for Social Enquiry (CASE); Mzi Shabangu, Arul Naidoo, and Statistics South Africa for enumeration area maps; and Rose Page, Richard Payne, and Gail Wilkins at the CSAE. This project was funded by Private Enterprise Development for Low-Income Countries (PEDL), a joint research initiative of the Centre for Economic Policy Research (CEPR) and the Department For International Development (DFID). We thank Chris Woodruff and the PEDL team for their support.

[†]Department of Economics, Duke University; robert.garlick@duke.edu.

[‡]Department of Economics, Centre for the Study of African Economies and Merton College, University of Oxford; kate.orkin@merton.ox.ac.uk.

[§]Department of Economics, Centre for the Study of African Economies and St Antony's College, University of Oxford; simon.quinn@economics.ox.ac.uk.

1 Introduction

We run the first randomized controlled trial to compare microenterprise data from interviews of different frequency (weekly versus monthly) and medium (phone versus in-person). We use this trial to understand the effect of interview frequency and medium on microenterprise outcomes and owners' reporting of these outcomes, and to improve our understanding of outcome dynamics. We draw a representative sample of microenterprises in the city of Soweto in South Africa and randomly assign them to three groups. The first group is interviewed in-person at every fourth week for 12 weeks, to mimic a standard method of collecting data from microenterprises. The second group is interviewed in-person every week for 12 weeks. This allows us to test the consequences of collecting data at higher frequency, holding the interview medium fixed. The third group is interviewed every week by mobile phone for 12 weeks. All surveys use an identical questionnaire, which takes approximately 20 minutes to administer and measures 17 enterprise outcomes. We then conduct a common in-person endline with all microenterprises.

We conclude that high-frequency mobile phone surveys offer comparable data quality, measure dynamics better, and cost less than conventional lower-frequency in-person surveys. We draw this conclusion for three reasons. First, we find that measurements from mobile phone and in-person interviews are not systematically different. We compare reported outcomes for weekly interviews conducted by phone and in person. Responses are similar on most measures and at most quantiles of the distributions. Owners interviewed by phone report different stock and inventory levels (due to some outliers), working fewer hours, and taking less money from the enterprise for their household. The latter two results may reflect variation in susceptibility to social desirability bias by medium. We are powered to detect relatively small differences between interview methods so our findings are not explained by low power.

Second, we find little evidence that 'real' microenterprise outcomes differ by interview frequency or interview medium. We conduct all endline interviews in person, so endline differences are more likely to reflect frequency- or medium-induced behavioural change than differences in measurement. We find slightly more statistically

significant differences than we would expect from sampling variation. But the pattern of differences is not consistent with simple behavioural change explanations such as more frequent reminders inducing more comprehensive reporting. Instead, we find that phone respondents report slightly lower household takings and are more likely to use written records than weekly in-person respondents, with some other small differences in asset ownership and money kept by the owner and household.

Third, we show that missed interviews are more common for weekly than monthly interviews but not systematically different for phone and in-person interviews. Most microenterprises miss multiple scheduled interviews, but the fraction of interviews missed is stable throughout the panel and reasons for attrition are generally balanced across groups. Missed interviews are predicted by few baseline microenterprise characteristics and the results reported above are robust to adjustment using inverse probability weights and Lee bounds. We conclude that missed interviews are common and reflect the logistical challenges of high-frequency interviews, rather than lower willingness to participate in phone interviews or frequency-induced respondent fatigue.

To our knowledge, this is the first experimental comparison of interview frequency and medium for microenterprises. Our findings imply that microenterprise researchers can use high-frequency mobile phone interviews rather than low-frequency in-person interviews to collect more useful data, at lower cost, without substantially compromising data quality. Our findings build off existing literatures that use high-frequency data, use mobile phone interview data, or compare polling or household survey data collected at different frequencies and with different media. Recent work has explored the feasibility of mobile phone-based data collection in developing country contexts ([Dabalen et al., 2016](#); [Gallup, 2012](#)); we go on to compare the quality of data from mobile phone interviews to an in-person benchmark.

Six factors motivate our interest in high-frequency mobile phone surveys: three potential benefits, one potential cost, and two factors that may be either benefits or costs. First, high-frequency surveys allow us to study dynamic processes. This facilitates rich description of volatility in economic experiences of both enterprises and households ([McKenzie and Woodruff, 2008](#); [Collins et al., 2009](#)). This description can inform economic models of intertemporal optimization, such as responses to in-

come or expenditure shocks in the presence of credit constraints (Banerjee et al., 2015; Rosenzweig and Wolpin, 1993). High-frequency data can also enhance impact evaluations by illustrating the time path of treatment effects (Jacobson et al., 1993; Karlan and Valdivia, 2011), allowing researchers to average over multiple measures and improve power (Frison and Pocock, 1992; McKenzie, 2012), and potentially informing dynamic treatment regimes (Abbring and Heckman, 2007; Robins, 1997).

Second, existing research shows that different recall periods yield different measures of consumption, labour supply, and investment in human and physical capital (Beegle et al., 2012; Das et al., 2012; De Nicola and Giné, 2014; Heath et al., 2016). Shorter recall periods are generally regarded as more accurate but may miss important but infrequent experiences, such as microenterprises' purchases of fixed assets. High-frequency surveys allow researchers to use short recall periods but still obtain more comprehensive time-series coverage. We compare interview frequency and medium effects on questions with different recall periods and find no systematic differences.

Third, mobile phone surveys can make high-frequency data collection cheaper, easier, and feasible in difficult environments. Researchers have also used mobile phones to conduct rapid data collection in areas affected by conflict (van der Windt and Humphreys, 2013; Bauer et al., 2013) or disease outbreaks (Turay et al., 2015). Croke et al. (2014) and Dillon (2012) both show that mobile phone follow-up surveys are substantially cheaper than in-person baseline surveys, particularly in the rural setting Dillon studies. We can directly compare the cost of mobile phone and in-person surveys. We find the former are substantially cheaper, even though we work in an urban area where enterprises are fairly close together.

Fourth, however, these advantages may be countered if mobile phone and in-person surveys yield different reporting errors. An extensive literature concluded that household surveys and political opinion polls in developed countries are only slightly sensitive to the choice of survey medium (De Leeuw, 1992; Groves, 1990; Körmendi, 2001). These differences are concentrated in responses to open-ended questions and questions that would ideally use visual aids. There is also some evidence of differential susceptibility to social desirability bias by interview medium, though not in a

consistent direction (Holbrook et al., 2003).¹

The small literature in developing countries has found some differences between mobile phone and in-person measurements in labour force surveys (Heath et al., 2016) and between phone and text message measurements in household surveys (Gallup, 2012).² We find that differences in microenterprises' reported outcomes by interview medium are small, suggesting that interview medium is no more important for microenterprise surveys in developing countries than for household or individual surveys in developed countries.

Fifth, high frequency surveys may change behaviour by reminding respondents about some experiences or making these experiences more salient. Zwane et al. (2011) document increases in health investments in children due to more frequent surveys but find no effect on borrowing. Beaman et al. (2014) find that more frequent surveys do not affect microenterprise profits or sales but improve management of small change. Franklin (2015) finds no effects of survey frequency on job search but Stango and Zinman (2013) find that more frequent surveys about overdrafts reduce incurred overdraft fees. The authors of these papers argue that behaviour change is less likely for already-salient outcomes such as enterprise profits or the amount of loans outstanding. We find very little evidence of behaviour change from interview frequency or medium. This may occur because the outcomes we measure (sales, costs, labour supply) are already salient for microenterprise owners.

However, we find one possible positive effect of salience induced by more frequent reporting. Weekly interviews reduce data inconsistencies, measured by the absolute value of sales minus costs minus profits, compared to monthly interviews. Respondents interviewed weekly report higher profits than respondents interviewed monthly, though reported sales and costs are not significantly different. Weekly interviews may

¹ Cognitive psychologists and survey methodologists have theorised that phone surveys might also perform worse on particular data quality metrics such as item non-response, non-differentiation (Krosnick, 1991), acquiescence (Smith and Fischer, 2008) or response order effects (Schwarz et al., 1992). However, experimental comparisons of phone and in-person surveys find little empirical evidence of systematic differences between data collection methods (De Leeuw, 1992; Groves, 1979; Jackle et al., 2006).

² A parallel literature finds that recorded survey responses are somewhat different for paper- and tablet-based surveys (Caeyers et al., 2012; Fafchamps et al., 2012; Lane et al., 2006).

increase the salience of profits and hence reduce the incidence of very inaccurate profit reports.

Sixth, attrition may vary by interview medium and frequency.³ Phone surveys may allow respondents to complete interviews at more convenient times and locations, but they may also reduce rapport between enumerators and respondents. For example, [Gallup \(2012\)](#) finds that panel attrition is higher with text message or ‘robocall’ interviews than enumerator-administered phone surveys. High-frequency surveys impose higher time costs on respondents but the regular contact may reduce the probability of losing respondents who change contact information. We compare the level and time path of attrition by interview medium and frequency, and examine whether reported microenterprise outcomes are sensitive to differential attrition.

We describe the experimental design and data collection process in section 2. In section 3 we discuss interview completion and attrition rates. We compare reported outcomes by data collection medium and frequency in section 4. We compare data collection costs in section 5 and conclude in section 6. Online appendices A to J report a variety of background data and robustness checks.

2 Design and data

2.1 Context

The study takes place in Soweto, the largest and oldest ‘township’ near Johannesburg, in South Africa.⁴ Soweto’s population in October 2011 was approximately 1.28

³ The early US literature comparing landline telephone and in-person cross-section interviews emphasizes the scope for differential refusal and coverage. This consideration is less relevant for studies that use an in-person baseline, followed by phone surveys. See [Groves \(1990\)](#) and [Groves et al. \(2001\)](#) for more discussion on this issue and [Leo et al. \(2015\)](#) for a discussion of how closely mobile phone surveys in developing countries approximate random samples.

⁴ ‘Townships’ are low-income urban areas designated as Black African living areas under apartheid’s residential segregation laws. They typically consist of formal and informal housing and located on the outskirts of cities.

million people. Residents are almost all Black Africans (99%).⁵ Of the 0.9 million residents aged 15 or older, 41% engage in some form of economic activity (including occasional informal work) and 78% of these adults work primarily in the formal sector. 19% of households report receiving no annual income and another 42% report receiving less than \$10 per day.⁶

2.2 Sample definition and sampling strategy

We define an eligible microenterprise as any enterprise that: (i) has at most two full-time employees (in addition to the owner); (ii) does not provide a professional service (e.g. medicine); (iii) operates at least three days each week; and (iv) whose owner has a mobile phone which uses prepaid airtime. The first two conditions are consistent with definitions of ‘microenterprises’ in the development economics literature. The third condition excludes microenterprises that are seasonal, occasional (e.g. selling food at soccer games), or run over weekends in addition to wage employment. We impose this condition to ensure week-to-week variation in the outcomes of interest. The fourth condition is necessary to allow phone surveys and to allow us to pay respondents in airtime for completing surveys. We observe no microenterprises that are eligible on the first three but not the fourth criterion.⁷

We use a three-stage clustered sampling scheme to gather a representative sample of the population of households who own eligible microenterprises and live in ‘low-income’ areas of Soweto. We discuss this scheme in detail in Online Appendix A. In brief, we randomly selected small geographic units from the 2011 population census. Between September 2013 and February 2014, we conducted a *screening survey* with all households in the sampled area to identify whether anyone in the household owned a microenterprise and, if so, whether the microenterprise and owner met the eligibility criteria. The screening process realized a sample of 1081 eligible microenterprises.

⁵ We follow the terminology of Statistics South Africa, which asks population census respondents to describe themselves in terms of five racial population groups: Black African, White, Coloured, Indian or Asian, and Other.

⁶ Authors’ own calculations, from the 2011 Census public release data.

⁷ This is unsurprising, as 87% of South Africans aged 18 or older own a mobile phone (Mitullah and Kama, 2013).

In households which owned multiple eligible microenterprises, we randomly selected one for the final sample, leaving a sample of 1046.

2.3 Data collection and assignment to interview frequency and medium

Between December 2013 and February 2014, we approached all 1046 eligible microenterprise owners identified in the screening stage to conduct a *baseline survey* of 30 questions. These interviews were conducted in-person at the enterprise premises to verify that the enterprises existed, whereas the screening survey was conducted at the owners' homes. All respondents who completed the interview were given a mobile phone airtime voucher of 12 South African rands (approximately USD0.97).⁸ We completed the baseline questionnaire with 895 of the 1046 microenterprise owners (85%) identified in the screening stage. Of the remaining 183 owners, 67% could not be contacted using phone calls or home visits, 18% closed their enterprise between screening and baseline, 8% relocated outside Soweto, 6% refused to be re-interviewed, and 1% did not answer key questions in the baseline survey.

We then *randomised* the 895 baseline microenterprises into three data collection groups: monthly in-person surveys (298 microenterprises), weekly in-person surveys (299 microenterprises), and weekly phone surveys (298 microenterprises). Following [Bruhn and McKenzie \(2009\)](#), we first created strata based on (i) gender, (ii) number of employees, (iii) microenterprise sector and (iv) enterprise location.⁹ This yielded 149 strata with one to 51 microenterprises each. We then split each stratum randomly between the three data collection groups.¹⁰

We randomly assigned fieldworkers to data collection groups to ensure no sys-

⁸ We use an exchange rate of USD1 to ZAR10.27 throughout the paper, the South African Reserve Bank rate at the start of the survey on 31 August 2013).

⁹ We used the census subplace in which the microenterprise was located as the location block. This generally differed from the census subplace in which the household was located, which we used for the initial sampling scheme.

¹⁰ This generated some residual microenterprises in each stratum, as not all strata contained multiples of three. We randomly assigned residual microenterprises to data collection groups with a restriction that a pair of residual microenterprises in a stratum would always go into separate groups.

tematic differences between data collection groups.¹¹ However, within groups, fieldworkers were not randomly assigned to microenterprises. We assigned fieldworkers so each owner would be interviewed in her or his preferred language (English, seSotho, seTswana, or isiZulu) and to minimize fieldworkers' travel time between microenterprises.

