PEDL Research Papers

[PEDL Twitter](#)

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

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

April 24, 2017

Abstract

We run the first randomised controlled trial to compare microenterprise data from surveys of different frequency and medium. We randomly assign enterprises to monthly in-person, weekly in-person, or weekly phone panel surveys. Higher-frequency interviews have little effect on enterprise outcomes or reporting behaviour. They generate more data, capture short-term fluctuations, and may increase reporting accuracy. They result in lower response rates for each interview but do not increase permanent attrition from the panel. Conducting high-frequency interviews by phone, not in-person, lowers costs and has little effect on outcomes, attrition, or response rates, but induces slightly different reporting on some measures.

JEL codes: C81, C83, D22, O12, O17

1 Introduction

We run the first randomised controlled trial to compare microenterprise data from surveys of different frequency and medium. We use this trial to understand the effect of interview frequency and medium on microenterprise outcomes and owners' reporting of these outcomes, the relative cost of phone and in-person interviews, and the scope for high-frequency interviews to measure short-term fluctuations and outcome dynamics. We study a representative sample of microenterprises in the city of Soweto in South Africa. We 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. We compare the monthly and weekly in-person groups to test the effects of collecting data at higher frequency, holding the interview medium fixed. The third group is interviewed every week by mobile phone for 12 weeks. We compare the weekly phone and in-person groups to test the effect of data collection medium on responses, holding the interview frequency constant. All interviews 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 to test if the medium or frequency of prior interviews has affected behaviour or outcomes.

We depart from the assumptions that high-frequency data is useful for many purposes and that phone surveys are a more affordable way to collect high frequency data. We provide evidence to support these assumptions but view them as non-controversial. Researchers can use high-frequency data to study volatility and dynamics in enterprise and household outcomes (McKenzie and Woodruff, 2008; Collins et al., 2009), inform economic models of intertemporal optimization (e.g. responses to income or expenditure shocks) (Banerjee et al., 2015; Rosenzweig and Wolpin, 1993), illustrate the time path of treatment effects (Jacobson et al., 1993; Karlan and Valdivia, 2011), explore dynamic treatment regimes (Abbring and Heckman, 2007; Robins, 1997), or average over multiple measures to improve power (Frison and Pocock, 1992; McKenzie, 2012). High-frequency surveys also allow the use of shorter, more accurate recall periods while obtaining comprehensive time-series coverage.[1] Indeed, high-frequency household surveys and phone surveys are increasingly used in developing countries (Beaman et al., 2014; Dabalen et al., 2016; Das et al., 2012; Franklin, 2015; Zwane et al., 2011). Phone surveys are likely to be cheaper and useful for highly mobile populations (Dillon, 2012). They can also be used during conflict or disease outbreaks when in-person surveys are not possible (Bauer et al., 2013; Turay et al., 2015; van der

---

[1] The literature finds shorter recall periods yield more accurate measures of consumption, health, profits, and investment in human and physical capital, but may miss important but infrequent experiences like asset purchases (Beegle et al., 2012; Das et al., 2012; De Mel, 2009; De Nicola and Giné, 2014; Heath et al., 2016). We do not test the effect of different recall periods here.

Windt and Humphreys, 2013).

We focus on testing if high-frequency phone surveys lead to lower data quality that might offset their potential advantages and cost savings. Other evidence suggests high-frequency surveys might affect respondent behaviour. For example, more frequent surveys increase whether respondents chlorinate water (using objective measures of water chlorination), possibly by reminding respondents to treat water (Zwane et al., 2011). Different survey media can also affect aspects of data quality, such as survey responses or rates of attrition. For example, paper- and tablet-based surveys have systematically different responses (Caeyers et al., 2012; Fafchamps et al., 2012; Lane et al., 2006) and shorter recall periods improve reporting accuracy (Beegle et al., 2012; Das et al., 2012; De Nicola and Giné, 2014; Heath et al., 2016). To our knowledge, this paper presents the first experimental comparison of either interview frequency or medium for microenterprises. We adopt adopt both a broad empirical strategy, by estimating mean and distributional effects of interview frequency and medium on a wide range of enterprise characteristics, and a targeted empirical strategy, where we test specific hypotheses based on the existing economic and survey methodology literature, to determine if either more frequent surveying or surveying over the phone affects data quality.

We find that higher-frequency interviews yield useful and equally accurate data but have lower response rates, while phone interviews allow cheaper high-frequency interviews without large compromises in data quality or response rates. Neither high-frequency nor phone interviews change reporting dramatically: high-frequency interviews are somewhat more accurate than monthly interviews on cognitively demanding variables requiring computation, while people are more likely to report socially desirable answers in person and to be less careful over the phone. Permanent attrition from the panel does not differ by interview frequency or medium, although respondents assigned to weekly interviews miss a higher fraction of interviews than those assigned to monthly interviews. In addition, neither high-frequency nor phone interviews change behaviour, which we test for by comparing measures from a common, in-person endline across enterprises previously interviewed with different frequencies and media. Thus both methods are thus viable alternatives to monthly in-person interviews. Unsurprisingly, phone interviews are substantially cheaper than in-person interviews.

We study microenterprises but our results can inform enterprise and household surveys more generally. We measure standard enterprise characteristics such as profits, sales, and line-item costs, which are relevant to microenterprises, larger enterprises, and households engaged in agricultural or home production. We also measure individual characteristics such as hours worked and transfers to household members, which are commonly measured in household surveys.

We test six general hypotheses about frequency and medium effects, which are not necessarily specific to microenterprises. These are motivated by the prior economic and survey methodology literature. In the remainder of the introduction we discuss the prior literature, these six hypotheses, and our specific findings. The first three hypotheses test the effect of being interviewed at weekly, rather than monthly, frequency. First, we test if more frequent measurement changes enterprise outcomes or respondent behaviour. Previous research has shown that more frequent interviews affect behaviour that is not already salient to respondents such as handwashing, but not behaviour that is highly salient, such as enterprise sales (Beaman et al., 2014; Zwane et al., 2011). In our common, in-person endline, we find little evidence that "real" enterprise outcomes differ by interview frequency, both for salient and obvious measures such as the number of employees and for less salient and obvious measures such as household takings from the enterprise.

Second, we test if more frequent measurement changes reporting behaviour, rather than true behaviour or enterprise outcomes. High-frequency interviews can decrease data quality if respondents become fatigued or anchor on past responses, artificially reducing variance. In contrast, they might report more accurate values of cognitively-demanding outcomes as they practice computing variables (Beegle et al., 2012). We find little evidence that interview frequency alters reporting: reported outcomes and the variance of outcomes by respondent are similar on most measures and at most quantiles of the distribution for monthly and weekly in-person interviews. The clearest difference we observe is that monthly respondents are more likely to report inaccurate profit data, which we measure as the deviation between reported profits and reported sales minus costs. We find no differences by interview frequency on many variables that are clearly easier to report, such as number of employees, hours worked, and fixed assets. Enumerators report that respondents give more honest and careful answers when they are interviewed monthly instead of weekly and in person instead of by phone, but these assessments are subjective and it is unclear how much weight to place on them.

Third, we test if interview frequency changes attrition from the panel and the rate of missed interviews. We find mixed evidence. There is no difference in the probability of permanent attrition from the panel between monthly and weekly groups. Weekly respondents do miss interviews at a higher rate but are still more likely to be interviewed in each four-week period than monthly respondents. The optimal trade-off between the greater information flow from high-frequency interviews and the higher nonresponse rate is likely to be application-specific.[2]

The remaining hypotheses test the effect of being interviewed over the phone, rather than

---

[2] The other findings we report here are robust to multiple strategies to adjust for nonresponse. Nonresponse is typically high in enterprise interviews, so we designed the study to be well-powered despite some nonresponse. Missed interviews are predicted by few baseline characteristics and reasons for attrition are mostly balanced across groups.

face-to-face. We test, fourth, if interview medium affects enterprise outcomes and respondent behaviour. We find some differences in responses to the in-person endline between respondents previously interviewed by phone and in person, but these are not robust to accounting for outliers or non-response. We conclude that there is little evidence of persistent medium effects on behaviour.

Fifth, we test if interview medium affects reporting behaviour, rather than true behaviour or enterprise outcomes. We find no differences for most measures at the means and most quantiles of the distribution. This finding is consistent with prior research showing that responses to household interviews and political opinion polls in developed countries are only slightly sensitive to the choice of interview medium (De Leeuw, 1992; Groves, 1990; Körmendi, 2001). In particular, we compare measures of easily stolen items (profit, sales, stock/inventory) to test if respondents trust enumerators less in either medium. We find limited evidence of this, although phone respondents understate stock and inventory. We do find that phone respondents appear to be less careful than in-person respondents: they are less likely to use written records to respond questions and more likely to report large discrepancies between profits and sales minus costs. Responses on questions subject to social desirability bias such as hours worked also differ slightly between phone and in-person interviews.

Sixth, we test if interview medium changes attrition from the panel and the rate of missed interviews. We find no difference by medium. In contrast, previous research from the US finds lower response rates to phone interviews as it is easier to avoid phone interviews and interviewers build less rapport with respondents (Groves, 1990; Groves et al., 2001). In developing countries, Gallup (2012) finds that panel attrition in Honduras and Peru is higher with text message or robocall interviews, which are even easier to avoid than enumerator-administered phone surveys. It is possible that these factors are offset by greater convenience of phone interviews, which can be conducted when the respondent is away from their enterprise (Croke et al., 2014). Most of the US studies are cross-sectional, suggesting that our result might be specific to panel studies where phone interviews follow an in-person baseline interview.[3]

We describe the experimental design and data collection process in section 2. We describe the estimation strategy in section 3. We compare reported outcomes by data collection medium and frequency in section 4. We compare data collection costs in section 5. Online appendices A to J report a variety of background data and robustness checks.

---

[3] Another consideration, whether mobile phone surveys approximate random samples, is less relevant for studies like ours that start with an in-person baseline. Leo et al. (2015) discuss this issue in several developing countries.

2 Design and data

2.1 Context

The study takes place in Soweto, the largest and oldest "township" near Johannesburg, in South Africa. Soweto's population in October 2011 was approximately 1.28 million people. Residents are almost all Black Africans (99%).[4] 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.[5]

2.2 Sample definition and sampling strategy

We use a three-stage clustered sampling scheme to gather a representative sample of the population of households who own eligible enterprises and live in low-income areas of Soweto. We discuss this scheme in detail in Appendix A. In brief, we randomly selected small geographic units ("subplaces") from the 2011 census. Between September 2013 and February 2014, we conducted a screening interview with all households in these randomly selected subplaces. The screening interview identified if any member of the household owned an enterprise that: (i) had at most two full-time employees (in addition to the owner); (ii) did not provide a professional service (e.g. medicine); (iii) operated at least three days each week. The first two conditions are consistent with definitions of "microenterprises" in the development economics literature. The third condition excludes enterprises that are seasonal, occasional (e.g. selling food at soccer games), or operate only over weekends as supplements to wage employment. We impose this condition to ensure week-to-week variation in the outcomes of interest. We also planned to limit the sample to household members with mobile phones to allow phone interviews and payment of respondent incentives but this condition was never binding.[6] The screening process realised a sample of 1081 eligible enterprises. In households which owned multiple eligible enterprises, we randomly selected one for the final sample, leaving a sample of 1046.

