

Testing the Effectiveness of Mobile Phone Data Collection for Microenterprises in Africa: **PRE-ANALYSIS PLAN**

Rob Garlick, Kate Orkin and Simon Quinn

April 4, 2014

1 Introduction

We are conducting a Randomised Controlled Trial to test the effectiveness of mobile phone data collection for microenterprises. This is funded under a PEDL Exploratory Grant. We will test whether mobile phones are a viable method for collecting accurate high-frequency data on microenterprises. We will further explore the structure of the high-frequency time-series data.

This document summarises the main estimations that we intend to run. We do not rule out the possibility of running other estimations. However, we will make clear in the paper which estimations are specified in this pre-analysis plan and which are not (Casey, Glennerster, and Miguel, 2012).

2 Experimental design

The trial is running in Soweto, the main township in the Greater Johannesburg area. We use a random sample of 900 microenterprises in selected census enumeration areas. We focus exclusively on enterprises that (i) have zero, one or two full-time employees (in addition to the entrepreneur), (ii) whose owner has a mobile phone, (iii) do not provide professional services, and (iv) operates at least three days a week, every week.

We will use a control method and two treatment methods. Our **control method** involves face-to-face interviewing at monthly intervals. (In this document, ‘monthly’ is used as shorthand for ‘every four weeks’.) This is intended to mimic a standard method of collecting data from microenterprises, to provide a benchmark against which alternative methods can be compared. Our **first treatment method** involves face-to-face interviewing at weekly intervals. This preserves one key feature of the benchmark method — that respondents are interviewed in person — but allows us to test the consequences of collecting data at a much higher frequency. Our **second treatment method** involves mobile phone interviews at weekly intervals. By trifurcating our design in

this way, we are able to test directly (i) whether high-frequency phone interviews are equivalent to more traditional methods (mobile phone versus monthly face-to-face interviews); (ii) whether high-frequency collection changes data patterns (weekly face-to-face versus monthly face-to-face interviews) and (iii) whether face-to-face collection changes data patterns (weekly face-to-face versus weekly mobile phone interviews).

We have assigned microenterprises to treatments using a method adapted from [Bruhn and McKenzie \(2009\)](#).¹ Specifically, we created strata based on (i) gender, (ii) number of employees, (iii) enterprise sector and (iv) ‘subplace’ (a geographical unit larger than an Enumeration Area). This yielded 149 strata with 1-51 businesses each. We then split each stratum randomly between the three treatment groups. This generated some residual firms in each stratum (due to strata sizes not all being multiples of three); these firms were then randomly assigned to treatment groups, with the restriction that a pair of residuals in a stratum would always go into separate groups.

Each respondent is receiving a gift of 27 rand each month for participating in the study. This is paid in all cases in the form of a mobile phone airtime voucher.

3 Effect on responses throughout the survey

We will begin by testing effects on reported responses throughout the survey — that is, for the questionnaires *after* the baseline and *before* the endline (not including the baseline and endline). For respondents in the first and second treatment groups, this refers to the 12 surveys conducted at weekly intervals. For respondents in the control group, this refers to the three surveys conducted at monthly intervals.

In this section, we use t to index the week of interview. For firms in the first and second treatment groups, this is $t \in \{1, \dots, 12\}$. For a firm in the control group, this refers *either* to $t \in \{1, 5, 9\}$, *or* to $t \in \{2, 6, 10\}$, *or* to $t \in \{3, 7, 11\}$, *or* to $t \in \{4, 8, 12\}$ (depending on the random survey timing for that firm). We index firms by i and outcome variables by k ; y_{kit} therefore refers to the response of firm i to outcome k in week t . We use T_{1i} and T_{2i} respectively as dummies for being assigned to treatment method 1 and treatment method 2.

