De Weerdt, Joachim; Gibson, John; Beegle, Kathleen

Working Paper

What can we learn from experimenting with survey methods?

LICOS Discussion Paper, No. 418

Provided in Cooperation with:
LICOS Centre for Institutions and Economic Performance, KU Leuven

Suggested Citation: De Weerdt, Joachim; Gibson, John; Beegle, Kathleen (2019) : What can we learn from experimenting with survey methods?, LICOS Discussion Paper, No. 418, Katholieke Universiteit Leuven, LICOS Centre for Institutions and Economic Performance, Leuven

This Version is available at:
https://hdl.handle.net/10419/230505

LICOS Discussion Paper Series

Discussion Paper 418/2019

What can we learn from experimenting with survey methods?

Joachim De Weerdt, John Gibson and Kathleen Beegle

Faculty of Economics And Business

LICOS Centre for Institutions and Economic Performance
Waaistraat 6 – mailbox 3511
3000 Leuven
BELGIUM
TEL:+32-(0)16 32 65 98
FAX:+32-(0)16 32 65 99
http://www.econ.kuleuven.be/licos

What can we learn from experimenting with survey methods? *

Joachim De Weerdt[1,2]

John Gibson[3]

Kathleen Beegle[4]

November 2019

This review covers a nascent literature that experiments with survey design to measure whether the way in which we collect socio-economic data in developing countries influences the data and affects the results of subsequent analyses. We start by showing that survey methods matter and the size of the effects can be nothing short of staggering, affecting basic stylized facts of development (such as country rankings by poverty levels) and conclusions drawn from econometric analyses (such as what the returns to education are or whether small farm plots are more productive than large ones). We describe some of the emerging best-practices for conducting survey experiments, including benchmarking against the truth, delving into the error-generating mechanisms, and documenting the costs of different survey approaches.

Keywords: Experiments, Measurement error, Socio-economic surveys

* This research was supported by the Excellence of Science (EOS) Research Project 30784531 at the Research Foundation – Flanders (FWO).

[1] IOB, Institute of Development Policy, University of Antwerp, Belgium. E-mail: Joachim.DeWeerdt@uantwerpen.be

[2] LICOS Centre for Institutions and Economic Performance, KU Leuven, Belgium

[3] Department of Economics, University of Waikato, New Zealand. E-mail: jkgibson@waikato.ac.nz

[4] World Bank, United States of America. E-mail: kbeegle@worldbank.org

1. INTRODUCTION

Which country has the largest number of poor people, in terms of absolute poverty in monetary terms? According to current World Bank numbers, it is India. But there is a catch: an impending change in India's survey method will, overnight, reclassify 50 million Indians from poor to non-poor and make Nigeria the country with the highest number of poor people, with India dropping to third place (World Bank, 2018). Moreover, the methodological change is not dramatic, and is typical of what many survey agencies and researchers may consider tinkering with in a questionnaire between survey rounds. The current Indian household survey methodology has respondents reporting consumption over the previous 30-day period. This recall period is quite long for frequently consumed items compared to that used in other countries and compared to the capacity of respondents to remember all of these consumption occasions. On the other hand, it is short for capturing infrequently purchased items. The reform in India will change recall periods of some items to either seven days (for the frequently consumed items) or to 365 days (for the infrequently purchased items) and with it, very probably, India's official poverty numbers.[1]

Such examples of how seemingly minor changes in survey design can have big impacts on results – combined with the realization that much data-driven research is presented without any recognition of the potential data flaws (Jerven and Johnson, 2005) – have contributed to an increased interest in measurement issues. There is a strong demand-driven impetus: whether it is measuring progress towards the Sustainable Development Goals, impact evaluations, or empirical academic research, the activities of many development economists have in common a reliance on data. In the past 15 years or so there has been an explosion of survey work in developing countries, with researchers increasingly collecting primary data on individuals, households, firms, farms, networks and so forth. As more development researchers are directly involved in survey design and the data collection process, there has been an increasing interest in understanding how decisions about survey methods influence the results of subsequent analyses with the data. As such, the evidence from developing countries is slowly catching up to a much older and robust literature in the U.S. (see, among many others that could be cited, the classic work of Sudman and Bradburn, 1974).

In this review we cover this still-growing literature on experiments with survey methods, be it questionnaire design or survey implementation, in development contexts. The review draws insights from papers that look at how to measure consumption, agricultural production, household size, household business income, plot size, household labor, skills, asset ownership, and so forth. While many studies explore implications of varying how questions are phrased, there are also studies of the implications of survey mode, respondent selection, overall questionnaire length and interviewer-respondent interactions.[2] This review does not aim to exhaustively document the

[1] When similar changes occurred two decades ago, in smaller rounds of the survey that were less influential in public debate, it cut measured poverty rates in half (Visaria, 2000), and a 'halfway house' approach of recalling frequent items over both seven days and 30 days which was used in one of the major survey rounds gave results that were inconsistent with anything from the past. And so, just as poverty numbers became very sensitive in political debates about liberalization, the data became much less reliable (Deaton, 2003).

[2] There are some specific 'survey methods' areas which this review does not cover. We do not consider measurement developments from laboratory settings (such as eye-tracking) and nor do we consider surveys that deal primarily with hypothetical choice, such as contingent valuation and discrete choice experiments. We also do not cover studies that delve into how to collect data on highly sensitive topics, like sexual or risky or illicit behaviors.

universe of survey experiments in developing country contexts, but hopefully covers well the scale and scope of work, much of it recent, to highlight the following emerging lessons.

First, the effects of survey design variation on headline results that are documented in some studies are often staggering. Commonly observed variations in survey design can result in Gini coefficients jumping by 12 points, poverty rates going up or down by 20 percentage points, hunger rates varying by up to 50 percentage points, and four-fold differences in reports of family labor use on household farms. It is obvious from the size of these effects that survey methods may matter substantially.

Second, survey experiments ideally anchor to the truth or some approximation of it. Some of the studies reviewed are validation studies that have two independent measurements of the same event. Typically one (true) measure is taken from a data source that is assumed to contain no or close to no errors (e.g. GPS measures of plots). However, this restricts the topics that validation studies are suitable for, as there is no unimpeachable source of true data on things like household consumption or household composition. In addition to this restriction on topics, validation studies often involve special circumstances or limited contexts that may not easily generalize; for example, the PSID validation study that informs about measurement errors in retrospective self-reports of wages is based on the records from just a single firm (Gibson and Kim, 2010).

