OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 80, 2 (2018) 0305–9049
doi: 10.1111/obes.12196

Multiple Visits and Data Quality in Household Surveys*

Matthias Schündeln

Goethe University Frankfurt, 60629 Frankfurt am Main, Germany,
(e-mail: schuendeln@wiwi.uni-frankfurt.de)

Abstract

In order to increase data quality some household surveys visit the respondent households several times to estimate one measure of consumption. For example, in Ghanaian Living Standards Measurement surveys, households are visited up to 10 times over a period of 1 month. I find strong evidence for conditioning effects as a result of this approach: In the Ghanaian data the estimated level of consumption is a function of the number of prior visits, with consumption being highest in the earlier survey visits. Telescoping (perceiving events as being more recent than they are) or seasonality (first-of-the-month effects) cannot explain the observed pattern. To study whether earlier or later survey visits are of higher quality, I employ a strategy based on Benford's law. Results suggest that the consumption data from earlier survey visits are of higher quality than data from later visits. The findings have implications for the value of additional visits in household surveys, and also shed light on possible measurement problems in high-frequency panels. They add to a recent literature on measurement errors in consumption surveys (Beegle *et al.*, 2012, Gibson *et al.*, 2015), and complement findings by Zwane *et al.* (2011) regarding the effect of surveys on subsequent behaviour.

I. Introduction

Consumption data are of central importance to the study of a large range of empirical questions that are of interest to both academics and policy makers. They are frequently used, for example, to study levels and distributions of welfare, poverty, or vulnerability, and their determinants. Collecting accurate data on consumption and expenditure, however, is not straightforward, and approaches differ, sometimes substantially, across countries and over time. Specifically, Beegle *et al.* (2012, p. 4) list four 'primary dimensions [in which] the main methods of consumption data collection' vary: 'diary vs. recall, the level of aggregation or detail in the commodity list, the reference period and the level of respondent'. Therefore, a number of papers have recently used experimental methods to assess the role

JEL Classification numbers: C81, O12, I32, D12.
*I thank Ghana Statistical Service for sharing the data. I gratefully acknowledge very helpful discussions with Luc Christiaensen and valuable inputs on an earlier version of this paper from Ahmed Ragab.

of different aspects of survey design for measuring consumption and possible biases that result from non-classical measurement error in consumption measures (Beegle *et al.*, 2012, Caeyers, Chalmers and De Weerdt, 2012, Gibson *et al.*, 2015, Friedman *et al.*, 2016). The present paper adds to this literature in two dimensions. First, it considers in depth one aspect of survey design that seems to be largely unexplored as of yet, namely the number of interviews that is used to calculate a single measure of consumption. Second, the paper suggests a novel approach to deal with the 'fundamental problem in assessing survey bias' (Meyer, Mok and Sullivan, 2015, p. 200), namely the lack of a benchmark measure of the true outcome. To deal with this, the paper proposes Benford's law as an analytical tool to investigate the quality of consumption data over time and applies it to the data at hand.

More specifically, the paper studies nationally representative surveys conducted in Ghana (the Ghana Living Standards Surveys 3, 4 and 5), in which households are interviewed repeatedly – responding to the *same* consumption module up to 10 times over a period of 1 month – with the purpose of obtaining *one* precise measure of consumption by combining data from a number of interviews. I find that the consumption measures obtained from any individual interview drop significantly over the interview period. Thus, the frequency of visits affects measured annualized consumption, and as a consequence, will also affect estimates of poverty and inequality, among others. The implications are economically significant: for example, as shown in section III, a measure related to food poverty in rural Ghana increases by about 13 percentage points if data from all visits are used instead of only using data from the first visit.[1]

Thus, the question arises: which consumption data should be considered the one that best reflects the true consumption of interviewed households? Theoretically, a number of arguments, which are reviewed below, would predict that data quality is higher in earlier interview visits. However, there are also arguments that suggest that later visits provide higher quality data. Finally, a combination of arguments could imply that the data quality is (inversely) U-shaped. To answer this question, ideally, a measure of the true outcome would be available. One possible approach is to compare survey results with administrative data (e.g. Meyer *et al.*, 2015). In the absence of administrative data, as is common in many developing country contexts, Beegle *et al.* (2012) and Gibson *et al.* (2015) use a 'personal diary' with daily visits by a local assistant and visits every two days by a survey enumerator to establish a benchmark.[2] This paper introduces a new approach to study the question of how data quality changes over time, based on Benford's law.

Benford's law describes a statistical regularity for the frequency distribution for first digits of numerical data.[3] The following analysis is based on the hypothesis that first significant digits of true (precisely measured) data will be more likely to conform to the distribution predicted by Benford. On the other hand, false – or, more generally, low quality – data will

[1] Variation in the number of interviews used across surveys will therefore also affect comparability of poverty measures over time and place (Lanjouw and Lanjouw, 2001).

[2] The paper by Beegle *et al.* (2012) also features two treatments that allow for a comparison of frequent (daily) and infrequent (a total of three) visits over a period of 14 days. But that paper's interest is in the total difference, and does not ask the question whether earlier or later visits provide more accurate data. Moreover, in both cases diaries are used, as opposed to interviews with recall.

[3] Benford's law also has implications for the distribution of second- and higher-order digits. This paper will only make use of first significant digits.

conform less to the predicted distribution. Statistical explanations of this regularity exist (e.g., Hill, 1995, see below) and Benford's law has been applied to a number of questions.[4] Indeed, a paper by Judge and Schechter (2009) uses Benford's law to study a related question, namely to identify false data in household surveys. Existing papers typically look at one data point (e.g. a measure of the distance between two distributions of first digits based on one variable, derived from one survey question), or a handful of data points (e.g. one data point for each of a small number of enumerators that were involved in a survey). But it cannot be guaranteed that any given set of data indeed conforms to the law, even if measured without error. Therefore, a large distance between the measured distribution and the expected distribution of first digits cannot necessarily be interpreted as evidence of poor data, and vice versa. Still, most of the literature relies on the differences between the expected (Benford's) distribution of first digits and the actual distribution. One of the innovations of the present application of Benford's law is that (i) I do not rely on this absolute difference being large or small, but rather observe responses to the same questionnaire items repeatedly over time, and will investigate changes in the distribution of first digits over time and (ii) I repeatedly observe a large number of data points, namely the – more than 200 – frequently consumed items that are recorded in the consumption modules of the Ghanaian surveys. Consequently, I can use regression analysis to systematically assess determinants of differences between Benford's distribution and the observed distribution of first digits. Most importantly, I can keep the average difference between Benford's predicted distribution of first digits and the observed distribution constant. That is, through the use of fixed effects I can account for the fact that data will differ in how much it conforms to Benford's law, even if they are of high quality. Thus, the central question is then whether this distance increases or decreases over the course of the – up to 10 – interviews. An increase in the distance between the distribution expected according to Benford's law and distribution observed in the actual data is then interpreted as a worsening of data quality.

The main results of the analysis based on Benford's law suggest that data quality is the highest for the first interviews and falls monotonically over the course of the interviews. The results are confirmed for three different surveys from Ghana (GLSS 3, 4 and 5).[5] However, the approach to collecting data through frequent visits or frequent diaries is not specific to Ghana, but used in many other nationally representative household surveys. One extreme example comes from Rwanda, where the 2010–11 Integrated Household Living Conditions Survey records consumption and expenditure in 10 different visits. Another example comes from Bangladesh, where – in the Household Income and Expenditure Survey of 2010 (HIES) – repeated consumption-related interviews happen every second day for 14 days, that is, in seven visits total. Indeed, in recent work Engle-Stone, Sununtnasuk and Fiedler (2017) found that food consumption in HIES decreases over time, with a drop in total energy consumed (in kcal) from the first to the last visit that is about 5%, thus mirroring stylized facts that motivate the analysis of the present paper.[6] One may be concerned that

[4] Examples of questions analysed with Benford's law range from analyses of tax returns (Nigrini, 1996) to brain electrical activity (Kreuzer *et al.*, 2014).

[5] The most recent Ghana Living Standards Survey (GLSS 6), carried out in 2012–13, also recorded consumption data during six visits, at 5-day intervals, but is not investigated here.

[6] The absolute decline in the HIES survey is smaller than the one found in the present data from Ghana. Note that the urban data collection in Ghana covers a period that is twice as long as in Bangladesh, namely about a month. Another

the high-frequency of a survey is particularly influential in surveys that are collecting consumption data based on interviews and less relevant for diary-based surveys. Note, however, that for illiterate households diary-based surveys become *de facto* interview-based surveys, often with a higher frequency than for the literate households. Several surveys state this explicitly, that is, they use mainly diaries, but visit illiterate households more frequently, to interview them in person. In Tanzania, the 2011–12 Household Budget Survey covers consumption and expenditure during 28 days. According to available survey documents, illiterate households are visited daily during this period, i.e. 28 times, while literate households receive diaries and are visited every 2–3 days. In Sierra Leone, the 2011 Integrated Household Survey covers 30 days of expenditure and consumption, and illiterate households are visited daily, while literate households receive diaries and are visited 6 times by enumerators to transfer records from their diaries.[7] Further, even for surveys that are run solely as high-frequency diaries, it may be that for unmotivated respondents a diary survey effectively becomes a high-frequency recall survey (Beegle *et al.*, 2012). This suggests that surveys in developed country that are based on repeated diaries can also suffer from similar measurement problems, e.g. the Consumer Expenditure Survey (CEX), by US Bureau of Labor Statistics, or the Canadian Food Expenditure Survey. Finally, even if consumption data are collected in only one visit, multiple visits – sometimes also with a high frequency – also occur when collecting panel data, and similar issues may arise in that situation.[8]

Most immediately, the findings of this paper are relevant for a better understanding of the precision of the measurement of consumption, poverty and inequality based on household surveys. Beyond the concern about the mere imprecision of consumption measures, a further concern is about biases in econometric analyses as a result of non-classical measurement error (Gibson *et al.*, 2015). Further, the findings are also of interest for practitioners because of the budgetary implications of a large number of repeated visits to the same household.[9] Finally, the findings also complement recent work by Zwane *et al.* (2011). While the present paper suggests that more frequent surveying affects subsequent measurement of the items surveyed, Zwane *et al.* (2011) find that surveying affects actual subsequent behaviour, and more frequent surveying affects later behaviour stronger – in their case a higher frequency of surveys leads to lower levels of child diarrhoea and cleaner water.