For each outcome y_{kit} , we will normalise first by the mean and standard deviation of the control group (*i.e.* those interviewed face-to-face at monthly intervals). That is, define \tilde{y}_{kit} as the raw data for firm i for outcome k in week t . Let $\hat{\mu}_k$ be the sample average of \tilde{y}_{kit} for all control firms (that is, firms interviewed at weekly intervals); let $\hat{\sigma}_k$ be the sample standard deviation for the same. Then we will use:

$$y_{kit} \equiv \frac{\tilde{y}_{kit} - \hat{\mu}_k}{\hat{\sigma}_k}. \quad (1)$$

¹ Specifically, we implement the recommendations of Bruhn and McKenzie in their June 2011 blog post, “[Tools of the trade: Doing stratified randomization with uneven numbers in some strata](#)”.

3.1 Outcomes

We are primarily interested in testing the effect of different measurement techniques on the following outcomes (where question numbers refer to the repeated questionnaire):

- (i) Whether the respondent reports that (s)he is **still operating the business** (question 2);
- (ii) The **total stock and inventory** held by the business (question 3);
- (iii) The **total fixed assets** held by the business (question 4);
- (iv) The **total income (profit)** of the business last week (question 5);
- (v) The **total number of employees** of the business (question 6), the number of **full-time employees** (question 7) and the total number of **paid employees** (question 8);
- (vi) The **total sales** of the business last week (question 9) and **total sales** of the business in the past four weeks (question 14);
- (vii) The **total costs** of the business last week (where total costs are calculated as the sum total of cost subcategories for the previous week, namely (a) purchase of stock/inventory (question 10(a)), (b) wages/salaries (question 10(b)), (c) rent/rates for the property where the business is based (question 10(c)), (d) repayments on business loans (question 10(d)), (e) equipment purchases (question 10(e)), (f) fixing/maintaining equipment (question 10(f)), (g) transport costs for the business (question 10(g)), (h) telephone and internet costs for the business (question 10(h)), and (i) all other business expenses (question 10(i)));
- (viii) The value of **money kept for yourself** from the business last week (question 11) and value of goods and services **consumed by the household** last week (question 12);
- (ix) Number of **hours the business was open yesterday** (question 16);
- (x) We will construct several measures of **consistency in reporting**. First, we will construct the **absolute value of (sales minus costs, minus profits)**. In principle, this should be zero for all firms; we can therefore interpret this as a measure of accounting error and/or misunderstanding of these concepts (as they are commonly applied in microenterprise surveys). Second, accuracy of recalling the **date of last airtime payment** (question 17). We will calculate this in terms of *absolute number of days in error from the true date, recorded through our independent administrative data*. Third, for participants in the weekly treatments, we will construct a measure of **the absolute value of (sales in week t reported in week t , minus sales in week t reported in week $t + 1$)**.
- (xi) We will use two measures of **enumerator perceptions**: whether the enumerator believes that the respondent answered honestly (question 21) and whether the enumerator believes that the respondent answered carefully (question 22). Each question is recorded on a five-point Likert scale; for each question separately, we will code a dummy variable for (i) whether the response is at or above the sample median for the question, or (ii) whether the response is below that sample median.

- (xii) We will use a dummy variable for **whether the respondent reports having referred to written financial records or notes in answering the questions** (question 15).

3.2 Testing average levels

We will estimate the effect of interview technology on average levels reported by estimating:

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

where η_g are dummy variables for the matched randomisation blocks, and ϕ_t are dummy variables for the survey week. We will cluster errors by firm. We will test whether either treatment induces different average reporting. We will do this by testing $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$.

3.3 Testing dispersion

For each outcome k , we will also test for difference in dispersion. We will calculate the sample standard deviation for each outcome as follows:

$$s_{ik} = \sqrt{\frac{\sum_{t=1}^{N_{ik}} (y_{kit} - \bar{y}_{ki})^2}{N_i - 1}}, \quad (3)$$

where N_{ik} is the number of non-missing observations for outcome k and firm i and \bar{y}_{ki} is the average for outcome k and firm i . We will then estimate:

$$s_{ik} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \mu_{ki}, \quad (4)$$

using heteroscedasticity robust standard errors. We will then test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$.

3.4 Quantile regressions

We will then test the effect of interview method on the quantiles of y_{kit} . For each outcome k , we will re-estimate equation 2 as a quantile regression. We will estimate for quantiles 0.05, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.95, reporting significance tests for each level. We will do this using simultaneous-quantile regression, and we will also jointly test for coefficient equality across all quantiles.