In most cases, no such 'true' measure is available. Tempting as it seems, conducting two survey measurements on the same respondent (a within-subject approach) raises concerns that the one survey process influences the other, through conditioning either the respondent or the interviewer. Consequently, a comparison between the results of the two types of measurement may not be informative about their performance under usual field conditions when both respondents and interviewers are exposed to only one approach to obtaining the measures of interest. Researchers have addressed this problem by experimenting with variations in survey methods across units (a between-subjects approach). This experimental design lets them infer the effect of survey methods by comparing moments of the distribution across the groups given different survey assignments. If the assignment was random, and on a sufficiently large sample, one should not expect any between-group differences to arise. Still, it remains important to know which method gets us closer to the truth. Knowing that two moments are different without being able to assess which survey method gives more accurate results is generally not helpful, except if one is willing to settle for the more limited goal of cost-effectiveness analyses. That is, in cases where two survey methods give similar results but one is far more costly than the other, then even though the comparison is not informed by knowing 'truth' it may still be useful in an operational sense for enabling survey agencies to use funds more effectively (see Sharp et al., 2019).

Third, some papers dig deeper to look at the nature of the resulting errors. Researchers typically assume, out of convenience rather than conviction and usually implicitly rather than explicitly, that measurement error is classical in nature. This means that errors in measuring a particular variable are uncorrelated with the true value of that variable, with the true values of other variables in the model, and with any errors in measuring those variables (Bound et al., 2001). Validation studies allow one to assign the true and measured variable to each observational unit and comparing the two allows for an accurate description of the measurement error. In contrast, the between-subjects survey experiments do not allow this, but nevertheless can give insights into the nature and consequences of measurement error. In this review we find that measurement error is not usually

classical, and instead has mean-reverting characteristics (Gibson et al 2015; Abay et al, 2019) and we discuss the implications of this finding for analyses in which the error-prone variable is being explained or does the explaining.

Fourth, some papers address external validity by making use of insights from psychology to better appreciate the cognitive processes underlying survey responses.[3] For example, in their study on measuring household farm labor, Arthi et al. (2018) show how people struggle to answer questions on the number of days they worked in the past agricultural season, but are not troubled in recalling the number of hours per day conditional on working that day. In their setting, in the Mara region in Tanzania, days worked per week are erratic, but, conditional on working, people tend to work a fixed number of hours per day. Understanding the cognitive mechanism that contributes to the error-ridden survey data helps us to hypothesize about what may happen in other contexts. For example, a motorbike taxi driver in town may work every day, except Sunday, but for a highly variable number of hours a day. We then would expect that in such a setting, the days worked should be accurately recalled, but the hours per day not. With an improved appreciation of these cognitive mechanisms and of the underlying process being measured, economists may be able to design more accurate surveys that are also less burdensome for respondents. More generally, a better understanding of whether respondents are using 'enumeration strategies' of actually counting (or adding up) each occurrence – such as spending on the specified item over the specified period or the number of days worked over the past season – or instead are using 'estimation strategies' to give a rule-of-thumb response, and also knowing what triggers any switch between these two ways of answering survey questions can help to design better questionnaires (Brown, 1995; Gibson and Kim, 2007).[4]

2. MAGNITUDES

The first thing that is striking about the literature on experiments with survey methodology is that the effect sizes can be very large. For example, Das et al. (2012) found that about one third of acute illness episodes are not reported when increasing the recall period from one week to one month – a large gap if one is trying to understand health and wellbeing. Beegle et al. (2012) show how commonly observed variations in consumption modules (changing the length of the recall period, or the length of the list of consumption items or the mode of information acquisition—recall or diary) can result in poverty rates going up or down by 20 percentage points in Tanzania. An even larger effect is found by De Weerdt et al. (2016), who find that hunger rates vary by up to 50 percentage points when using one randomly assigned consumption module versus another, drawing on the same data from Tanzania. Backiny-Yetna et al. (2017) find that annual per capita consumption from a 7-day recall method was, on average, 28 percent higher than that from the 7-day diary method in Niger. Di Maio and Fiala (2019) find that 30% of the variation revealed by questions on political preference is accounted for by enumerator effects. Using data on microenterprises in Sri Lanka de

[3] Sudman and Bradburn (1973), Sudman and Bradburn (1974), Blair and Burton (1987), Schwarz (1999) and Tourangeau et al. (2000) provide inspiration for quite a few studies of survey methods.

[4] According to Chang and Krosnick (2003, p.58) respondents are especially likely to retrieve and count episodes when the time frame is very short or very recent but otherwise rate-based rule-of-thumb estimation strategies are more commonly used for answering questions in surveys.

Mel et al. (2009) conclude that surveyed firms under-report revenues by 30%. It is obvious from these examples that survey methods may matter substantially to headline results.

Survey errors can manifest themselves in places other than just the first moment of the distribution, so it would be a mistake to consider the equality of the measured mean to the benchmark mean as a sufficient sign of correct measurement. For example, Garlick et al. (2019) found that weekly phone-based enterprise surveys in South Africa generate higher within-establishment variation in several variables compared to the variance coming from otherwise similar but in-person interviews. In Caeyers et al. (2012), paper-based surveys yielded much higher Gini coefficients compared to surveys conducted on hand-held devices. One reason for this is that the built-in checks within the software of the hand-held device flagged the most egregious errors to the interviewer during the interview, at a time and place where the respondent could be queried and an error could still be corrected. Many of these errors are random, and may stem from sources that tend to cancel out, on average, such as incorrect units (e.g. tens as hundreds and hundreds as tens, grams as kilograms and *vice versa*) even while they can greatly inflate variance-based measures.

Similarly the most cognitively abstract of the recall modules used in the study of Beegle et al. (2012) – based on a synthetic 'usual month' approach that requires recalling the months in the year that an item is consumed, the usual number of times per month and the typical amount or value per occasion – yielded the highest Gini coefficients. This was partly due to errors in reporting related to education levels, suggesting that some of the error is due to the cognitively demanding calculations required of the respondents. A final example comes from de Nicola and Gine (2014) who conclude that self-reported income data from their sample of fishermen in India are well-suited to estimating average income, even with longer recall periods. But when estimating the volatility of the income process, recall data yield an increasingly lower variance as the recall period increases. Specifically, when the recall period is two months the variance is the same as with the benchmark, but the variance is 13% lower than the benchmark when the recall period increases to 24 months.[5]

Survey errors may move in different directions such that they offset each other. With this comes the danger of erroneously concluding that a particular survey design is quite accurate, while in fact these findings are the result of offsetting errors. There are several manifestations of this phenomenon, and a common one arises through aggregation. For example, Arthi et al. (2018) and Gaddis et al. (2018) show how farmers over-report hours worked per plot per household member in Tanzania and Ghana respectively. But counteracting that over-reporting error is the pattern where recall questionnaires result in an under-report of the number of agricultural workers and the number of plots. The end result can be that the total hours worked at the household level (across all plots and all people) is, on average, correct: three erroneously measured variables have been aggregated to provide a roughly correct total. In a different context, Friedman et al. (2017) decompose errors in household consumption reporting into the reporting of incidence (the response to the yes/no question on whether someone within the household consumed an item) and the reporting of consumption value (the amount spent, or the quantity consumed, conditional on saying that the household consumed the item). They show how the results from a 7-day consumption recall module mimic the benchmark (based on intensively monitored 14-day individual diaries) quite well, but this

[5] And in a developed country context, retrospective recall of wage income in the United States for an even longer recall period of six years is quite accurate for the mean, but understates the variance due to the tendency for transitory earnings fluctuations to be ignored (Gibson and Kim, 2010).

is due to the happenstance of off-setting errors: negative errors in incidence offset by the positive errors in value. The latter effect, of positive errors in value, is hypothesized to be due to telescoping where consumption that occurred before the recall period was misdated in the mind of the respondent and they added it in to the amount or value that they reported as occurring during the last 7-days.

