Call Me Maybe: Experimental Evidence on Frequency and Medium Effects in Microenterprise Surveys*

Robert Garlick†, Kate Orkin‡, Simon Quinn§

February 9, 2019

PRE-ANALYSIS PLAN · QUESTIONNAIRES

Abstract

We study the effect of differences in survey frequency and medium on microenterprise survey data. We randomly assign enterprises to monthly in-person, weekly in-person, or weekly phone surveys for a 12-week panel. We find few differences across groups in measured means, distributions, or deviations of measured data from an objective data quality standard provided by Benford's Law. However, phone interviews generate higher within-enterprise variation through time in several variables and may be more sensitive to social desirability bias. Higher-frequency interviews do not lead to persistent changes in reporting or increase permanent attrition from the panel but do increase the share of missed interviews. These findings show that collecting high frequency survey data by phone does not substantially data quality. However, researchers who are particularly interested in within-enterprise dynamics should exercise caution when choosing survey medium.

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

*This project was funded by Exploratory Research Grant 892 from Private Enterprise Development for Low-Income Countries, a joint research initiative of the Centre for Economic Policy Research (CEPR) and the Department for International Development (DFID). The authors thank Bongani Khumalo, Thembela Manyathi, Mbuso Moyo, Mohammed Motala, Egines Mudzingwa, and fieldwork staff at the Community Agency for Social Enquiry (CASE); Mzi Shabangu and Arul Naidoo at Statistics South Africa; Rose Page and staff at the Centre for Study of African Economies; and Chris Woodruff and the PEDL team. Our thanks to editor David McKenzie, three anonymous reviewers, Markus Eberhardt, Simon Franklin, Markus Goldstein, David Lam, Murray Leibbrandt, Ethan Ligon, Owen Ozier, Duncan Thomas and seminar audiences and conference participants for excellent comments. Our pre-analysis plan can be viewed at https://www.socialscienceregistry.org/trials/346.

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

‡Blavatnik School of Government, Centre for the Study of African Economies and Merton College, University of Oxford; kate.orkin@merton.ox.ac.uk

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

1 Introduction

Researchers designing surveys must choose the interview frequency and medium that generate the optimal quality and volume of data given budget constraints. Alternatives to traditional in-person, low-frequency surveys are increasingly widely used. Phone surveys offer cost savings and the ability to reach mobile populations and to collect data during periods of conflict or disease.[1] High-frequency surveys enable better measurement of short-term fluctuations and outcome dynamics.[2] However, high-frequency or phone surveys may generate systematically different measurements. This might offset the advantages of richer, cheaper data. Experimental comparisons of data collected using different survey methods help researchers to evaluate these tradeoffs.

We run the first randomised controlled trial to compare microenterprise data from surveys of different frequency and medium. We study a representative sample of microenterprises in the city of Soweto in South Africa. We randomly divide them into three groups. The first group is interviewed in person at every fourth week for 12 weeks. This group is most similar to traditional panel surveys. 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 effects of data collection medium, holding the interview frequency constant. All interviews use an identical questionnaire measuring 14 enterprise outcomes in approximately 20 minutes.

We find three main results. First, for most outcomes there are few frequency or medium effects on the means, on prespecified quantiles of the distribution, or on the frequency of outliers. There

[1] For examples, see Bauer et al. (2013), Dillon (2012), Pape (2018), Turay et al. (2015), and van der Windt and Humphreys (2013).

[2] Researchers can use high-frequency data to study volatility and dynamics in enterprise and household outcomes (Dupas et al., 2018; Collins et al., 2009; McKenzie and Woodruff, 2008), inform models of intertemporal optimisation in response to shocks (Banerjee et al., 2015; Rosenzweig and Wolpin, 1993), illustrate the time path of treatment effects (Jacobson et al., 1993), 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 researchers to use shorter recall periods without sacrificing comprehensive time-series coverage (Beegle et al., 2012; Das et al., 2012; De Nicola and Giné, 2014; Heath et al., 2017). See Abebe et al. (2016), Beaman et al. (2014), Carranza et al. (2018), Dabalen et al. (2016), Franklin (2017), Leo et al. (2015), and Zwane et al. (2011) for other examples of high-frequency or phone-based surveys.

are no substantial differences for key outcomes like enterprise closure, profit, sales, costs, fixed assets, or numbers of employees. The largest difference is that phone surveys generate lower reported labour supply. This seems to arise because our in-person interviews take place at enterprises and disproportionately miss respondents who work few hours. Phone respondents also report holding less stock and inventory; transfering more money, stock, or services to the household; and using more often written records to answer survey questions. All outcomes with medium effects except the use of written records are "estimating outcomes," high-valued outcomes where responses are likely to be estimates rather than precise counts (Blair and Burton, 1987; Gibson and Kim, 2007). We see no frequency or medium effects on our smaller list of "counting outcomes," low-valued outcomes where respondents can feasibly give a precise answer by counting.

Second, objective indicators of data quality do not differ systematically between interview frequencies or media. We measure data quality by comparing the digit distribution in survey responses to Benford's Law, a statistical regularity often used to test for data manipulation.[3] No one method performs consistently better on this metric. We similarly find few differences between groups in a measure of internal consistency between survey answers: the difference between directly and indirectly elicited profit. These comparisons show that there are limited cross-sectional quality differences in data collected monthly or weekly, by phone or in person.

Third, however, we find phone surveys yield more dispersion in within-enterprise data through time than in-person surveys. Phone surveys yield lower one-week autocorrelations and higher within-enterprise standard deviations on roughly half our 14 outcomes, including some flow and some stock outcomes. We conclude that using phone or high-frequency surveys does not systematically raise or lower the quality of data used for cross-sectional or static panels models. However, researchers particularly interested in within-enterprise dynamics should exercise caution when choosing survey medium.

We also document four secondary results that may inform researchers' choice of survey frequency and medium. First, autocorrelations are higher for outcomes collected at higher frequen-

[3]We thank an anonymous reviewer for this suggestion. See Judge and Schechter (2009) for a review of multiple survey datasets from developing countries against this benchmark. Schündeln (2018), Mahadevan (2018), and Garlick (2019) use this approach to assess quality of administrative and survey data in development applications.

cies, so the new information generated by additional surveys may be smaller when surveys are closer together in time. Second, phone surveys are cheaper than in-person surveys. Third, respondents miss a higher share of high-frequency interviews, but high-frequency interviews are still more likely to capture respondents in any month. Fourth, the few frequency effects on means and distributions we observe during our panel do not persist in an in-person endline survey we conduct several weeks later. This shows that higher-frequency surveys in this setting do not generate large persistent changes in behaviour or reporting. This may be useful for researchers interested in the data quality implications of conducting high-frequency panels with subsamples of a larger sample.

These findings come with two caveats. First, non-response in our sample is relatively high: we complete slightly more than half the scheduled interviews. This occurs partly because we imposed a maximum of three attempts to contact each respondent for an interview. This avoided a backlog developing across multiple weeks and kept the number of contact attempts consistent across methods. Patterns of non-response might differ in panels that take advantage of low phone costs to make more attempts. Second, our panel lasts for only three months with at most 12 interviews per respondent. More frequent interviews over a longer time period might induce different patterns in reporting, attrition or non-response.

Our results contribute to a literature in development economics exploring effects of survey frequency or data collection mode (Caeyers et al., 2012; Lane et al., 2006; Fafchamps et al., 2012).[4] To the best of our knowledge, this and concurrent work by Heath et al. (2017) are the first papers experimentally comparing both survey frequency and medium in a developing country context. While we study microenterprises, our results may be relevant for other types of surveys. Some microenterprise outcomes correspond to outcomes in other surveys; for example, item-specific and total costs in a survey of small enterprises may behave similarly to item-specific and total expenditure in a household survey. A comprehensive mapping of our outcomes to outcomes in other types of surveys is outside the scope of this paper. Instead, we report outcome-specific information and allow the reader to decide if and how these map to their outcomes of interest.

[4] A related literature uses experimental variation to test if different questionnaire designs, recall periods, or survey incentives affect reported outcomes, response rates, or data quality (Arthi et al., 2018; Beegle et al., 2012; Beaman and Dillon, 2012; Das et al., 2012; Dillon et al., 2012; Friedman et al., 2017; Gibson and Kim, 2007; Scott and Amenuvegbe, 1991; Stecklov et al., 2017).

Our work also relates to research on survey media in household surveys and opinion polls, mostly from the US. We find that survey medium has limited effects on reported outcomes, consistent with De Leeuw (1992), Groves (1990), Groves et al. (2001), and Körmendi (2001). Unlike these studies, we find that response rates do not differ by medium. This difference may occur because we analyse medium effects in a panel that has been recruited in person, while the US studies often analysed medium effects in "cold-called" cross-sectional samples. Like this work, we find evidence consistent with higher social desirability bias in phone interviews (Holbrook et al., 2003). Phone respondents, whose actions cannot be seen by enumerators, are more likely to report using written records to help them answer our survey questions.

Our work also relates to the literature on panel conditioning, which shows that being surveyed or being surveyed more frequently sometimes changes behaviour (Beaman et al., 2014; Crossley et al., 2017; Stango and Zinman, 2014; Zwane et al., 2011). We find little evidence that differences in interview frequency over our three-month panel generate persistent changes in reported outcomes. This may arise because we study enterprise outcomes that are already salient to respondents. Bach and Eckman (2018) and Franklin (2017) similarly find no persistent effects of interview frequency on the already salient outcome of employment.

We describe the experimental design and data collection processes in Section 2 (and Appendices A and B). In Sections 3, 4, and 5 (and Appendices C – F), we present the three main results of the paper: frequency and medium effects on respectively outcome means and distributions, data quality, and within-enterprise data patterns through time. Section 6 brings these results together to categorize outcomes based on the pattern of frequency and medium effects. Section 7 (and Appendices G – I) presents our four secondary results on autocorrelations, costs, non-response, and persistence. Section 8 concludes.