---

[4] 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. "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 outskirt of cities.

[5] Authors' own calculations, from the 2011 Census public release data.

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

2.3   Data collection and assignment to interview frequency and medium

Between December 2013 and February 2014, we approached all 1046 eligible enterprise owners identified in the screening stage to conduct a baseline interview of 30 questions. These interviews were conducted in person at the enterprise premises to verify that the enterprises existed, whereas the screening interview 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).[7] We completed the baseline questionnaire with 895 of the 1046 enterprise owners (85%) identified in the screening stage.[8]

We then randomised the 895 baseline enterprises into three data collection groups: monthly in-person interviews (298 enterprises), weekly in-person interviews (299 enterprises), and weekly phone interviews (298 enterprises). Following Bruhn and McKenzie (2009), we first created strata based on gender, number of employees, enterprise sector and enterprise location.[9] This yielded 149 strata with one to 51 enterprises 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 systematic differences between data collection groups. 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. Within groups, fieldworkers were not randomly assigned to enterprises. We assigned fieldworkers so each owner would be interviewed in her or his preferred language (English, seSotho, seTswana, or isiZulu) and to minimise fieldworkers' travel time between enterprises.

We then conducted repeated interviews with each enterprise 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 enterprises were interviewed each week, providing a comparison group for each week when the weekly enterprises were interviewed. In the repeated interview phase, enterprise owners received a ZAR12 (USD0.97) mobile phone airtime voucher for every fourth interview

---

[7] 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.

[8] 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 interview.

[9] We used the census subplace in which the enterprise 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 left some residual enterprises in strata whose size was not divisible by 3. We randomly assigned residual enterprises to data collection groups with a restriction that a pair of residual enterprises from the same stratum would always go into separate groups.

they completed. This equates the per-interview payout across data collection groups.[11]

Finally, we conducted an endline interview in person with each enterprise owner at the enterprise location, starting one to two weeks after enterprises had finished the repeated interviews. This common endline format, irrespective of the assigned data collection medium for the repeated interviews, means that observed endline differences across randomly assigned data collection groups must reflect persistent effects of the data collection frequency or medium on enterprise behaviour or outcomes, rather than measurement effects.

2.4   Baseline data description

Table 1 shows baseline summary statistics for the final sample of 895 enterprises. We draw two conclusions from this table. First, the random assignment succeeded in balancing baseline characteristics across the three groups. We fail to reject joint equality of all three group means across all 40 characteristics. Only 2 of 40 reported baseline variables are significantly different at the 5% level. 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 similar to samples of microenterprises in urban areas of other middle-income countries. We compared the sample to 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 enterprises are slightly older and slightly more concentrated in the food and retail/trade sectors but are otherwise similar to most of the other samples.

The enterprise owners' households have a mean monthly income of ZAR4050 (approximately US$380 at the time of the survey) across all sources.[12] This falls in the fourth decile for all households across South Africa, a country with extremely unequal income distribution. Enterprise owners' households have an average of 3.8 other members, though this is widely dispersed with an interdecile range of 1 to 7. In 55% of households, the enterprise accounts for half or more

---

[11] This design choice prices respondent time equally across the data collection groups, reducing one possible cause for differential response rates. It does allow respondents in the weekly groups to earn more for completing interviews but the difference in the total potential payout is approximately 0.2% of mean annual household income. We believe this is too small to induce income effects. Recent phone surveys in developing countries find only small differences in attrition when incentives are randomly varied, suggesting small differences in incentive structure would have had limited impact (Croke et al., 2014; Gallup, 2012; Leo et al., 2015; Romeo, 2013).

[12] This is the average across the 87% of enterprise owners who 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 (ZAR)	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 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.

of household income. The enterprises in this sample are relatively well-established (average age seven 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: 61% have no employees other than the owner and 28% have only one other employee. 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.

2.5   Outcomes of interest

Our repeated and endline interviews 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 by other household members.[13] The interview also asked respondents if they used written records to help complete the interview and several tracking questions. At the end of the interview, the enumerator assessed whether the respondent answered questions honestly and carefully, recorded on a five-point Likert scale. We show summary statistics for all repeated and endline interview measures in Appendix B.

We elicit profits directly, following De Mel (2009), using the question "What was the total income the business earned last week, after paying all expenses (including wages of any employees), but not including any money that you paid yourself? That is, what were the profits of your business for last week?" This measure is more computationally intensive for the respondent. We compare this to sales minus costs as a measure of consistency in reporting.

Costs are calculated from nine cost subcategories for the previous week: purchase of stock or inventory, wages or salaries, rent and rates for the property where the enterprise is based, repayments on enterprise loans, equipment purchases, fixing and maintaining equipment, transport costs for the enterprise, telephone and internet costs for the enterprise, and all other enterprise expenses.

All flow measures used a one-week recall period except hours of operation (last day) and sales (both last week and last 4 weeks). The two sales measures allow us to test if the effects of medium or frequency differ by recall period. 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

---

[13] Each interview 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.

the exact value.[14]

2.6 Response rates and attrition

We impose a maximum number of attempts because the high-frequency panel did not allow enu-
merators to continue trying to contact respondents for long periods of time, as is feasible in low-
frequency panels (Thomas et al., 2012). Fieldworkers in all data collection groups were trained to
make three attempts to complete each repeated and endline interview, where an "attempt" was a
visit to the enterprise premises or an answered phone call.[15] If the interview was not completed
by the third attempt, that week/month was marked as missing and the respondent was contacted
in the next scheduled week or month. If the respondent was contacted but refused to complete the
interview, that week/month was marked as missing and the respondent was contacted in the next
scheduled week or month. If the respondent was contacted but refused to complete any further
interviews, that week/month and all subsequent weeks/months were marked as missing.

We choose to equate the number of interview attempts across data collection groups rather
than trying to the equate the response rates or the total cost of attempts (as the per-attempt cost is
lower for phone interviews). Equating the response rates would have obscured useful information
about the effects of interview medium and frequency on response rates and permanent attrition.
Equating the costs would only have been possible if we correctly predicted the per-attempt cost
and number of attempts required to successfully complete an interview, in each data collection
group.

In the remainder of the paper, we distinguish between *nonresponse*, where a respondent
misses an interview at time t and completes an interview at time $s > t$, and *permanent attrition*,
where a respondent misses all interviews from time t onward. Permanent attrition could occur
because the respondent refused any further interviews, changed contact information, or moved
away from the city.[16]

We completed 4070 of 8058 scheduled repeated interviews (51%) and 591 of 895 sched-
uled endline interviews (66%). We discuss the patterns of nonresponse and attrition in detail in
subsection 4.4 and show in subsections 4.1 - 4.2 and appendices C - E that our other findings are
robust to accounting for nonresponse. As a brief preview, the response rate is slightly higher for

---

[14] Friedman et al. (2016) discuss how two-step questions can lead to measurement error at both extensive and intensive margins.

[15] 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 (Grosh and Munoz, 1996; McKenzie, 2015b).

[16] We did not track respondents who moved away from the greater Johannesburg region, as we could not complete in-person interviews with these respondents. We did continue to interview respondents who closed or sold their enterprises using a different questionnaire.

enterprises interviewed monthly but is not different across phone and in-person interviews. The rate of permanent attrition does not differ by interview medium or frequency.

3   Estimation

This section discusses our empirical methods, which largely follow our pre-analysis plan.[17] Our experiment and empirical strategy are designed to answer these questions:

1. Do interview frequency or medium affect enterprise outcomes or owner behaviour?

2. Do interview frequency or medium affect reporting?

3. Do interview frequency or medium affect response rates and attrition?

The endline is conducted in person for all groups, so there are unlikely to be differences in reporting between the three treatment groups. So we interpret differences in endline measures as real changes, which reflect persistent effects of the data collection frequency or medium on enterprise owner behaviour, rather than measurement effects. In contrast, differences in repeated interview measures may be real changes or due to differences in reporting. Given the literature discussed in Section 1, we expect that enterprise outcomes and owner behaviour are more likely to be affected by interview frequency than medium. Interview responses, response rates, and attrition may be affected by either interview frequency or interview medium.

3.1   Do interview frequency or medium affect enterprise outcomes or owner behaviour?

Prior research has shown that participation in panel interviews can change respondents' behaviour, even over relatively short panels. We first test if endline measures differ between respondents who were previously interviewed at different frequencies or using different media. All endline interviews are conducted in person by an enumerator whom the respondent has not met before. Differences will mainly reflect real behaviour change which is caused by the prior interview frequency or medium that the respondent experienced during the repeated interviews. Behaviour change has been documented mainly in domains that are not already salient to respondents, including small change management for enterprise owners, rather than domains that are already salient, including enterprise profit and sales (Beaman et al., 2014; Franklin, 2015; Stango and Zinman, 2013; Zwane et al., 2011). We test for interview frequency effects on all endline measures, rather than classifying some measures as more or less salient ex ante. We also test for interview medium effects on

---

[17] Our pre-analysis plan is available at https://www.socialscienceregistry.org/trials/346.

all endline measures, although we are not aware of prior research into behaviour change induced by phone versus in-person interviews.

First, we explore mean differences by interview frequency and medium by estimating

$$Y_{ki} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \varepsilon_{ki}, \tag{1}$$

where i and k index respectively respondents and outcome measures, T_{1i} and T_{2i} are indicators for the monthly in-person group and the weekly phone group, respectively, and η_g is a stratification block fixed effect. If Y_{ki} is a continuous variable, we normalise it using the mean and standard deviation of the monthly in-person interview group. Categorical measures (# employees) and binary measures (enterprise closure, written records, enumerator assessments) are not normalised. The categorical variables seldom have values greater than one, so we discuss differences in these categorical variables in percentage point terms. We use heteroscedasticity robust standard errors and test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, and $H_0^3 : \beta_1 = \beta_2 = 0$.