In the following chapters, I first introduce the Ghanaian data and show that consumption measures decrease over the course of the interviews. The next section spells out some

possibly quite important difference between these surveys is that HIES only asks food consumption repeatedly (133 items), while GLSS data also contain repeated measures of non-food consumption, and overall a longer repeatedly asked consumption module (208 items).

[7] Further examples of countries that currently run household surveys that use frequent visits to collect consumption and/or expenditure data either through interviews or diaries or a combination, starting with the most recent ones, include: Turkey (most recently 2015), South Africa (most recently 2014–15), Tajikistan (2013), Seychelles (2013), Iraq (2012), Albania (2012), Tanzania (2011–12), West Bank and Gaza (2011–12), Chad (2011), Zimbabwe (2011), Cambodia (2010), Tuvalu (2010), Burkina Faso (2009–10), Syria (2008–09), Niger (2007–08), Cameroon (2007), Kenya (2005–06), and Yemen (2005–06).

[8] This could in principle apply to any panel data set. An extreme example is the ongoing Thai Survey by Robert Townsend, which has interviewed some household since 1998, initially with weekly recording of consumption, later with interviews every 2 weeks.

[9] For example, in United Nations (2005) a sample budget is presented, in which the cost of one additional round of household data collection is estimated to be around US$ 200,000.

hypotheses about why consumption measures might change over time. Then, Benford's law is introduced and applied to the data. The last section concludes.

II. Data

The data used for this paper are from different rounds of the Ghana Living Standards Survey (GLSS). Most data used are from the third round of this survey, the GLSS 3, which was conducted between September 1991 and September 1992. Robustness checks are performed using data from GLSS 4 (1998/9) and GLSS 5 (2005). GLSS are nationally representative surveys that are comparable to the typical Living Standard Measurement Surveys. These surveys also usually contain detailed consumption and expenditure modules, which is most relevant for our purposes. As is common in household surveys of this type, total consumption in GLSS is measured as the sum of consumption of own-produced goods and the expenditure for purchased goods.[10] For simplicity, I will henceforth mostly refer only to 'consumption' instead of consumption and expenditure. This paper exploits the fact that the consumption modules in GLSS 3–5 require that households are visited up to 11 times within a relatively short time period of 2 weeks to 1 month (in GLSS 3, 11 visits occurred in urban areas and 8 times in rural areas).[11] On the first day, households were only interviewed about basic demographics, while in each of the following interviews, the households were repeatedly asked to report their consumption using the same consumption module covering a large set of frequently consumed items. Interview 2 focused exclusively on the consumption module, while in the following interviews other modules were covered in addition to the consumption module.[12] The Data User's Guide for the GLSS 3 describes the purpose of these frequent visits as follows: 'To reduce the recall error in GLSS3, much more detailed information was collected by means of frequent visits to each household. Households were visited eight times at two-day intervals in rural areas, and 11 times at three-day intervals in urban areas. By reducing the recall period from two weeks to two or three days, much improved estimates of household consumption and expenditure should be obtained' (Ghana Statistical Service, 1995, p. 3).

For an easier reading of the remainder of the paper, which focuses exclusively on the consumption-related interviews, I will refer to the very first interview, in which no consumption-related data are collected, as the 'base interview', and I will refer to all the consumption-related interviews as 'visits'. So the 'first visit' will be the first consumption-related interview, which is the second interview overall.

It is worth noting that one particular problem in collecting consumption data is related to the boundedness of recall periods. If consumption questions are asked in the first (or only) interview, the absence of a bound might lead to telescoping (see below). But here the consumption questions are first asked during the second interview, which one could

[10] More specifically, the questionnaire asks (in section 8h) 'How much of home produced...was consumed by the household since my last visit?' and (in sections 9A2 and 9B) 'How much was spent on...since my last visit?'.

[11] Note that all data used is observational, that is, variation in the frequency of visits is due to variation between urban and rural areas and across surveys, but not due to random or quasi-random assignment.

[12] Non-consumption related modules (e.g. covering health or employment) were spread out over the remaining visits, so as to keep the total interview time during each visit roughly constant.

argue implies a bounded recall, namely bounded by the first interview. Also, as a preview, note that all results reported below are qualitatively similar if the first visit is omitted.

In GLSS 3, two variations between rural and urban areas need to be considered, namely with respect to recall period and with respect to the use of diaries: (i) Interviews use different recall periods in rural and urban areas, with 2-day recall in rural areas and 3 days in urban areas. The underlying reason for these differences in recall periods is (ii) that a diary was used to collect consumption data in urban areas, while in rural areas interviews were used. Here the argument was that, because of high illiteracy levels, the use of a diary was not always possible in rural areas. Because of these differences, most of the analyses in this paper are performed separately for urban and for rural areas.

The consumption module for GLSS 3 is very detailed. Each household was asked to report consumption levels in each one of 208 categories during each of the visits in which the consumption module was covered (i.e. in all but the base interview). For the present paper, most of the analysis is performed at the level of each visit and each of the 208 items separately, resulting in up to 2,080 visit-item observations. The analysis below is performed based on the value of an item. For own-produced consumption goods, I obtain a value by multiplying the number of units by the value of a unit, as reported in the questionnaire. The GLSS 3 sample consists of 4,521 households, of which 3,040 are identified as rural and 1,481 urban. Attrition is low. For example, in rural areas, only nine out of 3,040 households (0.3%) have no recorded consumption for any of the last two survey visits.

III. Consumption measures across survey visits

Figures 1 and 2 show the central motivation for this study. Figure 1 analyses each visit separately and shows how estimated mean daily consumption, estimated based on data from only one visit, varies over the course of the visits. The main message of this figure is that the reported level of consumption is highest in the first visit and is significantly lower in later rounds. The largest drop is between the first and the second visit, but the average consumption that is reported in each visit continues to drop over time (with one exception, namely visit 5 in urban areas). A drop from the first to the second visit could generally be due to the well-known phenomenon of telescoping, which will be discussed in more detail below. For example, for their analysis of the GLSS data, Coulombe and McKay (2008) omit the first consumption visit, based on the argument that recall is not bounded. However, as alluded to above, a reference point that bounds the recall period is available, namely, the first overall interview, in which households were not asked about consumption.[13] Further, even if we believe that telescoping explains the difference between the first and the second visit, telescoping cannot explain the drop in later visits.

Next, I show how the differences in estimated consumption for individual visits translate into estimates that a researcher who aggregates data for many visits would obtain. To this end, I first calculate mean daily consumption (over all households) based on only one (the first) visit. I then aggregate data for the first and the second consumption-related visit and again calculate mean daily consumption. This is the number that a researcher would work

[13] The questions explicitly make use of this bound: in the second interview, when consumption is first covered, the question is 'How much was spent on ... since my first visit?'. In later visits, the question is '... since my last visit?'.

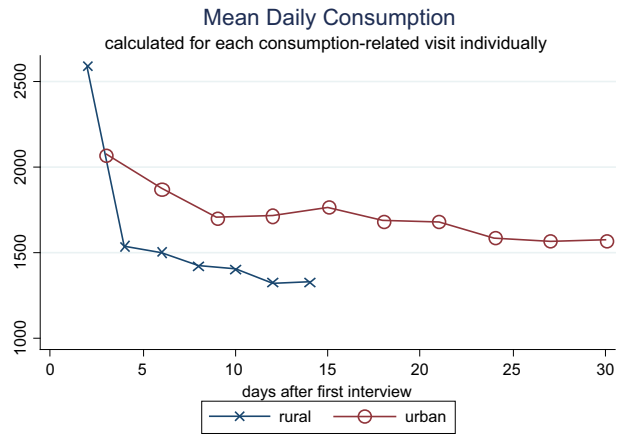

Figure 1. Mean daily consumption for each individual consumption-related visit, based on GLSS 3. Values are in contemporaneous Ghanaian Cedis

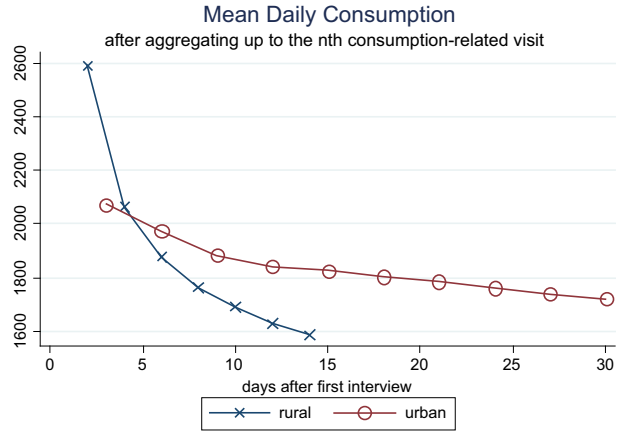

Figure 2. Mean daily consumption after aggregating up to the n^{th} consumption-related visit, based on GLSS 3. Values are in contemporaneous Ghanaian Cedis

with if she conducted two interviews to collect data on consumption. I continue up to the seventh consumption-related visit (in rural areas) and the 10^{th} visit (in urban areas). The figure plots the results versus the days elapsed after the base interview (which does not cover the consumption module). The last data point aggregates all available consumption data, and is the data that might be used by a researcher who conducts seven consumption-related interviews in rural areas and 10 in urban areas, and uses all available data to estimate consumption. The results in Figure 2 clearly show that estimates of consumption monotonically decrease in the number of visits used to calculate mean consumption. This drop is larger for rural areas, but clearly exists and is economically significant in both rural and urban areas.