2 Sample, Experimental Design, and Data

2.1 Sampling and Randomisation

We work in Soweto, near Johannesburg, South Africa. This is a city of approximately 1.28 million people in 2011, of whom 99% are Black Africans. 41% of adults 15 and older engage in some

form of economic activity, including occasional work. 19% of households reported receiving no annual income and another 42% reported receiving less than $10 per day.[5]

We recruited a representative sample of 1,046 households who owned eligible microenterprises and lived in low-income areas of Soweto. We recontacted 895 of these households several months later to complete our baseline survey. We describe the sampling scheme in Appendix A. We use a common definition of microenterprises: enterprises with at most two full-time employees (in addition to the owner) that do not provide a professional service (e.g. medicine). We excluded any enterprise which did not operate at least three days each week, to exclude seasonal or occasional enterprises where there would be limited intertemporal variation in outcomes.

Most of the 895 enterprises operated in food services (43%) or retail (32%). They were relatively well-established (mean age seven years) and had a diversified client base (mean and median client numbers of respectively 34 and 20, varying substantially by sector). However, they were relatively small: 61% had no employees other than the owner and 28% had only one other employee. Very few were formally registered for payroll or value-added tax, but 20% reported keeping written financial records. The sample is similar to five microenterprise samples from the Dominican Republic, Ghana, Nigeria, and Sri Lanka (De Mel et al., 2008; Drexler et al., 2014; Fafchamps et al., 2014; Karlan et al., 2012; McKenzie, 2017), though our enterprises are slightly older and more concentrated in food and retail/trade. Appendix Table A1 shows detailed summary statistics.

The enterprise owners' households had mean monthly income of US$394 across all sources, falling in the fourth decile for all households across South Africa.[6] The households had an average of 3.8 other members, with an interdecile range of 1 to 7. In 55% of households, the enterprise accounted for half or more of household income and 63% of owners perceived pressure within their households to share profits. Only 15% had less than some secondary education. All sampled enterprise owners owned mobile phones.[7]

[5] Authors' own calculations, from the 2011 Census public release data. We follow the terminology of Statistics South Africa, which asks census respondents to describe themselves in terms of five population groups: Black African, Coloured, Indian or Asian, Other, and White.

[6] We use an exchange rate of US$1 = ZAR10.28, the market exchange rate on the first day of data collection in August 2013.

[7] 87% of South Africans aged 18 or older own a mobile phone and the rate is higher in cities (Mitullah and Kama, 2013).

After a baseline survey, we divided the 895 enterprises into three data collection groups using stratified random assignment: monthly in-person interviews (298 enterprises), weekly in-person interviews (299 enterprises), and weekly phone interviews (298 enterprises). We describe the randomisation scheme in Appendix A and show in Table A1 that the groups are balanced on 38 of 40 measured baseline characteristics.

2.2 Survey Protocols

We conducted repeated interviews with each enterprise owner between March and July 2014. We attempted to survey microenterprises in the weekly group once per week for 12 weeks, in person or by phone. We attempted to survey enterprises in the monthly group three times every fourth week for 12 weeks, in person. We randomly split the monthly group into four, so 75 enterprises were interviewed each of weeks 1/5/9, 2/6/10, 3/7/11, and 4/8/12, providing a comparison group for each week when the weekly enterprises were interviewed.

We standardised all survey protocols across arms, except for the variations in survey frequency and medium we study. We used the same questionnaire in all rounds, which lasted roughly 20 minutes (see Section 2.4 for a detailed description).[8] All respondents received similar incentives: a mobile phone airtime voucher worth US$1.17 transferred to their phone for every fourth interview they completed, as well as after the baseline and endline interviews. The maximum individual payment is 0.3% of mean annual household income, so income effects should be negligible. The incentives were designed to encourage participation, not to precisely equalise compensation for respondent time across arms.[9] South African mobile phone users are not charged for calls received, so respondents pay no pecuniary cost for completing the surveys.

Enumerators surveyed the same enterprise each week or month to simplify tracking. We randomly assigned enumerators to data collection groups, conditional on languages spoken. We assigned two, eight and four enumerators respectively to the monthly in-person, weekly in-person

[8] We did not use real-time panel data consistency checks to query responses that changed a lot from previous weeks. However, Fafchamps et al. (2012) find that "the overall impact of these consistency checks on the full sample is rather limited."

[9] We do not test the effect of variation in incentive size. See Singer and Ye (2013) for a review of research into survey incentives, response rates, and data quality.

and weekly phone groups. Enumerator age, gender, experience, and language are balanced across groups. Within each group, enumerators were assigned to enterprises to allow interviews in owners' preferred language (English, seSotho, seTswana, or isiZulu) and to minimise enumerators' travel time between enterprises.

Surveys were conducted at similar times of day, during working hours. Enumerators set up an appointment time to contact their set of respondents before the first repeated interview and tried to use that time each week or month for the remainder of the panel. Enumerators confirmed the time for the next interview at the end of each interview. Enumerator assignments are collinear with treatment groups. We used only 14 enumerators, so readers may be concerned that treatment and enumerator effects are confounded. However, differences in reported outcomes across enumerators appear small. Conditioning on enumerator fixed effects increases the centered R^2 by only 0.004 to 0.078 for our main specifications, except for three variables we discuss below. All our findings are robust to controlling for enumerators' age, gender, experience, and language.

We conducted in-person interviews at the enterprises. If a respondent was scheduled to close their business, enumerators usually moved the interview to another day. All in-person interviews and 86% of phone interviews were conducted at the enterprise location (or respondent home for home-based enterprises). This difference highlights a useful feature of phone surveys – more flexibility in tracking respondents – but might induce differences in selection. We return to this issue in Section 3.

We also conducted an in-person endline interview with each enterprise owner at the enterprise location, 1-4 weeks after the repeated interviews finished. For this interview, we randomly re-assigned enumerators to enterprises. Because all enterprises are interviewed face-to-face for the endline, any differences observed in the endline data must be due to persistent frequency or medium effects from the repeated interviews.

2.3 Tracking Protocols

In this section, we describe our tracking protocols. We describe the patterns of non-response and attrition in Section 7.3 and Appendix H. We show in Section 3 and Appendices C – F that our main

results are robust to accounting for non-response.

We standardised our tracking protocol across groups to ensure that differences between groups reflect frequency and medium effects, rather than tracking effects.[10] Enumerators made three attempts to contact each respondent in each scheduled week/month, as in some Living Standards Measurement Studies and Demographic and Health Surveys (Grosh and Munoz, 1996; McKenzie, 2015). The high frequency of our panel meant that we had to impose a maximum number of contact attempts. Some low-frequency panels continue to attempt to contact respondents for many months (Thomas et al., 2012). Few economic studies report detailed tracking rules so we cannot measure the prevalence of different tracking protocols.

Enumerators were supposed to complete second attempts later on the same day as the first attempt and third attempts 1-2 days after the second attempt. A contact attempt for the in-person groups meant a visit to the enterprise premises. A contact attempt for the phone groups meant talking to the respondent, so a missed call or talking to another person did not count as an attempt. After failing to interview a respondent on the third attempt, enumerators marked them as missing for that week/month. Respondents who missed an interview were always contacted in the next scheduled week or month (except if they asked not to be recontacted).[11]

2.4 Outcome Measures

We used the same questionnaire for all repeated and endline interviews in all groups.[12] The questionnaire covered 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. The questionnaire also asked respon-

[10] Stecklov et al. (2017) run a survey experiment adopting a similar strategy on non-response. We could have instead used group-specific tracking protocols that aimed to equate the response rate across groups. However, this would have required strong prior evidence about frequency, medium, and tracking effects on response rates.

[11] We continued to interview respondents who closed or sold their enterprises using a different questionnaire. We did not track respondents who left the greater Johannesburg region, as we could only interview them by phone and did not want to break comparability between groups.

[12] The questionnaire first asked if the respondent still operated their enterprise. If not, the questionnaire asked what happened to the enterprise and asked about the respondents' current economic activities. Only 2% respondents stopped operating their enterprise during the survey period so we do not analyse data on closed enterprises.

dents if they used written records during the interview and several tracking questions. At the end of the interview, the enumerator assessed whether the respondent answered questions honestly and carefully. We show summary statistics for all outcomes in Appendix B and questionnaires are available in the supplementary materials. 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 frequency or medium effects differ by recall period.

We elicit profits directly, following De Mel et al. (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 total 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.

3 Few Frequency or Medium Effects on Outcome Means or Distributions

In this section, we estimate frequency and medium effects on reported mean outcomes and the distribution of outcomes. We pool observations through time across enterprises and do not yet examine patterns in within-enterprise outcomes through time. Most core enterprise outcomes do not differ by frequency or medium – enterprise closure, sales, costs, and various measures of employment – or differ by small margins in the upper tails – stock/inventory, assets, and profit. There are substantial frequency and medium effects on resources taken from the enterprise by the owner or her family and on hours worked, though the latter effect may be driven by medium-induced sample selection. Only medium effects on stock and inventory, hours worked and household takings survive corrections for multiple testing. Our findings do not differ by recall period and we find little heterogeneity in treatment effects by baseline covariates.

We estimate mean effects of interview frequency and medium using

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

where Y_{ki} is an outcome variable, winsorised at the 95$^{\text{th}}$ percentile; i, k and t index respectively enterprises, outcomes, and weeks; T_{1i} and T_{2i} are indicators for respectively the monthly in-person group and the weekly phone group; η_g is a stratification block fixed effect; and ϕ_t is a calendar week fixed effect to capture common shocks.[13] We cluster standard errors by enterprise and test $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_1 = \beta_2 = 0$. In Appendix C, we adjust our results to account for multiple testing. We estimate sharpened q-values that control the false discovery rate across all outcomes (Benjamini et al., 2006).

