DOI: 10.1111/rode.12583

WILEY

REGULAR ARTICLE

Using paradata to collect better survey data: Evidence from a household survey in Tanzania

Johanna Choumert-Nkolo | **Henry Cust** | **Callum Taylor**

Economic Development Initiatives (EDI) Limited, High Wycombe, United Kingdom

Correspondence
Johanna Choumert-Nkolo, Economic Development Initiatives (EDI) Limited, 38 Crendon Street, High Wycombe, Buckinghamshire HP13 6LA, United Kingdom.
Email: j.choumert.nkolo@surveybe.com

Funding information
This survey benefited from the financial support of the UONGOZI Institute.

Abstract

Data are a key component in the design, implementation, and evaluation of economic and social policies. Monitoring data quality is an essential part of any serious, large-scale data collection process. The purpose of this article is to show how paradata should be used before, during, and after data collection to monitor and improve data quality. To do this we use timestamps, global positioning system (GPS) coordinates, and other paradata collected from an 800-household survey conducted in Tanzania in 2016. We demonstrate how key paradata can be used during each phase of a research project to identify and prevent issues in the data and the methods used to collect it. Our results corroborate the importance of collecting and analyzing paradata to monitor fieldwork and ensuring data quality for micro data collection in developing countries. Based on these findings we also make recommendations as to how researchers can make better use of paradata in the future to manage and improve data quality. We argue for an expansion in the understanding and use of varied paradata among researchers, and a greater focus on its use for improving data quality.

KEYWORDS
data quality, face-to-face interview, paradata, timestamp, GPS, interviewer

1 | INTRODUCTION

Data quality is a public good. In recent years there has been a sharp rise in the availability of high-quality data relating to development economics. This has helped foster the growing importance of data

in the design, implementation, and evaluation of development programs and policies. This increasing use and importance of data to inform policy decisions requires that the data underlying those decisions is of high quality. Data quality is thus the focus of much attention within the field of development economics (Jerven, 2016; Jerven & Johnston, 2015[1] ; Tasciotti & Wagner, 2017). Generally, however, there has been relatively little research examining the quality of data and the methods used to collect it. As pointed out by Jerven and Johnston (2015), "much academic work on Africa regularly uses flawed data, but not all researchers demonstrate awareness of the flaws."

Recent developments to the techniques and methods used during data collection have helped in the struggle for high-quality data. This includes the increasingly widespread use of electronic surveys, and innovative research designs in the field of impact evaluation, among others (for example, randomized control trials and field experiments). Such improvements to research methods can only contribute positively to decision-making by helping to ensure that decisions are based on data acquired using the most rigorous and accurate methods. Here there is still much room for improvement, particularly in developing-country contexts, to ensure that decisions are based on accurate and reliable data.

Issues such as measurement errors, nonresponse bias, coverage bias, and sampling errors are key for researchers, and have been studied in detail in the literature (e.g., Caeyers, Chalmers, & De Weerdt, 2012; Grosh & Glewwe, 2000; Landry & Shen, 2005; United Nations, 2008). Yet, despite their potential as a powerful tool for improving data quality, "paradata" have so far been widely underused and there are very few studies highlighting their uses. The concept of paradata belongs to a longer list of data types that can be collected and used by researchers doing field work. According to Nicolaas (2011), survey data include questionnaire data, metadata, paradata, and auxiliary data. Questionnaire data are the respondents' answers; metadata include sample design and questionnaire coding instructions; auxiliary data include external data such as census data or other administrative data; and paradata include data about the data collection process, such as timestamps to capture the length of interviews or specific modules of the questionnaire, global positioning system (GPS) coordinates to track where interviews take place, and interviewers' characteristics to investigate interviewer trends.