We then conducted *repeated surveys* with each microenterprise owner between March and July 2014. These were conducted in-person or on mobile phones either every week or every four weeks. We randomly split the 'monthly' group, who were interviewed every four weeks, into four. Thus 75 of the monthly microenterprises were interviewed each week, providing a control group for each week when the 'in-person weekly' and 'phone weekly' microenterprises were interviewed.¹² In all treatment arms, enumerators were trained to make at most three attempts to complete each interview, where an "attempt" is a visit to the enterprise premises or an answered phone call.¹³ This design deliberately equalizes the respondents' opportunities to complete each interview. But it does not take advantage of the fact that mobile phone interview attempts are less costly than in-person interview attempts. Our results may thus understate the rate of interview completion that researchers can achieve with mobile phone surveys relative to in-person surveys. We successfully completed 4070 of 8058 repeated surveys. We discuss the pattern of missed interviews in detail in section 3. In the repeated interview phase, microenterprise owners received a ZAR12 (USD0.97) mobile phone airtime voucher for every fourth interview they completed. This equates the per-interview payout across data collection groups.¹⁴

Finally, we conducted an *endline survey* in person with each microenterprise owner at the microenterprise location. This common endline format, irrespective of

¹¹ We assigned two fieldworkers to the monthly in-person interview group, eight fieldworkers to the weekly in-person interview group, and four fieldworkers to the weekly phone interview group.

¹² We also staggered the start dates for this stage of the data collection. We randomly assigned 25% of enterprises in each group to be interviewed in weeks 1-12, 2-13, 3-14, and 4-15.

¹³ The Living Standards Measurement Study recommends a minimum of three attempts to contact each respondent and at least some Demographic and Health Surveys follow a similar rule (Grosch and Munoz, 1996; McKenzie, 2015b).

¹⁴ This design prices respondent time equally across groups but the total potential payout is higher for respondents assigned to weekly interviews. We predicted that the total income gain from answering all interviews was too small to generate meaningful income effects.

the assigned data collection method for the repeated surveys, means that observed endline differences across randomly assigned data collection groups must reflect persistent effects of the data collection method. We interpret differences at endline as ‘real’ differences in microenterprise outcomes, rather than measurement effects. We successfully completed 591 of 895 endline interviews and discuss the pattern of missed interviews in detail in section 3.

2.4 Baseline data description

Table 1 gives baseline summary statistics for the final sample of all 895 microenterprises. We draw two conclusions from this table. First, columns 3 to 6 show that the random assignment succeeded in assigning the microenterprises to three groups which are balanced on baseline characteristics. We fail to reject joint equality of all three group means across all 40 characteristics. The group means shown in columns 3 – 5 differ significantly at the 10% level for only 4 of 40 reported baseline variables. Differences are also small in magnitude. For each variable, we calculate the maximum pairwise difference between any two group means and divide this by the standard deviation of the variable, following [Imbens \(2015\)](#). This measure is 0.08 on average and exceeds 0.2 for only 2 of the 40 variables.

Second, our sample is broadly similar to samples of microenterprises in urban areas of other developing countries.¹⁵ The households in which the microenterprise owners in our sample live accrue a mean monthly income of ZAR4050 (approximately US\$380 at the time of the survey) across all sources.¹⁶ This falls in the fourth decile for all households across South Africa, a country with extremely unequal income distri-

¹⁵ We use five microenterprise samples from the Dominican Republic, Ghana, Nigeria, and Sri Lanka for which similar baseline variables are measured as benchmarks ([De Mel et al., 2008](#); [Drexler et al., 2014](#); [Fafchamps et al., 2014](#); [Karlan et al., 2012](#); [McKenzie, 2015a](#)). Our sample is more concentrated in the food and retail/trade sectors but is otherwise similar to at least one of the other samples on all common measures.

¹⁶ This is the average across the 87% of microenterprise owners who are willing to answer this question. There are essentially no missing values for the other variables.

Table 1: Sample Description and Balance Test Results

	(1)	(2)	(3)	(4)	(5)	(6)
	Full Sample		Monthly	Weekly	Weekly	p-value for
	Mean	Std Dev.	In-person	In-person	Phone	balance test
Panel A: Variables Used in Stratification						
Owner age	44.8	12.7	44.5	44.7	45.2	0.805
% owners female	0.617	0.486	0.601	0.629	0.621	0.769
# employees at enterprise	0.498	0.685	0.510	0.492	0.493	0.937
% enterprises in trade	0.318	0.466	0.312	0.311	0.332	0.824
% enterprises in food	0.426	0.495	0.423	0.438	0.416	0.857
% enterprises in light manufacturing	0.103	0.304	0.104	0.100	0.104	0.985
% enterprises in services	0.088	0.284	0.094	0.084	0.087	0.904
% enterprises in agriculture/other sector	0.065	0.246	0.067	0.067	0.060	0.929
Panel B: Other Owner Demographic Variables						
% owners Black African	0.993	0.082	0.990	0.997	0.993	0.576
% owners another race	0.007	0.082	0.010	0.003	0.007	0.576
% owners from South Africa	0.923	0.267	0.916	0.936	0.916	0.533
% owners from Mozambique	0.046	0.209	0.047	0.037	0.054	0.597
% owners from another country	0.031	0.174	0.037	0.027	0.030	0.778
% owners who speak English	0.065	0.246	0.064	0.087	0.044	0.096
% owners who speak Sotho	0.213	0.410	0.211	0.217	0.211	0.979
% owners who speak Tswana	0.084	0.277	0.077	0.087	0.087	0.876
% owners who speak Zulu	0.482	0.500	0.493	0.482	0.470	0.849
% owners who speak another language	0.156	0.363	0.154	0.127	0.188	0.124
# years lived in Gauteng	40.2	16.7	39.9	40.2	40.3	0.956
# years lived in Soweto	39.2	17.2	39.3	39.3	39.1	0.990
Panel C: Other Owner Education & Experience Variables						
% with at most primary education	0.152	0.359	0.124	0.181	0.151	0.157
% with some secondary education	0.469	0.499	0.487	0.482	0.440	0.450
% with completed secondary education	0.304	0.460	0.322	0.244	0.346	0.015
% with some tertiary education	0.075	0.263	0.067	0.094	0.064	0.353
% financial numeracy questions correct	0.511	0.264	0.513	0.508	0.512	0.970
Digit recall test score	6.271	1.489	6.333	6.220	6.260	0.632
% owners with previous wage employment	0.760	0.427	0.785	0.773	0.721	0.169
Panel D: Other Owner Household Variables						
Owner's HH size	4.785	2.683	4.745	4.756	4.856	0.852
# HH members with jobs	0.720	0.979	0.728	0.716	0.715	0.984
Owner's total HH income	4049	4285	3994	3957	4191	0.799
% owners whose enterprise supplies $\leq 1/2$ of HH income	0.554	0.497	0.581	0.515	0.567	0.238
% owners with primary care responsible for children	0.544	0.498	0.493	0.542	0.597	0.038
% owners perceive pressure within HH to share profits	0.634	0.482	0.607	0.635	0.658	0.444
% owners perceive pressure outside HH to share profits	0.565	0.496	0.581	0.605	0.510	0.053
Panel E: Other Enterprise Variables						
Enterprise age	7.187	7.511	7.302	7.278	6.980	0.842
% enterprises registered for payroll tax or VAT	0.079	0.270	0.081	0.060	0.097	0.232
% owners who keep written financial records for enterprise	0.196	0.397	0.195	0.167	0.225	0.207
% owners who want to grow enterprise in next five years	0.762	0.426	0.752	0.766	0.768	0.876
% owners who do business by phone at least weekly	0.568	0.496	0.554	0.579	0.570	0.823
# clients for the enterprise	33.7	71.4	28.9	40.8	31.3	0.189
Sample size	895		298	299	298	
Joint balance test statistic over treatment groups			70.9 (0.380)			
Joint balance test statistic over fieldworkers			793.1 (0.000)			

Notes: This table shows summary statistics for 40 variables collected in the screening and baseline interviews in columns 1 and 2. Columns 3 – 5 show the mean values of the variables for each of the three data collection groups. Column 6 shows the p -value for the test that all three groups have equal means. The first eight variables are used in the stratified random assignment algorithm and so are balanced by construction.

bution.¹⁷ The microenterprises in this sample are relatively well-established (average age 7 years) and have a diversified client base (mean and median numbers of clients are 34 and 20 respectively, though this varies by sector). However, they have remained relatively small. By design, we sample enterprises with at most 2 employees beside the owner: 61% have no other employees, while 28% and 11% have respectively one and two other employees. Most operate in food services (43%) or retail (32%). Very few are formally registered for payroll or value-added tax, but 20% keep written financial records. In 55% of households, the surveyed microenterprise accounts for half or more of household income.

2.5 Outcomes of interest

Our repeated and endline surveys cover both stock variables – replacement costs for stock and inventory and for fixed assets, number of employees, number of paid employees, number of full-time employees – and flow variables – total profit, total sales, nine cost items, hours of enterprise operation, money taken by owner, goods or services used by owner or other household members.¹⁸ All flow measures used a one-week recall period except hours of operation (last day) and sales (last week and last four weeks). The survey also asked respondents if they used written records to help complete the survey and several tracking questions. At the end of the survey, the enumerator assessed whether the respondent answered questions honestly and carefully. We use one-step questions that ask directly for values (for example, “How much did you spend last week on stock or inventory for your business?”) instead of two-step questions that first ask respondents whether the value is positive and then for the exact value.¹⁹

¹⁷ Microenterprise owners live in households with an average of 3.8 other people, though this is widely dispersed with an interdecile range of 1 to 7. Thus average monthly per capita income is roughly US\$100.

¹⁸ Each survey began by asking whether the respondent still operated her/his enterprise. If not, the respondent was asked what happened to the enterprise and what s/he was now doing. Only 2% respondents stopped operating their enterprise during the survey period so we do not use these data.

¹⁹ Friedman et al. (2016) discuss how two-step questions can lead to both extensive and intensive margin measurement error.

This survey design allows us to compare reporting behaviour across measures that we expect to be more or less volatile (for example, profits or sales versus fixed assets). We can also compare reported profits against profits calculated from sales and costs, following De Mel (2009). We show summary statistics for all repeated and endline interview measures in Appendix B.

3 Attrition and missed interviews

3.1 Attrition and missed interviews by interview medium and frequency

Interview medium and frequency may affect the rate of missed interviews, as different media entail different costs and experiences for respondents. Missed interviews are thus an outcome of interest for this study and we discuss missed interview rates in some detail. We also present robustness checks in section 4 and Appendices C to E that adjust our main results to account for missed interviews. We distinguish between *missed interviews*, which occur when a respondent misses a scheduled interview but may complete an interview later in the study, and *attrition*, which occurs when a respondent permanently leaves the study.²⁰ Enterprises can, and often do, miss one scheduled interview and then complete subsequent interviews. We focus on missed interviews in this section and provide more information about attrition in Appendix I.

We show missed interview rates for the repeated and endline interviews in Table 2 (columns 1 and 2 respectively). Missed interviews during the repeated interview phase are fairly common for all groups and more common for weekly than monthly interviews. Missed interviews in the endline phase are less common than in the repeated phase and higher for enterprises assigned to phone than both monthly and weekly in-person repeated interviews. The former pattern is consistent with weekly interviews causing respondent fatigue or simply being more logistically challenging to schedule. The latter pattern does not have an obvious interpretation. Endline and repeated inter-

²⁰ We observe very few cases where a respondent is successfully contacted but completes only part of the survey. We classify these as completed interviews.

Table 2: Rates of Missed Repeated and Endline Interviews by Data Collection Group

	(1)	(2)
	Repeated	Endline
Monthly in-person	0.427 (0.021)	0.336 (0.027)
Weekly in-person	0.485 (0.020)	0.274 (0.026)
Weekly phone	0.522 (0.020)	0.409 (0.029)
# enterprises	895	895
p-value for H ₀ : equal missed monthly & weekly in-person interviews	0.045	0.104
p-value for H ₀ : equal missed phone & in-person weekly interviews	0.2	0.0
p-value for H ₀ : equal missed interviews in all three groups	0.004	0.002

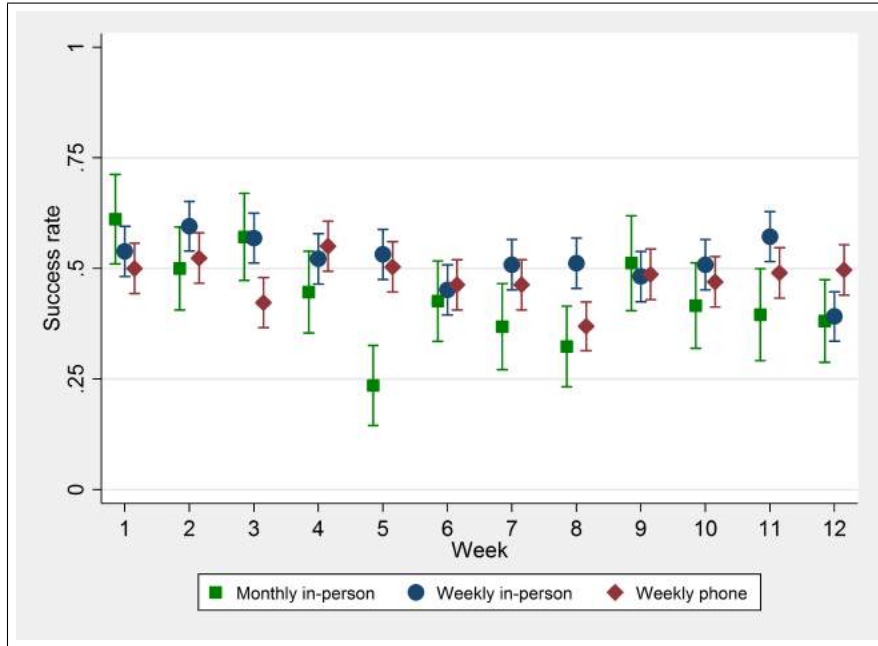
Notes: This table shows missed interview rates for repeated and endline interviews from linear regressions with heteroskedasticity-robust standard errors. ***, **, and * denote significance at the 1, 5, and 10% levels.

views are conducted by different enumerators so this pattern is unlikely to be due to respondents who were interviewed by phone having less rapport with survey staff.