We report the mean effects in Table 1, along with two measures of their reliability. First, we report minimum detectable mean differences to show that these comparisons are well-powered. The medians of the minimum detectable mean differences across the binary and continuous outcomes are respectively 8 percentage points and 0.06 standard deviations.[14] Second, we estimate bounds on mean effects that adjust for differences across groups in response rates, following Lee (2009). The median bounds across all binary and continuous measures allow us to rule out differences of, respectively, 11 percentage points and 0.16 standard deviations. These bounds account for differences in response rates across groups but not for the high overall level of non-response and not for any systematic relationship between baseline covariates and non-response. In Appendix C, we show that the results in this section are robust to adjusting for non-response using inverse probability of non-response weights.

We estimate distributional effects of interview frequency and medium in two ways. First, we estimate the empirical CDFs by group and show the results of quantile regressions testing for differences at prespecified quantiles $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. We test for frequency

[13] We prespecified trimming outcomes but we subsequently chose to winsorise to reduce the loss of information from real outliers. The trimmed and winsorised results are similar. We standardise all continuous outcomes to have mean zero and standard deviation one in the monthly in-person group. We do not standardise categorical and binary measures. The categorical variables seldom have values greater than one, so we discuss treatment effects on them in percentage point terms.

[14] See Appendix D for an explanation of how we calculate minimum detectable effects using the observed experimental data. Note that MDEs calculated using our approach may be smaller than coefficient estimates from the sample data that are not significant at the chosen test size because we aim for 80%, rather than 100% power.

and medium effects at each quantile, using the false discovery rate to control for multiple testing across quantiles (Benjamini et al., 2006). Second, we examine effects on the outcome tails by constructing indicators for observations above the 95th percentile and using these indicators as outcomes in model 1.[15] These distributional measures are important for some research questions, such as analyses of high-performing or fast-growing enterprises.

We report all mean effects in Table 1 and summarize these results in Figure 1. We display CDFs and quantile test results in Figure 2 for outcomes with significant differences at any quantile and Figure A1 for all other outcomes. We report tail effects in Table 2.

There are no frequency or medium effects on means, distributions, or shares of outliers for half our outcomes: enterprise closure; number of total, full-time and paid employees; sales over two recall periods; total costs; and enterprise money kept by the respondent. There are small and marginally statistically significant frequency effects at some higher quantiles of two outcomes: fixed asset value and profit.

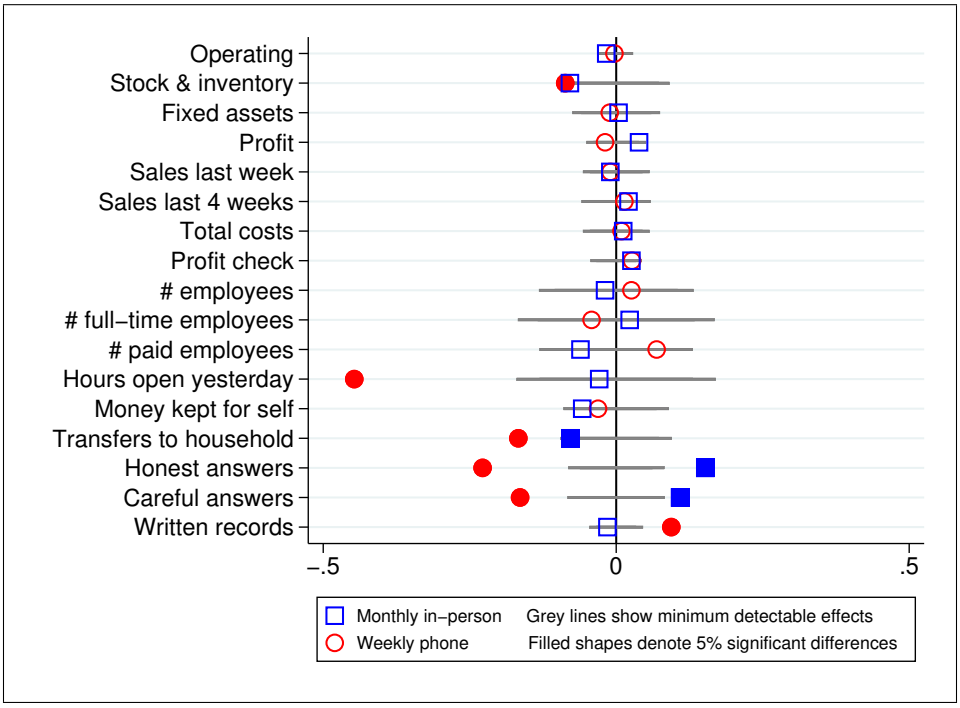
The few substantial differences we find are mostly medium, rather than frequency, effects. There are large medium effects but no frequency effects on two outcomes: phone respondents report working fewer hours and more often using written records to answer our survey. These are robust to corrections for multiple testing. The former effect is driven entirely by a higher probability of working zero hours. This result partly reflects selection induced by the location flexibility of phone surveys.[16] 14% of phone interviews were completed when respondents were away from their enterprises, while in-person interviews all took place at enterprises. 44% of respondents interviewed away from their enterprises reported working zero hours the previous day, more than 20 percentage points more than respondents interviewed at their enterprises. This shows that phone interviews catch respondents who work fewer hours and were more likely to be missed by in-person interviews at enterprise locations. The higher self-reported rate of using written records by phone respondents is surprising given that they are less likely to be interviewed at the enterprise. The effect is also large, an increase of 9 percentage points from an 8 percentage point base, and

[15] We focus on the right tails of the outcome distributions because all our measures are truncated below at zero and have substantial numbers of zeros. Results are similar when we focus on the top 10 or 1% of the outcome distributions.

[16] We thank the editor for suggesting this explanation.

Figure 1: **Frequency and Medium Effects on Mean Outcomes**

Panel A: Repeated Interviews



Panel B: Endline Interviews

Coefficients are from regressions of each outcome, winsorised at the 95th percentile, on a vector of data collection group indicators, randomisation 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. Significance tests are based on heteroskedasticity-robust standard errors, clustering by enterprise (repeated interviews only). The lines for each variable show the minimum detectable differences (MDEs) between weekly and monthly in-person interviews in Panel A; the MDEs between weekly in-person and phone interviews are approximately 25% smaller. The MDEs are between weekly in-person and phone interviews in Panel B; the MDEs between weekly and monthly in-person interviews are approximately 5% smaller.

Table 1: **Frequency and Medium Effects on Mean Outcomes in Repeated Interviews**

	(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**	**-0.079**	**0.004**	**0.039**	**-0.010**	**0.021**	**0.012**	**0.026**
	(0.010)	(0.040)*	(0.032)	(0.022)*	(0.025)	(0.025)	(0.028)	(0.023)
Weekly by phone	**-0.003**	**-0.087**	**-0.011**	**-0.019**	**-0.010**	**0.014**	**0.009**	**0.027**
	(0.006)	(0.029)***	(0.027)	(0.018)	(0.021)	(0.022)	(0.022)	(0.018)
Observations	4070	3989	3987	3986	3985	3987	3987	3984
All treatments equal (p)	0.262	0.011**	0.867	0.032**	0.880	0.677	0.882	0.248
MDE: Monthly in-person	0.029	0.091	0.075	0.051	0.057	0.059	0.057	0.044
MDE: Weekly by phone	0.023	0.073	0.060	0.039	0.045	0.047	0.045	0.034
Lee bound: Monthly in-person (lower)	-0.039	-0.125	-0.045	-0.002	-0.052	-0.023	-0.019	0.000
Lee bound: Monthly in-person (upper)	-0.013	0.175	0.163	0.157	0.125	0.144	0.156	0.148
Lee bound: Weekly by phone (lower)	-0.002	-0.116	-0.040	-0.041	-0.033	-0.001	-0.006	0.020
Lee bound: Weekly by phone (upper)	-0.001	-0.003	0.022	0.011	0.030	0.049	0.024	0.035

	(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	**-0.019**	**0.023**	**-0.061**	**-0.029**	**-0.058**	**-0.078**	**0.152**	**0.109**	**-0.015**
	(0.057)	(0.069)	(0.058)	(0.066)	(0.033)*	(0.031)**	(0.028)***	(0.030)***	(0.017)
Weekly by phone	**0.026**	**-0.042**	**0.069**	**-0.447**	**-0.031**	**-0.167**	**-0.228**	**-0.164**	**0.094**
	(0.051)	(0.063)	(0.056)	(0.058)***	(0.030)	(0.029)***	(0.031)***	(0.030)***	(0.022)***
Observations	3987	3984	3973	3987	3986	3986	4056	4056	3987
All treatments equal (p)	0.736	0.620	0.096*	0.000***	0.201	0.000***	0.000***	0.000***	0.000***
MDE: Monthly in-person	0.132	0.168	0.131	0.170	0.090	0.095	0.082	0.083	0.046
MDE: Weekly by phone	0.105	0.134	0.105	0.131	0.069	0.073	0.062	0.063	0.034
Lee bound: Monthly in-person (lower)	-0.081	-0.042	-0.126	-0.333	-0.125	-0.131	0.038	0.002	-0.031
Lee bound: Monthly in-person (upper)	0.254	0.359	0.213	0.202	0.155	0.160	0.226	0.200	0.052
Lee bound: Weekly by phone (lower)	0.015	-0.073	0.044	-0.621	-0.063	-0.209	-0.303	-0.229	0.069
Lee bound: Weekly by phone (upper)	0.131	0.064	0.171	-0.378	0.011	-0.026	-0.222	-0.144	0.095

Coefficients are from regressions of each outcome on a vector of data collection group indicators, randomisation stratum fixed effects, and survey week fixed effects. Continuous outcomes are standardised to have mean zero and standard deviation one in the monthly in-person group and winsorised at the 95th percentile. 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.