In this paper, we focus on face-to-face surveys that are still the dominant form of interview in developing countries, although there is an increasing use of mobile phone surveys with growing mobile phone penetration rates (Demombynes, Gubbins, & Romeo, 2013). In the last decade there has been a surge in the use of electronic surveys for face-to-face interviews. This can largely be explained by the increasing awareness of the need to collect data of the highest quality, the availability of cheaper and more efficient ultramobile PCs and tablets, the availability of several computer-assisted-personal-interview (CAPI) software programs, and by the significant savings in time and costs of data collection when using CAPI (see Banks & Laurie, 2000; Caeyers et al., 2012; Carletto, Jolliffe, & Banerjee, 2015; King et al., 2013; Leeuw, 2008; Leisher, 2014; MacDonald et al., 2016; Rosero-Bixby, Hidalgo-Céspedes, Antich-Montero, & Seligson, 2005). Using CAPI technology allows researchers to access data almost instantly and provides data of better quality compared with traditional paper-based surveys (Pen-And-Paper Interviewing, PAPI) (Caeyers et al., 2012).

When researchers collect primary data, they mainly focus on the survey questionnaire data, that is, the actual responses given by the individuals interviewed. Researchers often complement these data with auxiliary data, such as administrative data or census data. Survey paradata and metadata, which are less known to development economists, are an invaluable source of information given the implications of poor quality data on the results of research and thus on decision-making.

The collection and use of paradata is not widespread when compared with the overall amount of data collected. In this paper, we present an introduction to paradata and demonstrate how they can be used: (i) during fieldwork preparation (e.g., piloting) to manage time and resources more effectively; (ii) during fieldwork to monitor data quality on a day to day basis; and (iii) after fieldwork to evaluate

data quality and potential biases. We use timestamps, coordinates, and interviewers' characteristics collected from an 800-household survey conducted in Tanzania in 2016 and explore the possibilities they offer to improve data quality.

This article is organized as follows. Section 2 presents an overview of paradata that can be used by researchers implementing face-to-face surveys in developing countries. Section 3 describes the sample, study, and methods used in the present research. In Sections 4, 5, and 6, we present examples of the paradata we collected during an 800-respondent face-to-face survey in southern Tanzania. We show how this data can be analyzed to improve the quality of all three phases of data collection and discuss the results. Section 7 concludes the paper with a review of the findings and a view to further research.

2 | WHAT ARE PARADATA?

Paradata were first introduced to the literature on surveys by Couper (1998). Simply put, paradata are data about the data collection process, such as survey timings, locations, and response rates. As such, paradata can be used to investigate measurement error, and to understand the question-answering process and usability issues with CAPI (Yan & Olson, 2013)[2] . Using paradata to monitor fieldwork also allows researchers to identify issues or idiosyncrasies developed by specific interviewers and to take actions while fieldwork is on-going. Examples of paradata are provided in Table 1.

Paradata are well-known and widely used in the field of survey methodology but are much less familiar to development economists, despite the challenges they face when collecting primary data. Indeed, development economics journals have published very few articles on data quality at the micro level, despite data being the primary working tool of most development economists. Some exceptions are for instance Caeyers et al. (2012) who compare PAPI and CAPI surveys with a randomized survey experiment among 1840 Tanzanian households and find that PAPI surveys lead to more measurement errors. Yet recently, the topic of collecting data quality and evaluating the quality of secondary data has started to receive more attention (e.g., Beegle, Christiaensen, Dabalen, & Gaddis, 2016; Jerven, 2016; Jerven & Johnston, 2015; Sandefur & Glassman, 2015). Moreover, measurement issues in surveys have been the subject of relatively more research in developing countries, for instance in the fields of agriculture (e.g., Arthi, Beegle, De Weerdt, & Palacios-López, 2018; Carletto et al., 2015; Christiaensen, 2017), consumption (e.g., Caeyers et al., 2012; De Weerdt, Beegle, Friedman, & Gibson, 2016), recall bias (e.g., Beegle, Carletto, & Himelein, 2012), questionnaire design (e.g., Oya, 2015; Randall & Coast, 2015; Rizzo, Kilama, & Wuyts, 2015), and many others.

Among the list of available paradata, timestamps are one of the most commonly collected and analyzed. Timestamps refer to questions within the questionnaire that record the time at the point when the question is selected, for example, at the start and end of a questionnaire[3] . In most CAPI software, timestamps cannot be re-entered or changed by interviewers, thereby preventing any tampering with such variables.