Missed interviews during the repeated interview are moderately correlated within enterprises. The intra-enterprise correlation coefficient is 0.23 and enterprises surveyed in week t are 39 percentage points more likely to be surveyed in week $t + 1$ than enterprises not surveyed in week t . The rate of missed interviews increases slightly over the course of the survey period (Figure 1, Panel A) but a substantial fraction of microenterprises complete all their assigned interviews (Figure 1, Panel B). Interview misses are also correlated between the repeated and endline interview phases: missing half the repeated interviews is associated with a 27 percentage point increase in the probability of missing the endline interview. However, missing the endline interview is not differentially correlated with missing early versus late repeated interviews. These results provide weak evidence of ‘momentum effects’, in which interviews in rapid succession change the probability of missing interviews. Appendix I shows the rate of permanent attrition through the repeated interview phases. The rate of attrition increases slightly through the repeated phase, though not differentially by interview frequency. This may reflect right-censoring: respondents missing week t interviews have fewer opportunities to complete later interviews as t increases, artificially driving

Figure 1: Interview Success Rate by Treatment Arm

PANEL A: INTERVIEW SUCCESS RATE BY SURVEY WEEK AND TREATMENT ARM



PANEL B: TOTAL NUMBER OF COMPLETED INTERVIEWS BY TREATMENT ARM

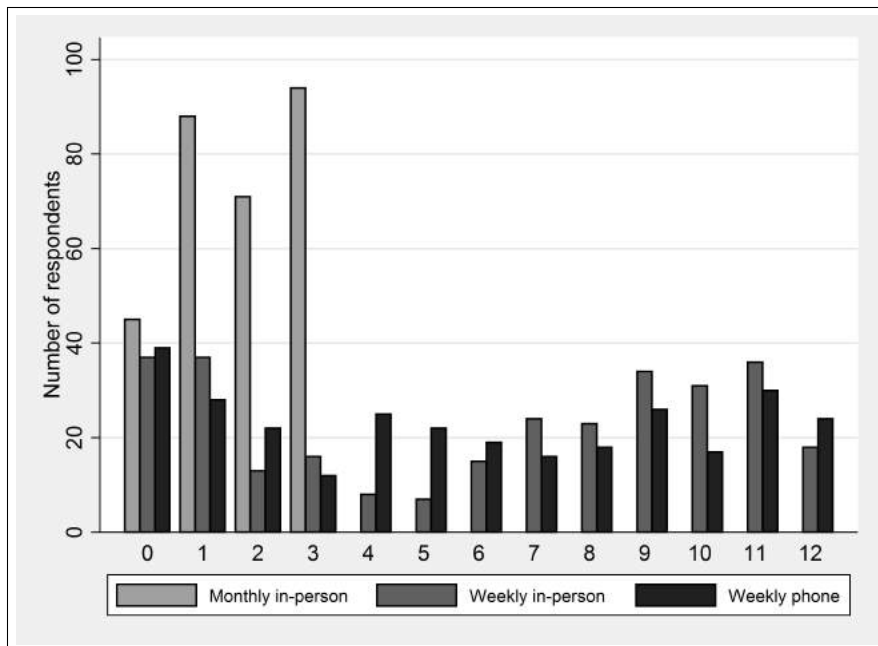


Table 3: Reasons for Missed Interviews by Treatment

	(1)	(2)	(2)
	Control mean	Weekly in-person	Weekly phone
Refused	0.077*** (0.015)	0.023 (0.023)	0.060** (0.025)
Wrong contact information	0.003 (0.003)	0.007 (0.007)	0.020** (0.009)
Moved from city	0.040*** (0.011)	-0.000 (0.016)	0.020 (0.018)
Business closed	0.037*** (0.011)	0.027 (0.018)	0.007 (0.016)
Owner ill	0.017** (0.007)	-0.000 (0.011)	-0.003 (0.010)
Owner traveling	0.034*** (0.010)	-0.027** (0.011)	-0.007 (0.014)
Owner too busy to talk	0.000 (0.000)	0.007 (0.005)	0.017** (0.007)

Notes: This table shows reasons for missed interviews across different data collection groups. Estimates are from linear regressions with monthly in-person interviews as the excluded category and with heteroskedasticity-robust standard errors shown in parentheses. ***, **, and * denote significance at the 1, 5, and 10% levels.

up the rate of attrition.

Reasons given for missing interviews differ by interview medium, but not by interview frequency, as shown in Table 3. Refusals and owners reporting that they are too busy to complete the interviews are more common for phone interviews than in-person interviews. This may be due to a weaker rapport between fieldworkers and respondents without in-person contact. Missed interviews due to enterprise closure and owner relocation do not differ by data collection method, consistent with the fact that these phenomena should not be influenced by the survey medium. Wrong contact information leads to more missed interviews in the phone than in-person interviews. This occurs because enterprise locations probably change less often than phone numbers, and when enterprises move, enumerators can call owners to find their new locations. This provides reassuring evidence that phone interviews are feasible even in a context

where phone numbers change often.²¹

3.2 Missed interviews by baseline characteristics

We report estimates of the relationship between missed interviews and baseline characteristics in Table 4. We report marginal effects from a fractional logit regression of the proportion of missed repeated interviews in column 1 and from a logit regression of missed endline interviews in column 2 (Papke and Wooldridge, 1996). The probability of missing interviews in both the repeated and endline data collection stages does vary systematically by baseline characteristics, but is predicted by a relatively small number of characteristics. Missed repeated stage interviews are more common for microenterprise owners with more education and who do not answer the baseline survey question about total household income. The same set of variables predict missed endline interviews (though with different coefficients), as does home language, number of employees, owners' growth plans for their microenterprise, and whether owners regularly conducted business over the phone before the survey period.²²

The results in Table 4 show that some enterprise and owner characteristics systematically predict missing interviews. However, we do not believe that these results are consistent with the most obvious economic models. For example, we might expect that microenterprise owners whose time is (self-perceived to be) more valuable or who find it more difficult to intertemporally shift their time expenditure from other activities to answer survey questions will miss more interviews. But the probability of missing interviews does not vary systematically with childcare responsibilities, the number of other employees in the microenterprise, or microenterprise sector. The probability also does not vary systematically with variables that might proxy for the difficulty owners face in determining the answers to our survey questions: microenterprise age, the presence of written records, registration for tax, financial literacy test results, or digit span

²¹ We collect multiple phone numbers for each enterprise owner at baseline (own, family members', etc.). We also restrict the sample to enterprises with a fixed location. So missed interviews or permanent attrition may be different in other settings, such as a survey of mobile street vendors that collects only one contact number per respondent.

²² Owners who are not Black African are more likely to miss both repeated and endline interviews. But there are only 6 of 895 owners in this group, so this result should be interpreted with caution.

Table 4: Predictors of Missing Endline and Repeated Interviews

	Repeated		Endline	
Owner's age	-0.003	(0.002)	0.003	(0.002)
Owner female (d)	-0.026	(0.028)	-0.035	(0.039)
Other is another race (d)	0.292**	(0.118)	0.353 *	(0.206)
Owner was born in Mozambique (d)	0.055	(0.066)	0.112	(0.100)
Owner was born in another country (d)	0.104	(0.079)	0.145	(0.116)
Owner speaks Sotho (d)	0.024	(0.051)	0.148	(0.097)
Owner speaks Tswana (d)	-0.011	(0.060)	0.081	(0.109)
Owner speaks Zulu (d)	0.005	(0.049)	0.182**	(0.084)
Owner speaks another language (d)	0.054	(0.057)	0.053	(0.099)
Owner has some secondary education (d)	-0.108***	(0.037)	-0.08*	(0.049)
Owner has finished secondary education (d)	-0.053	(0.043)	-0.036	(0.058)
Owner has some tertiary education (d)	-0.006	(0.062)	-0.091	(0.075)
Years owner has lived in Gauteng	0.002	(0.003)	0.004	(0.004)
Years owner has lived in Soweto	-0.002	(0.003)	-0.008**	(0.003)
% of financial literacy questions owner correctly answers	-0.02	(0.015)	0.018	(0.021)
Owner's digit recall test score	-0.002	(0.008)	-0.005	(0.012)
Owner has ever held regular paid employment (d)	-0.005	(0.030)	-0.025	(0.044)
Owner's household size	0.002	(0.005)	-0.008	(0.007)
Owner's household's total income	0.000	(0.000)	0.000	(0.000)
Missing value for owner's household's total income (d)	0.129***	(0.040)	0.090	(0.059)
Enterprise provides at most half of household income (d)	0.027	(0.026)	0.016	(0.036)
Owner has primary responsibility for childcare (d)	-0.005	(0.026)	-0.017	(0.036)
Owner perceives pressure within HH to share profits (d)	0.017	(0.027)	0.049	(0.037)
Owner perceives pressure outside HH to share profits (d)	-0.025	(0.026)	0.011	(0.037)
Food sector (d)	0.03	(0.028)	0.001	(0.040)
Light manufacturing sector (d)	0.034	(0.047)	0.082	(0.065)
Services sector (d)	-0.002	(0.046)	0.031	(0.065)
Agriculture/other sector (d)	-0.014	(0.056)	0.014	(0.073)
# employees	0.015	(0.019)	0.061**	(0.025)
Enterprise age	0.002	(0.002)	-0.001	(0.002)
Owner keeps written financial records (d)	0.003	(0.032)	0.036	(0.045)
Enterprise is registered for payroll tax or VAT (d)	0.064	(0.046)	0.023	(0.068)
Owner plans to grow enterprise in next five years (d)	-0.001	(0.029)	-0.089**	(0.042)
Owner conducts business by phone at least weekly (d)	0.011	(0.025)	0.082**	(0.035)
# clients	0.000	(0.000)	-0.000	(0.000)
# enterprises	895		895	
Pseudo-R2			0.071	
χ^2 test stat for H_0 : no covariates predict missed interviews	67.3		79.5	
p-value for H_0 : no covariates missed interviews	0.001		0.000	

Notes: This table shows missed interview rates for repeated and endline interviews. The first two columns show marginal effects from a fractional logit regression of the proportion of missed repeated interviews. The next two columns show marginal effects from a logit regression of a missed endline interview indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Marginal effects for continuous variables are evaluated at the sample mean of the relevant variable with standard errors calculated using the delta method. Discrete marginal effects are calculated for binary variables, denoted by (d). Omitted categories are Black African for race, South Africa for country of birth, English for home language, incomplete primary for education and trade/retail for business type. The unconditional probability of missing the endline interview is 34.0% and the unconditional mean and interquartile range of missed repeated interviews are respectively 47.8% and [16.7%,75.0%]. ***, **, and * denote significance at the 1, 5, and 10% levels.

recall test results.²³

3.3 Benchmarks for missed interviews

We compare the rate of missed interviews in our study to other high-frequency panel studies in developing countries, as well as more traditional low-frequency panels. Our rate of missed interviews is comparable to studies in the former group. [Croke et al. \(2014\)](#) complete 55% of scheduled mobile phone panel interviews with households in Dar es Salaam.²⁴ [Gallup \(2012\)](#) completes approximately 60% of scheduled phone/text panel interviews with households in Honduras and Peru.²⁵ In contrast, [Dillon \(2012\)](#) completes 98% of scheduled mobile phone surveys of rural Tanzanian households. The author notes that the sample was highly clustered in small villages, so that other villagers helped enumerators find respondents who did not answer calls or whose phones were lost or not charged. [Beaman et al. \(2014\)](#) complete 93% of scheduled mobile phone interviews with microenterprises in small Kenyan towns. [Heath et al. \(2016\)](#) complete 91% of scheduled mobile phone interviews in a labour market panel study in Ghanaian cities.

Some of these panels, however, use samples that have been selected for their willingness to participate in interviews. [Beaman et al. \(2014\)](#) begin with a baseline of 1195 microenterprises but include only 508 these in the panel; the other 57% either decline to participate in the panel or cannot be relocated after the baseline survey. If they had included all 1195 microenterprises in their panel and only completed interviews with the 508 actually included in the panel, their missed interview rate would have risen from 7% to 60%. [Heath et al. \(2016\)](#) interview respondents who have already

²³ We also estimate group-specific regressions of missed repeated or endline interviews on baseline characteristics, shown in Appendix I. The coefficients are jointly significantly different for missed interviews in the repeated stage but not the endline stage. The difference in missed repeated interviews is explained by race, language, digit recall score, and previous employment history. These patterns are again not consistent with any obvious economic explanation.

²⁴ They completed 55% of the planned interviews with 550 respondents. However, they focused interview efforts on the 84% of baseline respondents who also complete the first round of the panel survey. Within this group, they complete 66% of the planned interviews.

²⁵ In Peru, they completed 33% of interviews in the first wave of the panel. This fell monotonically to 25% by the sixth round. In Honduras, they completed 59% of interviews in the first wave of the panel and 50% in the sixth wave.

completed up to eight annual waves of a panel survey. These differences highlight a potential trade-off between selected samples with low attrition versus representative samples with high attrition. Some studies use random digit dialing to construct cross-sectional samples that are representative (conditional on high phone penetration) but typically have very high refusal rates (Groves, 1990; Groves et al., 2001). Our sample is slightly selected, as our panel includes only the 86% of the 1046 screened enterprises that we recontacted in the baseline. But the degree of selection is low compared to the other studies discussed here.

The rate of missed interviews in our study, and most high-frequency panels, is high relative to that in more traditional low-frequency panel studies. The Indonesian Family Life Survey, Kagera Health and Development Survey, and Kenya Life Panel Survey all successfully reinterview more than 80% of the target respondents at least ten years after the baseline (Baird et al., Forthcoming; Beegle et al., 2011; Thomas et al., 2012). However, these studies use tracking protocols that are unlikely to be feasible in a high-frequency panel. Each round of these surveys may take several months to complete, with several more month spent tracking difficult-to-locate respondents or respondents who have migrated (Thomas et al., 2012). High-frequency panels inherently have a short timeframe for completing each interview, increasing the risk of missed interviews. However, it should be possible to reduce the missed interview rate in high-frequency panels by requiring more than the three attempts per interview rule used in our study. We applied the same attempts-per-interview rule in in-person and phone arms. More contact attempts are particularly feasible in phone surveys, where each additional attempt to recontact a respondent is very low-cost.

4 Results

This section discusses our estimation methods, largely following our pre-analysis plan, and results.²⁶ We consider the following questions:

1. Does interview frequency and medium affect responses throughout the survey?

²⁶ Our pre-analysis plan is available at <https://www.socialscisceregistry.org/trials/346>.