Figure 2: **Frequency and Medium Effects on Outcome Distributions in Repeated Interviews**

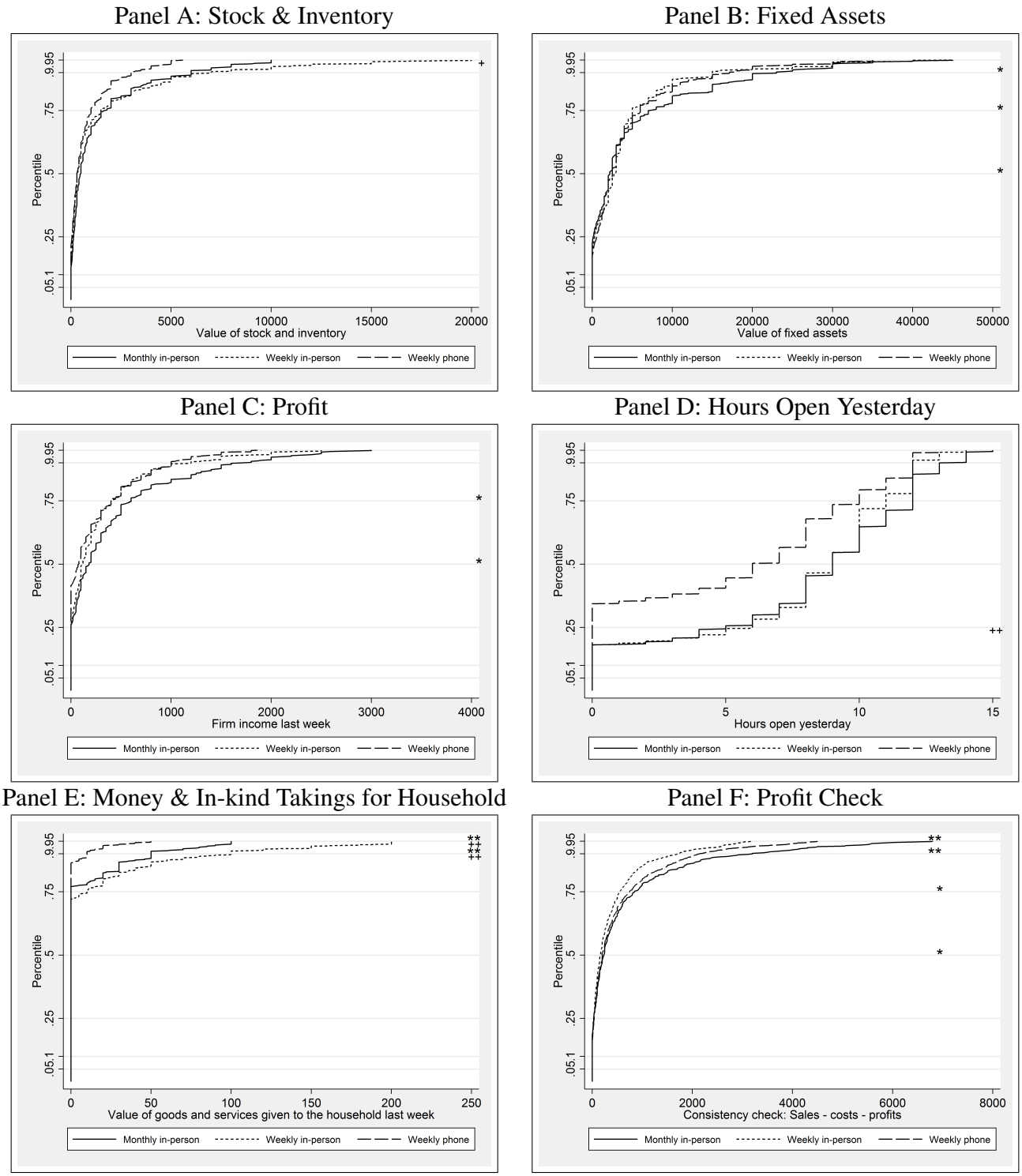


Figure shows empirical CDFs of *outcomes for which there are significant differences across groups at any prespecified quantile*. Empirical CDFs for all other outcomes – sales last week, sales in the last 4 weeks, total costs, money kept by respondent, number of employees, full-time employees and paid employees – are shown in Appendix Figure A1. We use quantile regression to test for differences at each of the quantile shown on the y-axis. We cluster by enterprise (Parente and Silva, 2016) and use the false discovery rate (Benjamini et al., 2006) to control for multiple testing across quantiles. $^{+}$ indicates a medium effect: rejection of the null hypothesis that the coefficients for weekly in-person and phone interviews are equal. * indicates a frequency effect: 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.

Table 2: **Frequency and Medium Effects on Share of Outliers in Repeated Interviews**

	(1) Operating	(2) Stock & inventory	(3) Fixed assets	(4) Profit	(5) Sales last week	(6) Sales last 4 weeks	(7) Total costs
Monthly in-person	-	-0.046**	-0.005	0.009	0.009	0.026	0.013
		(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Weekly by phone	-	-0.044***	-0.022	-0.014	-0.018	-0.005	0.010
		(0.015)	(0.018)	(0.014)	(0.016)	(0.016)	(0.015)
Observations	-	3989	3987	3986	3985	3987	3987
All groups equal (p)	-	0.879	0.429	0.218	0.157	0.124	0.878
	(8) Profit check	(9) Employees	(10) Full-time	(11) Paid	(12) Hours yesterday	(13) Money kept	(14) Household takings
Monthly in-person	0.029	-0.011	-0.012	-0.017	0.017	-0.006	-0.039***
	(0.018)	(0.017)	(0.011)	(0.015)	(0.018)	(0.014)	(0.014)
Weekly by phone	0.019	-0.012	-0.000	-0.002	-0.008	-0.003	-0.063***
	(0.013)	(0.015)	(0.013)	(0.015)	(0.012)	(0.013)	(0.012)
Observations	3984	3987	3984	3973	3987	3986	3986
All groups equal (p)	0.614	0.934	0.295	0.320	0.140	0.776	0.078

Coefficients are from regressing an indicator for being in the top ventile of the distribution on treatment indicators, randomization stratum fixed effects, and survey week fixed effects. Bootstrap standard errors from 1000 iterations are shown in parentheses, resampling by enterprise. ***, **, and * denote significance at the 1, 5, and 10% levels.

predicted by enumerator fixed effects. This result is consistent with social desirability bias and lack of verifiability: respondents may report using written records to please the enumerator and phone interviews make this claim less verifiable. This is consistent with work from the US showing more social desirability bias in phone interviews (Holbrook et al., 2003). This outcome is not completely verifiable even for in-person interviews, as respondents can claim to have used written records to prepare before the interview.

There are substantial frequency and medium effects on the means and distributions of two outcomes: stock/inventory and household takings of money/goods from the enterprise. For both outcomes, weekly in-person interviews yield higher winsorised means and more right-tail outliers than monthly in-person interviews or weekly phone interviews. However, only the medium effects are robust to adjustment for multiple testing (Appendix C). The stock/inventory effect is driven by a longer right tail for weekly in-person interviews. The lower stock/inventory value in the phone group might arise if respondents avoid reporting high values when enumerators cannot visually verify the values. The stock/inventory differences are only 0.08-0.09 standard deviations but this is large in value: roughly US$22 or 15% of mean winsorised stock/inventory value. The medium effect on household takings is driven by fewer zero values for weekly in-person interviews. This is

consistent with a social desirability bias explanation: enumerators may directly observe household members taking/receiving money/goods from the enterprises during in-person interviews but not during phone interviews. This is also consistent with respondents in phone interviews understating household takings of goods because they have just reported a lower value of stock/inventory, noted above.

There are large frequency and medium effects on two binary variables: enumerators' assessments of respondents' honesty and carefulness. These may show that respondents are most engaged in low-frequency in-person surveys and least engaged in high-frequency phone surveys. But they may also reflect enumerators' subjective impressions of the data collection methods. Consistent with the latter explanation, we find that enumerator effects, conditional on data collection method, strongly predict these two assessments. Enumerator assessments of the quality of a respondent's answers are also weakly related to the objective data quality measures we discuss in Section 4. Hence we place low weight on these two outcomes.

We explore why frequency and medium effects may differ across types of outcomes by aggregating outcomes into two indices based on two strategies for answering questions (Appendix Table A6). Respondents may give an actual count for rare events or outcomes they can easily count ('episodic enumeration') but estimate for higher-frequency events or higher-valued outcomes (Gibson and Kim, 2007). We construct a *counting index* based on number of total, full-time, and permanent employees and an *estimating index* based on values of stock/inventory, fixed assets, profit, sales, costs, money kept for the owner, household takings, and hours worked. Both indices are inverse-covariance weighted averages of the underlying variables, following Anderson (2008).

We find no frequency or medium effects on the counting index. Previous work finds that reported counting measures may be sensitive to factors like the length of the recall period (Blair and Burton, 1987; Gibson and Kim, 2007). Our non-result for the counting index may occur because neither frequency nor medium changes respondents' willingness and ability to count or because our only three counting measures are low-valued stock measures and require little counting. We find a large medium effect on the estimating index, which is 0.3 standard deviations lower for phone respondents. Half of this difference is due to the hours worked measure, discussed above.

Most of the remaining difference is due to stock/inventory and household takings, also discussed above. To the extent that respondents are estimating responses, their estimates are on average lower in phone-based interviews.

We are able to compare frequency and medium effects over different recall periods. Many survey responses are sensitive to recall periods: shorter recall periods can cause undercounting as they miss infrequent events, can cause overcounting as respondents compress events over a longer time period into the recall period ('telescoping'), or can avoid undercounting as respondents forget fewer events in short recall periods (Beegle et al., 2012; Friedman et al., 2017). Theory does not provide a clear guide to how these factors differ by survey frequency and medium. Most of our flow measures use a one-week recall period. We measure one variable, sales, over both one- and four-week recall periods. We find that the relationship between the one- and four-week sales measures does not substantially differ by medium or frequency and neither frequency nor medium effects on the two sales measures are not significantly different. We report a more detailed analysis in Appendix C. For sales at least, our conclusions are not sensitive to the recall period used.