Timestamps provide extremely useful information that can be used to check interviewers' behavior and identify individual trends. Short interview times may imply that an interviewer is rushing, not reading all instructions, consent notes and transition statements, not reading all response options when prompted to do so, not allowing the respondent time to think, or not probing sufficiently for responses. Conversely, long interview times may imply that an interviewer is struggling to smoothly read questions, is not keeping respondents on track, or may have been interrupted during the interview (for example, by the respondent having to temporarily attend to other tasks). In either case, further monitoring, investigation and training would then need to be considered. Timestamps can be programmed at any point within a questionnaire, such as at the beginning and end of certain sections. This allows researchers to check the length of important modules and detect any interviewers that could be cutting

TABLE 1 Examples of paradata

Paradata	Measure
Timestamps	Date and time of contact
	Number of interviews per day, average interview length
	Time per question, time per section
	Interviewers' performance
	Analysis of responses according to the day or time in the day
	Field teams' workload (budgeting, human resources)
	Time between interviews
	Measurement errors (respondents or interviewers who rush/low understanding of the questionnaire resulting in a long interview)
	Interview interruptions (time gaps between sections/disturbing the flow of the questionnaire)
GPS coordinates	Track the movements of interviewers during and between interviews
	Identify coverage bias, e.g., in random walk sampling
Data correction, data entry, keystrokes	Navigation throughout the questionnaire (e.g., time, change of answers)
Counts of household visits/ contact attempts	Level of effort among interviewers
	Cost/response rate analysis
	Inform on the best time to visit respondents for future surveys and follow-up surveys
Nonresponse rate	Acceptability of the survey overall or for specific populations
	Interviewer trends
	Nonresponse bias (completed interviews, reasons for refusal, interviewer's observations, etc.)
Audio recording[a]	Audio audit, number of interruptions
Interviewers' characteristics (gender, age, experience, etc.)	Interviewers' trends on various outcomes
Random number generator	Respondent selection, order of response list, order of questions

Note. [a]Audio-recording should be used carefully and only with informed consent of respondents.

corners. Section timestamps can also help researchers identify any particularly time-consuming sections when trying to reduce questionnaire length during testing and piloting. If a significantly long section contains fewer essential variables, this can be identified as a possible section to eliminate, allowing the focus to be on the sections of the survey more relevant to the research questions.

Paradata can thus be used throughout all stages of fieldwork, during the preparation phase (fieldwork preparation, budgeting, interviewers' training, piloting and field practice), during fieldwork for in-field monitoring and in-field quality control, and post fieldwork to evaluate data quality. They provide timely and useful data on survey implementation allowing researchers to swiftly identify problems and immediately correct them.

For example, during the piloting phase, researchers can use timestamps, augmented by GPS coordinates and the number of contact attempts, to estimate the time taken to interview and travel between respondents. This would help researchers assess whether they will be able to reach their sample and

foresee difficulties such as poor road conditions, respondents not being available during field teams' working hours. During the data collection phase, paradata can be used to identify and investigate problems and unexpected situations: for example, timestamps can help identify interviewers who do not read entirely the consent note which would be against research ethics principals.

3 | PRESENTATION OF THE DATA

In November and December 2016, we implemented a field survey[4] in Tanzania with 800 respondents for a study gathering information about the perceptions of the natural gas industry (see Choumert-Nkolo, 2018).

3.1 | Sample description

We first selected the two closest regions to the gas discoveries and extraction activities in Tanzania, namely the Mtwara and Lindi regions on the southeast coast. Within the Lindi and Mtwara regions, districts were chosen if the gas pipeline runs through them or if the entire district lies to the east of the pipeline, that is, between the pipeline and the coast. The districts included in the sample are Kilwa, Lindi Rural, Lindi Urban, Mtwara Rural, and Mtwara Urban. From these districts, we randomly selected 20 wards,[5] listed all the eligible villages (or mtaa in urban wards)[6] and randomly selected one village/mtaa per ward. The final level of division was to the cluster level (subvillage level). In rural areas, this is the kitongoji/subvillage. The research design was cluster-based with a target of 640 respondents.