The results of the above aggregate analysis are confirmed in a regression framework at the household level (results not shown), which shows that the drop in consumption is statistically significant. In addition, the effects shown above are also economically signi-

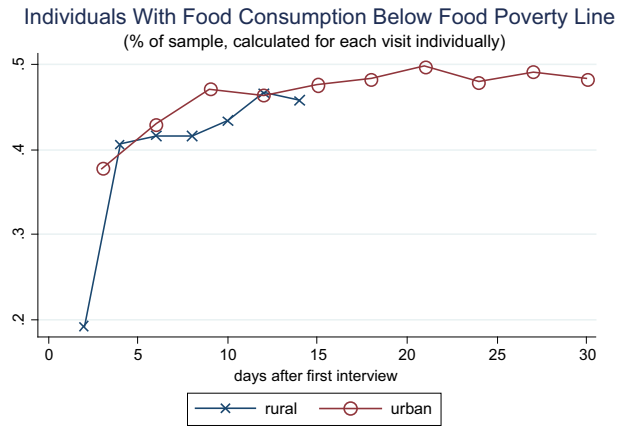

Figure 3. Individuals with food consumption below the food poverty line, separately for each visit, based on GLSS 3. For details about the calculation, see the text

ficant: the average consumption in rural areas drops from the first to the last visit by more than 40%. The corresponding number for urban areas is about 25%.

To further illustrate the economic significance of the effect of the number of visits, I also calculate a measure that is related to food poverty. Specifically, I calculate whether annualized per capita adult-equivalent food consumption per household falls below the food poverty line that is reported by Ghana Statistical Service (2000).[14] Figure 3 reports the share of individuals that fall below this food-poverty line, calculated separately for each visit. It is important to note that this is not equal to food poverty (which would take other consumption into account), but is used to illustrate that variation in food consumption occurs in a region of the consumption distribution that matters for poverty estimates.[15] According to this measure, if only data from the first visit are used, 19% of individuals in rural areas have food consumption that falls below the level of food consumption that is deemed necessary to cover basic needs. When only data for the last visit are used, this number is more than twice as high, namely, 46%. When data from all seven visits are used (results not shown in the figure), the resulting measure is 32%. When the first visit is omitted and data from the second to the last (seventh) visit are used, rural food consumption falls below the food poverty line for 38% of individuals in the sample. Overall, the results suggest that the changes in the consumption estimates occur in parts of the distribution that also affect poverty numbers, with a particular large effect in rural areas.

[14] Ghana Statistical Service (2000) provides a food poverty line, which is derived by calculating the food expenditure required to provide the minimum calorie requirements (based on the consumption basket of the bottom 50% of the distribution of total consumption). I inflate the data from 1991 to be in line with the poverty line, which is in 1999 Ghanaian Cedis, also taking into account regional price differences.

[15] I am not aiming at providing a comprehensive analysis of poverty in Ghana, but the goal is to illustrate the importance of the parts of the survey which are captured through repeated visits. Thus, the focus is on food consumption and the food poverty line; neither food poverty nor a more comprehensive measure of poverty are calculated, which would include other aspects of consumption, many of which are only collected in one visit. For simplicity of this illustration I also do not use sampling weights. For comparability across visits, I only use households that report positive consumption during all visits. Finally, I omit 117 observations (about 2.5% of all households) that also report receiving an in-kind wage in the form of food, because this part of food consumption cannot be attributed to consumption during a specific visit.
For an in-depth analysis of poverty in Ghana see, for example, Ghana Statistical Service (2000).

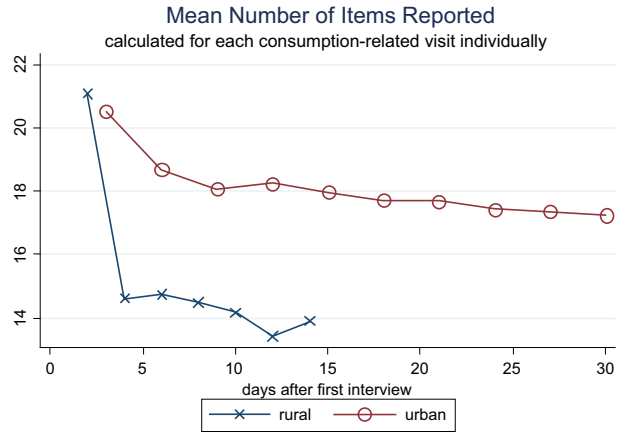

Figure 4. Mean number of items reported in each individual consumption-related visit, based on GLSS 3

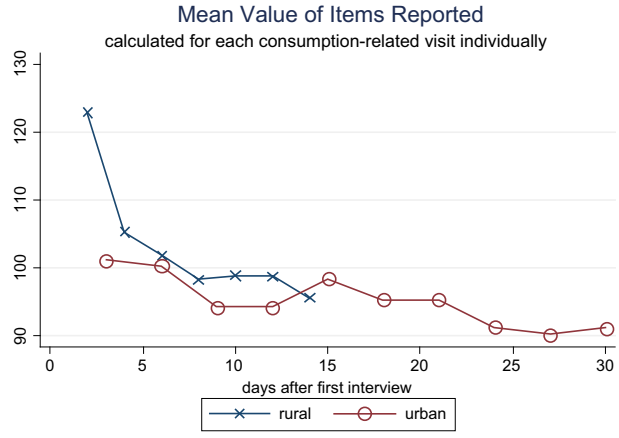

Figure 5. Mean value of items in each individual consumption-related visit, based on GLSS 3. For comparability items are reported per day, that is, dividing by 3 in rural areas and dividing by 2 in urban areas. Values are in contemporaneous Ghanaian Cedis

Mechanically, the drop in consumption over time could be due to a drop in the quantity of items reported, or to a drop in the average value reported for each item (conditional on reporting non-zero consumption for an item), or both. Figures 4 and 5 show that both effects are at work. First, the average urban household records consumption in roughly 20 of the 208 categories in the first consumption-related interview, but this number drops to <18 in the last visit. The drop is even more pronounced in rural areas, from about 21 in the first to approximately 14 items in the last visit. Second, when considering the average value of items with non-zero consumption, the mean also goes down, although in particular in urban areas the pattern is less clean over time.[16] The observed patterns are not unique to GLSS 3. Qualitatively similar results are found in robustness checks using GLSS 4 and GLSS 5.

[16] Because rural and urban areas have different recall periods, the mean value is reported as the value per day, in order to make the data comparable across rural and urban areas.

IV. Why might measures of consumption decrease over time?

A simple, innocuous explanation for the above-reported stylized facts might be first-of-the-week/month and other seasonality-related effects (e.g. Hastings and Washington, 2010). However, the survey starts interviewing households throughout the year from September 1991 to September 1992 and starting-weekdays are also fairly uniformly distributed. And indeed, through regression analyses I confirm that the results hold even after controlling for day-of-interview fixed effects (results not shown).[17] Therefore, seasonality can be ruled out as an explanation and the drop in estimated consumption over time must be related to survey design and not due to actual changes of consumption over those visits.

For a discussion of possible sources of error, Neter (1970) provides a useful categorization: He mentions the following broad categories of errors: recall errors, telescoping, reporting load effects, prestige errors, conditioning effects, respondent effects, interviewer effects and reporting instrument effects.

Several of these error sources are plausibly at work in any given survey, including GLSS 3. But, considered in isolation, these cannot explain the changes in survey errors over time, because the relevant characteristics of the survey do not change. These are, in particular: (i) Recall errors, because the recall period remains constant over the visits. Similarly, (ii) reporting load effects cannot explain the decrease in quality, because the consumption modules stay constant and (iii) reporting instrument effects, because the instruments and interview techniques (i.e. diary, interview) remain the same over time.

On the other hand, the following could potentially explain changes over time: (i) Prestige errors: For example, initially a respondent might not want to appear poor. However, over time, as she gets to know the interviewer, she is less concerned about this. (ii) Respondent effects: For example, the interviewed person might not know all the details about her household's consumption. As she gets to know the survey setup, she might be better prepared to respond to the questions in later interviews. (iii) Interviewer effects: The interviewer might also change her behaviour over time, as she gets to know the respondent better. For example, she could condition questions on previous responses, leaving out items that the household did not consume in the last visit.