2. Does interview frequency and medium affect microenterprise performance (as measured through an in-person endline)?

Given the prior findings reviewed in section 1, we expect that microenterprise responses may be affected by either interview frequency or interview medium. Microenterprise performance is more likely to be affected by interview frequency than medium. We differentiate between stock (fixed assets, employment) and flow (profit, sales, costs, stock/inventory) variables. We also evaluate consistency in reporting, measured by the absolute value of sales minus costs and profits. We also construct measures of enumerators' perceptions of respondents' honesty and carefulness.

4.1 Does interview frequency and medium affect survey responses?

We begin by estimating differences in reported responses during the repeated interviews – the interviews after the baseline and before the endline. Respondents in the weekly interview groups complete up to 12 repeated interviews, while those in the monthly group complete up to 3 repeated interviews. We use t to index the calendar week of the interview.²⁷ We index microenterprises by i and outcome variables by k ; Y_{kit} therefore refers to the response of microenterprise i to outcome k in week t . We use T_{1i} and T_{2i} respectively as indicators for the weekly in-person interview group and the weekly mobile phone interview group. If Y_{kit} is a continuous variable, we normalise first by the mean and standard deviation of the monthly in-person interview group, which we treat as the 'control' group.

Survey responses may differ by interview method in several ways. Distributions may be left- or right-shifted or may be more or less dispersed around a common mean. If responses only differ in dispersion (e.g. one introduces more classical measurement error), then some summary statistics will be robust across different interview methods. If responses differ in either dispersion or mean, then multivariate analy-

²⁷ We use indicator variables for actual calendar weeks to capture common shocks. Given the staggered start dates of the repeated interviews, there are 15 indicators but at most 3 and 12 indicators equal one for each enterprise assigned to respectively monthly and weekly interviews. The start dates are cross-randomized with the treatment assignments so the two sets of indicators are independent of each other.

ses including regression will generally be sensitive to the choice of interview method (Bound et al., 2001). We explore differences in three stages. We begin by examining the empirical CDFs of survey responses by method, pooling observations across microenterprises. This provides the most general overview of the differences in responses. We then run mean regressions; these test only for differences in mean survey responses but allow us to use the panel structure more effectively. We finally test if the microenterprise-specific standard deviations through the panel differ by interview method. Taken together, these provide a flexible and comprehensive description of possible differences in survey responses by interview method. Differences by survey method in these responses may reflect differences in reporting or in underlying microenterprise outcomes. We return to the distinction between these issues in section 3.2.

We begin by inspecting empirical CDFs; for each empirical CDF, we superimpose * and + to indicate significant differences in the quantiles of Y_{kit} .²⁸ For most measures of microenterprise performance – fixed assets, sales for the last four weeks, sales for the last week, total costs, total employees, paid employees, full-time employees and money kept by the entrepreneur – we do not find significant differences between the empirical CDFs for most measures.²⁹ We show these empirical CDFs together in Figure 2. Reported stock and inventory is systematically higher for weekly phone than weekly in-person interviews (Figure 3) but this effect is concentrated in

²⁸ Formally, for each outcome k , we estimate the quantile regression

$$Q_{\theta}(Y_{kit} | T_{1i}, T_{2i}) = \beta_0^{\theta} + \beta_1^{\theta} \cdot T_{1i} + \beta_2^{\theta} \cdot T_{2i} \quad (1)$$

for quantiles $\theta \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We use + to indicate rejection of the null hypothesis $\beta_1^{\theta} = \beta_2^{\theta}$ (i.e., weekly in-person and weekly phone interviews are equivalent) and we use * to indicate rejection of the null hypothesis $\beta_1^{\theta} = \beta_2^{\theta} = 0$ (i.e. all three treatments are equivalent). + and * denote $p < 0.1$; ++ and ** denote $p < 0.05$; +++ and *** denote $p < 0.01$. We report significance tests for each level of θ , where we cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. In the pre-analysis plan we planned to use stratum and week indicator variables; this proved computationally infeasible. We also planned to use simultaneous-quantile regression and to jointly test for coefficient equality across all quantiles. However, we do not believe an estimator has been proposed for systems of simultaneous quantile regression models with clustered standard errors.

²⁹ We occasionally reject at one of the quantiles, but never to indicate any pattern that suggests anything but random noise.

the right tail and is not robust to trimming (Appendix C.1). We conclude that, on most measures, weekly phone interviews do not induce different reporting than either weekly in-person interviews or monthly in-person interviews, so weekly phone interviews are a viable way for tracking microenterprise performance at high frequency.

We find some evidence that weekly interviews reduce data inconsistencies, measured by the absolute value of sales minus costs minus profits. Respondents in weekly interviews are more likely than those in monthly interviews to give answers for sales and costs such that the absolute value of sales minus costs is close to zero (Figure 4, Panel A). This occurs because respondents interviewed weekly report higher profits than respondents interviewed monthly, though reported sales and costs are not significantly different (Figure 4, Panel B). The difference occurs mainly in the right tail of the distribution. The weekly interviews may increase the salience of profits and hence reduce the incidence of very inaccurate profit reports. This improvement in reporting is not explained by shorter recall periods, as the recall period for profits is one week in both weekly and monthly interview groups.

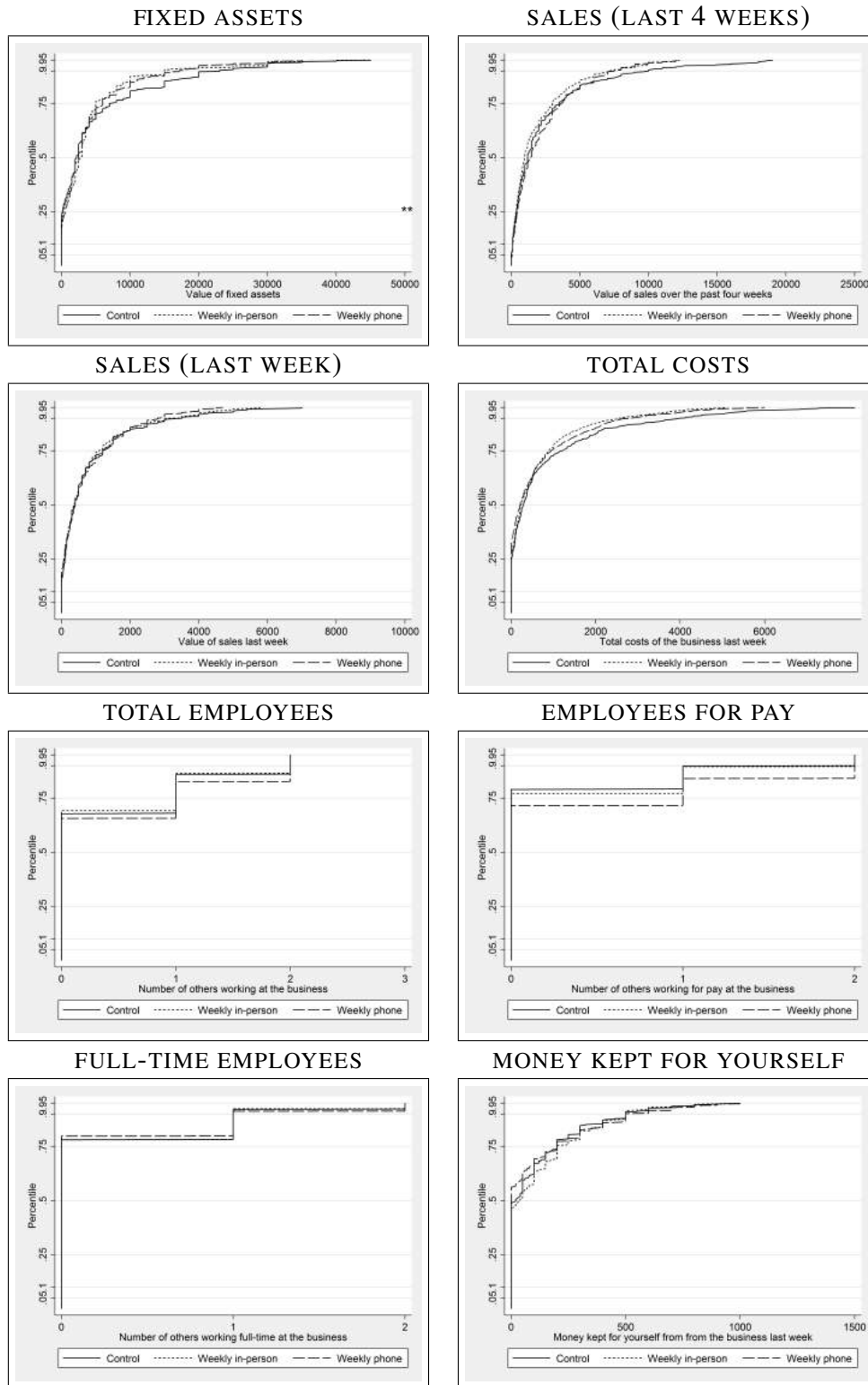
We ask respondents to report sales in both the last week and last four weeks. These measures are moderately correlated within enterprise-week ($\rho = 0.79$) and the median ratio of one-week sales to four-week sales is 0.36. The differences by interview frequency and medium are relatively small for both sales measures and are almost identical across measures. At least for sales, interview frequency and medium do not seem to differentially affect measures with different recall periods, reinforcing the potential value of high-frequency questions using short recall periods.³⁰

Respondents' answers to questions about hours worked and money taken from the enterprise differ by medium. This pattern is weakly consistent with differential social desirability bias. Respondents interviewed by phone report that their enterprises were open for fewer hours in the previous day (Figure 5 Panel A).³¹ The reported value of goods and services taken by the respondent and their household members is higher

³⁰ All respondents receive both questions, in the same order. So respondent-specific measurement errors may be correlated across the two measures.

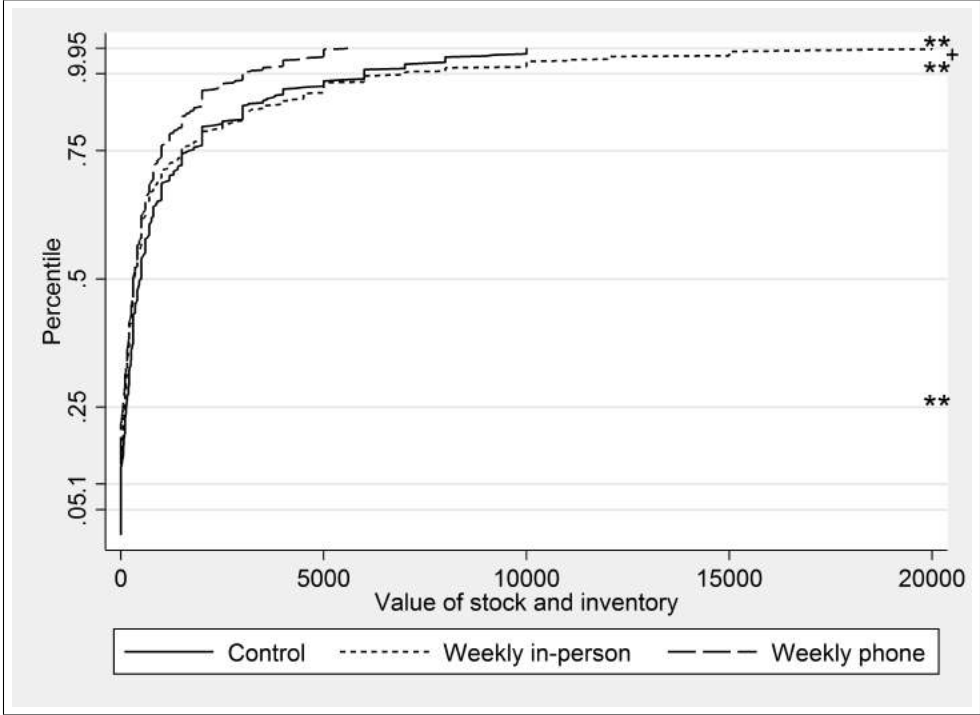
³¹ This result may arise if respondents who are assigned to in-person interviews and who work few hours are more difficult to interview and hence more likely to miss interviews. However, the result is robust to two different strategies to adjust for missed interviews, which we discuss in subsection 4.3.

Figure 2: Empirical CDFs in Repeated Interviews



Notes: We test for differences at each of the levels shown on the y -axis. We cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. + indicates rejection of the null hypothesis that weekly surveys in-person and by phone are equivalent; * indicates rejection of the null hypothesis that all three treatments are equivalent. +++/***, ++/**, and +/* denote significance at the 1, 5, and 10% levels.