Households[7] were selected via a random walk methodology, following a rigorous protocol. First, field supervisors would sketch the boundary of the cluster, with the help of the local guide, and draw a grid of four evenly spaced horizontal lines and four evenly spaced vertical lines. Starting from the top left, they would number the points of intersection that fall within the boundary of the cluster 1 to 16. Each interviewer would then be randomly allocated one of the starting points. If the interviewer's starting point number was even, he/she would begin their random walk by walking in a direction towards the center of the cluster. If the interviewer's starting point number was odd, he/she would begin their random walk by walking in a direction away from the center of the cluster. To determine the number of houses to skip, interviewers used an electronic software that would calculate a skip number of houses (between one and three) using the estimated number of households in the cluster.

Once a household was selected, the interviewer listed every household member in order to determine eligibility and selected a unique respondent using a random number generator. For a member of the household to be eligible, they had to satisfy the following criteria: be over 18 years old, be knowledgeable about the household, have heard about the country's natural gas sector, and speak Swahili. If there was more than one household member who met these criteria, the respondent was selected at random[8] . We selected only one respondent per household.

The final sample contains 783[9] complete interviews. In total, field teams visited five districts, 20 wards, 20 villages and 40 subvillages/sub-mtaa. We used two teams of eight interviewers per village with the target of performing five 1-hour interviews per interviewer, per day.

3.2 | Study description

The survey aimed to understand households' perceptions of Tanzania's nascent natural gas industry; provide an overview of their awareness and knowledge of local natural gas activities; and, offer recommendations for implementing socially inclusive and sustainable policies in the gas sector in

Tanzania, in line with the principles of sustainable development, corporate social responsibility, and community engagement (Choumert-Nkolo, 2018).

The 1-hour questionnaire contained nine sections with questions on household characteristics (household roster, food consumption, asset ownership, dwelling characteristics, energy use), perceptions of main issues faced in the community, perceptions of gas operations (environmental, economic, social, and governance impacts), use of fiscal revenues from gas activities, knowledge of natural gas activities, environmental concerns and networks. In total, there were 210 questions including respondent selection, availability of respondents, and consent to the interview.

3.3 | Timestamps and other paradata collected

Throughout the interview, we used 24 timestamps to give us a detailed picture of the time taken to complete certain groups of questions. This included three visible timestamps to be entered upon arrival at a household, and at the start and end of the main questionnaire. There were also 21 hidden timestamps (automatically triggered upon answering of specified questions) used to record the times of certain sections and questions. These timestamps were included in the questionnaire to capture paradata intended for a number of uses, as detailed in Table 1.

We also captured GPS coordinates both at the start and at the end of an interview. Additionally, we collected information on interviewers' characteristics (age, gender, education level, previous surveys experience).

In the following sections we present our main contribution to the emerging literature on paradata in development economics. Using the data collected, we analyze a variety of paradata, used across all three phases of a survey: planning and preparation, data collection fieldwork, and data cleaning and analysis. This includes analysis of timestamps, GPS coordinates, and interviewers' characteristics.

4 | EMPIRICAL ILLUSTRATION OF THE USE OF TIMESTAMPS

4.1 | Planning and preparation

As part of the preparation for this survey we arranged a pilot to test the survey tool and perfect our field protocols. The pilot took place on 19 November 2016. Appendix Table A1 shows the average time taken for each interview, the average time taken for each section of the interview, and the average time taken per question for each section. Figure 1 shows the average length of interviews and the average length of selected sections[10] throughout fieldwork, excluding piloting. During the pilot, we found an average length of 113 minutes per interview, which is almost double the intended target time of 60 minutes. To reach our target number of respondents, we needed to cut the questionnaire to a more realistic length. Once we looked at the section by section breakdown of the interview, we were able to see where cuts to the question count should be made.

From Appendix Table A1 it can be seen how the length of the questionnaire changed from the piloting phase to the end of fieldwork, how the time to complete each section changed, and from which sections a total of more than 100 questions were removed. Our target of 60 minutes per interview was achieved through careful restructuring and elimination of questions, informed by section-level breakdowns of the time taken to answer.