Many of the hypotheses spelled out above could be summarized by what Neter calls conditioning effects. At a general level, '[panel] conditioning arises if having been interviewed previously causes differences in knowledge, behaviour, or attitude, affecting the answers in later interviews' (Toepoel, Das and van Soest, 2009, p. 73). This could, for example, be due to raised consciousness, an improved understanding of questions and questionnaire set-up, or a change in motivation to participate in the survey. As indicated above, these could plausibly lead to both overreporting or underreporting of consumption items over time.

Finally, telescoping – incorrectly reporting consumption that does in fact not belong into the recall period – could explain especially the large drop from the first to the second visit. However, two arguments weaken the effect of telescoping. First, telescoping is typically

[17] More precisely, for each household I add up all the reported consumption items for each visit (considering relatively frequently reported items only, using the cutoff of 50 as in the main text). The data contain the precise day at which each visit was conducted, and I regress the sum of reported consumption on household fixed effects, visit dummies, and day-of-interview dummies (about 360).

assumed to be particularly strong for unbounded recall periods. For this reason, Deaton and Grosh (2000) argue for questions that refer to 'since the last visit', to establish a bound. And in fact, this is what the GLSS questionnaires do. The first visit in which information about consumption items is being asked, is the second overall interview. Thus, the first interview, in which only information about demographics is collected, is used as a bound.[18] Second, telescoping is most relevant for larger, 'non-routine' events (Gibson and Kim, 2007). For both of these reasons, the effect of telescoping in the first consumption visit is expected to be muted, although it cannot be ruled out. But, in any case, telescoping cannot explain the continued decrease in the reported consumption after the first consumption-related visit.

In sum, there are both hypotheses that suggest that later visits are of higher quality, as well as a hypothesis favouring earlier visits. Finally, it is conceivable that combinations of the above might imply a U-shape or inverted U-shape of the quality function. To study which consumption measure is more likely to be the correct one, Benford's law is employed in the next section.

V. Benford's law and the empirical strategy

Benford's law describes a regularity regarding the first significant digit (FSD) of numbers, where the FSD is the first non-zero digit (e.g. $\text{FSD}(-0.341) = 3$). Newcomb (1881) and Benford (1938) noticed that in 'collections of data', the FSD is approximately distributed with probability

$$p(\text{first digit is } d) = \log_{10}(1 + \frac{1}{d})$$

for $d = 1, 2, \ldots, 9$. Thus, Benford's law implies monotonically decreasing probabilities of the appearance of FSD, with 1 appearing approximately 30% of the time, 2 approximately 17.6% of the time, down to 9, which appears as the FSD approximately 4.6% of the time.[19]

Several formal explanations exist, with a particular prominent one coming from Hill (1995). He summarizes his result as follows: 'If distributions are selected at random (in any 'unbiased' way) and random samples are then taken from each of these distributions, the significant digits of the combined sample will converge to the logarithmic (Benford) distribution' (Hill, 1995, p. 354). An important property of Benford's law is that it is base and scale invariant, and thus, the unit of measurement does not matter. On the other hand, not all data will conform to Benford's distribution, for example binary or categorical data, or truncated data sets.

A large number of empirical examples for which the law holds have been described. Many of these papers then go on to interpret data that show deviations from the expected distribution as evidence of data manipulation. A particularly influential example is the work of Nigrini (1996), who shows that the data on interest payments and interest received on US-tax returns conforms very well to Benford's law, and argues that tax returns that do not confirm to Benford's law are possibly manipulated. Indeed, there is evidence that the

[18] The precise question in the second interview, when consumption is first covered, is 'How much was spent on ... since my first visit?'. In later visits, the question is '... since my last visit?'.

[19] Precise values are reported in Table 1.

Internal Revenue Service and other tax agencies now use Benford's law to flag potentially fraudulent tax returns (IRS, 2014).[20]

Most closely related to the present paper is a paper by Judge and Schechter (2009). These authors also apply Benford's law to study data quality in household surveys. In particular, they compare data collected by several different enumerators to identify suspected cheating by enumerators. One major difference from the present paper is that Judge and Schechter (2009) analyse their data with a focus on quality differences due to a specific interviewer effect (fraud on the part of the interviewer), and a focus on a small number of variables, for which one could argue that they should be measured with more or less error. The resulting data sets of distances between Benford's distribution and the distribution of FSDs in the data are consequently fairly small (they typically report only a handful of values for each distance measure), and further statistical work on that data is not possible. Additionally, it is unclear whether the data does indeed conform to Benford's law for a given variable or collection of variables, even in the absence of fraud or any other data quality problems. On the other hand, in the present paper, I will generate a large number of distance measures between Benford's distribution and the distribution of FSDs in the data, which allows me to perform further statistical analysis. First, I will exploit a large number of frequently purchased consumption items. Secondly, I will observe the same item several times over the course of up to 10 interviews. Together this will allow me to investigate changes over time, while controlling for fixed differences between Benford's distribution and the distribution of FSDs in data for a given consumption item. The focus of the present analysis will therefore not be on the average distance between Benford's distribution and the distribution of FSDs in the data – which may or may not be zero, even for very high quality data – but on the changes in this distance over repeated interviews, which should not change in the presence of high quality data. If Benford's distance measures change over time, this is used as an indicator for changes in data quality over time.

To identify how well FSDs from a given data set conform to Benford's distribution, several test statistics and distance measures can be used. First, a chi-squared statistic can be calculated to test the null hypothesis that the FSDs from a given data set are distributed according to Benford's law. However, the chi-square statistic is very sensitive to sample size. To analyse the goodness of fit with a measure that is not sensitive to sample size, I base the analysis on the normalized Euclidean distance measure, d^*. This measures the square root of the sum of squared differences between the observed and the expected percentage and normalizes this by the square root of the sum of squared maximum possible differences (which occurs if all FSD are equal to 9).[21] Alternatively, I also use $m = \max_{x=1,\ldots,9}\{Prob|(observed = x) - Prob(expected = x)|\}$, that is, the maximum difference between the observed and the expected (according to Benford) percentages across the FSDs 1–9 (e.g. Leemis, Schmeiser and Evans, 2000). In any case, the important point is that I do not test whether the FSDs of a given variable are distributed as Benford's law suggests

[20] The Internal Revenue Manual, Part 4, Examining Process, Section 4.1.10.3.1.L is titled 'Benford's Law Analysis – used to identify problematic preparers' (IRS, 2014). Benford's law has also been applied to study, for example, the possible manipulation of the Libor rate (Abrantes-Metz, Villas-Boas and Judge, 2011), to investigate measurement issues in macroeconomic statistics (Nye and Moul, 2007), or to study fraud in elections (Weidmann and Callen, 2013).

[21] Other papers that use this measure include Cho and Gaines (2007) and Judge and Schechter (2009).

TABLE 1

Observed distributions of FSD for pooled data, separately for urban and rural households

(First significant) digit	*Frequency of digit (%) according to Benford's law*	*Observed frequency of digit (%) in reported consumption values Rural*			*Observed frequency of digit (%) in reported consumption values Urban*		
		All visits	*First visit*	*Last visit*	*All visits*	*First visit*	*Last visit*
1	30.10	31.47	31.57	31.89	33.13	32.52	33.37
2	17.61	21.72	22.40	21.78	19.70	19.73	19.65
3	12.49	7.12	7.93	6.39	10.94	11.42	10.87
4	9.69	10.65	9.96	10.96	9.17	8.89	9.26
5	7.92	17.87	16.69	18.07	12.96	13.05	12.72
6	6.69	6.28	6.13	6.26	7.16	7.35	7.17
7	5.80	1.35	1.71	1.15	2.24	2.51	2.18
8	5.12	2.77	2.73	2.84	3.01	2.79	3.05
9	4.58	0.76	0.88	0.66	1.69	1.75	1.73
n		323664	60068	42283	267996	30384	25548
d*		0.132	0.122	0.138	0.079	0.077	0.079
m		9.9	8.8	10.1	5.0	5.1	4.8

– most likely FSDs of each individual variable are not exactly distributed according to Benford's law – but I test whether, on average, the distance between Benford's distribution and the actually observed distribution of FSDs for each consumption item changes over the course of the visits.

VI. Results based on Benford's law

Analysis based on all consumption items pooled

This section first considers pooled data, that is, combining all values of consumption of own produced food products, and frequently purchased food and non-food items for all households. Based on all values for a visit (i.e. all reported positive consumption values for all items listed), I calculate the distribution of FSDs. Table 1 shows that, for the most part, the observed percentage of FSDs indeed monotonically decreases from digit 1 to digit 9, as in Benford's distribution. However, there is a clear overrepresentation of digit 5, consistent with a tendency to round numbers. There is also a clear difference between rural and urban households. Both d^* and m, which summarize the distance between the observed and the expected distribution, are much smaller for urban households. Looking at changes across visits, in rural areas the distance measures clearly increase from the first to the last visit, while in urban areas the effect is much less clear. While these results are a first indication of data quality changing over visits, in particular for rural households, they only provide a starting point. In particular, it is obvious that the observed distributions do not conform to Benford's law, even for urban areas (a test based on the chi-squared statistic rejects the null of equal distributions with a P-value <0.00001 in all cases). However, rather than trying to say something about the absolute quality of data, the goal of the analysis is to analyze whether data quality changes over time, and if so, in which direction.