3.5 Trimming

We will then test the robustness of our conclusions to trimming. Specifically, for each outcome k separately, we will:

- (i) Generate $T_{kit} = 1$ if y_{kit} is either (i) at or below the 5th percentile of the pooled distribution of y_{kit} or (ii) at or above the 95th percentile. We will set $T_{kit} = 0$ otherwise.

- (ii) We will drop all observations for which $T_{kit} = 1$, and rerun all of our estimations on equations (2) and (4) (including recalculating s_{ik} using equation (3), where N_{ik} will count the number of non-missing observations for outcome k and firm i *after* dropping observations with $T_{kit} = 1$.)

3.6 Attrition

We will then test whether the probability of attrition differs by treatment type. We will define a dummy variable a_{it} for whether firm i is missing in week t . (For firms in the control group, we will code a_{it} as missing if firm i was not scheduled to be interviewed in week t .) We will then estimate:

$$a_{kit} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \phi_t + \varepsilon_{kit}, \quad (5)$$

where we will cluster errors by firm. We will then test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$. (We will then explore further, by estimating equation 5 separately for each week, to test when any attrition occurred. We anticipate showing these results graphically, displaying point estimates and confidence intervals. We anticipate showing graphs in the style of Figures 6 and 7 of [Croke, Dabalén, Demombynes, Giugale, and Hoogeveen \(2012\)](#), disaggregated by treatment group.)

If we find significant differences in rates of attrition from estimating equation 5 (pooled across all time periods), we will use [Lee \(2009\)](#) bounds to test the robustness of our earlier conclusions in estimating equations 2 and 4.

3.7 Variation across subgroups

We will repeat this entire exercise while interacting with (i) gender, (ii) education, (iii) digit recall, (iv) financial literacy, (v) numeracy, (vi) whether the firm commonly uses a phone for business (at baseline), (vii) whether the firm reported keeping financial records (at baseline), and (viii) the number of employees (at baseline), to test whether parameters vary across those subgroups. For education, digit recall, financial literacy, numeracy and the number of employees, we will interact with dummy variables, coded such that ‘0’ will refer to ‘*at or below the median value for the pooled sample*’, and ‘1’ will refer to ‘*above the median value for the pooled sample*’.

4 Effect on responses at endline

The previous section described tests for whether responses differ during the repeated surveys, not including baseline and endline. We will conduct a common endline with all firms, face to face. We will test whether endline responses differ between treatment groups. We will interpret this as a test of whether our interview techniques changed actual firm outcomes (for example, by prompting firm managers to pay more attention to their capital stock). For this section, we retain the same variable definitions as earlier.

For each outcome y_{kit} , we will estimate:

$$y_{ki} = \beta_1 \cdot T_{1i} + \beta_2 \cdot T_{2i} + \eta_g + \varepsilon_{ki}, \quad (6)$$

where we will allow for heteroscedasticity robust standard errors. As before, we will then test $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_1 = \beta_2$, $H_0^4 : \beta_1 = \beta_2 = 0$.

For each outcome k , we will use a bootstrap method to test whether the variance of measured y_{ki1} varies between treatments.

5 Multiple testing

We are testing multiple separate outcomes. We therefore intend to report corrected p -values, to allow for this. We intend to report these corrections for:

- (i) Our estimations on equation 2,
- (ii) Our estimations on equation 4, and
- (iii) Our estimations on equation 6.

(Depending on results, we may also report corrected p -values for the trimmed sample, and for the interaction specifications. We will do this if we obtain significant results for those estimations that we interpret as evidence of significant differences between treatment.)

For these corrections, we will report three types of p -values:

- (i) Regular p -values;
- (ii) The ‘sharpened q ’ of [Benjamini, Krieger, and Yekutieli \(2006\)](#); and
- (iii) The Family Wise Error Rate, which we will calculate using the Stepdown Bootstrap of [Westfall and Young \(1993\)](#) (see also [Anderson \(2008\)](#) and [Casey, Glennerster, and Miguel \(2012\)](#)).