We test for heterogeneous effects by estimating Equation (1) with interactions between the group indicators and six prespecified baseline measures. We find limited evidence of heterogeneous interview frequency or medium effects on six dimensions: respondent education, score on a digit span recall test, score on a numeracy test, keeping written records at baseline, number of employees at baseline, and gender.[17] 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) are no more or less susceptible to interview frequency or medium effects. There are a few scattered differences – male respondents report holding more stock and taking more money from the enterprise for their own use when interviewed weekly, and there are some scattered differences in affect by medium. However, these differences are generally imprecisely estimated and do not follow a clear pattern. Given the number of dimensions of heterogeneity we test and the generally lower power of subgroup analyses, the heterogeneity we observe may simply reflect sampling variation.

[17]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 Few Frequency or Medium Effects on Objective Data Quality Measures

Comparing outcome means and distributions by frequency and medium, as in Section 3, does not show which survey methods deliver higher quality data. We therefore examine two measures of data quality in this section. First, we compare the distribution of first digits in our data to a benchmark derived from Benford's Law. Second, we compare direct and indirect measures of enterprise profit as a measure of internal consistency between survey answers. We find only small differences by frequency and medium in these two measures.

4.1 Comparing Data to Benford's Law

Benford's Law is a statistical regularity characterising variables in many datasets. Specifically, Benford's Law states that the probability that the first significant digit (FSD) of a value, $j > 0$, is approximately $\log_{10}(1 + j^{-1})$. Data seldom exactly follow the distribution, but statisticians routinely use the distance between the actual distribution of FSDs and the distribution under Benford's Law as a measure of data quality. See Judge and Schechter (2009) and Schündeln (2018) for examples comparing household surveys in developing countries to Benford's distribution.

We use Benford's Law to evaluate each continuous variable in our data in two ways. First, we calculate the difference between the observed FSD distribution in each data collection group and the distribution under Benford's Law. This allows us to rank the 'quality' of data produced by each frequency and medium. Following Cho and Gaines (2007) we estimate the Euclidean distance between the distributions, $d = \sqrt{\sum_{j=1}^{9}\left(e_j - \log_{10}(1 + j^{-1})\right)^2}$ where e_j is the observed share of observations with FSD j and rescale d to have maximum value 1. Second, we test for pairwise equality of the FSD distribution between data collection groups. This allows us to test if differences in data quality are statistically significant between groups. We regress nine indicators for having FSDs 1, 2, . . ., 9 on data collection group indicators using systems estimation, clustering standard errors by firm. We then test if the nine coefficients on each group indicator jointly equal their values implied by Benford's Law. We exclude categorical measures such as the number of employees as Benford's Law does not generally hold for low-valued integer measures.

Our data follow Benford's Law reasonably closely (Table 3, second panel). The 24 d-statistics,

Table 3: **Comparing Each Data Collection Group to Benford's Law**

	(1) Stock & inventory	(2) Fixed assets	(3) Profit	(4) Sales last week
Panel A: Comparison of first digits across groups				
Monthly = weekly (p)	0.57	0.01	0.89	0.19
Weekly = phone (p)	0.28	0.02	0.34	0.62
Monthly = weekly = phone (p)	0.19	0.00	0.29	0.18
Panel B: Distance of first digit distribution from Benford's Law				
Monthly in-person	0.07	0.12	0.07	0.06
Weekly in-person	0.05	0.16	0.06	0.03
Weekly phone	0.05	0.08	0.06	0.03
	(5) Sales last 4 weeks	(6) Total costs	(7) Money kept	(8) Household takings
Panel A: Comparison of first digits across groups				
Monthly = weekly (p)	0.43	0.60	0.51	0.36
Weekly = phone (p)	0.73	0.43	0.74	0.00
Monthly = weekly = phone (p)	0.22	0.59	0.16	0.00
Panel B: Distance of first digit distribution from Benford's Law				
Monthly in-person	0.04	0.05	0.13	0.17
Weekly in-person	0.09	0.04	0.10	0.10
Weekly phone	0.05	0.02	0.11	0.17

This table compares distributions of first significant digits (FSDs). The first three rows report p-values from Wald tests that the distributions of FSDs are equal across data collection groups. These statistics are obtained by regressing indicators for each of the nine possible FSDs on group indicators in a system of equations, clustering standard errors by enterprise, and testing if the nine coefficients on each group indicator are jointly equal across groups. The final three rows report Euclidean distances (rescaled to be $\in [0, 1]$) between the observed FSD distribution for each data collection group and the distribution under Benford's Law, following Cho and Gaines (2007).

one for each continuous variable for each data collection group have interquartile range [0.05,0.10]. To contextualise this range, the statistic is bounded between 0 and 1 by construction and the d-statistics for developing country surveys reviewed by Judge and Schechter (2009) have interquartile range [0.05,0.13].

No single data collection group follows Benford's Law more closely than the others (Table 3, first panel). We find substantial medium effects on fixed assets and household takings and a smaller frequency effect on fixed assets. The results for household takings should be interpreted with caution as this outcome is zero for most observations and only the positive values are used for the test against Benford's Law. There are no significant differences across groups for the other six variables. The monthly in-person group is farthest from Benford's Law for six of the eight variables but is never significantly farther than the weekly in-person group. Taken together, these results show that neither phone nor high-frequency interviewing leads to drops in data quality.

We also use Benford's Law to show that enumerators' assessment of respondents' honesty and carefulness should be treated with caution. We estimate the normalised Euclidean distance between

the observed FSD distribution and the distribution under Benford's Law separately for interviews where the enumerator classified the respondent as honest and not honest. We then test if the FSD distributions differ between interviews classified as honest and not honest. The 'honest' interviews do not generate data whose FSD distribution is closer to Benford's Law. We repeat this exercise for interviews where the enumerator regarded the respondent as careful and not careful. The 'careful' interviews do not generate data whose FSD distribution is closer to Benford's Law. These findings echo Judge and Schetchter's evaluation of enumerators' subjective assessments using Benford's Law. This contributes to our skepticism of these subjective assessments as a data quality measure, first raised in Section 3. See Table A7 for detailed results.

Finally, we use Benford's Law to show that data quality does not decline over the life of the panel. We split the sample into observations from the first and second halves of the panel, test if the FSD distribution differs between the first and second halves, and estimate the deviation of the FSD distribution from Benford's Law in each half of the panel. The FSD distribution in the first half of the panel is not systematically closer to Benford's Law for any of the three data collection groups. This result differs from related work by Schündeln (2018), who finds that data quality in a Ghanaian household survey declines as households are surveyed more often. Schündeln examines even higher-frequency interviews (up to 10 in a single month), so we should be cautious about generalising our result to higher frequencies. See Table A7 for detailed results.

4.2 Consistency Across Multiple Profit Measures

We examine one prespecified measure of reporting consistency within the survey, the difference between two profit measures. We directly elicit profit, sales, and costs, and use these to construct a 'profit check' outcome equal to the absolute value of (sales - costs) - profits.[18] This is not a direct measure of reporting accuracy because we do not observe true profits. However, consistency across two ways of eliciting profits may indicate more accurate reporting, in line with psychometricians' use of consistency across questions to measure construct validity (John and Benet-Martinez, 2014).

[18]The correlation between directly measured profit and sales minus costs is 0.29, similar to most studies reviewed in De Mel et al. (2009). This correlation is highest for the weekly in-person interviews but does not significantly differ by interview frequency or medium.

We find limited evidence of frequency and medium effects on reporting consistency. There are no differences across groups in the profit check means (Table 1) or shares of outliers (Table 2), though the right tail of the distribution is higher in the monthly group (Figure 2). The latter result is consistent with the idea that high-frequency surveys raise data quality by allowing respondents to practice calculating or estimating profit from sales and costs and hence avoid large discrepancies. This result does not persist after the repeated interviews end (see Section 7.4), casting some doubt on the practice hypothesis, although comparisons in the endline survey are also less well powered. There are also no differences across groups in the panel structure measures discussed in Section 5 except a slightly higher within-enterprise standard deviation of profit checks in the weekly phone group (Table A10). We conclude that the weekly in-person interviews deliver slightly more consistent measures of profits and (sales - costs), though the differences by frequency and particularly medium are small.

5 Phone Surveys Yield Higher Within-Enterprise Dispersion through Time

In Sections 3 and 4 above, we pool survey outcomes from different enterprises in the same data collection group to estimate group-specific means, distributions, and measures of quality. Researchers may also be interested in the behaviour of outcomes for the same enterprise through time. In this section, we examine the panel structure of outcomes within enterprises, using four measures of panel structure. We focus mainly on medium effects, as the monthly in-person surveys are too widely spaced to estimate frequency effects on some of these measures. We show that, on two of the four measures, phone surveys yield more dispersed data than in-person surveys. This difference is consistent with higher measurement error in phone surveys or better measurement of transient shocks in the phone surveys. We do show in Section 4.1 that this is not driven by greater fatigue-induced decline in data quality during the panel, as there is little evidence of fatigue-induced decline in data quality in either group.