Abay et al. (2019) find that measurement errors in self-reported plot area and in crop production are strongly correlated.[6] When they use self-reported area and production to estimate the long-debated relationship between farm size and productivity, their econometric estimates are not much different from what they get using benchmark data based on actual measurement of production and of land area. However, correcting for just one of these two measurement problems exacerbates the parameter bias in the inverse size-productivity relationship relative to what would be estimated if one ignored both measurement problems. Thus, dealing with only one measurement error problem at a time (such as equipping survey teams with GPS receivers or with measuring devices like compass-and-rope in order to more accurately record plot sizes while continuing to rely on self-reported production) can paradoxically compromise the subsequent analysis of the (partially) error-ridden data.

3. ANALYTICAL CONSEQUENCES

The consequences of measurement error go beyond uncertainty surrounding some of the basic stylized facts in development, like which country has the largest number of poor people, as discussed above, or what share of women own land. Measurement error also has a bearing on the analysis of some of the basic economic relationships that underpin our understanding of the developing world. In one example, differential recall effects across income groups reverse the sign of the gradient between doctor visits and per-capita expenditures; this sort of gradient influences discussion of whether public services are pro-poor. These gradients also lead to questions of whether inequality in economic indicators like household expenditures gives a distorted picture of inequality in a broader notion of living standards (such as one that also includes health access). The specific details are that the poor appear to use health care providers more than do the rich if one relies on weekly recall surveys, but the gradient flips and the poor seem to have lower utilization than the rich if one uses monthly recall surveys (Das et al., 2012).

Borrowing notation from Bound et al. (2001), consider a true linear model $y^*=X^*\beta+\varepsilon$, where y^* and ε are scalars and X and β are vectors. However, instead of observing the true y^* we observe $y=y^*+\phi$ and instead of observing the true X^* we observe $X=X^*+\rho$. There is a special case, termed classical measurement error, in which the error is uncorrelated with the true value of the measured variable, with the true values of other variables in the model, and with the stochastic disturbance term, that is ρ and ϕ are uncorrelated to X^*, y^* or ε. Classical measurement error will not bias coefficient estimates if only the dependent variable suffers from it. If the explanatory variable is measured with

[6] Abay et al. (2019) point out that we need to distinguish between measurement error caused by misreporting and that caused by misperception. They argue that when estimating behavioral parameters what the farmer perceives to be true may be just as relevant as what the truth actually is. They show how measurement error in plot size at least partly reflects farmers' misperceptions of the land area they manage, which then affects their input use decisions.

classical error this will lead to the standard attenuation bias in the OLS regression coefficient (in proportion to the 'reliability ratio' of the mis-measured explanatory variable, in the case of a simple regression).

The assumption of classical measurement error is often made and could then arguably be considered as a reason to be less concerned about the impacts of errors in survey data. After all, it suggests that the researcher is erring on the side of caution and being conservative in the sense that the bias would be in terms of reducing the estimated rate of response of y to X, so the true effect – if we knew it – would be even bigger than what the researcher claims to find. However, even if the assumption of classical measurement error is correct (and, as we will show below, it very often is not) attenuation bias is not necessarily as benign as it is may initially seem. For example, we may erroneously conclude that households are easily able to smooth consumption because the coefficient on income, measured with classical error, is attenuated when regressed on consumption; such a finding would therefore downplay the need for interventions that might assist households to smooth. Relatedly, Gillen et al. (2019) argue that classical measurement error in control variables can lead to the identification of non-existing effects in other variables. They replicate the influential study by Niederle and Vesterlund (2007) which found that men are more likely to choose to compete compared to women and that these differences remain significant even after controlling for overconfidence, risk and feedback aversion. Gillen et al. (2019) find that once measurement error in the controls is properly accounted for, risk attitudes and overconfidence explain the gender gap in competition and there is no special male appetite for competition over and above this. This leads them to argue that classical measurement error can lead to the erroneous identification of what looks like "new" effects and phenomena.

Another example where classical measurement error may suggest some spurious effects comes from Larsen et al. (2019), who show how random errors in the reported month of birth produce a nonrandom pattern of height-for-age z-scores. The reason is that children who are mistakenly reported as born later in the year but were actually born earlier, will therefore appear taller for their age than they actually are, and vice versa for those mistakenly reported as born earlier in the year. They argue that previous research had erroneously interpreted this pattern as being due to seasonal fluctuations in height-for-age – leading to flawed policy advice, such as focusing child nutrition and health interventions more on children born in certain seasons of the year.

Further complicating things is the fact that measurement error, when studied, is often (if not usually) found to be non-classical. The most commonly found error pattern is mean-reverting error, in which the error is negatively related to the true value. When errors have this pattern, they can cause coefficient bias even when it is the outcome variable that is error-ridden. Depending on the degree of mean-reversion, if it is the right-hand side variable(s) measured with error then regression coefficients could be exaggerated rather than attenuated. Mean reverting error has been well-documented and comes up repeatedly in the debate on the role of measurement error in explaining the inverse size productivity relationship in developing country agriculture. Plot size is typically found to be reported with mean reverting error, meaning large plots are underestimated, while small plots are overestimated. For example, Carletto et al. (2015) report that farmers overestimate plot size by 103 percent relative to GPS measurement, for plots smaller than 0.5 acres, and underestimate it by 33 percent for plots larger than 5 acres. Dillon et al. (2019) find that, on average, self-reported plot sizes do not differ from those measured by either compass-and-rope or by GPS

methods. However, this lack of between-method differences in the mean is because of off-setting (and mean-reverting) errors: the self-reports overestimate the size of the smallest tercile of plots by 83% and underestimate that of the largest tercile by 21%, relative to the benchmark measurements based on compass-and-rope estimates. Such error patterns are exactly the kind of mean-reverting error that Bound et al. (2001) reported as being so prevalent in earnings and income data in various contexts in developed countries. A recent tabulation of some estimates of the strength of the mean-reversion in survey data is provided by Abay et al (2019).

Errors are also correlated to other variables of interest in the model, and in this way also violate the conditions for classical measurement error. Studies have found that crop production (as well as plot size) is systematically over-reported on small plots and under-reported on larger ones (Desiere and Jolliffe, 2018; Gourlay et al. 2019). Abay et al. (2019) point out that measurement error in plot size is correlated with crop-cut production as well as with true plot size. Another example comes from Gibson and Kim (2007), who find that errors in data on recalled food consumption correlate with household size, impacting analyses of economies of scale within households and affecting conclusions about whether poverty is more concentrated in large households or small ones.