The final version of the questionnaire contained 210 questions, which is 60% of the initial questionnaire used during piloting. This is reflected in the average time of interviews falling from 113

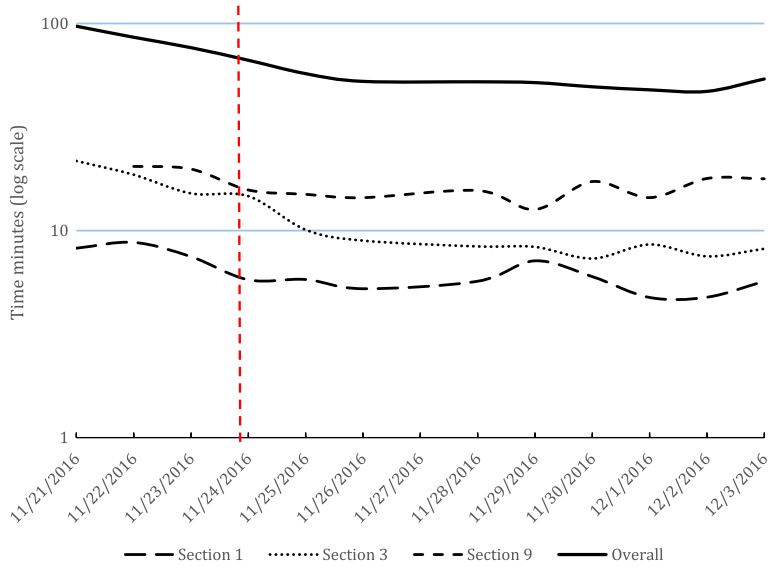

FIGURE 1 Average length over total interview and selected sections by date

Note. The red vertical line shows the date when the survey tool was finalized. $N = 783$ (completed interviews). Section 1: Introduction and randomized selection of respondent; Section 3: Household characteristics and assets; Section 9: Knowledge of natural gas, environment and networks

minutes to 51 minutes. This reduction of over 50% is evidence of two effects: the reduction in the number of questions and, the efficiency gains of interviewers.

Although we had anticipated the second effect, we knew that it would not be enough to meet our target. Question exclusion was considered on a section-by-section basis targeting those sections where there was the most to be gained in terms of time, and the least to be lost in terms of meeting our research aims. Table A1 details how the number of questions in each section changed over the first week of fieldwork.

We expected that, as interviewers became accustomed to using the questionnaire, the efficiency gains would bring the average below 60 minutes. This effect is evident when comparing interviews performed on 24 November, when the final changes were made, with interviews performed from 25 November until the end of fieldwork. On 24 November the average time per interview was 64 minutes. This figure declined to 51 minutes over the remaining days in the field as interviewers became more practiced with the questionnaire and became more efficient in asking the questions.

As can be seen in Figure 1, the steepest decline in interview duration came during the first 5 days when questions were being removed from the survey. However, even when the questionnaire became stable from 24 November, there were some time gains, particularly during the first few days of using this settled questionnaire, as the interviewers continued to improve their familiarity with the tool and their efficiency in using it. There do appear to be diminishing returns to this effect, and interviewers soon learnt the most time efficient way of conducting the questionnaire. This trend was remarkably consistent across interviewers, with every interviewer showing some decline in their average interview length between the questionnaire content being finalized and the end of fieldwork[11] .

There are many types of questions that can be asked as part of a survey and understanding how respondents answer different types can shed light on the quality of what data is being collected and can help with estimating proposed length of questionnaires before fieldwork begins. In our case we

asked a mixture of factual questions and perception questions where we found perception questions to take significantly longer than factual-based questions to answer. See subsection 4.4. for further explanation.

4.2 | In-field monitoring

Timestamps and other paradata can also be useful during fieldwork. If the paradata collected are immediately available to researchers, they can then potentially be used to identify areas for improvement, both in relation to the survey tool, and the interviewers using it.