First, we estimate one-week autocorrelations in outcomes and report these in Table 4. The autocorrelations are broadly consistent with the economic expectation that they should be higher for stock than flow measures: between 0.77 and 0.88 for stock measures such as assets and em-

Table 4: **Panel Structure of Repeated Interview Data**

	(1) Operating	(2) Stock & inventory	(3) Fixed assets	(4) Profit	(5) Sales last week	(6) Sales last 4 weeks	(7) Total costs
Panel A: Autocorrelations							
Weekly in-person	-	0.873	0.829	0.628	0.750	0.764	0.713
	(-)	(0.023)	(0.031)	(0.063)	(0.038)	(0.038)	(0.057)
Weekly by phone	-	0.665	0.861	0.473	0.589	0.737	0.555
	(-)	(0.070)	(0.035)	(0.054)	(0.049)	(0.038)	(0.054)
All groups equal (p)	-	0.004	0.488	0.057	0.008	0.625	0.039
Panel B: Pr(reporting identical value for two weeks)							
Weekly in-person	0.995	0.261	0.666	0.230	0.171	0.149	0.220
	(0.002)	(0.012)	(0.013)	(0.012)	(0.011)	(0.010)	(0.012)
Weekly by phone	0.995	0.215	0.675	0.285	0.179	0.098	0.220
	(0.004)	(0.030)	(0.029)	(0.029)	(0.025)	(0.022)	(0.028)
All groups equal (p)	0.956	0.123	0.747	0.054	0.767	0.020	0.994
Observations	2431	2414	2412	2412	2412	2414	2414
	(8) Profit check	(9) Employees	(10) Full-time	(11) Paid	(12) Hours yesterday	(13) Money kept	(14) Household takings
Panel A: Autocorrelations							
Weekly in-person	0.538	0.850	0.834	0.878	0.526	0.513	0.506
	(0.066)	(0.027)	(0.034)	(0.029)	(0.038)	(0.045)	(0.047)
Weekly by phone	0.513	0.771	0.800	0.865	0.515	0.458	0.293
	(0.046)	(0.031)	(0.042)	(0.024)	(0.037)	(0.051)	(0.069)
All groups equal (p)	0.758	0.050	0.523	0.726	0.840	0.420	0.011
Panel B: Pr(reporting identical value for two weeks)							
Weekly in-person	0.108	0.925	0.948	0.950	0.454	0.376	0.688
	(0.009)	(0.007)	(0.006)	(0.006)	(0.014)	(0.014)	(0.013)
Weekly by phone	0.106	0.837	0.930	0.915	0.384	0.445	0.815
	(0.020)	(0.020)	(0.015)	(0.015)	(0.032)	(0.032)	(0.033)
All groups equal (p)	0.907	0.000	0.211	0.021	0.029	0.033	0.000
Observations	2410	2414	2411	2393	2414	2413	2412

Panel A autocorrelations are correlations between week t and $t-1$ values for each measure, pooling observations across enterprises. Panel A standard errors in parentheses are from 1000 bootstrap iterations, resampling by enterprise. Panel B coefficients are from regressions of an indicator for no change in value between weeks t and $t-1$ on treatment group indicators, stratification block fixed effects & week fixed effects. Panel B standard errors in parentheses are heteroskedasticity-robust and clustered by enterprise. The still operating outcome is omitted from the autocorrelation analysis because the measure has little variation, with mean = 0.98.

ployment counts; and between 0.29 and 0.76 for flow measures such as profit, sales, costs, hours worked, money kept, and household takings.

Autocorrelations are significantly lower in the phone than in-person group for six outcomes: stock and inventory, profit, sales in the last week, costs, total employees, and household takings. These include both stock and flow outcomes and both estimating and counting outcomes. This may reflect higher measurement error in the phone group or, potentially, more anchoring on past answers in the in-person group. We are not aware of any aspect of the survey administration that would induce more anchoring specifically in the in-person group, so suspect the higher intertemporal variation in the phone group reflects slightly higher measurement error.

Second, we report the group- and outcome-specific probabilities that respondents report identical values for two consecutive weeks in Table 4. This provides a test for differential anchoring on past answers by medium. The probability of reporting identical values two weeks in a row is 0.10-0.29 for flow measures such as profit, sales, and costs where we expect frequent changes. The probability is much higher, 0.67-0.93, for stock measures such as assets and employment counts. Stock/inventory behaves more like the flow than stock variables, consistent with the fact that most enterprises are very small retailers that may restock regularly. The probability is 0.38-0.82 for hours worked, money kept, and household takings because these outcomes have many persistent zeros. These results are in line with economic expectations that stock outcomes should be more persistent that flow outcomes.

The probability of reporting the same value two weeks in a row is significantly lower in the phone group for four variables: sales in the last four weeks, total employees, paid employees, and hours worked. The probability is higher in the phone group for three variables: profit, money kept and household takings. The latter difference is driven by the higher share of zeros for household takings in the phone group. On this measure of panel structure, neither medium generates consistently higher or lower dispersion across outcomes.

Third, we calculate the within-enterprise standard deviation through time for each outcome, estimate treatment effects on the standard deviations, and report these in Table A10. This is the only measure of the panel structure we construct for the monthly in-person group. Phone interviews yield higher standard deviations than in-person surveys on five of thirteen outcomes (four of which are robust to adjustment for multiple testing, as shown in Table A12) and lower standard deviations only for household takings, a measure dominated by zero values. In contrast, we do not find large or robust frequency effects on within-enterprise standard deviations. There are frequency effects on the standard deviations of only three of thirteen variables, these do not have consistent signs, and they are not robust to adjustment for multiple testing.

Fourth, we characterize the panel structure of one flow variable – log profit – and one stock variable – log capital stock – following Blundell and Bond (1998). We estimate dynamic panel models separately for weekly phone and weekly in-person groups, assuming an AR(1) structure

on both error terms and using two lags for profit and four lags for capital stock. We fail to reject equality of the full set of parameters across the two groups. The full results are shown in Table A13. This exercise is driven by both the mean and panel structure of the outcomes, so it is possible that the lack of medium effects on the means offset any medium effects on the panel structure.

6 Classifying Outcomes Based on Interview Frequency and Medium Effects

Combining the results from Sections 3 – 5, we can divide outcomes into five categories. First, we see no frequency effects and at most small medium effects for enterprise closure, assets, profit, our profit consistency check, the number of full-time employees, and money taken from the enterprise by the respondent.[19] Second, there are no frequency or medium effects on the mean or distributions of sales over both recall periods, costs, and the numbers of total employees and paid employees, but phone surveys do generate higher within-enterprise dispersion through time. These eleven outcomes are robust to the frequencies and media we evaluate from the perspective of estimating means, distributions, or average or quantile treatment effects. But the second group of variables are more sensitive to medium choices from the perspective of estimating within-enterprise dynamics.

Third, there is a substantial medium effect, robust to multiple test correction, on the use of written records: phone respondent are more likely to self-report using written records to complete the survey. The same respondents report using written records less often when we interview them in person after the panel ends (Section 7.4). This pattern is most consistent with social desirability bias and lower verifiability in the phone interviews. This suggests researchers asking questions subject to social desirability bias and medium-specific verifiability should be cautious about medium choices. Fourth, there is a large medium effect but no frequency effect on hours worked. As discussed in Section 3, this reflects selection from our in-person interviews disproportionately missing respondents who were seldom working at their enterprises. We view this as an advantage of phone surveys relative to in-person surveys, though allowing in-person interviews at flexible locations may also achieve this.

[19]The value of fixed assets is difficult to classify. There are no frequency or medium effects on the mean and only small frequency effects on some quantiles of the distribution. But there are significant differences in digit distributions. The digit distributions for the two in-person groups are quite far from Benford's Law.

Fifth, there are substantial frequency and medium effects on the values of current stock/inventory and money/stock/services given to the household ('household takings') although only the medium effects are robust to multiple test adjustment. Their means, distributions, share of outliers, and within-enterprise dynamics are all sensitive to frequency and medium. Household takings also has a digit distribution that differs substantially from Benford's Law, though stock/inventory does not. It is unclear why these specific variables are the most sensitive. Both are estimating, rather than counting, measures but other estimating measures are less sensitive. Stock/inventory is a stock variable, with high intertemporal persistence, while household takings is a flow variable with most mass at zero and low intertemporal persistence otherwise. It is possible that household takings is subject to the same social desirability bias and differential verifiability as using written records. But we do not observe information about the presence of household members during the interview that might allow us to test this explanation.

7 Other Considerations when Choosing Survey Frequency and Medium

In this section, we discuss four remaining considerations for researchers choosing survey frequency and medium. First, we show that outcome autocorrelations are higher for very closely-spaced surveys — hence the precision gains from averaging multiple survey rounds are decreasing in survey frequency. Second, we document that phone surveys are substantially cheaper than in-person surveys. Third, we show that permanent attrition from the panel does not differ by survey frequency or medium, but that non-response in any given survey round is higher at higher frequencies. Fourth, we show that data collection at different frequencies or using a different medium does not have persistent effects on microenterprises or their owners after the panel has ended. We survey everyone in person at endline, holding the location and survey medium constant and randomly re-assigning enumerators to treatment groups, and find few differences in outcomes between treatment groups.

7.1 Precision Gains from Multiple Measures Are Lower at Higher Frequency

Researchers sometimes collect multiple measures of enterprise performance through time to improve precision by averaging out both transient real shocks and transient measurement error (McKen-

zie, 2012). We show in this section that there are substantial precision gains from averaging high-frequency measures, especially flow measures, but that the precision gains are larger when measures are spaced farther apart.

We focus on two outcomes of particular interest to microenterprise researchers: profit and the value of fixed assets. Both have substantial intertemporal variation: the within-enterprise coefficients of variation through time have interquartile ranges of [0.72,1.41] for standardised profit and [0.17,0.82] for standardised assets. This variation may reflect transient real shocks of interest to researchers or transient measurement error. Our experiment is not designed to separate these explanations but our data quality checks in Section 4 suggest this is not entirely measurement error.

In Appendix Table A9, we show large precision gains from repeated measures, particularly for flow outcomes such as profit.[20] Measuring profit and fixed assets two weeks in a row reduces outcome variance by respectively 22 and 8%, while measuring them four weeks in a row reduces outcome variance by respectively 33 and 12%.