We assess the risk of null results from low powered tests by estimating minimum detectable effect sizes (MDEs) for β_1 and β_2 for each outcome. We estimate the MDE for β_1 for outcome k using the sample of enterprises from the two in-person groups who completed the endline and the formula $MDE_{1k} = (\tau_{1-\kappa} + \tau_{\alpha/2}) \cdot \sqrt{\sigma_{Y_k}^2 / (\sigma_T^2 \cdot N)}$. Here $\sigma_{Y_k}^2$ is the variance of Y_k conditional on the stratification fixed effects, σ_T^2 is the variance of T_{1i}, N is the number of enterprises in the endline, and we set $\tau_{1-\kappa} + \tau_{\alpha/2} = 2.8$ for a test with 5% size and 80% power. We estimate the MDE for β_2 for each outcome k using an analogous formula. This approach simply updates ex ante power calculations using the realised values of $\sigma_{Y_k}^2$, $\sigma_{T_k}^2$, and N_k.[18]

Second, we examine the empirical CDFs of endline responses by interview frequency and medium. This provides a more general description of the differences in responses and can capture mean-preserving spreads that would not be visible in the mean regressions. The regression model in (1) is more restrictive than the quantile regressions but allows inclusion of stratification block fixed effects, reducing the residual variation. For each outcome k, we estimate

$$Q_\theta(Y_{ki} \mid T_{1i}, T_{2i}) = \beta_0^\theta + \beta_1^\theta \cdot T_{1i} + \beta_2^\theta \cdot T_{2i} \tag{2}$$

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

---

[18] We prefer this approach to calculating power for the observed coefficient estimates. 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 coefficient estimates 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 effects that change $\sigma_{Y_k}^2$.

hypothesis $\beta_1^\theta = 0$ (*i.e.* weekly and monthly in-person measures are equal) at the relevant quantile and we use $+$ to indicate rejection of the null hypothesis $\beta_2^\theta = 0$ (*i.e.*, weekly in-person and phone measures are equal). We report significance tests for each level of θ, using the False Discovery Rate to control for multiple testing across quantiles (Benjamini et al., 2006).[19]

3.2   Do interview frequency or medium affect reporting?

After testing if interview frequency or medium affects behaviour, we estimate 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 analyse these data in four ways. First, we estimate the effect of interview frequency and medium on mean response values

$$Y_{kit} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \phi_t + \varepsilon_{kit}, \tag{3}$$

where i, k and t index respectively enterprises, outcomes, and weeks; T_{1i} and T_{2i} are indicators for the monthly in-person group and the weekly phone group, respectively; η_g is a stratification block fixed effect; and ϕ_t is a week fixed effect.[20] We cluster standard errors by enterprise and test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, and $H_0^3 : \beta_1 = \beta_2 = 0$. We again report MDEs for β_1 and β_2 for each outcome k.[21]

Second, we test for difference in the dispersion of non-binary outcomes as different interview frequency and media may affect outcome dispersion. We calculate the standard deviation through the repeated interview phase for each enterprise i for each outcome k, denoted S_{ki}, and estimate

$$S_{ki} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \mu_{ki} \tag{4}$$

[19] This differs from our pre-analysis plan in two ways. We planned to use stratum and week indicator variables but this proved computationally infeasible. We planned to use simultaneous-quantile regression and to jointly test for coefficient equality across all quantiles. However, an estimator for systems of simultaneous quantile regression models with clustered standard errors was not developed at the time of preparing the paper.

[20] We use indicator variables for actual calendar weeks to capture common shocks. We stagger start dates of the repeated interviews so some enterprises are interviewed in weeks 1-12, some in weeks 2-13, some in weeks 3-14, and some in weeks 4-15. So there are 15 fixed effects but at most 3 and 12 indicators equal one for each enterprise assigned to respectively monthly and weekly interviews. The start dates are cross-randomised with the group assignments so the two sets of indicators are independent of each other.

[21] Here, to calculate the MDE for β_1, for example, we use the sample of enterprises assigned to in-person interviews and the formula $MDE_{1k} = (\tau_{1-\kappa} + \tau_{\alpha/2}) \cdot \sqrt{\frac{\sigma_{Y_k}^2}{\sigma_T^2} \cdot \frac{1}{N} \cdot \left(\rho_{Y_k} + \frac{1-\rho_{Y_k}}{N_W}\right)}$. ρ_{Y_k} is the intra-enterprise correlation in outcome Y_k conditional on the fixed effects, N is the total number of enterprises assigned to the in-person groups, N_W is the mean number of completed interviews per enterprise, and all other parameters are defined in section 3.1. We estimate the MDE for β_2 for each outcome k using an analogous formula.

using heteroskedasticity-robust standard errors. We then test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, and $H_0^3 : \beta_1 = \beta_2 = 0$. Third, we estimate quantile regression models as in (2), pooling observations from the repeated interviews and clustering standard errors by enterprise (Silva and Parente, 2013). This provides a richer description than the mean or standard deviation regressions but uses the panel structure less efficiently.

Fourth, we consider the implications of interview medium for estimating the dynamics of enterprise performance. Up to this point, we have focused on whether different interview methods have different implications for testing the level of different variables. However, panel data also allows researchers to model the dynamics of enterprise behaviour. This is potentially important for many reasons, such as understanding the trajectory of enterprise responses to shocks. We therefore explore the time series structure of enterprise outcomes by interview medium. To do this, we report the time series structure of a flow variable – log profit – and a stock variable – log capital – following Blundell and Bond (1998).

3.3 Robustness checks and subgroup analysis

In this subsection we describe robustness checks that account for outliers, missing data, and multiple testing. On outliers, our primary analysis of means, standard deviations, and outliers uses raw interview responses without adjusting outliers. We assess whether our findings are sensitive to outliers by rerunning all mean regressions with outcomes winsorized at the 5th and 95th percentiles. We report these results in appendices C - E. All CDFs shown in section 4 use raw measures. [22]

We then examine robustness to corrections for missing data. We observe some differences in the response rate – the fraction of scheduled interviews actually completed – by data collection method. We thus report if our results are robust to two methods of accounting for differential response rates. 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 assumptions generate respectively upper and lower bounds on the effects of interview frequency and medium (Lee, 2009). We use the width of the bounded set as a measure of the vulnerability of our findings to differential response rates by interview medium and frequency.

Second, we estimate the probability of completing each repeated or endline interview as a function of the baseline characteristics discussed in Section 2.4: $\hat{P} = \hat{P}r$ (Interview Completed$_{it}$|X_{i0}).

---

[22] In our pre-analysis plan we proposed trimming outcomes. We instead report results with winsorized outcomes to reduce the loss of information from real outliers, as opposed to measurement error. The trimmed and winsorized results are very similar.

15

We then construct inverse probability of interview completion weights, $\hat{W} = 1/\hat{P}$, and estimate equations (1), (3), and (4) using these weights. If nonresponse does not covary with latent outcomes after conditioning on these baseline characteristics, the weighted regressions fully correct for nonresponse.

Finally, we examine robustness to corrections for multiple testing. We also estimate sharpened q-values that control the false discovery rate across outcomes within variable families specified in our pre-analysis plan (Benjamini et al., 2006). Rather than pre-specifying a single q, we report the minimum q-value at which each hypothesis is rejected. We implement these adjustments separately for the repeated interview phase and the endline interviews.

We test for heterogeneous effects by estimating equations (1) and (3) with interactions between the group indicators and six pre-specified baseline measures: respondent education, score on a digit span recall test, score on a numeracy test, use of written records at baseline, number of employees at baseline, and gender. For education, digit span recall, numeracy and the number of employees, we interact the group indicators with indicator variables equal to one for values above the baseline median.

4 Results

4.1 Do interview frequency or medium affect enterprise outcomes or owner behaviour?

We first test if being interviewed at higher frequency changes behaviour or real outcomes by examining measures from the common in-person endline. We report all results in Table 2 and summarise the results in Figure 1. We see few differences. In particular, we see no evidence that the accuracy of enterprise outcomes is higher for the high-frequency group: the absolute difference between profits and sales minus costs is (insignificantly) smaller for the monthly in-person group. The only statistically significant difference is a slightly higher level of fixed assets for monthly in-person respondents. These respondents also report lower stock/inventory, profits, sales, and money taken from the enterprise, but none of these differences is statistically significant and all shrink close to zero with winsorization (Appendix E). We are powered to detect moderate differences in endline measures by previous interview frequency: the median MDEs for binary and continuous outcomes are respectively 11 percentage points and 0.24 standard deviations. The MDEs for continuous outcomes are sensitive to a small number of outliers and shrink substantially with winsorization.

These results hold up to various robustness checks. There is a moderate rate of nonresponse

in the endline but accounting for nonresponse does not change the general pattern. Most differences shrink when we reweight the data to account for differential nonresponse (only the difference in the number of employees becomes larger and statistically significant (Appendix E)). The Lee bounds are relatively narrow: the median intervals for binary and continuous outcomes are respectively 5 percentage points and 0.15 standard deviations wide. The absolute values of the bounds are small for all variables other than stock/inventory, profits, sales, and money taken from the enterprise; even these bounds shrink near to zero when we winsorize outliers. There is no evidence of mean-preserving shifts in the distribution of endline responses (Appendix H).

We also test if being interviewed by phone or in person changes behaviour or real outcomes, although the previous literature is silent on this question. We find few differences. Phone respondents are robustly more likely to use written records to answer questions. Phone respondents report lower stock/inventory, profits, full-time employees, and money kept for themselves. However, most of these differences disappear when we winsorize the upper and lower ventiles. Furthermore, most differences shrink substantially, when we reweight the data to account for nonresponse. Endline nonresponse is higher for phone respondents, so the Lee bounds on these medium effects are slightly wider than for the frequency effects: the median widths of the intervals across the binary and continuous outcomes are respectively 20 percentage points and 0.40 standard deviations.