The empirical literature on the analytical consequences of measurement error is somewhat skewed. Much more intellectual effort has been poured into mitigation efforts, of figuring out ways to derive non-biased econometric estimates when using existing error-prone data (e.g., by using instrumental variables or with bounds from reverse regression) compared to finding ways to avoid or reduce errors at the data collection stage. Yet in terms of magnitude it is possible that the effects of innovation at the data collection stage are at least as important as the mitigation strategies that use various econometric approaches. This point is illustrated by Serneels et al. (2017) for their study estimating the returns to education in Tanzania. The authors note how shortening the length of a labor survey module yields apparently higher returns to education, of five percentage points among the most highly educated men and 16 percentage points among the least educated women, using the standard Mincerian wage equation regressions. Theirs is the only study documenting these effects, yet the discrepancies they find are of a similar or larger magnitude to those found in the large literature that debates biases associated with the use of simple OLS estimation for Mincer equations in developed countries (Card, 1999).

4. BENCHMARKING THE TRUTH

Survey experiments can tell us a lot more when we can compare different measurement approaches to the true value. Some of the benchmark measurement methods that have been used in this way are compass-and-rope measures for land size, crop-cuts for agricultural output, highly supervised individual diaries for consumption, high frequency interviews throughout the agricultural season for measurement of labor, and administrative records for income from fish sales (Arthi et al. 2018; Beegle et al., 2012; Carletto et al., 2015; de Nicola and Gine, 2014; Fermont and Benson, 2011; Gaddis et al. 2018; Keita and Carfagna, 2009). These “gold standard” options are typically not implemented in larger scale surveys because they would be prohibitively expensive. For example, the intensively-monitored individual diary in the consumption survey experiment of Beegle et al. (2012) had variable costs that were ten times those of a household-level recall survey (which explains the interest in doing surveys by recall rather than diary). Moreover, to use these gold

standard approaches in large surveys would require infeasible levels of monitoring staff to ensure that the survey design does not devolve in unknown ways into simpler designs (e.g., interviewers doing a recall survey when they find a household with a blank diary in an income and expenditure survey but pretending that the data were appropriately entered in the diary). The gold standard approaches are possible for small samples and, of course, as part of survey experiments, but usually not for larger and multi-purpose data collection efforts. Most researchers using these approaches also acknowledge explicitly that their benchmarks may not be completely error-free.

The ideal data to characterize survey errors is to have, at the level of the unit of observation, the correctly measured data point and an independently measured (possibly) error-prone one. This is the basis of much of the validation work of Bound et al., who, in a Western setting compare survey reports of wages to tax or company records. The advantage of this setup is that the sign and magnitude of the survey recall error can be accurately described. A survey of this literature shows that reported wages typically display mean reverting error, which means that when we regress the true value on the measured value the regression coefficient will be smaller than one (see previous section). A paper that follows that approach is de Nicola and Gine (2014), who have access to company records of fish sales and of boat purchases and so are able to compare these to the self-reports made by the fishermen. The key assumption here, as in the studies described by Bound et al. (2001), is that the two measures are independent of each other. Yet as de Nicola and Gine (2014) point out in their paper, one caveat is that we cannot be sure that the responses given by the fishermen would be the same in the absence of the records being kept. Possibly a bigger problem for use of administrative records is that the survey populations of interest in development have very few administrative records keeping track of their economic lives that would be available to validate the survey reports. One possible source of such records, in countries where health systems are able to cover the population, is from child health records. For example, Sana et al. (2016) perform a validation check by asking survey respondents for formal documents, such as identity and child vaccination cards, right after the survey interview and then analyze the extent to which the information on those documents deviates from what had been recorded during the interview. Arguably, when available, this validation approach will be feasible in more formal environments (such as HR records with wages from garment factories used by Menzel and Woodruff 2019, though they do not compare with worker reports), and for information that is recorded (such as on child vaccination cards). However, in practice, at least in the near future, a reliance on independent measures for the same respondent will not be feasible for many important questions posed in the developing country context.

One possible way forward is to elicit, within a survey experiment, what should be the same measure in two different ways from the same respondent, for example with an (unbounded) recall survey given to respondents who are then given diaries to record their subsequent expenditures (Battistin et al, 2019). An obvious worry here is contamination between the two reports. If the survey participant is first asked to keep a diary in which he or she records all consumption expenditures or records how they allocated their time during the day and then, at the end of the diary-keeping period is also asked to report through recall on the same period then it seems plausible that the answers to the recall will have been informed by the fact the individual kept a diary over this same period. Instead, if they get recall first and then the diary, not only are the periods not overlapping, but having been exposed to the recall method the respondent may be less diligent at completing a

diary because they know that there is an easier (or fallback) way for the interviewer to gather the information, via a recall interview if the respondent does not comply with diary-keeping.

This problem of potential contamination comes in various levels of severity. For example, researchers engaged in the debate on the inverse farm size--productivity relationship typically have benchmark measures of land area (through compass-and-rope measurement) and/or of output (through crop cuts), which they contrast to the farmer's self-report. Here the chance of contamination is less severe than in the example of keeping a diary, because the farmer is less involved in the measurement.[7]

Another response to the contamination issue is to take multiple measures over the same household, but not for the same period. In Das et al. (2012), over several rounds of data collection, households were sometimes asked illness questions with a 4 week recall and sometimes with a 1 week recall. The periods covered by the questions do not overlap, reducing concerns of contamination.

Finally, we can randomly allocate different survey modules within the population of interest (as in Beegle et al., 2012, Gaddis et al. 2018, Garlick et al. 2019), in some cases with one module argued to be closest to the truth. The disadvantage of such randomization over time within the same respondent or across different respondents within the same enumeration area is that we cannot regress the measured outcome on the truth for the same person, in order to exactly characterize the measurement error. Therefore, other approaches need to be taken, like for example aggregating the data over household-level respondents to form village-level averages by module to study the measurement error using a within-village approach (Gibson et al., 2015).

One can also build a case that some survey methods or designs are closer to the truth, even if one does not explicitly benchmark to the truth. For example, Schündeln (2018) studies the decline in reported consumption over the course of 30 days of observation in Ghana (based on a sequence of short period recalls) from the perspective that the first weeks are closer to actual consumption than later reports that come from increasingly fatigued and non-compliant respondents (that is, the earlier reports are of "higher quality" which is also confirmed using Benford's Law). Beegle et al. (2012b) posit that farm input and production data do not get more accurate when reported several months from the harvest compared to one or two months after the harvest. Relatedly, self-reported data are taken as closer to the truth than proxy reports for labor in Bardasi et al (2011). This is also the approach, though not always explicitly stated, in several studies on asset ownership of men and women which compared proxy reports to own reports (Kilic and Moylan, 2016, and United Nations, 2019, as well as studies described in Doss et al., 2013).[8]

Some papers remain agnostic about the truth and just highlight differences in results that come with different survey designs. There can be good reason for this, for instance when there is no clear truth to benchmark against. Beaman and Dillon (2012) look at the effect of changing the definition of who

[7] Farmer involvement may still matter if farmers do not inform interviewers about far-away plots in order to shorten the interview time; if these are of different size or productivity to the nearby ones their omission may create a bias.