These methods require placing outcomes into families. We intend to use the following families (where question numbers again refer to the repeated questionnaire):

- (i) **Firm capital** (questions 3 and 4);
- (ii) **Firm outputs** (questions 5, 9 and 14);
- (iii) **Firm inputs** (questions 6, 7, 8 and 10);
- (iv) **‘Personal’ inputs and outputs** (questions 11, 12 and 16);
- (v) **Recall quality** (the constructed measures described in item (x) in section 3.1);
- (vi) **Enumerator perceptions and reliance on financial records** (questions 21 and 22, coded as we describe in item (xi) in section 3.1, and question 15).

6 Predictors of firm closure

We will use the weekly data to consider three questions regarding correlates of firm closure. We define a firm as closed if the owner is no longer operating the business and has (i) closed the business or (ii) sold the business (questions 2 and 24 from repeated questionnaire). We will pool data from the weekly phone and weekly in-person interviews if we conclude that outcomes listed in section 3 do not systematically differ across treatment groups.

First, we will test whether data from week $t - 1$ are better predictors of closure than week $t - 4$. This captures the idea that closure may be induced by short-term shocks that are not captured by data collected at conventional (i.e. monthly) frequency. Denote \mathbf{x}_{it} as a vector containing all variables listed in section 3.1, items (ii) to (ix) (i.e. questions 2 – 12 and 16). Denote c_i as the week that firm i closes — such that, if firm i is operating in week t but not in week $t + 1$, $c_i = t + 1$. If firm i does not close during the sample, we will denote $c_i = \infty$ (but this is without loss of generality; we could choose any value greater than 12).

Then (following [Cameron and Trivedi \(2005, p.602\)](#)), we estimate the following model of firm closure:

$$\Pr(c_i = t | c_i \geq t) = \Lambda(\gamma_t + \boldsymbol{\beta} \cdot \mathbf{x}_{i,t-1}), \quad (7)$$

where Λ represents the logistic link function. γ_t represent dummy variables for survey weeks (i.e. weeks 1 to 12 of data collection). We will interpret the estimates of $\boldsymbol{\beta}$ in descriptive terms, rather than causal terms. We will repeat the estimation replacing $\mathbf{x}_{i,t-1}$ with $\mathbf{x}_{i,t-4}$. We will compare these two models – one with $\mathbf{x}_{i,t-1}$ and one with $\mathbf{x}_{i,t-4}$ – using a standard Vuong test. If the former model dominates the latter, we will interpret this evidence that more recent data better predict firm closure. (For comparability, we will limit *both* estimations to the subsample of respondents for whom we observe $\mathbf{x}_{i,t-1}$ and $\mathbf{x}_{i,t-4}$.)

Second, we will test whether data from week $t - 2$ significantly predict closure conditional on data from week $t - 1$. This tests the hypothesis that business closure is a first-order Markov process against the null hypothesis the process is a second-order Markov process (which will hold if, for example, firms fully adjust to shocks within a single period). We will implement this test by estimating:

$$\Pr(c_i = t | c_i \geq t) = \Lambda(\gamma_t + \boldsymbol{\beta} \cdot \mathbf{x}_{i,t-1} + \boldsymbol{\delta} \cdot \mathbf{x}_{i,t-2}), \quad (8)$$

and testing the hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}$ (using a cluster-robust Wald test). We will again limit both estimations to the subsample of respondents for whom we observe $\mathbf{x}_{i,t-1}$ and $\mathbf{x}_{i,t-2}$.

Third, we will describe the precursors to business closure, in terms of both profits and sales. Closure should follow falling profits and/or sales if businesses face liquidity constraints or if owners are learning about their suitability for self-employment. Closure should follow rising profits and/or sales if owners view self-employment as a temporary state they will exit once they have

saved enough to finance migration, further education, costly search for formal employment, *etc.* For sales, we will describe these paths as follows:

- (i) We will calculate mean sales for each survey week for those firms that do not close during the period of our panel.
- (ii) For those firms that *do* close during our sample, we will subtract the mean sales for the relevant survey week for firms not closing.
- (iii) For firms closing, we will then re-index sales, counting as ‘sales in the week immediately before closure’, ‘sales in the week two weeks before firm closure’, and so on. We will graph this sales trend, with confidence intervals, for firms having closed (where, by construction, a level of zero shows the average normalised experience of firms not having closed).