We interpret these results as evidence against hypotheses 1 and 4 from section 1, that interview frequency or medium induce changes in real behaviour or real outcomes. However, the endline measures can also reflect differences in reporting behaviour from the different interview frequencies or media that persist to the common endline. In the latter case, the coefficients on the monthly in-person indicator variable (and the weekly phone indicator variable) should be similar for the endline and repeated interview data. We plot these coefficients in Figure 1. Inspection reveals little relationship between the endline and repeated interview differences and only 20 of 32 coefficients have the same signs in the endline and repeated interviews. More formally, we test if each coefficient in the endline interviews is significantly different to the corresponding coefficient in the repeated interviews and report the test results in Appendix F. We reject equality for only 12 of 32 differences but this reflects the relative imprecision of many endline differences. The clearest pattern is that enumerators view respondents as giving more careful and honest answers during lower-frequency and in-person repeated interviews, but this difference pattern largely disappears in the endline. We interpret this as evidence that interview frequency and medium do not exert large persistent effects on the nature of respondents' interaction with enumerators.
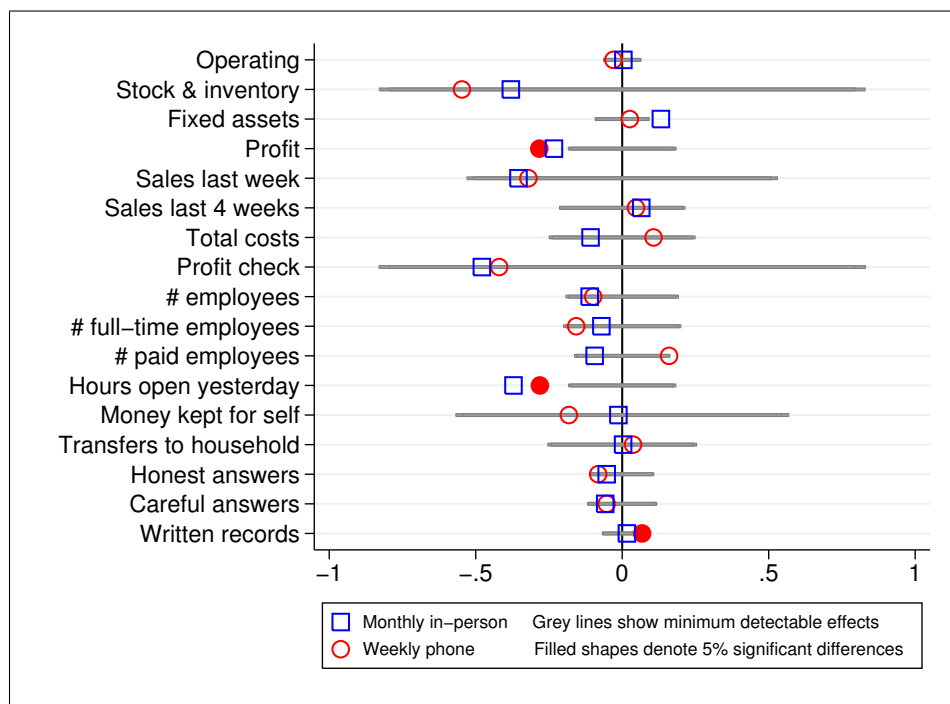
Table 2: **Mean Differences for Endline Interviews (equation 1)**

	(1) Operating	(2) Stock & inventory	(3) Fixed assets	(4) Profit	(5) Sales last week	(6) Sales last 4 weeks	(7) Total costs	(8) Profit check
Monthly 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	**-0.030** (0.028)	**-0.547** (0.314)*	**0.026** (0.072)	**-0.283** (0.125)**	**-0.320** (0.201)	**0.047** (0.122)	**0.107** (0.152)	**-0.420** (0.279)
Observations	594	546	546	546	546	546	546	546
All groups equal (*p*)	0.437	0.217	0.241	0.033**	0.276	0.773	0.293	0.260
MDE: Monthly in-person	0.062	0.794	0.088	0.176	0.508	0.205	0.238	0.794
MDE: Weekly by phone	0.064	0.829	0.092	0.184	0.530	0.214	0.249	0.830
Lee bound: Monthly in-person (lower)	0.002	-0.531	0.023	0.013	-0.464	0.075	-0.157	-0.614
Lee bound: Monthly in-person (upper)	0.004	0.022	0.123	0.053	0.052	0.082	0.042	0.019
Lee bound: Weekly by phone (lower)	-0.100	-0.781	-0.050	-0.256	-0.473	-0.042	0.004	-0.593
Lee bound: Weekly by phone (upper)	-0.017	0.214	0.219	0.225	0.286	0.405	0.555	0.253

	(1) Employees	(2) Full-time	(3) Paid	(4) Money kept	(5) Household takings	(6) Hours yesterday	(7) Honest	(8) Careful	(9) Written records
Monthly 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	**-0.099** (0.105)	**-0.157** (0.090)*	**0.160** (0.285)	**-0.281** (0.143)**	**-0.182** (0.175)	**0.037** (0.127)	**-0.082** (0.046)*	**-0.053** (0.049)	**0.068** (0.029)**
Observations	546	546	546	546	546	546	593	593	546
All groups equal (*p*)	0.402	0.219	0.340	0.113	0.203	0.955	0.191	0.397	0.049**
MDE: Monthly in-person	0.183	0.192	0.156	0.175	0.544	0.243	0.104	0.114	0.065
MDE: Weekly by phone	0.192	0.200	0.163	0.183	0.568	0.254	0.108	0.118	0.068
Lee bound: Monthly in-person (lower)	-0.144	-0.084	-0.067	-0.443	-0.081	-0.065	-0.100	-0.094	0.002
Lee bound: Monthly in-person (upper)	-0.044	0.009	-0.048	-0.013	-0.010	-0.009	-0.042	-0.040	0.032
Lee bound: Weekly by phone (lower)	-0.210	-0.260	0.107	-0.397	-0.278	-0.150	-0.203	-0.174	-0.063
Lee bound: Weekly by phone (upper)	0.282	0.282	0.465	0.194	0.037	0.276	-0.015	0.020	0.100

Notes: Coefficients are from regressions of each outcome on a vector of data collection group indicators and randomization stratum fixed effects. Continuous outcomes are standardised 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.

18

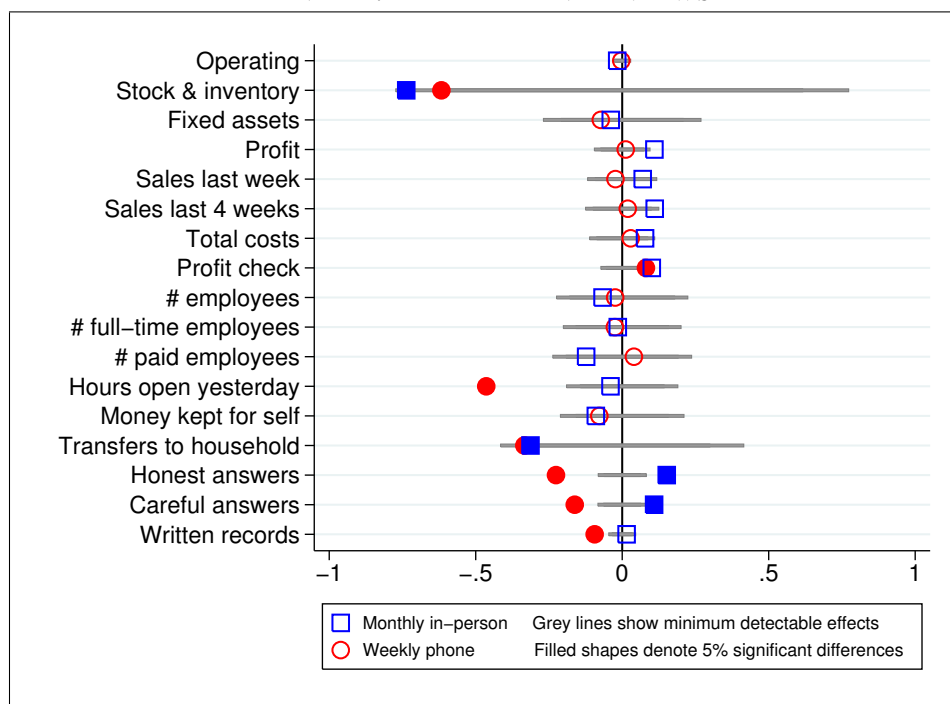Figure 1: **Mean Differences for Repeated and Endline Interviews**

Notes: Coefficients are from regressions of each outcome on a vector of data collection group indicators, randomization stratum fixed effects, and survey week fixed effects (repeated interviews only). Continuous outcomes are standardised 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. Significance tests are based on heteroskedasticity-robust standard errors, clustering by enterprise for the repeated interviews. Panel A shows the minimum detectable differences (MDEs) between weekly in-person and phone interviews; the MDEs between weekly and monthly in-person interviews are approximately 5% smaller. Panel B shows the MDEs between weekly and monthly in-person interviews; the MDEs between weekly in-person and phone interviews are approximately 25% smaller.

4.2 Do interview frequency or medium affect reporting?

We now test if interview medium or frequency affects responses during repeated interviews. Given the limited evidence of changes in behaviour or real outcomes we saw in section 4.1, we interpret differences during the repeated interviews as measurement effects. We begin by exploring differences by interview frequency and then turn to differences by medium. We report comparisons of means in Table 3, comparisons of standard deviations in Table 4, and comparisons of empirical distributions in Figures 2 and 3.

Mean responses are similar for monthly and weekly in-person respondents on most measures: probability of enterprise closure, assets, sales, costs, employees, hours worked, and money taken from the enterprise. The within-respondent standard deviations of these measures through time are also similar across the monthly and weekly groups, as are the empirical distributions of these measures, except assets. These comparisons are fairly well-powered: the medians of the minimum detectable differences across the binary and continuous outcomes are respectively 8 percentage points and 0.16 standard deviations.

The mean differences remain small and insignificant when we reweight the data to account for nonresponse or winsorize the top and bottom ventiles (Appendix C). The Lee bounds for binary measures are quite wide (median width of 20 percentage points) because the nonrespondents are not balanced on number of employees. Nonetheless, the median bounds for binary and continuous measures allow us to rule out differences of respectively 23 percentage points and 0.29 standard deviations.

We do observe three potentially informative patterns when we compare weekly and monthly respondents. First, monthly respondents report holding lower stock/inventory and giving fewer goods/services from the enterprise to household members (Table 3) and the within-respondent standard deviations of these measures are smaller for monthly groups (Table 4). These differences are explained by longer right tails for both measures (Figures 2 and 3) and are robust to winsorizing, reweighting, and adjusting critical values to account for multiple testing (Appendices C and G). These patterns may arise if some enterprises make large, irregular stock purchases and disbursements to household members, which are more likely to be missed by the monthly interviews but are picked up in the weekly interviews. This highlights the value of high-frequency measurement, particularly as stock management practices are robustly associated with measures of enterprise success (McKenzie and Woodruff, 2017).

Second, our main measure of reporting inaccuracy, the absolute value of sales minus costs and profits, is higher for monthly than weekly respondents (Table 3). This is explained by large differences in the right tail of the distribution (Figure 2). We interpret this as evidence that high-

20

frequency interviews can increase reporting accuracy, consistent with hypothesis 2 discussed in section 1.[23] This does not persist into the endline, which typically takes place more than a week after the final repeated interview.

Given this result, we go on to test if differences by interview frequency are compounded by recall bias. For example, gains in accuracy from being interviewed more frequently may be larger for measures with a longer recall period, where responses tend to be less accurate. We thus ask respondents to report sales in both the last week and last four weeks. However, the interview frequency (and medium) effects are small and almost identical for the one-week and four-week recall periods. At least for sales, interview frequency and medium do not seem to differentially affect measures with different recall periods.[24]

Third, enumerators believe that monthly respondents provide more careful and more honest answers (Table 3). This pattern is robust to reweighting and multiple test corrections (Appendices C and G) but these assessments are subjective, so the pattern should be interpreted with caution. It may indicate that high-frequency interviews lead to more respondent frustration and/or exhaustion. We return to this issue in section 4.4, when we discuss nonresponse and attrition.