Regression analysis at the consumption item level

I now consider each consumption item in each consumption-related visit as a separate observation. For each of the 208 consumption items in each visit I calculate the distribution of first significant digits and generate the normalized Euclidean distance as the preferred measure of distance between the observed distribution and the expected distribution of first digits. Therefore, I have a maximum possible number of 208 × 7 observations from rural areas and 208 × 10 in urban areas.[22] I then regress the Euclidean distance on dummy variables for each consumption-related visit as well as on item fixed effects:

$$d_{iv}^{*} = visit_v + consumption\_item_i + e_{iv}$$

where $i = 1, \ldots, 208$, and $v = 1, \ldots, 7$ (rural areas) or $v = 1, \ldots, 10$ (urban areas). This will allow me to test whether the distance between the expected and the actual distribution of first significant digits systematically changes over consumption-related visits, while controlling for fixed deviations of each item from the expected (Benford) distribution. Note that any distance measure is mechanically fairly large for items that are only consumed by a small number of households, even if the draws come from the Benford distribution. Therefore, I also show regressions that omit item-visit observations that are based on <50 observations. As an alternative, I also use weighted regressions, where I use all data points, using the number of households that consume the item in the first consumption-related visit as weight for that item across all visits.[23] The main results are in Table 2.

The results clearly show an increase in the Euclidean distance over the repeated survey visits. In rural areas, the increase is particularly pronounced between the first and the second visit, but it also increases in later visits. In urban areas, the statistical significance of the difference between first and second visit varies a bit across specifications. The difference between the first and the fourth visit (and later visits) is significant, independent of the specification. The table also reports the *P*-value of a test whether the coefficients for the second and the last visit are statistically different, which shows that indeed in all specifications the difference between the second visit and the last is also statistically significant.[24]

To show robustness of the results, I also use the maximum absolute distance between the expected distribution and the actual distribution of FSDs. See Table 3 for the results of regressions that use the maximum absolute distance measure instead of the Euclidean

[22] Since neither in rural nor in urban areas are all items that appear in the consumption module reported, the actual maximum number of observations in one visit is 205 and 201 respectively.

[23] The number of observations is slightly smaller in the weighted regressions, because some marginal items are not consumed in the first round, but in later rounds, so weights, which are based on first-round observations, for these items are missing.

[24] To illustrate the magnitude of these changes, note that a worsening in the normalized Euclidean distance measure of 0.062 (as is the case for the visit 7 relative to visit 1, in column 1 of Table 2), occurs if, in data with first significant digits that follow Benford's distribution, all the numbers that start with a 9 are changed such that they start with another digit (e.g. with a 1). In other words, the worsening of 0.062 is obtained if about 4.5% of data with 'correct' significant digit are changed to data with 'incorrect' first significant digit.

TABLE 2

Regression using Euclidean distance as the dependent variable

Dependent variable: Euclidean distance (normalized) between the distribution of first significant digits according to Benford's law and the distribution of first significant digits of the observed values of reported consumption in one item-visit pair	*Rural*			*Urban*		
	(1)	*Restricted to items with frequency ⩾ 50 (2)*	*Weighted by frequency (3)*	*(4)*	*Restricted to items with frequency ⩾ 50 (5)*	*Weighted by frequency (6)*
visit 2	0.039	0.014	0.024	0.020	0.003	0.004
	(0.008)***	(0.004)***	(0.003)***	(0.009)**	(0.004)	(0.003)
visit 3	0.047	0.024	0.028	0.038	0.006	0.007
	(0.009)***	(0.004)***	(0.003)***	(0.010)***	(0.004)	(0.003)**
visit 4	0.057	0.034	0.035	0.048	0.015	0.014
	(0.008)***	(0.004)***	(0.003)***	(0.010)***	(0.005)***	(0.003)***
visit 5	0.050	0.029	0.036	0.033	0.007	0.009
	(0.008)***	(0.004)***	(0.004)***	(0.009)***	(0.004)*	(0.003)***
visit 6	0.065	0.030	0.038	0.058	0.011	0.013
	(0.010)***	(0.004)***	(0.003)***	(0.011)***	(0.004)***	(0.003)***
visit 7	0.062	0.033	0.040	0.051	0.012	0.016
	(0.009)***	(0.005)***	(0.004)***	(0.010)***	(0.004)**	(0.003)***
visit 8				0.074	0.018	0.018
				(0.011)***	(0.005)***	(0.003)***
visit 9				0.084	0.017	0.022
				(0.013)***	(0.005)***	(0.003)***
visit 10				0.094	0.021	0.026
				(0.014)***	(0.005)***	(0.004)***
Item fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Constant	0.235	0.191	0.187	0.238	0.166	0.154
	(0.006)***	(0.003)***	(0.003)***	(0.008)***	(0.003)***	(0.002)***
Observations	1384	952	1371	1878	980	1869
Number of items	205	136	199	201	98	197
R^2	0.07	0.13	0.22	0.08	0.04	0.06
P-value for test of H_0:						
visit 2=visit 7 (rural)	< 0.01	< 0.01	< 0.01			
visit 2=visit 10 (urban)				< 0.01	< 0.01	< 0.01

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

Notes: Each observation is the normalized Euclidean distance between the distribution of first significant digits according to Benford's law and the distribution of first significant digits of the observed values of reported consumption for one consumption item in one consumption-related visit. 'frequency' in columns 2, 3, 5, and 6 refers to the number of household that mention a particular consumption item. In columns 2 and 5 only item-visit observations for items are used that are mentioned at least 50 times, i.e. by 50 households, in the first consumption-related visit. In columns 3 and 6 frequencies of an item in the first consumption-related visit are used as weights.

TABLE 3

Regression results using maximum absolute distance as the dependent variable

Dependent variable: maximum absolute distance	*Rural*			*Urban*		
	(1)	*Restricted to items with frequency ⩾ 50 (2)*	*Weighted by frequency (3)*	*(4)*	*Restricted to items with frequency ⩾ 50 (5)*	*Weighted by frequency (6)*
visit 2	3.393	1.140	2.319	1.512	0.146	0.310
	(0.936)***	(0.447)**	(0.404)***	(0.964)	(0.472)	(0.346)
visit 3	4.070	2.078	2.571	3.527	0.184	0.531
	(0.915)***	(0.427)***	(0.404)***	(1.104)***	(0.467)	(0.351)
visit 4	5.074	2.903	3.084	3.970	1.380	1.240
	(0.851)***	(0.447)***	(0.403)***	(1.053)***	(0.525)***	(0.350)***
visit 5	4.060	2.194	3.177	2.753	0.315	0.752
	(0.876)***	(0.419)***	(0.451)***	(0.993)***	(0.470)	(0.335)**
visit 6	5.994	2.717	3.434	5.209	0.661	1.053
	(1.063)***	(0.418)***	(0.391)***	(1.145)***	(0.435)	(0.341)***
visit 7	5.621	2.674	3.554	4.562	0.631	1.306
	(1.009)***	(0.484)***	(0.451)***	(1.056)***	(0.464)	(0.375)***
visit 8				6.976	1.122	1.387
				(1.147)***	(0.481)**	(0.388)***
visit 9				7.780	1.425	1.857
				(1.298)***	(0.511)***	(0.418)***
visit 10				8.869	1.406	1.991
				(1.464)***	(0.535)***	(0.427)***
item fixed effects	yes	yes	yes	yes	yes	yes
Constant	17.853	14.579	13.972	17.966	12.702	11.414
	(0.638)***	(0.302)***	(0.324)***	(0.780)***	(0.317)***	(0.259)***
Observations	1384	952	1371	1878	980	1869
Number of items	205	136	199	201	98	197
R^2	0.05	0.08	0.15	0.06	0.03	0.03
P-value for test of H_0:						
visit 2=visit 7 (rural)	0.02	< 0.01	< 0.01			
visit 2=visit 10 (urban)				< 0.01	0.03	< 0.01

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

Notes: Each observation is one consumption item in one consumption-related visit. 'frequency' in columns 2, 3, 5, and 6 refers to the number of household that mention a particular consumption item. In columns 2 and 5 only item-visit observations for items are used that are mentioned at least 50 times, i.e. by 50 households, in the first consumption-related visit. In columns 3 and 6 frequencies of an item in the first consumption-related visit are used as weights.

distance, but otherwise the same specifications as in Table 2. The results of Table 2 are confirmed.