We will then repeat the exercise by analogy for profits.

7 Characterising the time-series properties of the data

One of the key advantages of using mobile phones for data collection is that it provides a cost-effective method for obtaining firm data at a high frequency (*i.e.* weekly). We plan to exploit this by running a set of estimations designed to characterise the time-series properties of our outcome variables. Specifically, we intend estimating the autocorrelation structure of log profits and log capital.

For this part of the analysis, we will use only the data collected at weekly intervals — that is, data from respondents subject to what we earlier termed the ‘first treatment’ and the ‘second treatment’. For each of the estimations that follows, we will:

- (i) Estimate using just the ‘first treatment’ sample;
- (ii) Estimate using just the ‘second treatment’ sample; and
- (iii) Estimate using a sample that pools both treatments (if earlier tests of differences in measurement suggest that this is appropriate).

We anticipate using the following $AR(L)$ model:

$$y_{it} = \sum_{l=1}^L \rho_l \cdot y_{i,t-l} + \mu_i + \delta_t + \varepsilon_{it}. \quad (9)$$

We intend to use the [Arellano and Bond \(1991\)](#) estimator and the [Blundell and Bond \(1998b\)](#) estimators, after conducting exploratory analysis of unit root structures following the guidance in [Blundell and Bond \(1998a\)](#) and [Bond, Nauges, and Windmeijer \(2005\)](#). These estimations will inherently be more exploratory/descriptive than the experimental estimations described earlier. For this reason, we do not specify in detail in advance how we anticipate proceeding with this part of

the analysis. In our view, this approach strikes the appropriate balance in this experimental context between (i) being flexible/reasonable in selection of an appropriate linear panel model and (ii) ‘tying our hands’ against the possibility of data mining.

References

- ANDERSON, M. L. (2008): “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects,” *Journal of the American Statistical Association*, 103(484), 1481–1495.
- ARELLANO, M., AND S. BOND (1991): “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations,” *The Review of Economic Studies*, 58(2), 277–297.
- BENJAMINI, Y., A. M. KRIEGER, AND D. YEKUTIELI (2006): “Adaptive Linear Step-up Procedures That Control the False Discovery Rate,” *Biometrika*, 93(3), 491–507.
- BLUNDELL, R., AND S. BOND (1998a): “GMM Estimation with Persistent Panel Data: An Application to Production Functions,” *IFS Working Paper Series*, (W99/4).
- BLUNDELL, R., AND S. BOND (1998b): “Initial conditions and moment restrictions in dynamic panel data models,” *Journal of Econometrics*, 87(1), 115–143.
- BOND, S., C. NAUGES, AND F. WINDMEIJER (2005): “Unit Roots: Identification and Testing in Micro Panels,” *cemmap Working Paper*, (CWP07/05).
- BRUHN, M., AND D. MCKENZIE (2009): “In pursuit of balance: Randomization in practice in development field experiments,” *American Economic Journal: Applied Economics*, pp. 200–232.
- CAMERON, A. C., AND P. K. TRIVEDI (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press.
- CASEY, K., R. GLENNERSTER, AND E. MIGUEL (2012): “Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan,” *The Quarterly Journal of Economics*, 127(4), 1755–1812.
- CROKE, K., A. DABALEN, G. DEMOMBYNES, M. GIUGALE, AND J. HOOGEVEEN (2012): “Collecting High Frequency Panel Data in Africa Using Mobile Phone Interviews,” *World Bank Policy Research Working Paper*, (6097).
- LEE, D. S. (2009): “Training, wages, and sample selection: Estimating sharp bounds on treatment effects,” *The Review of Economic Studies*, 76(3), 1071–1102.
- WESTFALL, P. H., AND S. S. YOUNG (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*.