We then turn to examining differences by interview medium. Again, reporting is similar for high frequency phone-based and in-person interviews on most measures and at most quantiles of the distributions. In particular, we see little evidence that respondents interviewed by phone trust enumerators less than respondents interviewed in person. If this were true, respondents might give report lower values for easily stolen items (assets, stock/inventory, profits, and sales) or measures that can be verified in person but not over the phone (assets, stock/inventory, and employees). Respondents interviewed by phone report lower stock/inventory (Table 3) due to fewer very high values (Figure 2), but differences in the other measures are small and insignificant with mixed signs. These patterns are robust to winsorizing and reweighting (Appendices C and D).

However, we do see two important differences in reporting by interview medium, consistent with hypothesis 5 from section 1. Both differences are robust to winsorizing, reweighting, and multiple test corrections (Appendices C and G). First, phone respondents appear to be less careful than in-person respondents. They are less likely to use written records to respond to questions, although this may be partly because the in-person interviews always take place at the enterprise

---

[23] This pattern could also arise if respondents in the weekly group are more likely to report identical values in each interview, either to save time or because they remember their previous answers. However, we reject these explanations because we see no evidence that monthly respondents have more dispersed profit, sales, or cost measures (Table 4).

[24] The two measures yield fairly consistent information: they are strongly correlated within enterprise-week ($\rho = 0.79$) and the median ratio of one-week sales to four-week sales is 0.36. However, all respondents answer both questions in the same order. So respondent-specific measurement errors may be correlated across the two measures.

Table 3: Mean Differences for Repeated Interviews (equation 3)

	(1) Operating	(2) Stock & inventory	(3) Fixed assets	(4) Profit	(5) Sales last week	(6) Sales last 4 weeks	(7) Total costs	(8) Profit check
Monthly in-person	**-0.017** (0.010)	**-0.737** (0.223)***	**-0.039** (0.070)	**0.110** (0.061)*	**0.070** (0.068)	**0.111** (0.067)*	**0.078** (0.066)	**0.101** (0.056)*
Weekly by phone	**-0.003** (0.006)	**-0.617** (0.223)***	**-0.073** (0.086)	**0.012** (0.059)	**-0.023** (0.054)	**0.019** (0.058)	**0.029** (0.048)	**0.081** (0.037)**
Observations	4070	3989	3987	3986	3985	3987	3987	3984
All groups equal (p)	0.262	0.004***	0.695	0.191	0.429	0.251	0.475	0.027**
MDE: Monthly in-person	0.029	0.773	0.269	0.095	0.118	0.125	0.111	0.073
MDE: Weekly by phone	0.023	0.616	0.208	0.073	0.092	0.099	0.087	0.054
Lee bound: Monthly in-person (lower)	-0.039	-0.903	-0.139	0.057	0.012	0.056	0.043	0.072
Lee bound: Monthly in-person (upper)	-0.014	0.303	0.248	0.288	0.267	0.284	0.295	0.288
Lee bound: Weekly by phone (lower)	-0.002	-0.766	-0.151	-0.008	-0.059	-0.001	0.013	0.029
Lee bound: Weekly by phone (upper)	-0.001	-0.000	0.013	0.009	0.031	0.048	0.023	0.077

	(1) Employees	(2) Full-time	(3) Paid	(4) Hours yesterday	(5) Money kept	(6) Household takings	(7) Honest	(8) Careful	(9) Written records
Monthly 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	**-0.024** (0.072)	**-0.025** (0.079)	**0.040** (0.080)	**-0.464** (0.061)***	**-0.078** (0.056)	**-0.334** (0.128)***	**-0.226** (0.031)***	**-0.162** (0.030)***	**-0.094** (0.022)***
Observations	3987	3984	3973	3987	3986	3986	4056	4056	3987
All groups equal (p)	0.681	0.950	0.096*	0.000***	0.280	0.015**	0.000***	0.000***	0.000***
MDE: Monthly in-person	0.224	0.201	0.237	0.190	0.211	0.415	0.082	0.083	0.046
MDE: Weekly by phone	0.179	0.158	0.191	0.143	0.157	0.299	0.062	0.063	0.034
Lee bound: Monthly in-person (lower)	-0.152	-0.088	-0.193	-0.358	-0.178	-0.428	0.033	0.008	-0.055
Lee bound: Monthly in-person (upper)	0.318	0.367	0.261	0.231	0.251	0.261	0.225	0.203	0.028
Lee bound: Weekly by phone (lower)	-0.037	-0.055	0.019	-0.648	-0.120	-0.411	-0.307	-0.225	-0.099
Lee bound: Weekly by phone (upper)	0.123	0.065	0.162	-0.385	0.012	-0.034	-0.218	-0.140	-0.069

Notes: Coefficients are from regressions of each outcome on a vector of data collection group indicators, randomization stratum fixed effects, and survey week fixed effects. Continuous outcomes are standardised 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, clustering by enterprise. ***, **, and * denote significance at the 1, 5, and 10% levels.

Table 4: **Standard Deviation Differences for Repeated Interviews (equation 4)**

	(1) Stock & inventory	(2) Fixed assets	(3) Profit	(4) Sales last week	(5) Sales last 4 weeks	(6) Total costs	(7) Profit check
Monthly in-person	**-0.486** (0.129)***	**-0.076** (0.047)	**0.004** (0.050)	**0.042** (0.053)	**0.045** (0.052)	**-0.039** (0.040)	**-0.008** (0.051)
Weekly by phone	**-0.334** (0.134)**	**0.001** (0.066)	**0.065** (0.062)	**-0.016** (0.032)	**0.020** (0.030)	**0.084** (0.041)**	**0.143** (0.047)***
Observations	610	609	608	608	608	608	608
All groups equal (p)	0.001***	0.167	0.576	0.520	0.648	0.039**	0.009***
MDE: Monthly in-person MDE: Weekly by phone	0.372 0.338	0.095 0.086	0.077 0.069	0.080 0.072	0.055 0.049	0.075 0.068	0.070 0.063

|	(1) Employees	(2) Full-time	(3) Paid	(4) Hours yesterday	(5) Money kept	(6) Household takings
Monthly in-person	**0.031** (0.048)	**0.006** (0.055)	**0.015** (0.043)	**0.000** (0.064)	**-0.053** (0.093)	**-0.494** (0.171)***
Weekly by phone	**0.126** (0.040)***	**0.020** (0.045)	**0.067** (0.038)*	**0.087** (0.049)*	**-0.080** (0.074)	**-0.135** (0.190)
Observations	608	608	608	608	608	608
All groups equal (p)	0.006***	0.907	0.208	0.087*	0.560	0.010***
MDE: Monthly in-person MDE: Weekly by phone	0.099 0.090	0.112 0.101	0.086 0.078	0.165 0.149	0.220 0.199	0.462 0.418

Notes: Coefficients are from regressions of the standard deviation of each outcome on a vector of data collection group 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.

Figure 2: **Empirical CDFs in Repeated Interviews**
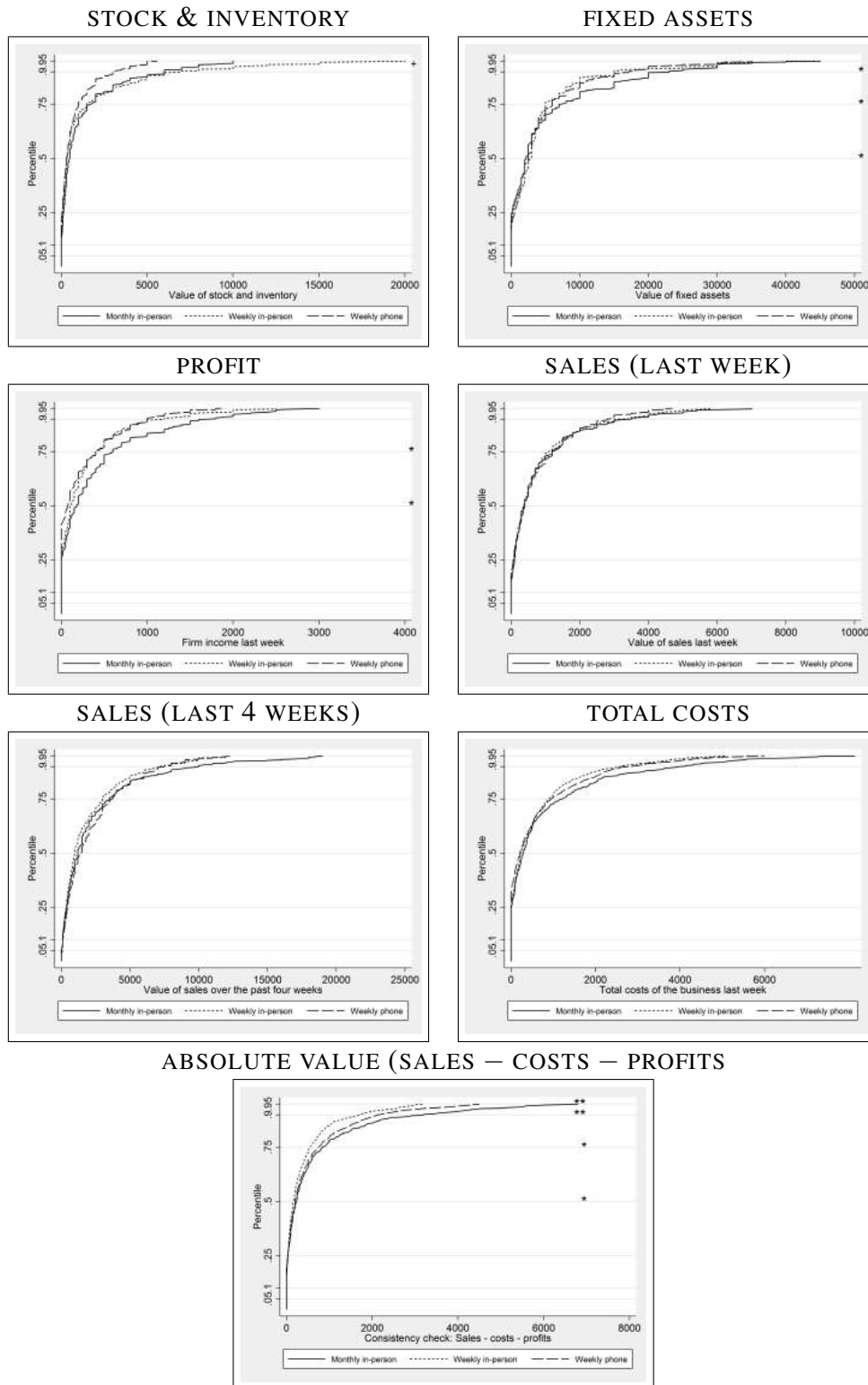
Notes: We test for differences at each of the levels shown on the y-axis. We cluster by enterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. $^+$ indicates rejection of the null hypothesis that the coefficients for weekly in-person and phone interviews are equal. * indicates rejection of the null hypothesis that the coefficients for weekly and monthly in-person interviews are equal. $^{+++}$/***, $^{++}$/**, and $^{+}$/* denote significance at the 1, 5, and 10% levels.