As argued above, false data or low-quality data will conform less to the expected (Benford's) distribution of first significant digits. Therefore, the results suggest that data quality is the highest in the first consumption-related visit, and significantly deteriorates over time. This is true even after restricting the analysis to frequently mentioned items or weighting them by the frequency mentioned. The drop in quality seems to be largest

TABLE 4

Comparing GLSS 3, 4 and 5 (using Euclidean distance as dependent variable)

Dependent variable: Euclidean distance (normalized) restricted to observations (items) which are mentioned at least 50 times	*Rural*			*Urban*		
	GLSS 3 (1)	*GLSS 4* (2)	*GLSS 5* (3)	*GLSS 3* (4)	*GLSS 4* (5)	*GLSS 5* (6)
visit 2	0.014	0.014	0.013	0.003	0.015	0.009
	(0.004)***	(0.003)***	(0.003)***	(0.004)	(0.003)***	(0.003)***
visit 3	0.024	0.019	0.014	0.006	0.019	0.014
	(0.004)***	(0.003)***	(0.003)***	(0.004)	(0.003)***	(0.004)***
visit 4	0.034	0.026	0.012	0.015	0.017	0.011
	(0.004)***	(0.003)***	(0.003)***	(0.005)***	(0.004)***	(0.004)***
visit 5	0.029	0.031	0.020	0.007	0.026	0.019
	(0.004)***	(0.003)***	(0.003)***	(0.004)*	(0.004)***	(0.004)***
visit 6	0.030	0.035	0.021	0.011	0.031	0.021
	(0.004)***	(0.003)***	(0.003)***	(0.004)***	(0.004)***	(0.004)***
visit 7	0.033		0.023	0.012		0.019
	(0.005)***		(0.003)***	(0.004)**		(0.004)***
visit 8			0.027	0.018		0.021
			(0.003)***	(0.005)***		(0.004)***
visit 9			0.029	0.017		0.026
			(0.004)***	(0.005)***		(0.004)***
visit 10			0.034	0.021		0.030
			(0.003)***	(0.005)***		(0.004)***
item fixed effects	yes	yes	yes	yes	yes	yes
Constant	0.191	0.135	0.163	0.166	0.136	0.170
	(0.003)***	(0.002)***	(0.002)***	(0.003)***	(0.002)***	(0.003)***
Observations	952	972	1630	980	792	1430
Number of items	136	162	163	98	132	143
R^2	0.13	0.17	0.10	0.04	0.13	0.07

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

between the first and the second visit in rural areas, but quality drops in both rural and urban areas, and continues to decrease over time in both areas.

A comparison to GLSS 4 and GLSS 5

To further show robustness of the results, I repeat the main analysis, based on the normalized Euclidean distance, for two further data sets, namely GLSS 4 and GLSS 5. These were collected by the Ghanaian Statistical Service in 1999 and 2005, respectively, and generally follow a setup that is broadly comparable to GLSS 3. The number of consumption-related visits varies, though: GLSS 4 in both urban and rural areas consumption data are collected based on seven interviews, while GLSS 5 uses 10 visits. The results (Table 4) confirm the central previous findings: The distance measures go up over visits and thus data quality seems to deteriorate monotonically over visits.

The analysis of GLSS 4 and 5 can additionally be used to investigate the pronounced differences between rural and urban areas that I found in GLSS 3. While rural and urban areas in GLSS 3 follow different data collection modes, there are no such differences between rural and urban areas in GLSS 4 and GLSS 5, both use diaries wherever possible. Comparing rural and urban areas, it is striking that the differences between them are much less pronounced in GLSS 4 and 5 than in GLSS 3. This strongly suggests that the differences between rural and urban areas in GLSS 3 are not based on some structural differences between rural and urban areas that are largely constant over time. Rather, the observed differences in GLSS 3 seem to be due to differences in the survey setup between rural and urban areas. This is further explored below.

Heterogeneity of the effect

Heterogeneity with respect to the respondent's education

In this section, I investigate two different dimensions of education: The distinction between literate and illiterate households is particularly relevant, because literacy determines the survey mode in urban areas. While in GLSS 3 in rural areas all households were interviewed, in urban areas households received a diary. However, households unable to fill out a diary, that is, those without a literate member, received frequent visits of enumerators to interview them. Thus, this analysis can not only identify the role of education, but also help to investigate whether differences between urban and rural surveys are solely due to the survey mode.

The survey allows me to define as illiterate a person who is recorded as neither able to read nor to write. For each consumption item in each visit, the Euclidean distance measure is then calculated separately for the households that are identified as having at least one literate household member and for those with no literate member. To simplify the following tables, I combine the visits as follows: The dummy variable *visit*_2_4 is equal to one for visits two through four, similarly defined are dummy variables *visit*_5_7 and *visit*_8_10. The omitted category is visit one. The results are shown in Table 5, where columns 1, 2, 4 and 5 report the results of regressions separately for those with at least one literate member, and those without literate members. Columns 3 and 6 show the results for a pooled regression in which interaction terms with an indicator for 'literate' are used. For rural areas (columns 1–3), the results show a large, and statistically significant difference between rural literate and illiterate households, with indications of lower data quality for illiterate households. The initial wedge in data quality increases in later surveys. On the other hand, for urban areas (columns 4–6), there are no statistically significant differences between literate and illiterate households.

To further analyse the role of education, while abstracting from the issue of different survey modes in urban areas, I also analyse the role of education conditional on having a literate household member. Among those households with a member who can read and write, I distinguish between two levels of education: higher levels of education, defined as having received education at the level of BECE (Basic Education Certificate Examination, which is typically taken after the third year of Junior High Schools, that is, typically implies at least 9 years of education), or above, and lower education, defined as below

TABLE 5

Heterogeneity with respect to whether the household has a literate member

Dependent variable: Euclidean distance (normalized) restricted to observations (items) which are mentioned at least 50 times						
	Rural			*Urban*		
	No literate household member (1)	*At least one literate household member* (2)	(3)	*No literate household member* (4)	*At least one literate household member* (5)	(6)
visit_2_4	0.030	0.023	0.030	0.021	0.009	0.021
	(0.006)***	(0.003)***	(0.006)***	(0.011)*	(0.004)**	(0.010)**
visit_5_7	0.051	0.030	0.051	0.020	0.011	0.020
	(0.006)***	(0.004)***	(0.006)***	(0.012)	(0.003)***	(0.011)*
visit_8_10				0.016	0.020	0.016
				(0.011)	(0.004)***	(0.010)
literate			−0.028			0.013
			(0.016)*			(0.041)
visit_2_4*literate			−0.007			−0.012
			(0.006)			(0.011)
visit_5_7*literate			−0.021			−0.009
			(0.006)***			(0.012)
visit_8_10*literate						0.004
						(0.011)
item fixed effects	yes	yes	yes		yes	yes
constant	0.216	0.186	0.215	0.200	0.165	0.157
	(0.005)***	(0.003)***	(0.011)***	(0.009)***	(0.003)***	(0.037)***
Observations	476	896	1372	130	970	1100
Number of items	68	128	149	13	97	105
R^2	0.20	0.10	0.12	0.03	0.04	0.03

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

BECE, including those with no formal education at all (but excluding the illiterate). I assign to a household the highest education level obtained by any member within that household.[25] The Euclidean distance measure is then calculated separately for each item for the households that are identified as having a respondent with education at the level of MSLC/BECE or above (labelled 'some education') and for those with less education. Results are in Table 6. They show no statistically significant differences in the deterioration of data quality over time between more and less educated individuals. However, they again show that the apparent decline in data quality across survey rounds is faster in rural areas than in urban areas.

[25] Unfortunately, the questionnaire does not identify the respondent in the sections related to frequently purchased and own consumed items. For frequently purchased items, the questionnaire only gives the instruction 'respondents are the persons mainly responsible for household purchases'. Therefore, to proxy the respondent's level of education, I use the person with the highest level of education. Alternatively, in results not shown, I have also used the person that is listed first on the household roster. Since these two variables are highly correlated, the results do not change much in this alternative approach.

TABLE 6

Heterogeneity with respect to the household's education (conditional on having a literate member)

Dependent variable: Euclidean distance (normalized) restricted to observations (items) which are mentioned at least 50 times	*Rural*			*Urban*		
	Education less than BECE	*Education BECE or above ('BECE+')*		*Education less than BECE*	*Education BECE or above ('BECE+')*	
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
visit_2_4	0.027	0.022	0.027	0.009	0.005	0.009
	(0.004)***	(0.004)***	(0.004)***	(0.006)	(0.004)	(0.006)
visit_5_7	0.037	0.033	0.037	0.013	0.006	0.013
	(0.005)***	(0.004)***	(0.005)***	(0.006)**	(0.004)	(0.006)**
visit_8_10				0.025	0.014	0.025
				(0.006)***	(0.004)***	(0.006)***
BECE+			0.001			0.014
			(0.009)			(0.014)
visit_2_4 * BECE+			−0.005			−0.004
			(0.005)			(0.007)
visit_5_7 * BECE+			−0.005			−0.007
			(0.006)			(0.007)
visit_8_10 * BECE+						−0.011
						(0.007)
item fixed effects	yes	yes	yes		yes	yes
Constant	0.189	0.190	0.189	0.155	0.165	0.152
	(0.003)***	(0.003)***	(0.005)***	(0.005)***	(0.004)***	(0.010)***
Observations	693	728	1421	450	820	1270
Number of items	99	104	141	45	82	97
R^2	0.12	0.10	0.07	0.05	0.03	0.03

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

Heterogeneity with respect to household asset wealth

Another important heterogeneity to investigate is heterogeneity with respect to consumption or wealth. If the difference in consumption data quality is correlated with true consumption, this could give rise to problems beyond simply noisy measures of mean consumption. In particular, Gibson *et al.* (2015) find evidence for a negative correlation between true consumption and the measurement error for consumption, that is, a 'mean-reverting measurement error' (Bound and Krueger, 1991). As a possible consequence Gibson *et al.* (2015) point out that this would imply that impacts may be understated if consumption is used as the dependent variable in an impact evaluation. They also illustrate the possible bias due to this kind of non-classical measurement error when consumption is used as an independent variable, in their example in a food Engle curve regression framework.