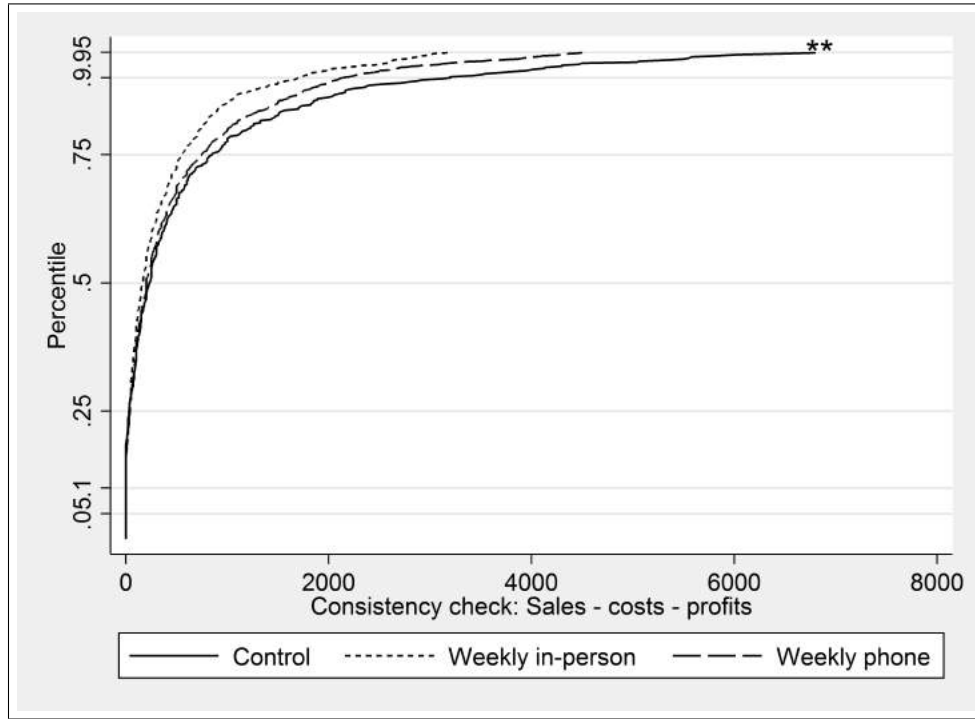
Figure 3: Empirical CDFs in Repeated Interviews: Stocks and inventories



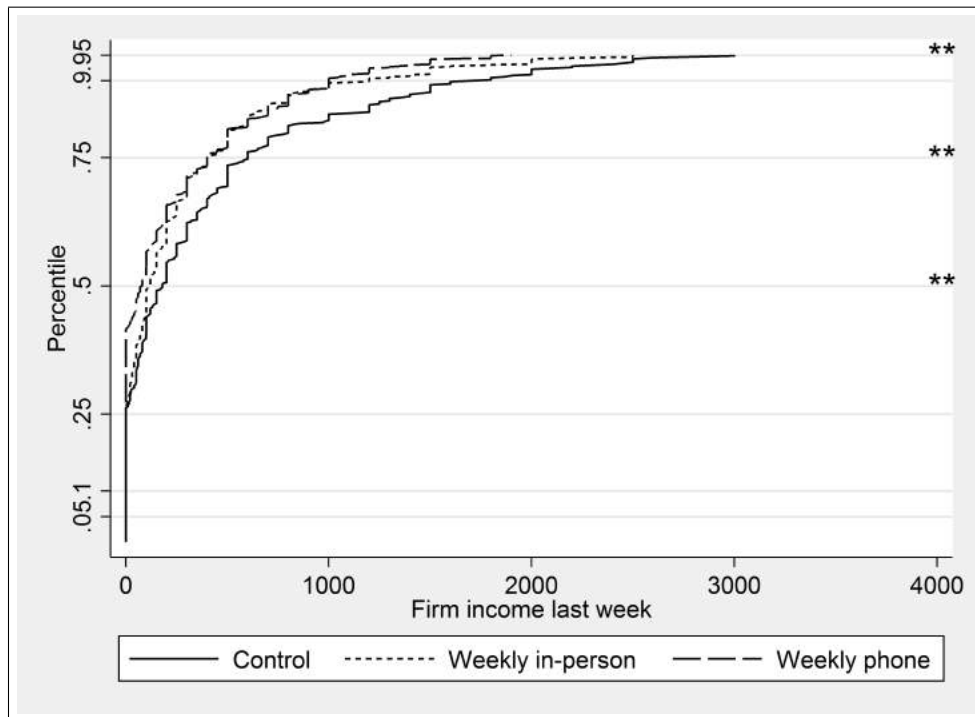
Notes: We test for differences at each of the levels shown on the *y*-axis. We cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. + indicates rejection of the null hypothesis that weekly surveys in-person and by phone are equivalent; * indicates rejection of the null hypothesis that all three treatments are equivalent. +++/**, ++/**, and +/* denote significance at the 1, 5, and 10% levels.

Figure 4: Empirical CDFs in Repeated Interviews

ABSOLUTE VALUE OF [(SALES - COSTS) - PROFIT]



PROFIT



Notes: We test for differences at each of the levels shown on the y -axis. We cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. + indicates rejection of the null hypothesis that weekly surveys in-person and by phone are equivalent; * indicates rejection of the null hypothesis that all three treatments are equivalent. +++/***, ++/**, and +/* denote significance at the 1, 5, and 10% levels.

for weekly in-person interviews than for monthly in-person interviews and higher for monthly in-person interviews than weekly phone interviews (Figure 5 Panel B). It is not clear whether taking enterprise goods and services for household use is seen as socially desirable or not, so we are cautious about overinterpreting this pattern.

We next estimate the effect of interview method on mean response values:

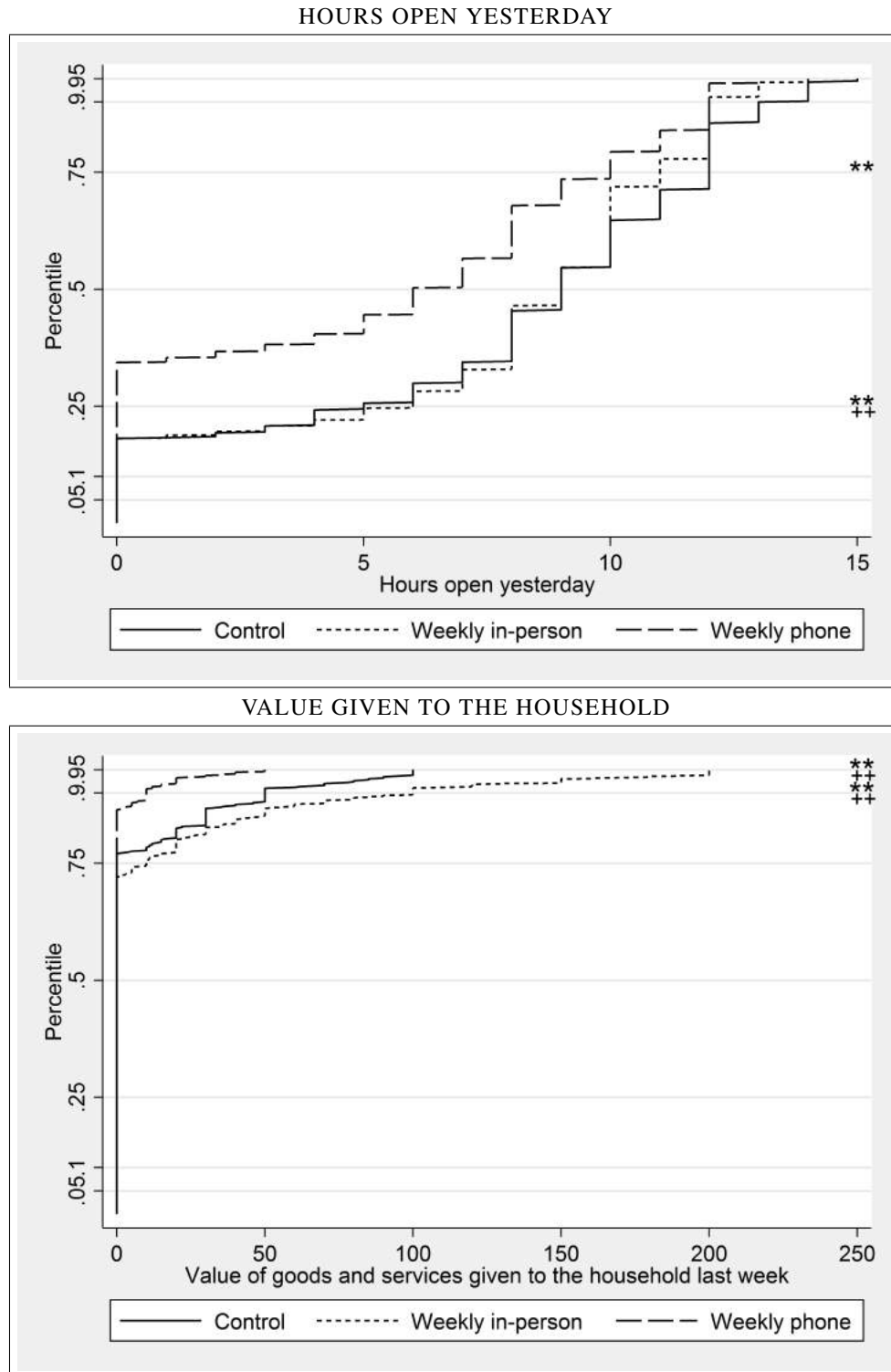
$$Y_{kit} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \phi_t + \varepsilon_{kit}, \quad (2)$$

where η_g are indicator variables for the randomisation strata, and ϕ_t are indicator variables for the survey week. We cluster errors by microenterprise. We test whether either treatment induces different average reporting by testing $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$. This model is more restrictive than the quantile regressions but the mean regressions have several desirable features. We can include randomisation stratum and survey week indicators, reducing the residual variation. We can also calculate exact minimum detectable effect sizes for all outcomes, which is more complex for quantile regression.³² We show all results in Table 5 and summarize the findings in Figure 6 Panel A.

We find some significant differences by interview medium and frequency in the repeated interviews, which are broadly consistent with the quantile regression results. The weekly interviews generate lower reported profits, different profit vs sales-costs

³² We calculate the minimum difference in the mean survey response we can detect for a test with 5% size and 80% power. We use the formula $MDE = (1.96 + 0.84) \cdot \sqrt{\frac{\sigma_Y^2}{\sigma_T^2} \cdot \frac{1}{N_M} \cdot \left(\rho_Y + \frac{1-\rho_Y}{N_W}\right)}$, where $1.96 + 0.84$ is derived from the chosen test size and power, N_M and N_W are respectively the number of microenterprises in the sample and the mean number of completed repeated interviews per enterprise, and the variance of each treatment indicator σ_T^2 is calculated from the sample of completed repeated interviews. The variance of the outcome σ_Y^2 and the intertemporal correlation coefficient ρ_Y is calculated for the monthly in-person interview group after conditioning on randomization block and survey week indicators. This approach updates prospective MDE calculations by using realized values of N_W , σ_T^2 , σ_Y^2 , and ρ_Y from the trial. But it calculates MDEs for chosen test size and power, rather than calculating power for the observed treatment effects. The latter approach is uninformative, as the retrospective power of a test is a one-to-one function of the p -value. See [Scheiner and Gurevich \(2001\)](#) for a detailed discussion on this issue. Note that MDEs calculated using our approach may be smaller than insignificant differences between treatment arms estimated from the sample data. This occurs because we set power at 80%, rather than 100%, and because the MDE formula does not account for heterogeneous treatment effects that increase σ_Y^2 .

Figure 5: Empirical CDFs in Repeated Interviews



Notes: We test for differences at each of the levels shown on the y -axis. We cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. + indicates rejection of the null hypothesis that weekly surveys in-person and by phone are equivalent; * indicates rejection of the null hypothesis that all three treatments are equivalent. +++/**, ++/**, and +/* denote significance at the 1, 5, and 10% levels.

Table 5: Regression Results: Average Treatment Effect for Repeated Interviews (equation 2)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Operating	Stock & inventory	Fixed assets	Profit	Sales last week	Sales last 4 weeks	Total costs	Profit check
Weekly in-person	0.017 (0.010)	0.737 (0.223)***	1.659 (1.598)	-0.110 (0.061)*	-0.070 (0.068)	-0.111 (0.067)*	-0.078 (0.066)	-0.101 (0.056)*
Weekly by phone	0.014 (0.010)	0.120 (0.155)	0.316 (0.431)	-0.098 (0.074)	-0.093 (0.072)	-0.092 (0.074)	-0.049 (0.069)	-0.020 (0.064)
Observations	4070	3989	3988	3986	3985	3987	3987	3984
Weekly treatments equal (<i>p</i>)	0.566	0.006***	0.287	0.843	0.673	0.748	0.538	0.029**
All treatments equal (<i>p</i>)	0.262	0.004***	0.568	0.191	0.429	0.251	0.475	0.027**
MDE: Weekly in-person	0.015	0.270	0.079	0.072	0.072	0.077	0.070	0.060
MDE: Weekly by phone	0.015	0.273	0.080	0.072	0.073	0.078	0.071	0.060
Lee bound: Weekly in-person (lower)	0.014	-0.303	-0.229	-0.287	-0.267	-0.284	-0.295	-0.288
Lee bound: Weekly in-person (upper)	0.039	0.909	2.106	-0.060	-0.010	-0.056	-0.045	-0.070
Lee bound: Weekly by phone (lower)	0.011	-0.303	-0.210	-0.278	-0.236	-0.237	-0.273	-0.259
Lee bound: Weekly by phone (upper)	0.038	0.138	0.434	-0.069	-0.065	-0.057	-0.029	0.008

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Employees	Full-time	Paid	Hours yesterday	Money kept	Household takings	Honest	Careful	Written records
Weekly in-person	0.067 (0.077)	0.015 (0.077)	0.123 (0.079)	0.040 (0.071)	0.090 (0.066)	0.313 (0.123)**	-0.152 (0.028)***	-0.109 (0.030)***	-0.015 (0.017)
Weekly by phone	0.043 (0.075)	-0.009 (0.081)	0.163 (0.078)**	-0.424 (0.070)***	0.012 (0.061)	-0.021 (0.113)	-0.378 (0.029)***	-0.271 (0.031)***	-0.109 (0.023)***
Observations	3987	3984	3973	3987	3986	3986	4056	4056	3987
Weekly treatments equal (<i>p</i>)	0.735	0.754	0.616	0.000***	0.162	0.009***	0.000***	0.000***	0.000***
All treatments equal (<i>p</i>)	0.681	0.950	0.096*	0.000***	0.280	0.015**	0.000***	0.000***	0.000***
MDE: Weekly in-person	0.104	0.113	0.118	0.082	0.063	0.118	0.038	0.036	0.025
MDE: Weekly by phone	0.105	0.114	0.119	0.082	0.064	0.119	0.039	0.036	0.026
Lee bound: Weekly in-person (lower)	-0.318	-0.361	-0.265	-0.235	-0.249	-0.261	-0.223	-0.203	-0.031
Lee bound: Weekly in-person (upper)	0.152	0.086	0.195	0.359	0.176	0.452	-0.033	-0.009	0.057
Lee bound: Weekly by phone (lower)	-0.201	-0.300	-0.110	-0.616	-0.236	-0.295	-0.446	-0.342	-0.129
Lee bound: Weekly by phone (upper)	0.093	0.037	0.216	-0.291	0.052	0.002	-0.338	-0.234	-0.012

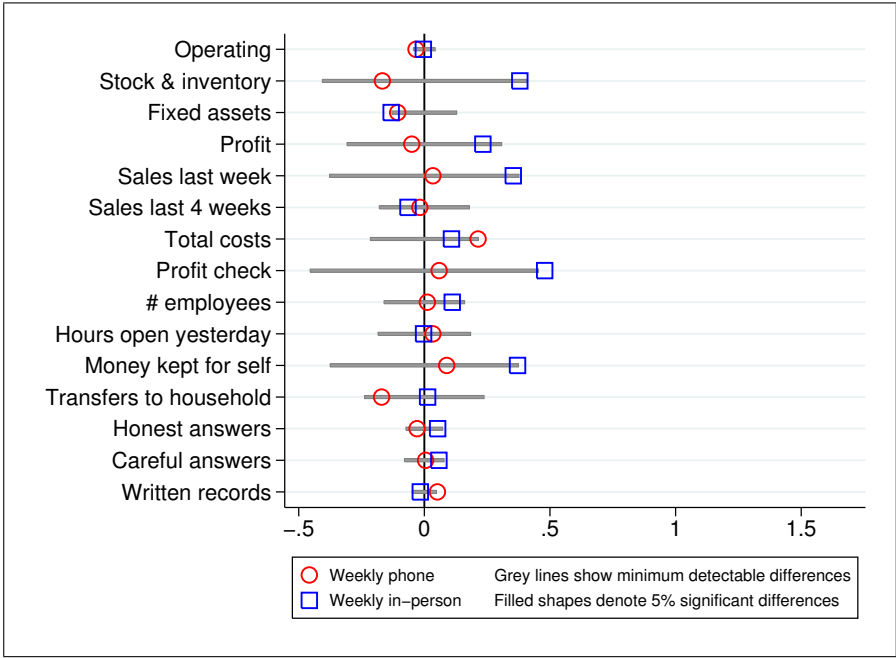
Notes: Coefficients are from regressions of each outcome on a vector of treatment indicators, randomization stratum fixed effects, and survey week fixed effects. Continuous outcomes are standardized to have mean zero and standard deviation one within survey week. Owners who close their enterprises are included in regressions only for panel A column 1 and panel B columns 7 and 8. heteroskedasticity-robust standard errors are shown in parentheses. ***, **, and * denote significance at the 1, 5, and 10% levels.