[8] These studies document the discordant reporting by husbands and wives on asset ownership and values, which then presents a dilemma for researchers when a survey collects contradictory information within the sample unit (in this case, two different reports for the same household). See the discussion in United Nations (2019) on the challenge of reconciliation.

is a household member, which is not something that has a single true answer. Lajaaj and Macours (2019) assess the reliability of skills measurement, a relatively fixed individual trait, by resurveying their respondents and analyzing the temporal consistency of various existing scales. In these cases simply documenting and quantifying the inconsistencies that arise is in and of itself interesting. The reverse – of different designs yielding similar results – can also be informative, particularly when survey organizations are contemplating changing the way they conduct their surveys, where such change is often to reduce field costs or to reduce data processing timelines (such as switching from diary to recall, or from pen-and-paper to tablet, as shown in Sharp et al., 2019). There is typically a lot of policy and public interest in trends for various development indicators (e.g. SDGs), and this can create concerns that a new survey design will create a break that prevents data analysts from making temporally consistent measurement of progress. Survey experiments that see if there are similar results using the "new" methods as compared to using the "old" methods can help maintain confidence in the trends across the break point. This allows a survey agency to not be forced – through concerns about maintaining temporal comparability – to stick with a historically used survey design especially when cost-saving alternatives are available.

5. INFORMING EXTERNAL VALIDITY

Survey experiments are typically rolled out at a relatively small scale, while the interest lies in applying the insights quite widely in other geographical, socio-cultural, and economic contexts. This raises concerns about external validity because the context of the experiment clearly matters. For example while Arthi et al. (2018) find that reports of the hours worked per plot per person in an end-of-season own farm labor recall module are exaggerated over 200 percent in Tanzania, compared to the benchmark in their survey experiment, a comparable experiment with an overlapping co-author (Gaddis et al., 2018) showed just a ten percent overstatement in self-reports in Ghana. We cannot conduct all experiments everywhere in order to give an experimental result for each context, so our best hope is to understand *why* respondent reports have errors, in order to translate such insights to other contexts.

In this regard, some of the recurring elements in discussions of the sources of error-prone survey reports are inference, cueing, whether respondents use an 'enumeration strategy' when they recall (that is, counting and adding in their head when answering a question), or instead they economize on cognitive resources by using an 'estimation strategy' to give a rate-based rule-of-thumb answer, the level of abstraction required in the answer, salience and interviewer and respondent incentives.

In terms of inference, a change in the recall period for a survey can change the inferred meaning of the question. Consider the phrase "illness episode". The length of the recall period may provide a cue to the respondent regarding what the interviewer means exactly by 'illness episode' and so changes in the recall period change the way that the respondent interprets the question. This phenomenon has been studied by Winkielman et al. (1998) who experimented with recall length in a question asking about episodes of anger in a sample of undergraduates at the University of Michigan. The authors found that for recall periods of one day respondents included minor irritations as 'episodes of anger', but when the recall period was extended to one year the respondents believed that the interviewer meant them to report much more serious bouts of anger.

An example of cueing is Strack et al. (1988). When they ask their sample of undergraduate students in the US first "how happy are you with life in general" and then "how often do you normally go out on a date?" there is no correlation between the answers to these two questions. But change the order of these questions and the two become correlated. An important advantage for future survey experiments is that many of them will be carried out using computer-assisted personal interviewing (CAPI), as survey efforts increasingly transition away from pen and paper surveys. It is much easier with CAPI to experiment with changing the place and order of questions, or introducing seemingly minor changes in wording to study whether it cues different responses.

Another important distinction is between recalling by counting versus estimating. Above we discussed that de Nicola and Gine (2014) find that reporting error in income increased as the period over which the income was to be recalled got longer. As a respondent is asked to reach further back into their memory for reports on events that are more distant in time, the self-reported income converges towards the mean. This pattern is hypothesized to reflect respondents relying less on explicit counting and adding the further back in time that they are asked about, and instead switching to some rate-based rule-of-thumb estimation strategy (Chang and Krosnick, 2003). This switch matters because these rule-of-thumb answers tend to understate transitory fluctuations, which generates the mean-reversion. A similar result is found by Gibson and Kim (2010) for retrospective recall of wages over even longer periods.

A similar switch in the cognitive strategy used to answer questions may occur in surveys that require a recall of expenditures; in one of the earliest papers from the developing country context that we know of, Scott and Amenuvegbe (1991) found that each additional day of recall resulted in about a three percentage point decline in reported daily expenditure, which plateaued at about 20–25% after a recall period of more than 7–10 days. One reason for this plateau, rather than for the recall error continuing to grow and reported expenditure linearly declining towards zero, is that respondents may cease trying to remember and count each expenditure occasion (what we have called an 'enumeration strategy') and instead switch to an 'estimation strategy' where they use a rule-of-thumb such as 'we eat one sack of rice a week so it must have been four sacks over the last month' (Gibson and Kim, 2007). Indeed, the use of estimation strategies may be one reason why surveys (typically these are general purpose surveys, like the PSID, rather than specialized expenditure surveys) that ask a single question about total spending, or food spending, over a long period like a month or a year still provide some meaningful data (Browning et al, 2003). It is not possible that respondents can add everything up over all the types of food or other expenditure over such a long period, so their answer must be the result of some rule-of-thumb estimate.

Often a simple heterogeneity analysis can uncover potential mechanisms for why respondent reports have errors. For example Bardasi et al. (2011) note how the large impacts of proxy reporting on male employment rates are attenuated when the proxy informants are spouses and individuals with some schooling. Heterogeneity analysis often points to the importance of formal education, especially for questions that are more abstract in nature. Beegle et al. (2012) consider a consumption module in which the respondent is asked to imagine a 'typical month' and asked to report on consumption patterns during this hypothetical period of time. Following a recommendation of Deaton and Grosh (2000) this module aims to measure permanent rather than transitory living standards, without interviewing the same households repeatedly throughout the year. In their experiment in Tanzania, Beegle et al. (2012) find that results with this 'typical month'

module deviate substantially from the benchmark results, but the accuracy of the module improves with the years of formal education of the respondent. In a stretch for relevance for surveys in rural Africa, once respondents reach PhD levels of formal education the module performs quite nicely.

It is helpful to take the salience of the event into account when designing questionnaires. Arthi et al. (2018) and Gaddis et al. (2018) both found that peripheral plots and family members who only help sporadically on the family farm are underreported in recall surveys. A survey designer who understands this could probe for these plots and individuals specifically.