Figure 3: **Empirical CDFs in Repeated Interviews**
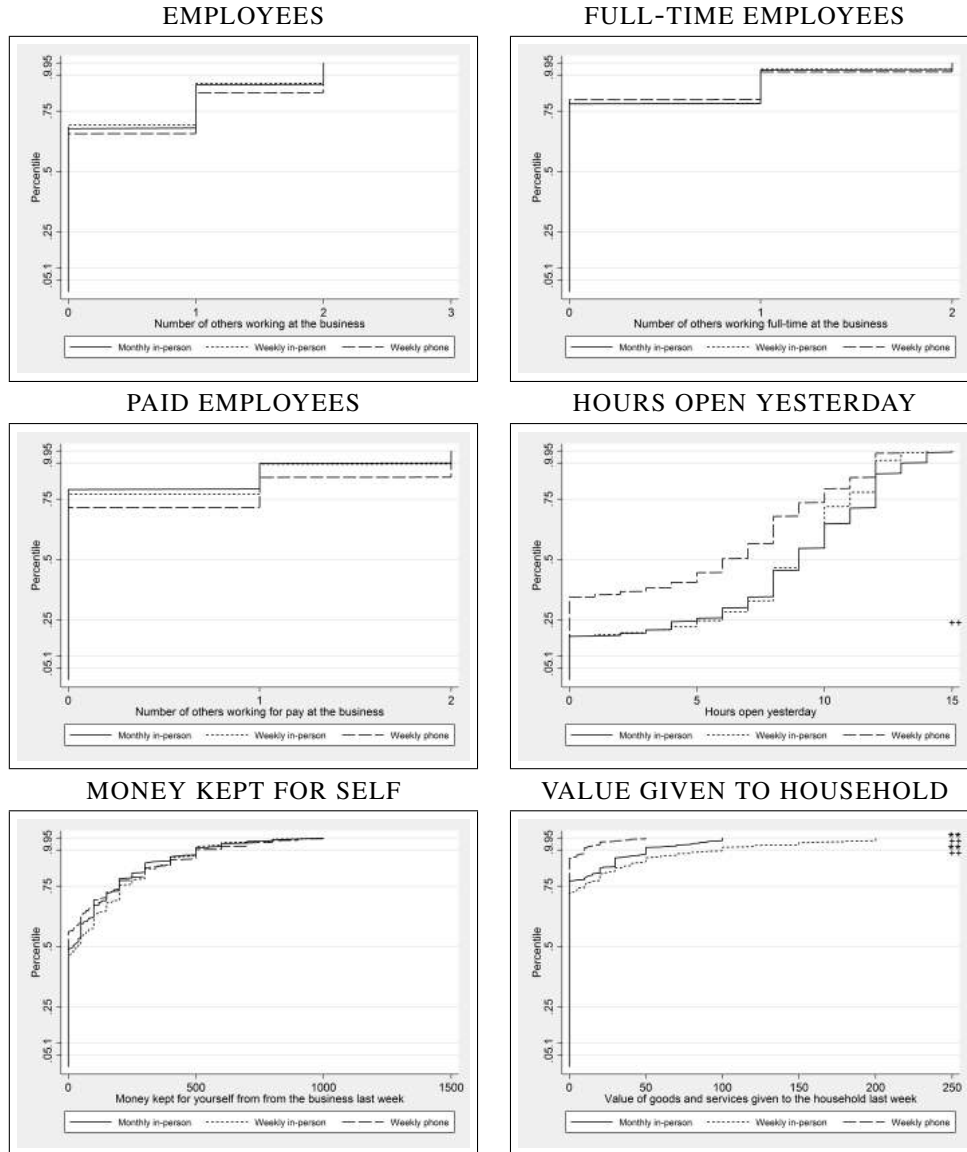
Notes: We test for differences at each of the levels shown on the y-axis. We cluster by enterprise (Silva and Parente, 2013) and use the False Discovery Rate (Benjamini et al., 2006) to control for multiple testing across quantiles. $^{+}$ indicates rejection of the null hypothesis that the coefficients for weekly in-person and phone interviews are equal. * indicates rejection of the null hypothesis that the coefficients for weekly and monthly in-person interviews are equal. $^{+++}$/***, $^{++}$/**, and $^{+}$/* denote significance at the 1, 5, and 10% levels.

while the phone interviews need not (Table 3). They are more likely to have large differences between profit and sales minus costs and this measure has higher within-respondent variance through time (Table 4). Their reported number of employees also has higher within-respondent variance through time, even though the true number of employees is unlikely to vary substantially through time. Enumerators believe phone respondents give less careful and less honest answers. This is consistent with prior evidence from developed countries that respondents in phone surveys are less likely to differentiate between response options (Krosnick, 1991).

Second, respondents are more likely to give socially desirable answers when interviewed in person. These respondents report working longer hours, entirely due to a lower probability of working zero hours (Figure 3).[25] Measures of money taken from the enterprise and goods and services given to household members may also be subject to social pressure in in-person interviews but the direction of the pressure is ambiguous. Respondents may want to display generosity by reporting large gifts of goods and services to their household or display business focus by taking less money and fewer goods and services from the enterprise. We find that in-person respondents report giving more goods and services from the enterprise to their household, mainly due to a lower probability of giving nothing, but do not take more money from the enterprise for themselves (Figure 3). Previous research has documented higher social desirability bias in phone interviews, though on very different outcomes (Holbrook et al., 2003).

Importantly, we do not find evidence that interview medium effects are systematically different for reported flow measures (profits, sales, costs) and stock measures (fixed assets, number of total/paid/full-time employees). We might expect the latter to be more strongly correlated through time, in which case more frequent interviews would yield less additional information. However, the flow and stock measures are not systematically more dispersed for one medium than the other (Figures 2 and 3) and the mean differences by medium do not have systematically different signs for flow and stock measures (Table 3). This arises in part because we see high dispersion even in fixed assets and stock/inventory, which may be closer to a stock measure in some enterprises and a flow measure in others. We show some of this dispersion in Appendix J.

We explore this idea more formally by modelling the time-series processes for one flow variable – log profit – and one stock variable – log capital stock. We emphasise that there is substantial variation within firm through time in both flow and stock variables, so modeling time-series processes is a potentially important topic. Figure 4 shows the CDF of the within-enterprise interquartile ranges of profit and capital stock over the panel. The range for profit exceeds one log

---

[25] This pattern 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 persists when we account for nonresponse using Lee bounds or reweighting (Appendix C).

Figure 4: **Within-Enterprise Dispersion of Repeated Interview Measures**
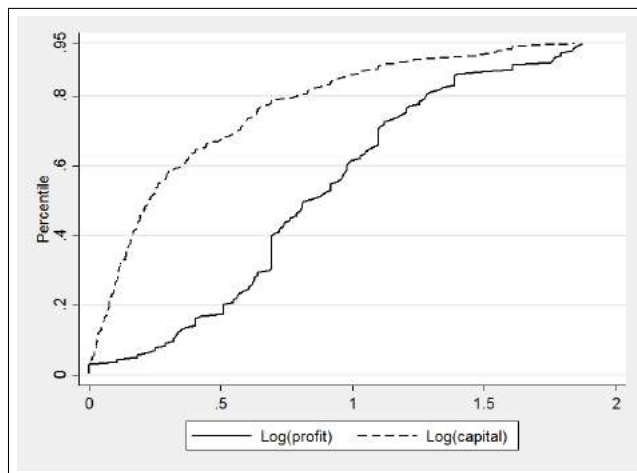
Notes: We calculate the interquartile range for each of log profit and log capital stock over the repeated interview phase for each enterprise. We show the CDF of this measure here. Capital stock includes both stock/inventory and fixed assets.

point for 40% of firms and the range for capital exceeds one log point for 15% of firms. Table 5 reports Blundell-Bond estimates for these two measures, separately for weekly phone and weekly in-person respondents (Blundell and Bond, 1998). We assume an AR(1) structure on both error terms and use two lags for profit and four lags for capital stock.[26] We report Wald tests for the null hypothesis that the estimated dynamics are equivalent between phone and in-person interviews at the bottom of the table. We fail to reject equality for both measures, showing that phone and in-person interviews provide similar information about time-series structure.

4.3 Do interview medium and frequency effects differ by subgroup?

We find limited evidence of heterogeneous interview frequency or medium effects on six pre-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 (Appendices C.4 and D.4). We might expect that owners with better record-keeping capacity (had multiple employees or kept written records at baseline) or better numerical skills (education, digit recall span, and numeracy scores) would be less susceptible to interview frequency or medium effects. However, we observe few differences by interview frequency for 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

---

[26] We justify this decision empirically in Appendix J.

Table 5: **Blundell-Bond Estimates: Disaggregating by Data Collection Group**

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

Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 5% levels respectively. The final row reports Wald tests of the 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 6: **Response Rates by Data Collection Group**

Interview completed	(1) % of repeated	(2) Any repeated	(3) All repeated	(4) Endline
Monthly in-person	0.573	0.849	0.315	0.664
	(0.021)	(0.021)	(0.027)	(0.027)
Weekly in-person	0.515	0.876	0.060	0.726
	(0.020)	(0.019)	(0.014)	(0.026)
Weekly phone	0.478	0.869	0.081	0.591
	(0.020)	(0.020)	(0.016)	(0.029)
# enterprises	895	895	895	895
p-value for monthly in-person = weekly in-person	0.045	0.334	0.000	0.104
p-value for weekly in-person = weekly phone	0.187	0.794	0.332	0.000

Notes: This table shows coefficients from respondent-level linear regressions on data collection group indicators with heteroscedasticity-robust standard errors. The dependent variables are calculated at the respondent level from the panel data.

weekly.

There is more evidence of heterogeneity in interview medium effects 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.