Figure 6: Average Treatment Effects for Repeated and Endline Interviews

PANEL A: REPEATED INTERVIEWS



PANEL B: ENDLINE INTERVIEWS



Notes: Coefficients are from regressions of each outcome on a vector of treatment indicators, randomization stratum fixed effects, and survey week fixed effects (repeated interviews only). Continuous outcomes are standardized to have mean zero and standard deviation one within survey week. Owners who close their enterprises are only included in the “Operating,” “Honest answers” and “Careful answers” regressions. Significance tests are based on heteroskedasticity-robust standard errors, clustering by enterprise for the repeated interviews. Minimum detectable differences are calculated using the approach discussed in footnote 32.

deviations, and higher reported fixed assets and stock/inventory.³³ We again see suggestive evidence of social desirability bias: phone respondents report opening their microenterprises for fewer hours yesterday but they do not keep more money for themselves or give more goods/services to their households. Respondents interviewed by phone are less likely to use written records to answer the survey, partly because the in-person interviews always take place at the enterprise while the phone surveys may not. Enumerator perceptions differ sharply by survey method: they report less carefulness and honesty for weekly in-person than monthly in-person respondents and even less for weekly phone respondents. This may reflect real differences in data quality or simply differences in enumerator perceptions. Phone interviews have slightly lower profit-sales-cost discrepancies than in-person interviews, which is our most direct test for data quality differences.

These mostly insignificant differences between reported enterprise outcomes do not reflect low power. The median minimum detectable difference across all outcomes between any two groups (monthly in-person, weekly in-person, weekly phone) is 0.075 standard deviations for continuous outcomes and 3.8 percentage points for binary outcomes.³⁴ All MDEs are smaller than 0.12 standard deviations/percentage points except for stock and inventory. This measure is relatively noisy and only weakly correlated with the randomisation block indicators.³⁵

We find limited evidence of heterogeneous interview method effects on six pre-

³³ The fixed assets and stock/inventory measures have long right tails, which explain the estimated differences. Neither set of differences is robust to trimming the right tail (Appendix C.1) and the differences in fixed assets are very imprecisely estimated.

³⁴ Recall that continuous outcomes are normalised to have mean 0 and standard deviation 1 in the monthly interview group. Categorical measures (# employees) and binary measures (enterprise closure, written records, enumerator assessments) are not normalized. The categorical variables can take values greater than one but this is very rare. Hence we discuss differences in these categorical variables in percentage point terms.

³⁵ In Appendix G Table A31 we show both p -values and sharpened q -values for all hypothesis tests in Table 5. The sharpened q -values account for multiple testing by showing the probability of incorrectly rejecting the null hypotheses of zero differences within pre-specified families of outcomes. Most of our q -values are larger than the corresponding p -values but the differences in stock and inventory, hours worked, and enumerators' perceptions of respondents honesty and care remain statistically significant. We also show in Appendix G the results from similar adjustments for the hypothesis tests in the next two subsections. The substantive conclusions from these subsections are robust to accounting for multiple corrections.

specified dimensions: gender, education, digit recall span score, numeracy score, use of written records at baseline, and having at least one employee other than the owner (see Appendix C.4). We might expect that enterprises with better record-keeping capacity (had multiple employees or kept written records at baseline) or owners with better numerical skills (education, digit recall span, and numeracy scores) would be less sensitive to interview medium. However, we observe few differences in response to interview frequency (weekly versus monthly in-person) by any of these five measures. The differences we observe are not consistent across the three measures of numerical skills or across the two measures of record-keeping capacity. The clearest differences are that male respondents report holding more stock and taking more money from the enterprise for their own use when interviewed weekly.

There is more evidence of heterogeneity in responses to interview medium (weekly phone vs weekly in-person) but these differences are generally imprecisely estimated and do not follow a clear pattern. Most notably, enumerators consistently assess phone respondents as less careful and less honest than in-person respondents, but there is substantial heterogeneity in this relationship. Higher numerical skills and using written records at baseline result in a smaller perceived carefulness/honesty penalty. These respondents may give more confident or more precise answers that partly offset enumerators' perception that phone interview respondents are less careful and honest. But given the number of dimensions of heterogeneity tested here, the heterogeneity we observe may simply reflect sampling variation rather than heterogeneity in reporting or behaviour.

We also test for difference in dispersion for each non-binary outcome. We calculate the within-enterprise standard deviation for each outcome through the repeated interviews and estimate

$$S_{ki} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \mu_{ki} \quad (3)$$

using heteroskedasticity-robust standard errors. We then test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$. We show in Table 6 that weekly reports of stock and inventory, household takings, and fixed assets are more dispersed than monthly

reports (the latter difference is large but imprecisely estimated). The higher-frequency interviews may pick up large short-term fluctuations in these measures. It is perhaps surprising that we see less evidence of differential dispersion in flow measures such as profits, sales, and costs. The phone respondents' responses show more dispersion in number of (paid) employees, hours of operation, costs, and hence our profit check measure.

We do not find any systematic differences in interview medium or frequency effects between reported flow measures (e.g. profits, sales) and stock measures (e.g. fixed assets, number of employees). We might expect the latter to be more strongly correlated through time, so more frequent interviews would yield less additional information. However, the flow and stock measures are not differentially dispersed in the pooled sample (Figure 2) and the interview frequency and medium mean effects are not systematically different (Table 5). This arises in part because even stock measures in our sample show high dispersion through time. We show some of this dispersion in Appendix J.

We also consider the implications of interview medium for estimating the dynamics of microenterprise performance. Up to this point, we have focused on whether different interview methods have different implications for testing the level of different variables. However, one key advantage of panel data is that researchers can model the dynamics of enterprise behaviour; this is potentially important for many reasons, including for understanding the trajectory of enterprise responses to shocks. We therefore compare estimated microenterprise dynamics by interview medium. To do this, we report Blundell-Bond estimates for $\log(\text{profit})$ and $\log(\text{capital})$, assuming an AR(1) structure on the error term (Blundell and Bond, 1998); we use two lags for the former and four lags for the latter.³⁶ Table 7 reports these estimates, disaggregated between weekly phone interviews and weekly in-person interviews. At the bottom of the table, we report Wald tests for the null hypothesis that the estimated dynamics are equivalent between phone and in-person interviews; in each case, the null hypothesis of

³⁶ We justify this approach empirically in Appendix J.

Table 6: Regression Results: Effect on Sample Standard Deviation for Repeated Interviews (equation 3)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Stock & inventory	Fixed assets	Profit	Sales last week	Sales last 4 weeks	Total costs	Profit check
Weekly in-person	0.486	5.241	-0.004	-0.042	-0.045	0.039	0.008
	(0.129)***	(5.050)	(0.050)	(0.053)	(0.052)	(0.040)	(0.051)
Weekly by phone	0.151	1.638	0.061	-0.058	-0.025	0.123	0.151
	(0.099)	(1.644)	(0.078)	(0.051)	(0.049)	(0.050)**	(0.065)**
Observations	610	609	608	608	608	608	608
Weekly treatments equal (<i>p</i>)	0.012**	0.308	0.293	0.617	0.512	0.038**	0.002***
All treatments equal (<i>p</i>)	0.001***	0.583	0.576	0.520	0.648	0.039**	0.009***
MDE: Weekly in-person	0.188	5.415	0.094	0.066	0.061	0.067	0.080
MDE: Weekly by phone	0.187	5.390	0.094	0.066	0.061	0.066	0.080

34

	(1)	(2)	(3)	(4)	(5)	(6)
	Employees	Full-time	Paid	Hours yesterday	Money kept	Household takings
Weekly in-person	-0.031	-0.006	-0.015	-0.000	0.053	0.494
	(0.048)	(0.055)	(0.043)	(0.064)	(0.093)	(0.171)***
Weekly by phone	0.095	0.014	0.052	0.087	-0.027	0.359
	(0.051)*	(0.055)	(0.047)	(0.052)*	(0.079)	(0.177)**
Observations	608	608	608	608	608	608
Weekly treatments equal (<i>p</i>)	0.001***	0.663	0.079*	0.076*	0.282	0.477
All treatments equal (<i>p</i>)	0.006***	0.907	0.208	0.087*	0.560	0.010*
MDE: Weekly in-person	0.075	0.078	0.069	0.085	0.122	0.304
MDE: Weekly by phone	0.075	0.078	0.068	0.085	0.121	0.302

Notes: Coefficients are from regressions of the standard deviation of each outcome on a vector of treatment indicators and randomization stratum fixed effects. Owners who close their enterprises are excluded from these regressions. Heteroskedasticity-robust standard errors are shown in parentheses. ***, **, and * denote significance at the 1, 5, and 10% levels.

equivalence comfortably passes.³⁷ We conclude that phone interviews and in-person interviews are equally good for this specific approach to modeling microenterprise dynamics.

4.2 Does interview method affect microenterprise performance?

We conduct an in-person endline with all microenterprises and compare endline responses by *previous* interview frequency and medium. We interpret this as a test of whether actual microenterprise outcomes change due to interview method. For example, more frequent interviews may prompt microenterprise managers to pay more attention to their capital stock.

We begin by repeating the analysis of empirical CDFs.³⁸ We then compare means by estimating

$$Y_{ki} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \varepsilon_{ki} \quad (4)$$

for each outcome Y_{ki} , using heteroskedasticity-robust standard errors. As before, we test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$.

We find few difference in endline enterprise outcomes based on previous interview method, as shown in Table 8 and Figure 6 Panel B. Weekly respondents report slightly lower values of fixed assets than monthly respondents. Phone respondents report slightly lower household takings and are more likely to use written records, flipping the pattern found in the repeated interviews. We show in Figure 7 that the former difference is driven by the upper tail of the distribution. Comparisons of the empirical CDFs for the other outcomes provide little useful information and are shown in Appendix H. We cannot rule out moderate differences based on the MDEs (median = 0.27 standard deviations for continuous outcomes and 7.9 percentage points for binary outcomes). In particular, weekly in-person respondents report higher stock/inventory,

³⁷ Since the sample collected by phone and the sample collected in person are non-overlapping, we run this test by imposing that the cross-equation covariance terms are zero.

³⁸ Our pre-analysis plan specified that we would test whether the variance of measured Y_{ki} differs by group. We later decided that quantile regression is more intuitive and informative approach to speak to the same issue.

Table 7: Blundell-Bond Estimates: Disaggregating by Treatment

	(1) Profit <i>In-person</i>	(2) Profit <i>Phone</i>	(3) Capital <i>In-person</i>	(4) Capital <i>Phone</i>
Lag 1	0.477*** (0.070)	0.363*** (0.079)	0.459*** (0.115)	0.380*** (0.103)
Lag 2	0.305*** (0.114)	0.261*** (0.076)	0.194*** (0.075)	0.195* (0.115)
Lag 3			0.158*** (0.048)	0.137*** (0.046)
Lag 4			0.175*** (0.065)	0.104*** (0.036)
Time dummies	✓	✓	✓	✓
Observations	520	289	420	418
Firms	132	87	116	89
Arellano-Bond: AR(1) (<i>p</i> -value)	0.000***	0.000***	0.024**	0.003***
Arellano-Bond: AR(2) (<i>p</i> -value)	0.302	0.168	0.325	0.140
Arellano-Bond: AR(3) (<i>p</i> -value)	0.340	0.319	0.783	0.967
Arellano-Bond: AR(4) (<i>p</i> -value)	0.242	0.392	0.323	0.497
Hansen test (<i>p</i> -value)	0.199	0.988	0.288	0.596
H_0 : Equal parameter estimates (<i>p</i> -value)		0.548		0.907

Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The final row reports Wald tests of the a null hypothesis of equal lag coefficients between in-person and phone data.

To run this test, we impose zero parameter covariance between estimations (since the estimation samples are non-overlapping).