It is also helpful to be aware of the respondent and enumerator incentives. One element here is the length of the survey and the extent to which it can be manipulated. With this in mind surveys often ask respondents to first enumerate (e.g. 'list all plots', 'list all household members') or go through a series of yes/no questions in a fixed list ('did you consume this item') before asking them further, follow-up, questions on each of the relevant items. The intuition is that if every additional item always comes with a series of follow-up questions then that would incentivize a time-constrained respondent to stop listing or answer 'no' and, at the margin, more so for items on the list they feel are unimportant, such as peripheral plots and workers. Without proper follow-up and supervision this phenomenon can also manifest itself through the interviewer trying to shorten the length of the interview.[9]

Survey length certainly seems to play a role. Analyzing a survey experiment in Malawi, Kilic and Sohnesen (2019) show how a short questionnaire yields different proxy-based poverty measures compared to a longer equivalent.[10] This leads them to caution against variation in length and complexity of the questionnaire across rounds in a panel or across treatment and control groups. However, not all misreporting relates to attempts to cut down time spent in the interview. Stecklov et al. (2018) show that providing a small one-time payment to respondents helped reduce unit non-response a little but also led to decreases in measured consumption. The authors interpret this as respondents presenting themselves as needy in order to justify the payment or because they expect (conditional) payments in future.

6. COST CONSIDERATIONS

A good survey experiment carefully spells out the costs of various alternatives. In some cases accurate measurement does not always need to be more expensive, or could even be cheaper. Switching from paper to CAPI (shown to yield greater accuracy) is one example, especially where tablets can be reused in subsequent surveys. Or consider the study by Schündeln (2018) which concludes that additional in-person recall interviews to collect consumption data in the 3rd and 4th week of the month-long observation period in Ghana increases mis-measurement, such that, in this setting at least, less field work would likely improve accuracy.

[9] Choumert-Nkolo et al. (2019) provide a nice overview of how monitoring the survey's paradata – those data collected semi-automatically as part of the survey process, like time stamps in CAPI surveys -- can help identify and prevent some of these issues during the fieldwork process.

[10] Gazeaud (2018) also studies the vulnerability of proxy-means testing to changes in survey design, although the focus is wider than only questionnaire length.

In other cases, accuracy comes at a cost and those additional costs need to be weighed against the benefits of increased accuracy. Consider, for example, the debate on whether plot size should be self-reported, measured with GPS, or, as the FAO (1988) puts forward as the gold standard, with a compass-and-rope (CR) measurement. Dillon et al. (2019) note that the use of GPS can, on average, require as little as 28% of the time needed for compass-and-rope measurement. Keita and Carfagna (2009) find that on small plots CR measurement can take up to 17 times as long as GPS-based measurement (noting that GPS measurement itself can be costly because it requires survey teams to be relocated to the plots and survey team time to trace the perimeter of each plot). Because of these time and travel costs, many surveys instead rely on a farmer's own estimate of land size, which avoids the cost of either the CR or GPS measurement approaches.

There are less extreme examples when what seems like a cost saver is both not much of one and also introduces errors. For example, it may be tempting to cut down on the list of food items over which a respondent is probed in consumption recall surveys, in order to shorten interview times. However, when Beegle et al. (2012) reduced the recall list from 58 food and drink groups to just 11 more broadly defined groups (designed to cover the same universe of items, and using a 7-day recall period in both cases) they found that it cut the average time of an interview by just seven minutes (requiring 42 minutes on average, rather than the 49 minutes with the 58 groups), but introduced large inaccuracies. This was in a context (Tanzania) with very little diet diversity so that questions on the quantity consumed and spending (or implicit values for non-purchases) did not apply to most food groups because the respondents did not consume any food from within that group. Whether the minor time saving from shortening the recall list also holds in other areas with more diverse diets is an open question.

Some survey experiments are able to explore using cost-saving approaches to mimic more costly and intensive survey methods that may provide a theoretical benchmark but are viewed as too expensive to scale-up outside of the experimental context. For example, Arthi et al. (2018) and Gaddis et al. (2018) both suggest that high-frequency visits during the agricultural season are a more accurate way to collect data on the farm labor cycle than is a retrospective recall after the harvest. However, such frequent visits are very expensive to field (and may cause declining compliance, as Schündeln (2018) found in the context of a ten-visit survey) and so are unlikely to be an option in most survey settings. Therefore, both studies included an arm to their survey experiments that used weekly mobile phone interviews, in addition to the most costly in-person weekly interviews, in order to explore the reliability of the phone surveys as a cheaper way of getting the desired high frequency in-season readings on farm labor use.[11]

7. CONCLUDING DISCUSSION

The increased involvement of researchers in the design and implementation of socio-economic surveys has brought with it an interest in how decisions made during these early design stages influence the quality of the data obtained and, through this channel, affect the results of subsequent analyses. In this review we have covered a nascent literature in the developing country context that

[11] They draw on lessons in Dillon (2012), who fielded high-frequency phone interviews with cotton farmers in Tanzania.

tries to gain insights into these matters by experimenting with survey methods. These experiments do this by collecting and contrasting data sets that vary by questionnaire design and survey implementation in a way that ensures that any differences in the resulting data should only be due to the choice of survey method.

This literature suggests that the effects of survey design can be huge. The large magnitude found in some studies for foundational measures of socioeconomic status, such as household consumption, labor use, and farm production, warrants a much deeper and wider appreciation by the profession on how to prevent such errors from occurring in the first place. While mitigating the impacts of the measurement error already present in existing data is better than ignoring it, it would be even better to develop surveys that should yield less error-ridden data in the future.[12] A particular challenge in this regard is the growing interest of practitioners and policy-makers in multidimensional measures of poverty and wellbeing – while we are beginning to learn about errors in survey reports on traditional welfare indicators like consumption, the variables used in multidimensional measures cover a much wider range across health, education, assets, housing conditions, sanitation, and access to water, yet little is known about the nature of the errors in some of these variables.[13]

A particular concern is that for the variables that have been studied, like household consumption and farm plot area, the error structures appear to be complex and are seldom classical in nature, limiting the usefulness of the typical econometric approaches to mitigating the impacts of errors in variables. Specifically, the studies that identify the structure of the errors have typically found them to be mean-reverting. Consequently, there are no simple adjustments that have emerged from this literature as a way to purge variables of their errors, such as using rescaling factors that differ by survey method. Nor do we expect these to become available in the future because the errors are likely to vary by survey respondent, not only in terms of their observed characteristics, such as education, but also by their unobserved characteristics that may correlate with the effects of interest. The limited scope to deal with errors when data reach the researchers desk (be it for use in traditional econometrics or in newer machine learning techniques) shifts the agenda to figuring out how to prevent these errors from happening in the course of data collection itself.

That agenda can greatly benefit from considering some of the best-practices that have emerged from the existing survey experiments. First, it is important to try to benchmark, so that when different methods yield different results we can learn something about which is closer to the truth and under what circumstances. Second, it is important to understand why errors are introduced in order to inform survey design more generally, outside the context of the particular experiment. While few of the survey experiments that we review have been able to do this, future work may usefully link to the large literature in psychology that studies the way that respondents answer questions. In particular, it would help if survey designers were more explicit about whether they have in mind that respondents are meant to count and add in order to answer their questions – what we have called 'enumeration strategies' for which designs with short and recent recall periods

[12] The current focus by applied economists on identifying causal effects has led to a relative neglect of measurement error. For example, Gibson (2019) notes that mentions of "measurement error" in economics papers declined 15% over the last decade while mentions of "identification strategy" rose 150%.