These precision gains are slightly larger with longer gaps between measures. Measuring profit and fixed assets two months in a row, rather than two weeks in a row, reduces outcome variance by respectively 27 and 11%, instead of 22 and 8%. Measuring them four months in a row rather than four week in a row reduces outcome variance by respectively 40 and 16%, instead of 33 and 12%. Precision gains are larger with longer gaps because the 4-week autocorrelations are lower than the 1-week autocorrelations for most flow and stock measures. With a fixed budget, researchers gain more power by averaging over three measures a month apart than over three measures a week apart. This is stronger evidence of non-stationarity than in the enterprise datasets reviewed in McKenzie (2012). We may find more evidence of non-stationarity because we evaluate higher-frequency panel data relative to most of the literature.

7.2 Phone Surveys Reduce Costs

Phone interviews reduce our per-interview costs by approximately 25% and larger cost savings should be possible in other settings. We calculate costs by analysing the survey firm's general

[20] These calculations use 1- and 4-week autocorrelations from pooling the weekly in-person and phone groups, shown in columns (7) and (8) of Appendix Table A9.

ledger entries, which break expenditure down by date and purpose. We exclude the costs of the screening, baseline, and endline interviews (conducted in person for all respondents); fixed costs (e.g. office costs and management salaries); and equipment costs. Each completed phone interview cost US$4.76 while each completed in-person interview cost US$7.30 in the monthly group and US$6.12 in the weekly group.[21] 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.

Each completed phone interview, relative to a completed interview in the weekly in-person group, saved US$1.94 on enumerator transport and US$0.91 on enumerator salaries but cost US$1.21 more in airtime. The remaining cost differences are due to data capture and respondent incentives, which depend entirely on medium-specific response rates. See Figure A3 for detailed breakdown. Our cost savings are relatively low because we worked in a dense urban area with low transport costs and high airtime costs (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.

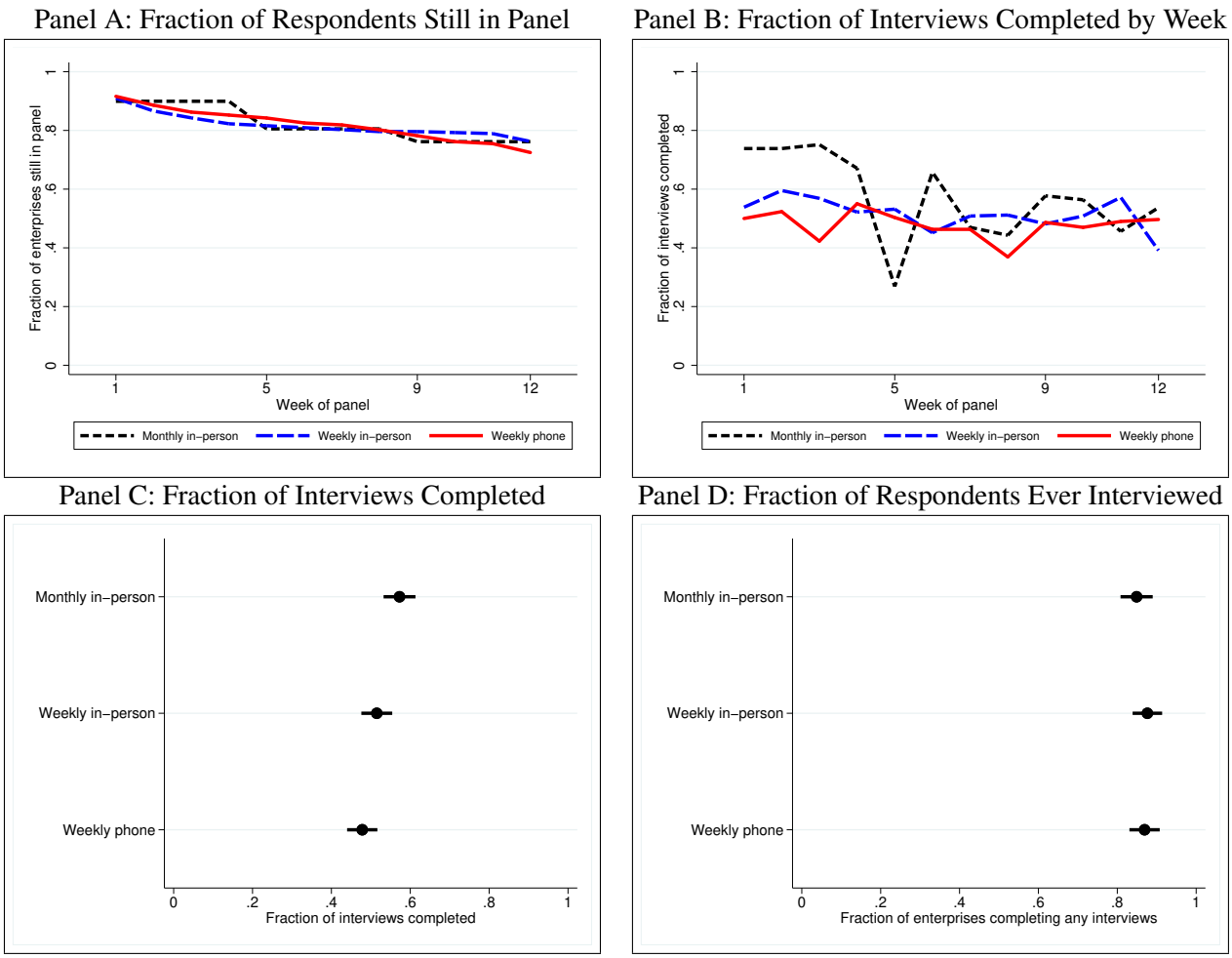
7.3 High-Frequency Measures Risk Higher Non-Response but Not Higher Attrition

Both interview frequency and medium may in principle change respondents' participation in interviews. We briefly outline differences in participation in this section and report more detailed results in Appendix H. We distinguish between two types of non-participation: *permanent attrition* and *non-response*. We define a respondent as a permanent attriter from round $t+1$ if she is interviewed in round t but not in any round $s \geq t$, including the endline interview. Roughly 20% of respondents attrit by week 12 of the panel. This rate does not differ by frequency or medium (Figure 3, Panel A).

We define non-response as missing an interview in a specific round. Non-response is fairly high: we completed 4070 of 8058 scheduled repeated interviews (51%). There are no medium effects on non-response or on the probability of completing any interviews (Figure 3, Panels C and D). There

[21] 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).

Figure 3: **Response Rates and Attrition by Data Collection Group**



Panel A 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. Panel B shows 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 due to the staggered start dates. Panel C shows the fraction of interviews completed by each respondent, separately by treatment group. The p-values for testing equality of this measure across groups are 0.045 for the monthly and weekly in-person groups and 0.187 for the weekly in-person and phone groups. Panel D shows the fraction of respondents that complete at least one interview, separately by treatment group. The p-values for testing equality of this measure across groups are 0.334 for the monthly and weekly in-person groups and 0.794 for the weekly in-person and phone groups.

is also little difference in the panel structure of responses: the autocorrelations in non-response in the weekly groups are -0.017 and 0.039 for, respectively, the in-person and phone groups (p-value of difference = 0.078), after conditioning on respondent-specific response rates. There is a substantial frequency effect on non-response. Respondents complete 6 percentage points more

interviews in the monthly in-person group than the weekly in-person group (Figure 3, Panel C). This difference occurs only in the first four weeks of repeated interviews (Figure 3, Panel A). This timing, and the fact that permanent attrition does not differ by frequency, shows that the frequency effect on non-response is not driven by survey fatigue or exhaustion.

Although respondents miss more weekly interviews, weekly interviews are more likely to find all respondents at least once in a given period. The fraction of respondents that are interviewed at least once in each x-week period is higher in both weekly groups than in the monthly group for all values of x (Table A15).[22] This presents a trade-off: weekly interviews deliver a higher volume of information, but this information may be less representative in some weeks. If the non-response in any one period is close to random, then the greater volume of information will more than offset the lower response rate in each week. We show in Appendix H that differences in non-response by frequency are weakly related to baseline characteristics and that the marginal respondents who are captured only by higher-frequency surveys are not systematically different to the inframarginal respondents who are captured by high and low-frequency surveys.

Non-response in our weekly panel is comparable to other high-frequency surveys with representative samples (e.g. Croke et al. 2014; Gallup 2012). However, non-response in our weekly panel is higher than in surveys of samples with revealed willingness to persist in panel surveys (e.g. Arthi et al. 2018; Beaman et al. 2014; Heath et al. 2017). This reflects a potential trade-off between high response rates in the panel and gathering a representative sample, though lower-frequency panel surveys are able to achieve both goals (e.g. Thomas et al. 2012). See Appendix H for more detail on these benchmarks.

7.4 Frequency and Medium Effects on Means Do Not Persist

We test if interview frequency or medium during the 12-week panel has persistent effects on data collected several weeks after the panel has ended. Persistent effects may occur for two reasons: survey methods may persistently change how respondents report real outcomes or may actually change real outcomes, potentially by changing behaviour through reminder or salience effects. We

[22] The coverage rate for monthly interviews is mechanically lower lower for $x < 4$. But the lower coverage rate in the monthly group over longer time periods is not mechanical and is informative.

expect that frequency effects are more likely than medium effects but we test for both. This issue is particularly important for researchers using high-frequency panels for a subsample of a broader survey sample who want to preserve comparability between the subsample and the full sample (e.g. Franklin 2017).

We conduct an endline survey one to four weeks after the panel ends, surveying respondents from all three groups in person and randomly re-assigning enumerators to respondents. Any differences in outcomes measured at this stage must reflect persistent effects of prior interview methods. We regress respondent-level outcomes on indicators for the monthly in-person and weekly phone groups, conditional on stratification block fixed effects and using heteroskedasticity-robust standard errors. We plot the estimates in Figure 1 and show the detailed results in Table A20.