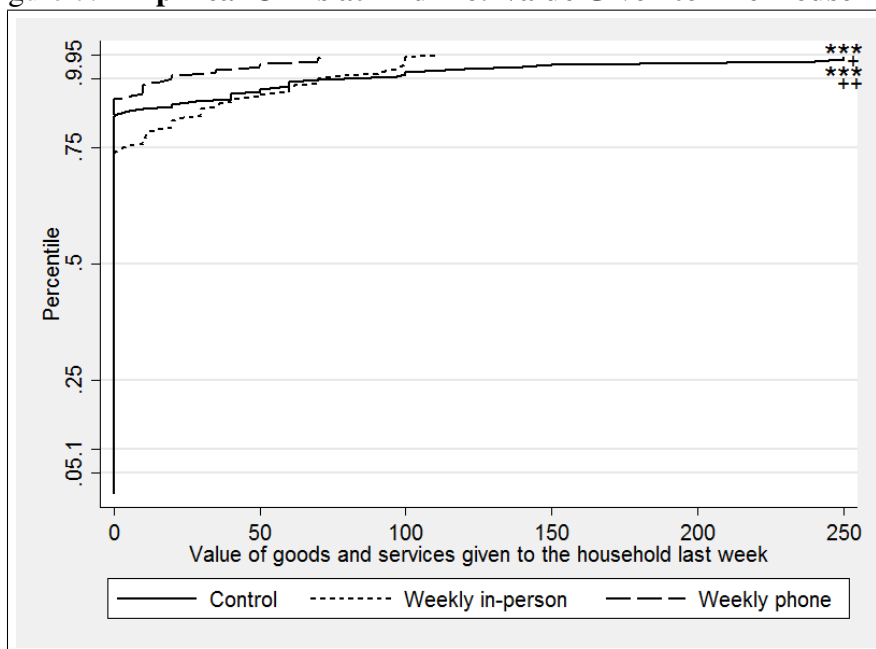
Table 8: Regression Results: Average Treatment Effect for Endline Interviews (equation 4)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Operating	Stock & inventory	Fixed assets	Profit	Sales last week	Sales last 4 weeks	Total costs	Profit check
Weekly in-person	-0.004 (0.026)	0.380 (0.243)	-0.132 (0.078)*	0.233 (0.222)	0.354 (0.258)	-0.065 (0.092)	0.108 (0.106)	0.479 (0.297)
Weekly by phone	-0.033 (0.028)	-0.167 (0.162)	-0.106 (0.091)	-0.050 (0.141)	0.034 (0.155)	-0.018 (0.124)	0.214 (0.147)	0.059 (0.169)
Observations	594	546	546	546	546	546	546	546
Weekly treatments equal (<i>p</i>)	0.281	0.082*	0.722	0.024**	0.110	0.703	0.481	0.131
All treatments equal (<i>p</i>)	0.437	0.217	0.241	0.033**	0.276	0.773	0.293	0.260
MDE: Weekly in-person	0.042	0.394	0.125	0.299	0.366	0.174	0.209	0.441
MDE: Weekly by phone	0.045	0.418	0.133	0.317	0.388	0.185	0.221	0.468
Lee bound: Weekly in-person (lower)	-0.008	-0.022	-0.121	-0.053	-0.051	-0.082	-0.042	-0.019
Lee bound: Weekly in-person (upper)	0.003	0.523	-0.024	-0.040	0.096	-0.075	0.120	0.607
Lee bound: Weekly by phone (lower)	-0.099	-0.264	-0.177	-0.273	-0.206	-0.117	0.127	0.013
Lee bound: Weekly by phone (upper)	-0.018	0.192	0.195	0.172	0.234	0.324	0.513	0.234

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Employees	Full-time	Paid	Money kept	Household takings	Hours yesterday	Honest	Careful	Written records
Weekly in-person	0.111 (0.090)	0.071 (0.092)	0.094 (0.091)	0.371 (0.272)	0.013 (0.138)	-0.003 (0.095)	0.053 (0.045)	0.058 (0.048)	-0.016 (0.030)
Weekly by phone	0.012 (0.113)	-0.086 (0.096)	0.254 (0.263)	0.089 (0.162)	-0.170 (0.095)*	0.034 (0.128)	-0.029 (0.048)	0.005 (0.050)	0.052 (0.030)*
Observations	546	546	546	546	546	546	593	593	546
Weekly treatments equal (<i>p</i>)	0.348	0.081*	0.574	0.050*	0.297	0.770	0.077*	0.273	0.018**
All treatments equal (<i>p</i>)	0.402	0.219	0.340	0.113	0.203	0.955	0.191	0.397	0.049**
MDE: Weekly in-person	0.156	0.145	0.353	0.364	0.231	0.179	0.073	0.077	0.047
MDE: Weekly by phone	0.165	0.153	0.374	0.386	0.245	0.190	0.077	0.081	0.050
Lee bound: Weekly in-person (lower)	0.070	-0.003	0.023	0.013	0.009	0.000	0.051	0.009	-0.009
Lee bound: Weekly in-person (upper)	0.144	0.106	0.036	0.465	0.046	0.043	0.119	0.107	-0.002
Lee bound: Weekly by phone (lower)	-0.049	-0.141	0.129	0.046	-0.191	-0.098	-0.101	-0.056	-0.066
Lee bound: Weekly by phone (upper)	0.348	0.265	0.514	0.207	0.047	0.269	0.033	0.058	0.082

Notes: Coefficients are from regressions of each outcome on a vector of treatment indicators and randomization stratum fixed effects. Continuous outcomes are standardized to have mean zero and standard deviation one. Owners who close their enterprises are included in regressions only for panel A column 1 and panel B columns 7 and 8. heteroskedasticity-robust standard errors are shown in parentheses. ***, **, and * denote significance at the 1, 5, and 10% levels.

Figure 7: **Empirical CDFs at Endline: Value Given to The Household**



Notes: We test for differences at each of the levels shown on the y -axis. We cluster by microenterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. + indicates rejection of the null hypothesis that weekly surveys in-person and by phone are equivalent; * indicates rejection of the null hypothesis that all three treatments are equivalent. +++/***, ++/**, and +/* denote significance at the 1, 5, and 10% levels.

sales, and profits than monthly in-person respondents but these differences are not statistically significant. Weekly phone respondents do not have the same reporting patterns so we do not regard this as strong evidence of reminder effects.

Figure 6 summarizes the differences between data collection methods for both repeated and endline interviews. We note two important patterns. First, differences in the endline interviews are smaller than in the repeated interviews: mean absolute value 0.12 versus 0.19. They are also less precisely estimated, as the differences in repeated interviews use up to 12 data points per enterprise. Second, there is little systematic relationship between the repeated and endline differences: the correlation coefficient for outcome differences across the two types of interviews is -0.08. We cannot reject that the repeated and endline differences are equal for most measures but

this partly reflects the imprecision of the endline differences.³⁹ This pattern of results is not consistent with sustained behavioural change due to differences in interview methods.

4.3 Accounting for differential missed interview rates

We observe some differences in the rates of missed interviews. We check the robustness of our findings in two ways. First, we estimate the effects of interview frequency and medium on responses assuming that the differential missed-interview rates are concentrated in the lower or upper tail of the observed outcome distribution. These generate respectively upper and lower bounds on the effects of interview method and frequency (Lee, 2009).⁴⁰ We report these bounds in Tables 5 and 8. Researchers typically interpret results as “robust” if the bounds exclude zero. This interpretation is inappropriate for our research question, where zero and non-zero coefficients are both of interest. Instead, we use the width of the bounded set as a measure of the vulnerability of our findings to differential interview miss rates across interview media and frequencies. For the repeated interviews, the median width is 0.28 standard deviations for continuous outcomes and 23 percentage points for binary outcomes. Sets are very wide only for the long-tailed stock/inventory and fixed asset measures. We can rule out differences for most other outcomes between interview media and frequencies of more than 0.4 standard deviations for continuous outcomes and 40 percentage points for binary outcomes. The bounded sets are tighter for the endline interviews, reflecting the lower and more balanced rates of missed interviews shown in Table 2. The median width is 0.30 standard deviations for continuous outcomes and 10 percentage points for binary outcomes.

Second, we estimate the probability of successfully completing each repeated or

³⁹ To conduct this test, we estimate the mean differences in repeated and endline interviews using a system of equations. We show the results in Appendix F. We reject the hypothesis of equal differences for enumerator perceptions and use of written records but do not reject for most enterprise outcomes.

⁴⁰ Lee’s original method applies only to binary comparisons. We use this method to calculate bounds for pairwise comparisons between monthly in-person interviews and each of weekly in-person and weekly phone interviews. We do not report bounds for comparing weekly in-person and weekly phone interviews.

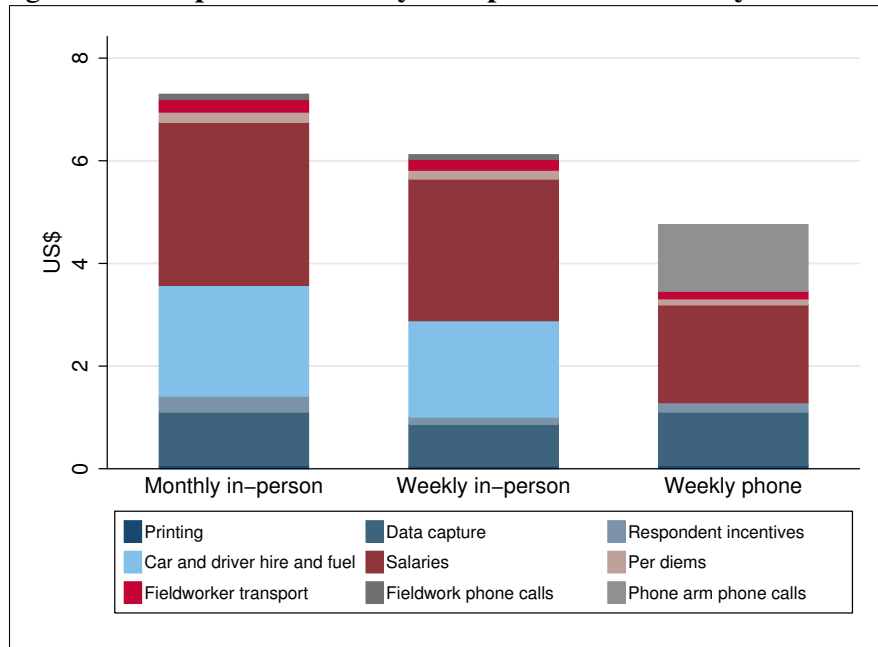
endline interview as a function of the baseline characteristics discussed in section 2.2: $\hat{P} = \hat{P}_r(\text{Interview Success}_{it}|X_{i0})$. We then construct inverse probability of interview success weights, $\hat{W} = 1/\hat{P}$, and estimate equations 2 and 4 using these weights. If interview misses do not covary with latent outcomes after conditioning on these baseline characteristics, the weighted regressions fully correct for missed interviews. The weighted results are shown in Appendices – E and differ only slightly from the unweighted results. They provide no evidence that calls for revising the conclusions reached above.

5 Cost effectiveness

Phone surveys may reduce per-interview costs, allowing larger samples or longer panels. We provide a detailed breakdown of our data collection costs to inform other researchers about the potential savings. We use the survey firm’s general ledger entries, which give a detailed breakdown of expenditure by date and purpose. We exclude the costs of the screening, baseline, and endline surveys – that were conducted in-person for all respondents – and fixed costs – such as the survey manager’s salary and the survey firm’s rent and office costs.

We split the costs of the repeated interviews into nine categories. Enumerator salaries, enumerator per diems and enumerator transport allowances are reported for each enumerator and enumerators worked in only one treatment arm, so we can easily allocate these costs to treatment arms. One car was hired for each of the weekly in-person and the monthly in-person interview teams, so these are also easily allocated. There are two types of phone calls, phone calls to arrange in-person interviews and phone interviews. However, the survey firm separated phone calls to arrange in-person interviews (conducted on a project mobile phone) from phone calls to survey firms in the phone arm (conducted on landlines), so these are also easily allocated. However, we observe only the total cost for the last three categories: printing, data capture, and respondent incentives. We allocate these to treatment arms by dividing the total in proportion to the number of successfully completed interviews.

Figure 8: Cost per Successfully Completed Interview by Treatment



Notes: This figure shows the average per-interview cost interview by treatment group. The average cost is constructed by summing the nine cost categories shown above, separately by treatment group, then dividing this by the number of successfully completed interviews in each group. This calculation excludes the costs of the in-person screening, baseline, and endline interview; and excludes fixed costs such as office rental and management salaries. US\$ values are calculated using at the South African Reserve Bank exchange rate on 31 August 2013: 1 US\$D = 10.27 ZAR.

Once a baseline was conducted to collect contact details, a phone interview conducted weekly cost roughly ZAR49 (US\$4.76 at August 2013 rates).⁴¹ An in-person interview conducted weekly cost ZAR63 (US\$6.12) and one conducted monthly cost ZAR75 (US\$7.30). Phone surveys reduced the per-interview cost of high-frequency data collection by approximately 25%. High-frequency in-person interviews cost approximately 15% less than low-frequency in-person interviews because a more flexible interview schedule allowed enumerators to spend less time and money on travel. All costs are per successfully completed interview. More phone than in-person interviews were missed, so this approach overstates the relative cost per attempted phone inter-

⁴¹ This is similar to the per-interview cost range of US\$4.10 - US\$7.10 for mobile phone surveys in a Dar es Salaam panel study (Croke et al., 2014).

view.

Cost savings from phone interviews will increase as the time and expense of travelling between interviews increase and as the costs of calling mobile phones decrease. Our cost savings are relatively low because we were interviewing in a high density urban area and South African call costs are relatively high (interviews cost roughly US\$1.30 per 15 minute interview). In contrast, a Tanzanian high-frequency household survey with farmers in remote rural areas spent US\$97 per in-person baseline interview and US\$7 per phone follow-up survey (Dillon, 2012).

Our largest cost saving came from transport costs. Enumerators doing phone interviews received an allowance for transport to the office, while enumerators doing in-person interviews met in a central place close to their houses and were transported to the enterprises. This yielded a total transport cost of US\$0.15 per phone interview and US\$2.09 per in-person interview. The cost saving may be even larger if enumerators require overnight travel and incur accommodation costs, which was not the case in our study.

Most of our remaining cost savings came from enumerator salaries and per diems (US\$2.02 for phone and US\$2.93 for in-person). We achieved this saving because we assigned respectively four and eight enumerators to conduct the weekly phone and weekly in-person interviews. We could have further reduced the staff costs for the phone interviews: enumerators were able to conduct more phone interviews per day than we had expected, but we did not change the original allocation. Enumerators were paid the same daily rate and per diem for phone and in-person surveys to avoid differences in motivation and incentives.

All our surveys were done on paper and data-captured to avoid differences in data quality between interview methods. But data capture can be easier, cheaper and more accurate for phone than in-person surveys. Both phone and in-person interviews could be conducted on tablets in the field, but phone interviews can be captured directly on computers using more advanced data capture software.