To investigate the correlation between the level of consumption and measurement error, I use household asset wealth as a proxy for the true, but unknown, consumption of a household. To measure household asset wealth in GLSS 3 I use the section on assets and durable consumer goods and aggregate for all 23 categories the self-reported resale value of

TABLE 7

Heterogeneity with respect to the household asset wealth

Dependent variable: Euclidean distance (normalized) restricted to observations (items) which are mentioned at least 50 times	*Rural*			*Urban*		
	Assets below median ('poor')	*Assets above median ('rich')*		*Assets below median ('poor')*	*Assets above median ('rich')*	
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
visit_2_4	0.030	0.023	0.030	0.007	0.010	0.007
	(0.004)***	(0.004)***	(0.004)***	(0.005)	(0.004)**	(0.005)
visit_5_7	0.035	0.038	0.035	0.014	0.012	0.014
	(0.004)***	(0.004)***	(0.004)***	(0.005)***	(0.004)***	(0.005)***
visit_8_10				0.030	0.018	0.030
				(0.006)***	(0.004)***	(0.006)***
"rich" (assets above median)			−0.012			−0.012
			(0.005)***			(0.006)*
visit_2_4 * 'rich'			−0.007			0.003
			(0.005)			(0.006)
visit_5_7 * 'rich'			0.002			−0.002
			(0.005)			(0.006)
visit_8_10 * 'rich'						−0.012
						(0.008)
item fixed effects	yes	yes	yes		yes	yes
Constant	0.199	0.185	0.198	0.168	0.160	0.170
	(0.003)***	(0.003)***	(0.003)***	(0.004)***	(0.003)***	(0.005)***
Observations	777	777	1554	680	790	1470
Number of items	111	111	121	68	79	83
R^2	0.11	0.13	0.12	0.08	0.03	0.08

Robust standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%

each reported category of asset / durable good. I then split the sample into two halves, with households with above-median asset holdings labelled 'rich'. Table 7 shows the results. Indeed, households with above median asset wealth have significantly lower values for the distance measures, that is, their survey data seem to have higher data quality. However, the difference is not changing significantly over time, and the problem of data quality for observed items does not appear to get worse as households receive more visits. The finding of a correlation between consumption, as proxied by asset wealth, and data quality, confirms one aspect of the earlier findings by Gibson *et al.* (2015). Whether this correlation is positive or negative cannot be established, as the criterion based on Benford's law will not allow me to make a statement about the direction of the measurement error.

Heterogeneity with respect to survey design

I also investigate possible heterogeneity in four other dimensions that are of particular interest for survey design. To economize on space, details of the hypotheses and results are not reported, but available upon request. First, the value of items might play a role. Indeed, I find that the decrease in data quality is smaller for items of higher value than for

lower value items, although this is significant only in rural areas. Second, data quality might decrease depending on how frequently an item is consumed. However, there is no significant difference for this. Third, I investigate whether the length of the questionnaire plays a role and find that data quality in rural areas, where interviews are conducted, deteriorates over time for items that appear later on the questionnaire. Finally, I also investigate whether the timing of the survey matters. I find statistically significant evidence for this in rural areas (surveys that start in the second and third quarter of 1992 show larger drops in data quality, as measured by Benford's law, than surveys that start in the last quarter of 1991).

VII. Implications and conclusion

This paper analyses the role of a somewhat neglected aspect of survey design, namely the frequency of visits for measurement of consumption and expenditure. Using data from Ghana, I first highlight that reported consumption decreases in economically and statistically significant magnitudes over the course of up to 10 visits. The goal of the subsequent analysis is then to identify which of the visits result in higher data quality. In the absence of a benchmark measure of the correct consumption and expenditure measures of the households under consideration, Benford's law is used as a diagnostic tool for this purpose. The analysis exploits the fact that the same consumption items are recorded repeatedly and heterogeneity in data quality across questions can be controlled for by consumption item-fixed effects. Thus, Benford's law is not used to identify individual items from the questionnaire that are of higher data quality, but the analysis focuses on the change in data quality over time.

The main results strongly suggest that data quality deteriorates over time if the household is asked to report consumption at a very detailed level repeatedly over a short time period. This results is also consistently found in later surveys from Ghana, GLSS 4 and 5.

Overall, the results are most consistent with conditioning effects, that is, 'having been interviewed previously causes differences in knowledge, behaviour, or attitude, affecting the answers in later interviews' (Toepoel *et al.*, 2009, p. 73). Telescoping and first-of-the-month-type seasonality effects cannot explain the findings. Various comparisons between rural and urban areas and between GLSS 3, GLSS 4 and GLSS 5 allow me to dig deeper into the underlying determinants of decreasing data quality. In the GLSS 3 data the deterioration in quality is particularly pronounced in rural areas, where interviews and shorter recall periods are being used, while the quality deteriorates less in urban areas, where diaries are used to record consumption and longer periods occur between visits. Structural differences between rural and urban areas cannot explain this finding, otherwise one would have to observe similar findings in GLSS 4 and 5. However, in GLSS 4 and 5, in which diaries are used whenever possible, and the spacing of visits is the same in rural and urban areas, the deterioration in data quality over visits is quite similar in rural and urban areas. Thus, the two dimensions (i) diary vs. interview and (ii) differences in recall periods are possible determinants of faster declines in data quality. An analysis of heterogeneity with respect to literacy allows me to say more about the role of diary vs. interview. I can exploit the fact that within urban areas diaries are used only for households with a literate member, while illiterate households are interviewed, which allows a within-urban area comparison of diary vs. interview mode. I do not find a significant difference in data quality between

households with only illiterate members and those with at least one literate member in urban areas. This therefore suggests that the survey mode, diary vs. interview, is not the main driver of the difference in the slope in the decrease of data quality over time between rural and urban areas. Instead, taken together the findings suggest that, for GLSS 3, the shorter recall periods in rural areas are driving the larger declines in data quality over time.

A limitation of the approach based on Benford's law is that it only allows statements about data quality for items with non-zero consumption values. It is impossible to say if the decrease in items mentioned over time reflects increasing or decreasing total data quality. This problem is similar to that described in Judge and Schechter (2009), who compare data from surveys that were conducted in Paraguay in 1999 and 2002. These authors find that the data from Paraguay 1999 looks better than the 2002 data, according to the Benford's law criterion. On the other hand, in 2002 the enumerators were encouraged to collect data more comprehensively, not just from the most important ones. The change in coverage might explain the decrease in data quality according to the Benford's law criterion (because in 2002 many smaller crops are covered, about which households might have less precise information) but might still make the 2002 data more suitable for analysis, since it is more comprehensive in terms of coverage. In related work, Beegle *et al.* (2012) find that all but one of their eight experimental survey modules result in lower consumption than the benchmark consumption. In particular, their frequent diary also yields lower total consumption values than the infrequent diary. In Friedman *et al.* (2016) the same experimental data is further analysed to understand consumption incidence (i.e. reports of non-zero consumption values). They find that for eight of the 12 food groups considered, consumption incidence is significantly lower in the frequent diary treatment than in the case of the benchmark, while for the infrequent diary it is significantly lower in five groups. Although these authors do not focus on changes over time, for the decrease in items reported in GLSS over time that this is more likely an indication of an increase in missing reports on items that were in fact consumed (i.e., a movement away from the true consumption value), rather than reducing the level of initial overstatements over time (i.e. a convergence to the true consumption value). Taken together, this suggests that more frequent surveying has two negative implications: First, the precision of the values reported for each item decreases, conditional on reporting an item. And, secondly, the number of items reported decreases below the true number.

The findings are relevant beyond Ghana. First, as pointed out in section I, the strategy to collect consumption and expenditure data through repeated visits is not uncommon in household surveys. Second, even in surveys where only one or few visits take place, and households record consumption and expenditure in diaries, there are often subgroups of the population who cannot fill out diaries and are visited frequently to collect the data. For example, in government-run surveys in Tanzania (HBS 2011–12) and Sierra Leone (SLIHS 2011) literate households receive diaries, while illiterate households are interviewed, 28 and 30 times respectively. This, in turn, points to a third concern. If a survey uses frequent interviews, but literate and illiterate household face different recall periods, the measurement error introduced through frequent surveys is likely correlated with other important characteristics of the population, such as literacy status and wealth. This would affect comparisons of poverty across sub-populations and measures of inequality even further (Gibson *et al.*, 2015). Finally, since from the perspective of the interviewed households it

does not matter whether repeated visits serve the purpose of estimating one consumption number or to collect panel data, this paper also suggests that panel surveys in which the households are repeatedly visited over short periods, will face similar data problems.

The findings are also relevant beyond developing countries and surveys that are based (at least in part) on interviews. In developed countries, surveys typically use diaries to measure consumption, and issues such as illiteracy and innumeracy and consequently a lack of ability to fill out a diary can be considered negligible. Although in well-known surveys, such as the US Consumer Expenditure Survey (CEX), diaries are not administered with high frequency, they are nevertheless administered more than once. The Canadian Food Expenditure Survey, for example, collects consumption expenditure through two diaries, that cover 1 week each. Reflecting the findings in the present paper, but at a lower frequency, Ahmed, Brzozowski and Crossley (2006) find a drop of about 10% in consumption between the first and the second week diary, similar effects are found for the CEX by Stephens (2003). They attribute this to 'diary fatigue', that is, a specific form of conditioning effects, although the alternative explanations mentioned in the present paper, including those that would imply higher data quality for later visits, are not ruled out by these authors.