We find few frequency or medium effects. We are powered to detect moderate differences: the median MDEs for binary and continuous outcomes are respectively 11 percentage points and 0.11 standard deviations. Monthly in-person and weekly phone respondents both report fewer employees than weekly in-person respondents, and weekly phone respondents report very slightly lower household takings and using written records less often. These differences are no longer statistically significant after adjusting for multiple testing (Table A22). These findings are not consistent with the most obvious prediction of a persistence model: frequency and medium effects during the panel should also be visible in the endline. Only the household takings result has the same sign during the panel and the magnitude in the panel is much higher. More generally, the estimated mean effects during the panel and in the endline are not similar. We have 34 mean estimates in total: phone and monthly estimates for each of 17 outcomes. The correlation between mean effects during the panel and in the endline across the 34 estimates is 0.004.

The largest persistent frequency and medium effects are on reported employment. This is driven by the share of respondents reporting zero versus one employees (Figure A4). This is a puzzling finding. It is unlikely that different survey methods induce large enough behavioural changes to shift real employment. It is possible that prior interaction with enumerators changes respondents' understanding of the definition of employee, inducing a persistent change in how they respond to this question. But the employment differences are driven by full-time and paid employees, which

are easier to define than part-time or unpaid employees. Given that these effects are not statistically significant after adjusting for multiple testing, they may simply reflect noise.

How do we reconcile these results with prior research, discussed in Section 1, which shows that participation in panel interviews can change respondents' behaviour, even over relatively short panels? A likely explanation is that behaviour change has been documented particularly in domains that are not already salient to respondents or where the surveys provide information about previously unknown options: small change management for enterprise owners (Beaman et al., 2014), savings and borrowing (Crossley et al., 2017; Stango and Zinman, 2014), water chlorination (Zwane et al., 2011), or participation in active labour market programmes (Bach and Eckman, 2018). When outcomes are already salient, such as whether a respondent has a job, being surveyed more frequently does not change reporting (Bach and Eckman, 2018; Franklin, 2017).

8 Conclusion

This paper reports the first randomised controlled trial to compare microenterprise data from surveys of different frequency and medium. To study a representative sample of microenterprises in Soweto, South Africa, randomly assigning enterprises to be interviewed in person each month, in person each week, or by phone each week.

We find three main results. First, we find few effects of frequency or medium on the means or distributions of reported outcomes. In particular, we find no substantial differences for enterprise closure, profit, sales, costs, fixed assets, or employment. We do find substantial medium effects on stock/inventory, money/goods/services given to the household, hours worked, and self-reported use of written records. Second, we use Benford's Law to show that data quality does not differ systematically between survey frequencies and media. Third, we find that phone interviews generate higher within-enterprise dispersion through time for some flow and some stock measures.

We conclude that using phone or high-frequency surveys does not systematically raise or lower the quality of microenterprise data used for cross-sectional or static panels models. However, researchers particularly interested in within-enterprise dynamics should exercise caution when choosing survey medium. These results can help researchers choose the interview frequency and

medium that generate the optimal quality and volume of data given budget constraints. In particular, our findings suggest researchers can use phone surveys to reduce costs and high-frequency surveys to collect richer panel data that captures transient shocks and informs models of intertemporal optimisation without substantially reducing data quality.

References

ABBRING, J. AND J. HECKMAN (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choices, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics Volume 6B*, ed. by J. Heckman and E. Leamer, Elsevier, 5145–5303.

ABEBE, G., S. CARIA, M. FAFCHAMPS, P. FALCO, S. FRANKLIN, AND S. QUINN (2016): "Curse of Anonymity or Tyranny of Distance? The Impacts of Job-Search Support in Urban Ethiopia," *NBER Working Paper No. 22409*.

ANDERSON, M. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Re-evaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

ARTHI, V., K. BEEGLE, J. DE WEERDT, AND A. PALACIOS-LOPEZ (2018): "Not Your Average Job: Measuring Farm Labor in Tanzania," *Journal of Development Economics*, 130, 160–172.

BACH, R. AND S. ECKMAN (2018): "Participating in a Panel Survey Changes Respondents' Labour Market Behaviour," *Journal of the Royal Statistical Society Series A*, forthcoming.

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*, 1–2.

BEAMAN, L. AND A. DILLON (2012): "Do Household Definitions Matter in Survey Design? Results from a Randomized Survey Experiment in Mali," *Journal of Development Economics*, 98, 124–135.

BEAMAN, L., J. MAGRUDER, AND J. ROBINSON (2014): "Minding Small Change: Limited Attention among Small Firms in Kenya," *Journal of Development Economics*, 108, 69–86.

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.

BLAIR, E. AND S. BURTON (1987): "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions," *Journal of Consumer Research*, 14, 280–288.

BLUNDELL, R. AND S. BOND (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

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.

CARRANZA, E., R. GARLICK, K. ORKIN, AND N. RANKIN (2018): "Job Search and Hiring with Two-sided Limited Information about Workseekers' Skills," Working paper.

CHO, W. AND B. GAINES (2007): "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance," *The American Statistician*, 61, 1–6.

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.

CROSSLEY, T., J. DE BRESSER, L. DELANEY, AND J. WINTER (2017): "Can Survey Participation Alter Household Saving Behaviour?" *Economic Journal*, 127, 2332–2357.

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, 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 MEL, S., D. MCKENZIE, AND C. WOODRUFF (2009): "Measuring Microenterprise Profits: Must We Ask How the Sausage is Made?" *Journal of Development Economics*, 88, 19–31.

DE NICOLA, F. AND X. GINÉ (2014): "How Accurate are Recall Data? Evidence from Coastal India," *Journal of Development Economics*, 106, 52–65.

DILLON, A., E. BARDASI, K. BEEGLE, AND P. SERNEELS (2012): "Explaining Variation in Child Labor Statistics," *Journal of Development Economics*, 98, 136–147.

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.

DUPAS, P., J. ROBINSON, AND S. SAAVEDRA (2018): "The Daily Grind: Cash Needs and Labor Supply," Working paper.

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. (2017): "Location, Search Costs and Youth Unemployment: Experimental Evidence from Transport Subsidies," *Economic Journal*, 128, 2353–2379.

FRIEDMAN, J., K. BEEGLE, J. DE WEERDT, AND J. GIBSON (2017): "Decomposing Response Errors in Food Consumption Measurement: Implications for Survey Design from a Randomized Survey Experiment in Tanzania," *Food Policy*, 72, 94–111.

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*.

GARLICK, R. (2019): "The Effects of Nationwide Tuition Fee Elimination on Enrollment and Attainment," Working paper.

GIBSON, J. AND B. KIM (2007): "Measurement Error in Recall Surveys and the Relationship Between Household Size and Food Demand," *American Journal of Agricultural Economics*, 89, 473–489.

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 (2017): "Measuring Employment: Experimental Evidence from Ghana," 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.

JACOBSON, L., R. LALONDE, AND D. SULLIVAN (1993): "Earnings Losses of Displaced Workers," *American Economic Review*, 83, 685–709.

JOHN, O. AND V. BENET-MARTINEZ (2014): "Measurement," in *Handbook of Research Methods in Social and Personality Psychology*, ed. by H. Reis and C. Judd, Cambridge University Press, 473–503.

JUDGE, G. AND L. SCHECHTER (2009): "Detecting Problems in Survey Data Using Benford's Law," *Journal of Human Resources*, 44, 1–24.

KARLAN, D., R. KNIGHT, AND C. UDRY (2012): "Hoping to Win, Expected to Lose: Theory and Lessons on Micro Enterprise Development," *National Bureau of Economic Research Working Papers*, 1–54.

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.

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.

MAHADEVAN, M. (2018): "The Price of Power: Costs of Political Corruption in Indian Electricity," Working paper.

MCKENZIE, D. (2012): "Beyond Baseline and Follow-up: The Case for More T in Experiments," *Journal of Development Economics*, 99, 210–221.

——— (2015): "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.

——— (2017): "Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition," *American Economic Review*, 107, 2278–2307.

MCKENZIE, D. AND C. WOODRUFF (2008): "Experimental Evidence on Returns to Capital and Access to Finance in Mexico," *World Bank Economic Review*, 22, 457–82.

MITULLAH, W. AND P. KAMA (2013): *The Partnership of Free Speech and Good Governance in Africa*, vol. 3, Cape Town: Afrobarometer, University of Cape Town.

PAPE, U. (2018): "Informing Rapid Emergency Response by Phone Surveys," Http://blogs.worldbank.org/developmenttalk/informing-rapid-emergency-response-phone-surveys.

PARENTE, P. M. AND J. M. S. SILVA (2016): "Quantile Regression with Clustered Data," *Journal of Econometric Methods*, 5, 1–15.

ROBINS, J. (1997): "Causal Inference from Complex Longitudinal Data," in *Latent Variable Modeling and Applications to Causality*, ed. by M. Berkane, Springer-Verlag, 69–117.

ROSENZWEIG, M. AND K. WOLPIN (1993): "Credit Market Constraints, Consumption Smoothing and the Accumulation of Durable Production Assets in Low-Income Countries: Investments in Bullocks in India," *Journal of Political Economy,*, 101, 223–244.

SCHÜNDELN, M. (2018): "Multiple Visits and Data Quality in Household Surveys," *Oxford Bulletin of Economics and Statistics*, 80, 380–405.

SCOTT, C. AND B. AMENUVEGBE (1991): "Recall Loss and Recall Duration: An Experimental Study in Ghana," *Inter-Stat*, 4, 31–55.

SINGER, E. AND C. YE (2013): "The Use and Effects of Incentives in Surveys," *Annals of The American Academy of Political and Social Science*, 645, 112–141.

STANGO, V. AND J. ZINMAN (2014): "Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees," *Review of Financial Studies*, 27.

STECKLOV, G., A. WEINREB, AND C. CARLETTO (2017): "Can Incentives Improve Survey Data Quality in Developing Countries? Results from a Field Experiment in India," *Journal of the Royal Statistical Society: Series A (Statistics and Society)*.

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.