6 Conclusion

We provide the first experimental evidence on the usefulness and viability of high-frequency phone surveys of microenterprises. We draw a representative sample of microenterprises in Soweto and randomly assign each respondent to one of three interview methods. The first group is interviewed in-person at monthly intervals, to mimic a standard method of collecting data from microenterprises. The second group is interviewed in-person at weekly intervals. This allows us to test the consequences of collecting data at a much higher frequency, holding interview medium fixed. The third group are interviewed at weekly intervals by mobile phone.

Our results show that mobile phone interviews are accurate: on most measures and at most quantiles of the distribution, data patterns are indistinguishable in interviews conducted weekly, whether by mobile phone or in-person). Using an endline administered in-person, we find little evidence that high-frequency data collection (either over a mobile phone or in person) alters microenterprise behavior. We find some differences in attrition (levels and reasons) across data collection methods but these are generally small and are not correlated with microenterprise characteristics. Mobile phone interviews are, however, substantially cheaper than in-person interviews and offer a far larger volume of data for the same price.

Our results also show that weekly phone interviews are useful: they capture extensive volatility in a number of measures that is not visible in less frequent data collection. Low-frequency data collection both fails to observe this volatility and captures measures that do not account for the past history of volatility. This can lead to measurement error problems that produce both biased and less precisely estimated coefficient estimates in regression analysis of microenterprise outcomes. We conclude that mobile phone data collection offers considerable cost savings with little reduction in data quality.

References

ABBRING, J. AND J. HECKMAN (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choices, and General

- Equilibrium Policy Evaluation,” in *Handbook of Econometrics Volume 6B*, ed. by J. Heckman and E. Leamer, Elsevier, 5145–5303.
- BAIRD, S., J. H. HICKS, M. KREMER, AND E. MIGUEL (Forthcoming): “Worms at work: Long-run impacts of a child health investment,” *Quarterly Journal of Economics*.
- BANERJEE, A., E. DUFLO, R. GLENNERSTER, AND C. KINNAN (2015): “The Miracle of Micro-finance? Evidence from a Randomized Evaluation,” *American Economic Journal: Applied Economics*, 7, 22–53.
- BAUER, J.-M., K. AKAKPO, M. ENLUND, AND S. PASSERI (2013): “A New Tool in the Toolbox: Using Mobile Text for Food Security Surveys in a Conflict Setting,” *Humanitarian Practice Network Online Exchange* (<http://www.odihpn.org/the-humanitarian-space/news/announcements/blog-articles/a-new-tool-in-the-toolbox-using-mobile-text-for-food-security-surveys-in-a-conflict-setting>), 1–2.
- BEAMAN, L., J. MAGRUDER, AND J. ROBINSON (2014): “Minding Small Change: Limited Attention among Small Firms in Kenya,” *Journal of Development Economics*, 108.
- BEEGLE, K., J. DE WEERDT, J. FRIEDMAN, AND J. GIBSON (2012): “Methods of Household Consumption Measurement Through Surveys: Experimental Results from Tanzania,” *Journal of Development Economics*, 98, 3–18.
- BEEGLE, K., J. D. WEERDT, AND S. DERCON (2011): “Migration and Economic Mobility in Tanzania: Evidence from a Tracking Survey,” *Review of Economics and Statistics*, 93, 1010–1033.
- BENJAMINI, Y., A. M. KRIEGER, AND D. YEKUTIELI (2006): “Adaptive Linear Step-Up Procedures that Control the False Discovery Rate,” *Biometrika*, 93, 491–507.
- BLUNDELL, R. AND S. BOND (1998): “GMM estimation with Persistent Panel Data: An Application to Production Functions,” Tech. Rep. W99/4, Institute for Fiscal Studies.
- BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): “Measurement Error in Survey Data,” in *Handbook of Econometrics Volume 5*, ed. by J. Heckman and E. Leamer, Elsevier, 3705–3843.
- BRUHN, M. AND D. MCKENZIE (2009): “In Pursuit of Balance: Randomization in Practice in Development Field Experiments,” *American Economic Journal: Applied Economics*, 200–232.
- CAEYERS, B., N. CHALMERS, AND J. DE WEERDT (2012): “Improving Consumption Measurement and Other Survey Data through CAPI: Evidence from a Randomized Experiment,” *Journal of Development Economics*, 98, 19–33.
- COLLINS, D., J. MORDUCH, S. RUTHERFORD, AND O. RUTHVEN (2009): *Portfolios of the Poor: How the World’s Poor Live on \$2 a Day*, Princeton: Princeton University Press.
- CROKE, K., A. DABALEN, G. DEMOMBYNES, M. GIUGALE, AND J. HOOGEVEEN (2014): “Collecting High Frequency Panel Data in Africa using Mobile Phone Interviews,” *Canadian Journal of Development Studies*, 35, 186–207.

- DABALEN, A., A. ETANG, J. HOOGEVEEN, E. MUSHI, Y. SCHIPPER, AND J. VON ENGELHARDT (2016): *Mobile Phone Panel Surveys in Developing Countries: A Practical Guide for Microdata Collection*, World Bank Directions in Development.
- DAS, J., J. HAMMER, AND C. SÁNCHEZ-PARAMO (2012): “The Impact of Recall Periods on Reported Morbidity and Health Seeking Behavior,” *Journal of Development Economics*, 98, 76–88.
- DE LEEUW, E. (1992): *Data Quality in Mail, Telephone and Face to Face Surveys*, Amsterdam: TT Publikaties.
- DE MEL, M. D. W. C. (2009): “Measuring microenterprise profits: Must we ask how the sausage is made?” *Journal of Development Economics*, 88, 19–31.
- DE MEL, S., D. MCKENZIE, AND C. WOODRUFF (2008): “Returns to Capital in Microenterprises: Evidence from a Field Experiment,” *Quarterly Journal of Economics*, 123, 1329–1372.
- DE NICOLA, F. AND X. GINÉ (2014): “How Accurate are Recall Data? Evidence from Coastal India,” *Journal of Development Economics*, 106, 52–65.
- DILLON, B. (2012): “Using Mobile Phones to Collect Panel Data in Developing Countries,” *Journal of International Development*, 24, 518–27.
- DREXLER, A., G. FISCHER, AND A. SCHOAR (2014): “Keeping it Simple: Financial Literacy and Rules of Thumb,” *American Economic Journal: Applied Economics*, 6, 1–31.
- FAFCHAMPS, M., D. MCKENZIE, S. QUINN, AND C. WOODRUFF (2012): “Using PDA Consistency Checks to Increase the Precision of Profits and Sales Measurement in Panels,” *Journal of Development Economics*, 98, 51–57.
- (2014): “Microenterprise Growth and the Flypaper Effect: Evidence from a Randomized Experiment in Ghana,” *Journal of Development Economics*, 106, 211–226.
- FRANKLIN, S. (2015): “Job Search, Transport Costs and Youth Unemployment: Evidence from Urban Ethiopia,” *Working paper: University of Oxford*.
- FRIEDMAN, J., K. BEEGLE, J. D. WEERDT, AND J. GIBSON (2016): “Decomposing Response Errors in Food Consumption Measurement,” Tech. Rep. 7646, World Bank Policy Research Working Paper.
- FRISON, L. AND S. POCOCCO (1992): “Repeated Measures in Clinical Trials Analysis using Mean Summary Statistics and its Implications for Design,” *Statistics in Medicine*, 11, Statistics in Medicine.
- GALLUP (2012): “The World Bank Listening to LAC (L2L) Pilot: Final Report,” *Gallup Report*.
- GROSH, M. AND J. MUNOZ (1996): *A Manual for Planning and Implementing the Living Standards Measurement Study Survey*, The World Bank.
- GROVES, R. (1990): “Theories and Methods of Telephone Surveys,” *Annual Review of Sociology*, 16, 221–240.
- GROVES, R., P. BIEMER, L. LYBERG, J. MASSEY, W. NICHOLLS, AND J. WAKSBERG (2001): *Telephone Survey Methodology*, Wiley.

- GROVES, R. M. (1979): “Actors and Questions in Telephone and Personal Interview Surveys,” *Public Opinion Quarterly*, 43, 190–205.
- HEATH, R., G. MANSURI, D. SHARMA, B. RIJKERS, AND W. SEITZ (2016): “Measuring Employment in Developing Countries: Evidence from a Survey Experiment,” Working paper.
- HOLBROOK, A. L., M. C. GREEN, AND J. A. KROSNICK (2003): “Telephone vs. Face-to-Face Interviewing of National Probability Samples With Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias,” *Public Opinion Quarterly*, 67, 79–125.
- IMBENS, G. (2015): “Matching Methods in Practice: Three Examples,” *Journal of Human Resources*, 50, 373–419.
- JACKLE, A., C. ROBERTS, AND P. LYNN (2006): “Telephone Versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project,” *Institute for Social and Economic Research Working Paper*, 41.
- JACOBSON, L., R. LALONDE, AND D. SULLIVAN (1993): “Earnings Losses of Displaced Workers,” *American Economic Review*, 83, 685–709.
- KARLAN, D., R. KNIGHT, AND C. UDRY (2012): “Hoping to Win, Expected to Lose: Theory and Lessons on Micro Enterprise Development,” Working paper, Yale University.
- KARLAN, D. AND M. VALDIVIA (2011): “Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions,” *The Review of Economics and Statistics*, 93, 510–27.
- KÖRMENDI, E. (2001): “The Quality of Income Information in Telephone and Face-to-Face Surveys,” in *Telephone Survey Methodology*, ed. by R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, and J. Waksberg, New York: John Wiley and Sons.
- KROSNICK, J. A. (1991): “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys,” *Applied Cognitive Psychology*, 5, 213–236.
- LANE, S. J., N. M. HEDDLE, E. ARNOLD, AND I. WALKER (2006): “A Review of Randomized Controlled Trials Comparing the Effectiveness of Hand Held Computers with Paper Methods for Data Collection,” *BMC Medical Informatics and Decision Making*, 6, 1–10.
- LEE, D. S. (2009): “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *The Review of Economic Studies*, 76, 1071–1102.
- LEO, B., R. MORELLO, J. MELLON, T. PEIXOTO, AND S. DAVENPORT (2015): “Do Mobile Phone Surveys Work in Poor Countries?” *Centre for Global Development Working Paper Series*, 398, 1–65.
- MCKENZIE, D. (2012): “Beyond Baseline and Follow-up: The Case for More T in Experiments,” *Journal of Development Economics*, 99, 210–221.
- (2015a): “Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition,” Tech. Rep. 7391, World Bank Policy Research Working Paper.

- (2015b): “Three strikes and they are out? Persistence and Reducing Panel Attrition among Firms,” [Http://blogs.worldbank.org/impactevaluations/three-strikes-and-they-are-out-persistence-and-reducing-panel-attrition-among-firms](http://blogs.worldbank.org/impactevaluations/three-strikes-and-they-are-out-persistence-and-reducing-panel-attrition-among-firms).
- MCKENZIE, D. AND C. WOODRUFF (2008): “Experimental Evidence on Returns to Capital and Access to Finance in Mexico,” *World Bank Economic Review*, 22, 457–82.
- MITULLAH, W. AND P. KAMA (2013): *The Partnership of Free Speech and Good Governance in Africa*, vol. 3, Cape Town: Afrobarometer, University of Cape Town.
- PAPKE, L. AND J. WOOLDRIDGE (1996): “Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates,” *Journal of Applied Econometrics*, 11, 619–632.
- ROBINS, J. (1997): “Causal inference from complex longitudinal data,” in *Latent Variable Modeling and Applications to Causality*, ed. by M. Berkane, Springer-Verlag, 69–117.
- ROSENZWEIG, M. AND K. WOLPIN (1993): “Credit Market Constraints, Consumption Smoothing and the Accumulation of Durable Production Assets in Low-Income Countries: Investments in Bullocks in India,” *Journal of Political Economy*, 101, 223–244.
- SCHEINER, S. AND J. GUREVICH (2001): *Design and Analysis of Ecological Experiments*, Oxford University Press, 3rd ed.
- SCHWARZ, N., H. HIPPLER, AND E. NOELLE-NEUMANN (1992): “A Cognitive Model of Response-Order Effects in Survey Measurement,” in *Context Effects in Social and Psychological Research*, ed. by N. Schwarz and S. Sudman, Springer-Verlag, 187–201.
- SILVA, J. S. AND P. M. PARENTE (2013): “Quantile Regression with Clustered Data,” *Economics Discussion Papers, University of Essex, Department of Economics*, 728.
- SMITH, P. B. AND R. FISCHER (2008): “Acquiescence, Extreme Response Bias and Culture: A Multi-level Analysis,” in *Multilevel Analysis of Individuals and Cultures*, ed. by F. J. R. van de Vijver, D. A. van Hemert, and Y. H. Poortinga, Taylor and Francis Group and Lawrence Erlbaum Associates, 285–314.
- STANGO, V. AND J. ZINMAN (2013): “Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees,” *Working paper: Dartmouth College*.
- THOMAS, D., F. WITOELAR, E. FRANKENBERG, B. SIKOKI, J. STRAUSS, C. SUMANTRI, AND W. SURIASTINI (2012): “Cutting the costs of attrition: Results from the Indonesia Family Life Survey,” *Journal of Development Economics*, 98, 108–123.
- TURAY, A., S. TURAY, R. GLENNESTER, K. HIMELEIN, N. ROSAS, T. SURI, AND N. FU (2015): “The Socio-Economic Impacts of Ebola in Sierra Leone: Results from a High Frequency Cell Phone Survey,” Note prepared by Statistics Sierra Leone, the World Bank, and Innovations for Poverty Action.
- VAN DER WINDT, P. AND M. HUMPHREYS (2013): “Crowdseeding Conflict Data,” *Working paper: Columbia University*.

ZWANE, A. P., J. ZINMAN, E. VAN DUSEN, W. PARIENTE, C. NULL, E. MIGUEL, M. KREMER, D. KARLAN, R. HORNBECK, X. GINÉ, E. DUFLO, F. DEVOTO, B. CREPON, AND A. BANERJEE (2011): “Being Surveyed can Change Later Behavior and Related Parameter Estimates,” *Proceedings of the National Academy of Sciences*, 108, 1821–1826.