4.4 Do interview frequency or medium affect response rates and attrition?

The interview response rate may differ by both interview medium and frequency, making this an important outcome of interest. We evaluate different data collection media and frequencies on three criteria: the *response rate*, the fraction of scheduled interviews completed; the *attrition rate*, the fraction of respondents who permanently attrit from the panel at each time period; and the *period coverage rate*, the fraction of respondents that are interviewed at least once in each x-week period for different values of x. We define a respondent as an attriter in week t if the respondent does not complete an interview in any week $s \geq t$, including the endline interview. This is technically an upper bound on the attrition rate, as these respondents might have been interviewed again if we had continued the study for longer. We observe very few cases of item nonresponse (i.e. a respondent is interviewed but answers only some questions) and so do not analyse this outcome.

Table 7: **Coverage Rates by Data Collection Group**

	(1) 2 weeks	(2) 4 weeks	(3) 6 weeks	(4) 8 weeks	(5) 10 weeks
Monthly in-person	0.283	0.543	0.628	0.723	0.792
	(0.010)	(0.021)	(0.022)	(0.022)	(0.022)
Weekly in-person	0.648	0.713	0.746	0.778	0.817
	(0.023)	(0.023)	(0.022)	(0.022)	(0.020)
Weekly phone	0.604	0.710	0.765	0.799	0.830
	(0.021)	(0.022)	(0.022)	(0.021)	(0.020)
# enterprises	895	895	895	895	895
p-value: monthly in-person = weekly in-person	0.000	0.000	0.000	0.080	0.397
p-value: weekly in-person = weekly phone	0.155	0.901	0.557	0.481	0.658

Notes: This table shows coefficients from respondent-level linear regressions on data collection group indicators with heteroscedasticity-robust standard errors. The dependent variable equals the fraction of periods of x weeks in which the respondent completes at least one interview, with different x in each column. For example, the dependent variable in column 1 equals one if the enterprise completes at least one interview in each of the following periods: weeks 1-2, weeks 2-3, weeks 3-4, etc. The design of the data collection means that the dependent variable is less than one for enterprises in the monthly in-person interview group when $x \leq 3$. Note that the coverage rate for $x = 12$ is identical to the values in column 2 of Table 6.

Table 8: **Reasons for Attrition by Data Collection Group**

|	(1) Monthly in-person	(2) Weekly in-person	(3) Weekly phone	(4) H_0: monthly = weekly	(5) H_0: phone = in-person
Enterprise closed & owner refused to continue	0.057	0.097	0.060	0.100	0.127
	(0.014)	(0.020)	(0.014)		
Owner dead or too sick to continue	0.017	0.017	0.013	0.996	0.741
	(0.007)	(0.007)	(0.007)		
No reason given for refusal	0.077	0.100	0.138	0.321	0.160
	(0.015)	(0.017)	(0.020)		
Incorrect contact/location information	0.003	0.010	0.023	0.318	0.201
	(0.003)	(0.006)	(0.009)		
Owner moved away	0.074	0.047	0.087	0.166	0.054
	(0.015)	(0.012)	(0.017)		
Total	0.228	0.271	0.322	0.314	0.261
	(0.027)	(0.033)	(0.031)		

Notes: This table shows reasons for attrition across different data collection groups. Estimates are from linear regressions of an indicator for each reason on data collection group indicators. Heteroskedasticity-robust standard errors are shown in parentheses. Columns 4 and 5 show the p-values from testing equality between respectively monthly and weekly in-person groups and weekly in-person and phone groups.

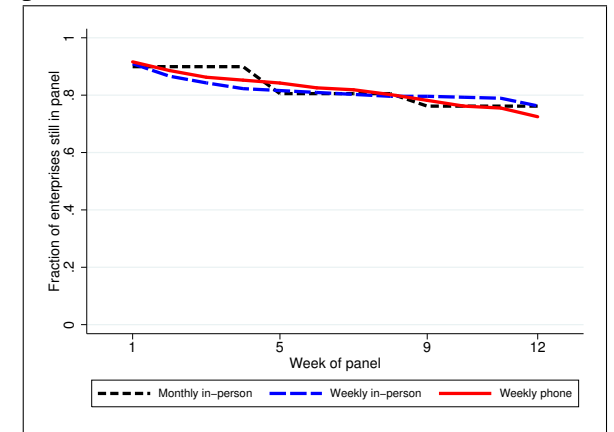Figure 5: **Response Rates and Attrition by Data Collection Group**

Panel A: Fraction of respondents completing each possible fraction of interviews
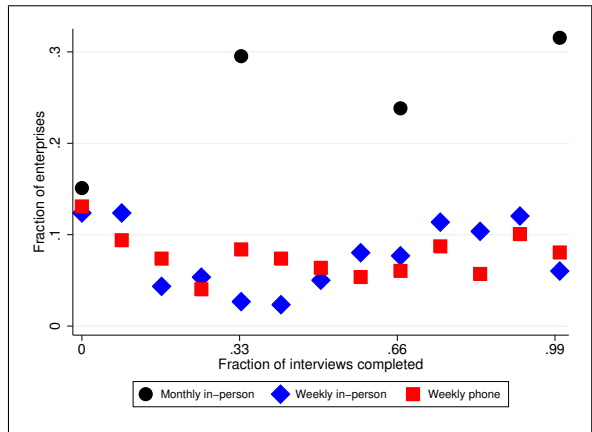
Panel B: Response rate by week
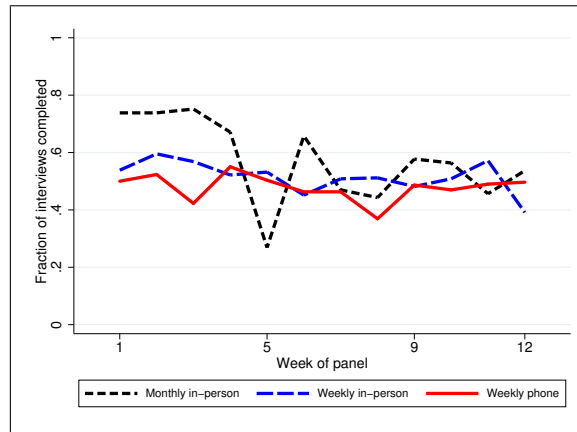
Panel C: Fraction of respondents still in panel after each week of the repeated interview phase

Notes: Panel A shows the fraction of respondents in each data collection group who complete each possible fraction of their scheduled repeated interviews. This fraction is in $\{0, 1/3, 2/3, 1\}$ for the monthly data collection group and in $\{0, 1/12, \ldots, 11/12, 1\}$ for the weekly data collection groups. Panel B shows the response rate: the fraction of respondents in each data collection group who are interviewed in each week. Note that the set of respondents in the monthly in-person group is different in weeks 1/5/9, 2/6/10, 3/7/11, and 4/8/12. Panel C shows the fraction of respondents in each data collection group in each week $t \in \{1, \ldots, 12\}$ who are interviewed in at least one week $s \geq t$. This equals one minus the rate of permanent attrition in the panel. We include the endline interview in the set of possible s but the differences between the data collection groups are robust to excluding the endline.

We observe four main patterns in our analysis of repeated interview response and attrition. First, monthly in-person interviews deliver higher response rates, similar attrition, and lower period coverage to weekly in-person interviews. Respondents in the monthly in-person group complete a higher fraction of scheduled interviews and are more likely to complete all interviews, but are equally likely to complete no interviews (Table 6 columns 1-3 and Figure 5 panel A). Most of this difference occurs because respondents in the monthly group are more likely to complete their first interview than respondents in the weekly group (see weeks 1-4 in Figure 5 panel B). The attrition rate is similar for weekly and monthly groups in almost all weeks (Figure 5 panel C).[27] However, period coverage is lower in the monthly group: these respondents are less likely to be interviewed in each x-week period for all possible values of x.[28] We conclude that weekly interviews deliver a higher volume of information (due to higher coverage) but that this information may be less representative (due to lower response rates in some periods). If the nonresponse is not driven by time-varying respondent characteristics, then the higher period coverage will be more useful than the higher response rate.

Second, the lower response rate in weekly interviews is partly explained by logistical constraints. The average response rate in each week in the weekly in-person group is 52% in the repeated phase, when interviews must be conducted within one week, but 73% in the endline phase, when interviews were were spread over four weeks. In the 7 weeks of the repeated phase with a public holiday, the response rate in the weekly in-person group was 5 percentage points lower (standard error 1.4 points) than in other weeks. The response rate is not lower in weeks with public holidays in the monthly in-person or weekly phone group, and fieldworkers in those groups reported feeling less time pressure. These patterns suggest that the logistical constraint of interviewing respondents in person each week was binding in many cases. We assigned two, eight, and four fieldworkers to the monthly in-person, weekly in-person, and weekly phone groups respectively.

Third, medium does not affect attrition. Weekly phone and weekly in-person interviews deliver very similar response rates, attrition, and period coverage. Respondents in both groups complete approximately 50% of scheduled repeated interviews, 87% of respondents in both groups complete at least one interview, and approximately 7% of respondents in each group complete all interviews (Table 6 columns 1-3, Figure 5 panel B). None of these differences is statistically significant. The response rate differs between the two groups in 4 of the 12 weeks but there is

---

[27] The difference in attrition is above 4 percentage points and statistically significant in only weeks 3 and 4 of the panel. In particular, the difference is negligible in the final weeks of the panel, suggesting this pattern might have continued in a longer panel.

[28] Table 7 shows coverage rates for selected values of x. The coverage rate is lower for all values of x for respondents in the monthly in-person group than the weekly in-person group. The difference is statistically significant for $x \leq 7$.

no clear time pattern to these differences (Figure 5, panel B). The attrition rate is not statistically significantly different for weekly phone and in-person interviews in any week (figure 5, panel C). Reasons for attrition are similar across the two groups, but respondents in the phone group are 4 percentage points more likely to report that they have moved away (Table 8).[29] This claim is easier to verify with in-person interviews, so this pattern may arise if respondents want to avoid interviews without explicitly refusing.[30] Period coverage is very similar: respondents in both groups are equally likely to be interviewed in each x-week period during the panel for all possible values of x (Table 7). There is little difference in the time-series structure of responses: the autocorrelations in responding are -0.017 and 0.039 for respectively the in-person and phone groups (p-value of difference = 0.078).[31] We conclude that researchers should not prefer in-person over phone interviews for response, attrition, or coverage reasons.

Fourth, switching from phone to in-person interviews during a panel survey can raise the nonresponse rate. The response rate for respondents in the weekly groups rose from the repeated phase to the endline phase. But the rise is substantially smaller for the weekly phone group than the weekly in-person group (11 versus 21 percentage points, p-value of difference = 0.007). This difference is largely explained by a higher probability that endline enumerators could not find the enterprise location for respondents in the phone group than in either in-person group.