Finally, the findings are not only relevant from a survey design point of view, but also because of the costs associated with a survey. If the value of additional survey rounds is limited, the money spent might not be spent most efficiently. Further, there is the opportunity costs of households, who spend considerable time responding to interviews (or keeping a diary). In fact, the results of this paper do not just suggest that money and time used on additional survey rounds are unnecessary, but that it may actually hurt precision. These findings are in line with recent work by Engle-Stone *et al*. (2017), which also provides evidence for the limited value of additional survey rounds. If the goal of governments is to obtain better data, while keeping the budget allocated to data collection constant, rather than adding more survey visits, it might be better to spread survey visits over the year to take into account within-year variation of consumption within a household. As shown theoretically by Scott (1992), and implemented empirically by Gibson (2001) and Gibson *et al.* (2003), because of limited correlation of levels of expenditure over different parts of the year, there is significant value in surveying households at various points during the year as opposed to only during one specific season.[26] McKenzie (2012) also provides related arguments. The finding regarding the timing of the survey reported in the section on heterogeneity, further support this.

Overall, the paper shows that the frequency of visits is yet another survey design decision that can have significant consequences for estimates that are based on those surveys. As such, this paper adds to the literature that points out that 'there is a risk of asking' (Zwane *et al.*, 2011, p. 1826). Results using Benford's law as an analytical tool suggest that survey quality deteriorates over time, and identify important dimensions of heterogeneity of the effect. Understanding the underlying reasons for deteriorating data quality and how to balance the need for high-frequency data collection and the possible downsides due to frequent visits is an important area of future research into optimal survey design.

Final Manuscript Received: April 2017

[26] Indeed, even the changes in the composition of households over time may affect estimates of annual consumption (Halliday, 2010).

References

Abrantes-Metz, R., Villas-Boas, S. and Judge, G. (2011). 'Tracking the libor rate', *Applied Economics Letters*, Vol. 18, pp. 893–899.

Ahmed, N., Brzozowski, M. and Crossley, T. (2006). Measurement Errors in Recall Food Consumption Data, Institute for Fiscal Studies Working Paper (W06/21).

Beegle, K., De Weerdt, J., Friedman, J. and Gibson, J. (2012). 'Methods of household consumption measurement through surveys: experimental results from Tanzania', *Journal of Development Economics*, Vol. 98, pp. 3–18.

Benford, F. (1938). 'The law of anomalous numbers', *Proceedings of the American Philosophical Society*, Vol. 78, pp. 551–572.

Bound, J. and Krueger, A. B. (1991). 'The extent of measurement error in longitudinal earnings data: do two wrongs make a right?', *Journal of Labor Economics*, Vol. 9, pp. 1–24.

Caeyers, B., Chalmers, N. and De Weerdt, J. (2012). 'Improving consumption measurement and other survey data through CAPI: evidence from a randomized experiment', *Journal of Development Economics*, Vol. 98, pp. 19–33.

Cho, W. and Gaines, B. (2007). 'Breaking the (Benford) law: statistical fraud detection in campaign finance', *The American Statistician*, Vol. 61, pp. 218–223.

Coulombe, H. and McKay, A. (2008). The estimation of components of household incomes and expenditures: a methodological guide based on the last three rounds of the Ghana Living Standards Survey, 1991/1992, 1998/1999 and 2005/2006, Ghana Statistical Service, Accra, Ghana.

Deaton, A. and Grosh, M. (2000). 'Consumption', in Margaret Grosh and Paul Glewwe (eds), *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of Living Standards Measurement Study*, Vol. 1, World Bank, Washington, DC, pp. 91–133.

Engle-Stone, R., Sununtnasuk, C. and Fiedler, J. L. (2017). 'Investigating the Significance of the Data Collection Period of Household Consumption and Expenditures Surveys for Food and Nutrition Policymaking: analysis of the 2010 Bangladesh Household Income and Expenditure Survey', *Food Policy*, Vol. 72, pp. 72–80.

Friedman, J., Beegle, K., De Weerdt, J. and Gibson, J. (2016). Decomposing Response Errors in Food Consumption Measurement: Implications for Survey Design from a Survey Experiment in Tanzania, World Bank, Policy Research Working Paper 7646.

Ghana Statistical Service (1995). Data User's Guide: Ghana Living Standards Survey Round Three (GLSS 3) 1991/92, Ghana Statistical Service, Accra, Ghana.

Ghana Statistical Service (2000). Poverty Trends in Ghana in the 1990s, Ghana Statistical Service, Accra, Ghana.

Gibson, J. (2001). 'Measuring chronic poverty without a panel', *Journal of Development Economics*, Vol. 65, pp. 243–266.

Gibson, J., Huang, J. and Rozelle, S. (2003). 'Improving estimates of inequality and poverty from Urban China's Household Income and Expenditure Survey', *Review of Income and Wealth*, Vol. 49, pp. 53–68.

Gibson, J. and Kim, B. (2007). 'Measurement error in recall surveys and the relationship between household size and food demand', *American Journal of Agricultural Economics*, Vol. 89, pp. 473–489.

Gibson, J., Beegle, K., De Weerdt, J. and Friedman, J. (2015). 'What does variation in survey design reveal about the nature of measurement errors in household consumption?', *Oxford Bulletin of Economics and Statistics*, Vol. 77, pp. 466–474.

Halliday, T. (2010). 'Mismeasured household size and its implications for the identification of economies of scale', *Oxford Bulletin of Economics and Statistics*, Vol. 72, pp. 246–262.

Hastings, J. and Washington, E. (2010). 'The first of the month effect: consumer behavior and store responses', *American Economic Journal: Economic Policy* Vol. 2, pp. 142–162.

Hill, T. (1995). 'A statistical derivation of the Significant-Digit Law', *Statistical Science*, Vol. 10, pp. 354–363.

IRS (2014). Internal Revenue Manual Part 4, Examining Process, Section 4.1.10.3.1. Internal Revenue Service. http://www.irs.gov/irm (accessed November 10, 2016).

Judge, G. and Schechter, L. (2009). 'Detecting problems in survey data using Benford's law', *Journal of Human Resources*, Vol. 44, pp. 1–24.

Kreuzer, M. Jordan, D. Antkowiak, B. Drexler, B. Kochs, E. and Schneider, G. (2014). 'Brain electrical activity obeys Benford's law', *Anesthesia and Analgesia*, Vol. 118, pp. 183–191.

Lanjouw, J. O., and Lanjouw, P. (2001). 'How to compare apples and oranges: poverty measurement based on different definitions of consumption', *Review of Income and Wealth*, Vol. 47, pp. 25–42.
Leemis, L., Schmeiser, B. and Evans, D. (2000). 'Survival distributions satisfying Benford's law', *The American Statistician*, Vol. 54, pp. 236–241.
McKenzie, D. (2012) 'Beyond baseline and follow-up: the case for more T in experiments', *Journal of Development Economics*, Vol. 99, pp. 210–221.
Meyer, B., Mok, W. and Sullivan, J. (2015). 'Household surveys in crisis', *Journal of Economic Perspectives*, Vol. 29, pp. 199–226.
Neter, J. (1970). 'Measurement errors in reports of consumer expenditures', *Journal of Marketing Research*, Vol. 7, pp. 11–25.
Newcomb, S. (1881). 'Note on the frequency of use of the different digits in natural numbers', *American Journal of Mathematics*, Vol. 4, pp. 39–40.
Nigrini, M. (1996). 'A taxpayer compliance application of Benford's law', *Journal of the American Taxation Association*, Vol. 18, pp. 72–91.
Nye, J. and Moul, C. (2007). 'The political economy of numbers: on the application of Benford's law to International Macroeconomic Statistics', *The B.E. Journal of Macroeconomics*, Vol. 7, pp. 1–14.
Scott, C. (1992). 'Estimation of annual expenditure from one-month cross-sectional data in a household survey', *Inter-Stat Bulletin*, Vol. 8, pp. 57–65.
Stephens, M. (2003). "'3rd of the month": do social security recipients smooth consumption between checks?', *American Economic Review*, Vol. 93, pp. 406–422.
Toepoel, V., Das, M. and van Soest, A. (2009). 'Relating question type to panel conditioning: comparing trained and fresh respondents', *Survey Research Methods*, Vol. 3, pp. 73–80.
United Nations (2005). Household Sample Surveys in Developing and Transition Countries. Department of Economic and Social Affairs, Statistical Division, Publication No. 96. United Nations Publications.
Weidmann, N. and Callen, M. (2013). 'Violence and election fraud: evidence from Afghanistan', *British Journal of Political Science*, Vol. 43, pp. 53–75.
Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D. S., Hornbeck, R., Giné, X., Duflo, E., Devoto, F., Crepon, B. and Banerjee, A. (2011). 'Being surveyed can change later behavior and related parameter estimates', *Proceedings of the National Academy of Sciences*, Vol. 108, pp. 1821–1826.