[13] For example, travel time estimates, such as self-reported access to markets, typically have non-classical errors (Gibson and McKenzie, 2007; Escobal and Laszlo, 2008) but whether the same occurs in reports of travel time to water is unknown.

and (potentially, although it remains untested) a prior visit to bound the recall period so as to reduce the impact of telescoping, may be helpful – or instead are meant to answer by using some sort of rate-based, rule-of-thumb estimation strategy, for which other design details may be helpful. Indeed, while the current trend in development economics is to take laboratory methods into the field in order to play incentivized games, there would be much to gain from also having a greater focus on what can be called field-in-lab experiments, where aspects of the field conditions that survey respondents face are studied in more controlled ways in laboratories, perhaps with the aid of measurements such as eye-tracking that can help us to understand the cognitive processes being used. The third key finding to emerge from existing survey experiments is that the cost differences of different survey methods should be carefully documented and quantified so that more informed cost-benefit choices can be made when accuracy comes at a cost.

REFERENCES

Abay, Kibrom, Gashaw Abate, Christopher Barrett and Tanguy Bernard. 2019. Correlated Non-Classical Measurement Errors, 'Second Best' Policy Inference, and the Inverse Size-Productivity Relationship in Agriculture. **Journal of Development Economics** 139(1): 171-184.

Arthi Vellore, Kathleen Beegle, Joachim De Weerdt and Amparo Palacios-Lopez. 2018. Not Your Average Job: Measuring Farm Labor in Tanzania. **Journal of Development Economics** 130: 160–172.

Backiny-Yetna, Prospère, Diane Steele, and Ismael Yacoubou Djima. 2017. The Impact of Household Food Consumption Data Collection Methods on Poverty and Inequality Measures in Niger. **Food Policy** 72(C): 7-19.

Bardasi, Elena, Kathleen Beegle, Andrew Dillon, and Pieter Serneels. 2011. Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania. **World Bank Economic Review** 25(3): 418–47.

Battistin, Erich, Michele De Nadai and Nandini Krishnan. 2019. The Insights and Illusions of Consumption Measurement: Evidence from a Large Scale Randomization. Paper presented at the IARIW-World Bank Conference: New Approaches to Defining and Measuring Poverty in a Growing World, Washington DC, November.

Beaman, Lori and Andrew Dillon. 2012. Do household definitions matter in survey design? Results from a randomized survey experiment in Mali. **Journal of Development Economics** 98(2012):124-135.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman and John Gibson. 2012. Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania. **Journal of Development Economics** 98:3-18

Beegle, Kathleen, Calogero Carletto, and Kristen Himelein. 2012b. Reliability of Recall in Agricultural Data. **Journal of Development Economics** 98(1): 34–41.

Blair, Edward, and Scot Burton. 1987. Cognitive processes used by survey respondents to answer behavioral frequency questions. **Journal of Consumer Research** 14(2): 280-288.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement Error in Survey Data. In Handbook of Econometrics, vol. 5, edited by James J. Heckman and Edward Leamer, 3705–3843. Amsterdam: Elsevier Science.

Brown, Norman. 1995. Estimation strategies and the judgement of event frequency. **Journal of Experimental Psychology** 21(6): 1539-1553.

Browning, Martin, Thomas Crossley, and Guglielmo Weber. 2003. Asking Consumption Questions in General Purpose Surveys. **Economic Journal** 113(491): 540-567.

Caeyers, Bet, Joachim De Weerdt and Neil Chalmers. 2012. Improving Consumption Measurement and other Survey Data through CAPI: Evidence from a Randomized Experiment. **Journal of Development Economics** 98:19–33.

Card, David. 1999. The Causal Effect of Education on Earnings. In O. C. Ashenfelter, & D. Card (Vol. Eds.), Handbook of labor economics. In Handbooks in economics series 5: Vol. 3A (pp. 1801–1863). Amsterdam: North-Holland, Elsevier Science.

Carletto, Calogero, Sydney Gourlay, and Paul Winters. 2015. From Guesstimates to GPStimates: Land Area Measurement and Implications for Agricultural Analysis. **Journal of African Economies** 24(5):593–628.

Chang, LinChiat, and Jon A. Krosnick. 2003. Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week". **Sociological Methodology** 33(1): 55-80.

Choumert-Nkolo, Johanna, Henry Cust and Callum Taylor. 2019. Using Paradata to Collect Better Survey Data: Evidence from a Household Survey in Tanzania. **Review of Development Economics** 23(2):598-618.

Das, Jishnu, Jeffrey Hammer, and Carolina Sánchez-Páramo. 2012. The Impact of Recall Periods on Reported Morbidity and Health Seeking Behavior. **Journal of Development Economics** 98(1): 76–88.

de Mel, Suresh, David McKenzie and Christopher Woodruff. 2009. Measuring Microenterprise Profits: Must We Ask How the Sausage is Made? **Journal of Development Economics** 88(1): 19-31.

de Nicola, Francesca and Xavier Giné. 2014. How Accurate Are Recall Data? Evidence from Coastal India. **Journal of Development Economics** 106(1): 52–65.

De Weerdt, Joachim, Kathleen Beegle, Jed Friedman and John Gibson. 2016. The Challenge of Measuring Hunger through Survey. **Economic Development and Cultural Change** 64(4):727–758.

Deaton, Angus. 2003. Adjusted Indian Poverty Estimates for 1999-2000. **Economic and Political Weekly** (January 25): 322-326.

Deaton, Angus and Margeret Grosh. 2000. Consumption. In Margaret Grosh and Glewwe, Paul (Eds.), Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study. World Bank,Washington, D.C.

Desiere, Sam, and Dean Jolliffe. 2018. Land Productivity and Plot Size: Is Measurement Error Driving the Inverse Relationship? **Journal of Development Economics** 130: 84–98

Di Maio, Michele and Nathan Fiala. 2019. Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. **The World Bank Economic Review** (forthcoming).

Dillon, Brian. 2012. Using Mobile Phones to Collect Panel Data in Developing Countries. **Journal of International Development** 24(4): 518-527

Dillon, Andrew, Sydney Gourlay, Kevin McGee, and Gbemisola Oseni. 2019. Land Measurement Bias and Its Empirical Implications: Evidence from a Validation Exercise. **Economic Development and Cultural Change** 67(3): 595-624.

Doss, Cheryl William Baah-Boateng, Louis Boakye-Yiadom, Zachary Catanzarite, Carmen Diana Deere, Hema Swaminathan, Rahul Lahoti, and Suchitra J.Y. 2013. Measuring personal wealth in developing

countries: interviewing men and women about asset values. Gender Asset Gap Project Working Paper Series, No.15. Bangalore: Centre of Public Policy, Indian Institute of Management.