We also explore the relationship between response rates and baseline characteristics, with results shown in Table 9. Missing interviews in both the repeated and endline data collection stages varies systematically with only a small number of baseline characteristics. Respondents with more education and who do not answer the baseline interview question about total household income are more likely to miss repeated interviews. Respondents who speak Zulu, moved more recently to Soweto, have more employees, do not plan to grow their enterprises, and regularly conduct business over the phone are more likely to miss the endline interview.[32]

---

[29] The attrition rate is higher in Table 8 than Figure 5 panel C because respondents had the option of refusing further repeated interviews but still participating in the endline. Few respondents did so.

[30] The fraction of respondents refusing to continue in the panel without specifying a reason is 4 percentage points higher in the phone than the in-person group (p-value of difference = 0.160), which also weakly suggests that phone interviews are more frustrating for respondents or build weaker respondent-fieldworker rapport. The fraction of respondents with incorrect contact information is similar in the phone and the in-person groups. This provides reassuring evidence that phone interviews are feasible even in a context where phone numbers change often. 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 response rates and attrition may be different in other studies, such as a survey of mobile street vendors that collects only one contact number per respondent.

[31] These autocorrelations are estimated conditional on the respondent-specific response rate. The unconditional autocorrelations for the weekly in-person and phone groups are respectively 0.487 and 0.489 (p-value of the difference = 0.953). The unconditional autocorrelations are dominated by variation in the response rate, so the conditional autocorrelations are more useful for understanding variation in the timing of responses.

[32] Black African are less likely to miss both repeated and endline interviews. But 889/895 respondents are Black African, so this result should be interpreted with caution.

Table 9: **Predictors of Nonresponse in Repeated and Endline 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 nonresponse rates for repeated and endline interviews. The first two columns show marginal effects from a fractional logit regression of the respondent-level nonresponse rate for repeated interviews (Papke and Wooldridge, 1996). The next two columns show marginal effects from a logit regression of an endline nonresponse indicator. Heteroskedasticity-robust standard errors are shown in parentheses. Marginal effects for continuous variables are evaluated at their sample mean value 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 enterprise type. The unconditional endline response rate is 34.0% and the unconditional mean and interquartile range of the repeated interviews response rate are respectively 47.8% and [16.7%,75.0%]. ***, **, and * denote significance at the 1, 5, and 10% levels.

The results in Table 9 show that some enterprise and owner characteristics systematically predict nonresponse. However, these results are not consistent with the most obvious economic models of nonresponse. For example, we might expect that enterprise respondents whose time is (self-perceived to be) more valuable or who find it more difficult to intertemporally shift their time expenditure from other activities will miss more interviews. But the response rate does not vary systematically with childcare responsibilities, the number of other employees in the enterprise (in the repeated phase), or enterprise sector. The probability also does not vary systematically with variables that might proxy for the difficulty respondents face in calculating answers to our interview questions: enterprise age, the presence of written records, registration for tax, financial literacy test results, or digit span recall test results. We conclude that nonresponse may be nonrandom, but that it is not explained by any obvious economic model that would cause serious concern about nonrepresentative responses.[33]

Our review of other high-frequency panel studies and other phone surveys suggests a potential trade-off between representative samples with lower response rates and higher attrition on the one hand versus more highly selected samples with higher response rates and lower attrition on the other hand. We use a population representative sample. Weekly nonresponse in our sample is very similar to that in other urban representative samples. Croke et al. (2014) complete 55% of weekly mobile phone interviews with a random sample of Dar es Salaam households. Gallup (2012) complete approximately 42% of scheduled phone/text interviews with nationally representative households in Honduras and Peru.

Alternatively, to maintain high power in experiments, some researchers prioritise low attrition and high response rates over representativity by selecting respondents for their willingness to participate in interviews. Beaman et al. (2014) begin with a baseline of 1195 microenterprises but include only 508 these in their panel study; the other 687 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 the 687 enterprises completed half as many panel surveys as the 508, their response rate would have fallen from 93% to 66%. Heath et al. (2016) obtain weekly response rates above 77% in a high-frequency panel, but work with respondents who have already completed up to eight annual waves of a household panel survey in Ghana and were easy to track. Arthi et al. (2016) conduct a high-frequency panel of rural households but replace the 8% of households that

---

[33] We also estimate group-specific regressions of response rates on baseline characteristics, shown in Appendix I. The coefficients are jointly significantly different for the repeated stage but not the endline stage. The difference in missed repeated interviews is explained by language, digit recall score, previous employment history, and race (which has almost no variation). These patterns are again not consistent with any obvious economic model of systematic nonresponse.

miss any interview in the first five weeks of the panel.[34]

5   Cost effectiveness

Phone interviews 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 interviews (since these 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 data collection group, so we can easily allocate these costs to data collection groups. One car was hired for each of the weekly in-person and the monthly in-person interview teams, so these are also easily allocated. Phone calls to arrange in-person interviews and phone interviews were conducted on different phone lines so these costs can be allocated to specific data collection groups. However, we observe only the total cost for the last three categories: printing, data capture, and respondent incentives. We allocate these to the three data collection groups by dividing the total cost by the number of successfully completed interviews in each arm.

A phone interview conducted weekly cost roughly ZAR49 (US$4.76 at August 2013 rates), excluding the cost of conducting a prior baseline to collect contact details.[35] An in-person interview conducted weekly cost ZAR63 (US$6.12) and one conducted monthly cost ZAR75 (US$7.30). Phone interviews 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.[36]

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

---

[34] Our sample is slightly selected, as our panel includes only 87% of the 1046 eligible enterprises that we recontacted in the baseline. But the degree of selection is low compared to some other samples.

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

[36] 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 interview.

Figure 6: **Cost per Successfully Completed Interview by Data Collection Group**
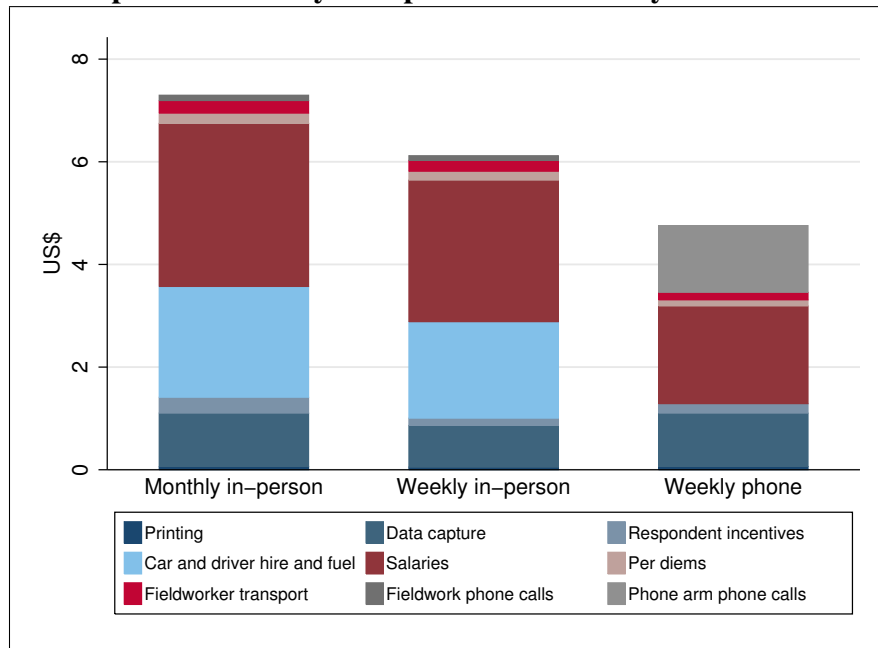
Notes: This figure shows the average per-interview cost interview by data collection group. The average cost is constructed by summing the nine cost categories shown above, separately by data collection group, then dividing this by the number of successfully completed interviews in each group. This calculation 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.

cost saving will be larger if enumerators require overnight travel and incur accommodation costs, which was not the case in our study.

Our cost savings are relatively low because we were interviewing in a high density urban area with low transport costs and South African call costs are relatively high (interviews cost roughly US$1.30 per 15 minute interview). 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. For example, 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 interview (Dillon, 2012).

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

interviews to avoid differences in motivation and incentives.

All our interviews were conducted on paper and data-captured to avoid differences in data quality between interview media. But electronic data entry can be easier and more accurate for phone than in-person interviews. 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 higher frequency, holding interview medium fixed. The third group are interviewed at weekly intervals by mobile phone. This allows us to test the consequences of collecting data by phone, holding interview frequency fixed

We show that weekly phone interviews are useful. Weekly interviews capture extensive volatility in a number of measures that is not visible in monthly interviews and the time-series structure of these measures does not differ by medium. More generally, researchers can use high-frequency phone interviews to describe this volatility, inform models of intertemporal optimization, illustrate the time path of treatment effects, inform dynamic treatment regimes, and average over multiple measures to improve statistical power.

We show that high-frequency mobile phone interviews do not yield lower-quality data than low-frequency in-person interviews. Data patterns do not differ by interview frequency or medium for most measures and at most quantiles of the distribution. Indeed, high-frequency interviews may be somewhat more accurate than monthly interviews – respondents in in-person interviews use written records more and fail a check on the profit calculation less often. Face-to-face interviews may suffer more social desirability bias – respondents overreport hours worked and goods/services given to household members compared to respondents in phone interviews. Using an endline administered in person, we find little evidence that high-frequency data collection (either over a mobile phone or in person) alters enterprise outcomes or owner behaviour. All three methods result in similar attrition from the panel. Respondents assigned to high-frequency interviews miss a higher fraction of interviews, but there is no difference between phone and in-person interviews.

Mobile phone interviews are, however, substantially cheaper than in-person interviews and offer a far larger volume of data for the same price.

We conclude that high-frequency mobile phone surveys offer considerable cost savings, little reduction in data quality, and scope for more informative data. While we study microenterprises, our results can inform enterprise and household surveys more generally.

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.

ARTHI, V., K. BEEGLE, J. DE WEERDT, AND A. PALACIOS-LOPEZ (2016): "Not Your Average Job: Irregular Schedules, Recall Bias, and Farm Labor Measurement in Tanzania," Tech. Rep. 21, Universiteit Antwerpen, Institute of Development Policy and Management.

BANERJEE, A., E. DUFLO, R. GLENNERSTER, AND C. KINNAN (2015): "The Miracle of Microfinance? 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.

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.

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. POCOCK (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.

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.

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.

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.

——— (2017): "Business Practices in Small Firms in Developing Countries," *Management Science*, forthcoming.

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.

ROMEO, G. D. P. G. A. (2013): "Challenges and Opportunities of Mobile Phone-Based Data Collection: Evidence from South Sudan," *World Bank Policy Research Paper Series*, 6321, 1–38.

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.

SILVA, J. S. AND P. M. PARENTE (2013): "Quantile Regression with Clustered Data," *Economics Discussion Papers, University of Essex, Department of Economics.*, 728.

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.