Escobal, Javier and Sonia Laszlo. 2008. Measurement error in access to markets. **Oxford Bulletin of Economics and Statistics** 70(2): 209-243.

FAO (UN Food and Agriculture Organization). 1982. Estimation of Crop Areas and Yields in Agricultural Statistics. Rome: FAO.

Fermont, Anneke and Todd Benson. 2011. Estimating Yield of Food Crops Grown by Smallholder Farmers: A Review in the Uganda Context. IFPRI Discussion Paper 01097.

Friedman, Jed, Kathleen Beegle, Joachim De Weerdt and John Gibson. 2017. Decomposing Response Errors in Food Consumption Measurement: Implications for Survey Design from a Survey Experiment in Tanzania. **Food Policy** 72(1): 94-111.

Gaddis, Isis, Gbemisola Oseni Siwatu, Amparo Palacios-Lopez and Janneke Pieters. 2018. Measuring Farm Labor: Survey Experimental Evidence from Ghana. World Bank Policy Research Working Paper No. 8717. **World Bank**, Washington, DC.

Garlick, Rob, Kate Orkin, and Simon Quinn. 2019. Call Me Maybe: Experimental Evidence on Using Mobile Phones to Survey African Microenterprises. **World Bank Economic Review** (forthcoming).

Gazeaud, Jules. 2018. Proxy Means Testing Vulnerability to Measurement Errors? Mimeo. CERDI, Université Clermont Auvergne.

Gibson, John. 2019. Are You Estimating the Right Thing? An Editor Reflects. **Applied Economic Perspectives and Policy** 41(3): 329-350.

Gibson, John, Kathleen Beegle, Joachim De Weerdt and Jed Friedman. 2015. What Does Variation in Household Survey Methods Reveal About the Nature of Measurement Errors in Consumption Estimates? **Oxford Bulletin of Economics and Statistics** 77(3): 466-474.

Gibson, John and Bonggeum, Kim. 2007. Measurement Error in Recall Surveys and the Relationship between Household Size and Food Demand. **American Journal of Agricultural Economics** 89(2):473-489.

Gibson, John, and Bonggeun Kim. 2010. Non-Classical Measurement Error in Long-Term Retrospective Recall Surveys. **Oxford Bulletin of Economics and Statistics** 72(5): 687-695.

Gibson, John, and David McKenzie. 2007. Using Global Positioning Systems in Household Surveys for Better Economics and Better Policy. **The World Bank Research Observer** 22(2): 217-241.

Gillen, Ben, Erik Snowberg and Leeat Yariv. 2019. Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. **Journal of Political Economy** 127(4):000-000 https://doi.org/10.1086/701681

Gourlay, Sydney, Talip Kilic and David Lobell. 2019. A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale - Productivity relationship in Uganda. **Journal of Development Economics** 141(2019):102376.

Jerven, Morten and Deborah Johnston. 2015. Statistical Tragedy in Africa? Evaluating the Database for African Economic Development. **The Journal of Development Studies**, 51(2).

Keita, N. & Carfagna, E. (2009). Use of modern geopositioning devices in agricultural censuses and surveys: Use of GPS for crop area measurement. In Bulletin of the International Statistical Institute, the 57th Session, 2009, Proceedings, Special Topics Contributed Paper Meetings (STCPM22), Durban

Kilic, Talip and Heather Moylan. 2016. Methodological Experiment on Measuring Asset Ownership from a Gender Perspective (MEXA): Technical Report.

Kilic, Talip and Thomas Sohnesen. 2019. Same Question but Different Answer: Experimental Evidence on Questionnaire Design's Impact on Poverty Measured by Proxies. **Review of Income and Wealth** 65(1):144-165.

Lajaaj. Rachid and Karen Macours. 2019. Measuring Skills in Developing Countries. **Journal of Human Resources**, forthcoming.

Larsen, Anna Folke, Derek Headey, and William A. Masters. 2019. Misreporting Month of Birth: Diagnosis and Implications for Research on Nutrition and Early Childhood in Developing Countries. **Demography** 56(2): 707-728.

Menzel, Andreas, and Christopher Woodruff. 2019. Gender Wage Gaps and Worker Mobility: Evidence from the Garment Sector in **National Bureau of Economic Research Working Paper** no 25982.

Niederle, Muriel, and Lise Vesterlund. 2007. Do Women Shy Away From Competition? Do Men Compete Too Much? **The Quarterly Journal of Economics** 122(3): 1067–1101.

Sana, Mariano, Guy Stecklov, and Alexander A. Weinreb. 2016. A Test of the Stranger-Interviewer Norm in the Dominican Republic. **Population Studies** 70(1): 73–92.

Schündeln, Matthias. 2018. Multiple Visits and Data Quality in Household Surveys. **Oxford Bulletin of Economics and Statistics** 80(2): 380-405.

Schwarz, Norbert. 1999. Self-reports: how the questions shape the answers. **American Psychologist** 54(2): 93-105.

Scott, Christopher, and Ben Amenuvegbe. 1991. Recall Loss and Recall Duration: An Experimental study in Ghana. **Inter-Stat** 4(1): 31-55.

Serneels, Pieter, Kathleen Beegle and Andrew Dillon. 2017. Do Returns to Education Depend on How and Whom you Ask? **Economics of Education Review** 60(1) :5-19.

Sharp, Michael, Bertrand Buffière, Kristen Himelein and John Gibson. 2019. Effects of data collection methods on estimated household consumption and poverty, and on survey costs: Evidence from an experiment in the Marshall Islands. Paper presented at the IARIW-World Bank Conference: New Approaches to Defining and Measuring Poverty in a Growing World, Washington DC, November.

Stecklov, Guy, Alexander Weinbred and Calogero Carletto. 2018. Can incentives improve survey data quality in developing countries?: results from a field experiment in India. **Journal of the Royal Statistical Society Series A** 181(4):1033-1056.

Strack, Fritz, Leonard Martin and Norbert Schwarz. 1988. Priming and communication: Social determinants of information use in judgments of life satisfaction. **European Journal of Social Psychology** 18:429-442. 10.1002/ejsp.2420180505.

Sudman, Seymour, Bradburn, Norman. 1973. Effects of Time and Memory Factors on Response in Surveys. **Journal of the American Statistical Association** 68(344):805–815.

Sudman, Seymour and Norman Bradburn. 1974. Response effects in surveys: A review and synthesis. National Opinion Research Center.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. The Psychology of Survey Response. **Cambridge University Press**, New York.

United Nations. 2019. *Guidelines for Producing Statistics on Asset Ownership from a Gender Perspective.* United Nations, New York.

Visaria, Pravin. 2000. Poverty in India during 1994-98: Alternative Estimates. *Mimeo*, Institute for Economic Growth, New Delhi.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. **Journal of Personality and Social Psychology**, 75(3), 719-728.

World Bank. 2018. Poverty and Shared Prosperity 2018: Piecing Together the Poverty Puzzle. **World Bank**, Washington, DC. License: Creative Commons Attribution CC BY 3.0 IGO