Looking at timestamps during fieldwork can reveal whether specific sections are taking longer than expected, or if certain interviewers are struggling with certain sections of the questionnaire. This could be particularly important when interviewers are expected to carry out tasks other than just asking questions and recording answers. For example, if interviewers are asked to count the stock of medicine at a health facility, a short interview time for this section could suggest that they are estimating the count, rather than counting exactly. In such cases it may be necessary to remind interviewers of their responsibilities, or to provide additional training.

As detailed in the previous section, during the early stages of our fieldwork, the use of timestamps enabled us to identify those sections that were taking a long time to complete, relative to their research value. This enabled us to make necessary changes to the questionnaire without significantly harming the overall research.

Another way we can monitor fieldwork at an interviewer level is through the analysis of individual interviewers. This allows researchers to pick out interviewers who are comparatively fast or slow with some sections. This information can be incorporated into interviewer-specific checks to isolate reasons for interviewer idiosyncrasies.

4.3 | Protocol adherence

In addition, we used timestamps to monitor the time between interviews in relation to our random walk protocol. Table 2 provides the average time between interviews. During preparation, estimates for the expected length of time for a random walk were made at around 10 minutes. Timestamps can then be used to monitor whether fieldworkers are performing the walks as expected when not being supervised. The median length of time between consecutive interviews on the same day was 11 minutes, suggesting that our estimate was correct and any anomalous cases below 10 minutes should be investigated. Additionally, the time between the start of interviewers' first interview and the end of their last interview each day can be used to monitor the overall length of time teams are in the field, and how long we are employing local guides for. In our case an average of 5 hours and 21 minutes was observed, which is typical for this type of survey.

By combining timestamp information with GPS information, we can further investigate adherence to field protocols. See Section 5 for further explanation.

4.4 | After field to evaluate data quality

Timestamps can also be used after the conclusion of fieldwork to assess the quality of the data and give insights into the behavior of respondents. First, we review the time taken for different types of question; quantitative, factual questions about the household and its members, and perception-based questions where the respondent may have to consider their answers. Second, we look at timestamps taken on every row of a roster section of the questionnaire on the topic of household assets.

TABLE 2 Average time between interviews

Interviewer ID	Average length of time between interviews (minutes)	Median length of time between interviews (minutes)	Average distance between interviews (meters)	Total number of interviews attempted by the interviewer
630617	14.60	11.63	169	39
631329	19.95	12.68	193	40
631405	12.57	10.63	105	36
631422	23.57	20.13	128	40
631515	14.41	13.02	131	37
631521	13.15	9.11	91	38
631525	17.15	9.31	109	40
631529	13.41	7.42	90	39
631532	15.73	9.48	202	42
631538	18.12	13.30	81	39
631579	11.02	9.72	146	37
631581	14.08	7.50	96	39
631583	19.48	15.19	128	38
631591	14.56	12.18	173	45
631606	14.75	12.12	200	38
631615	17.92	8.13	84	38
Total	**15.94**	**11.47**	**134**	**625***

Note. *This number is lower than the overall sample size because the gap between the last interview of one day, and the first interview of the next day is not calculated here. Only gaps between two interviews conducted by the same interviewer, on the same day are included.

In our study we compare the time taken to answer questions of a factual type from Section 3 of the questionnaire (household characteristics) and the time-taken questions on Section 5 of a perceptive type (perceptions of gas operations). Factual type questions include questions on the construction of the household and the assets owned by the household, or more broadly any questions about the tangible things about the household that we would expect most to recall adequately. Section 3 was made up of 70 factual questions based on the household and its members including questions on personal characteristics, asset ownership, and food consumption. Perception questions relate to all questions where we ask the opinion of the respondent on a certain subject and would require time to think in many cases. Section 5 contained 48 questions on perceptions of natural gas in the community[12] . Both these sections lie in the first half of the questionnaire meaning respondent fatigue should not factor into the calculations.

The average length of factual questions was 7 seconds compared with 10 seconds for perception questions—around 45% longer for perception questions. This difference is statistically significant at the 1% significance level. Differences in the length of time taken to answer different types of questions are important when planning and preparing data collection projects with different types of question taking up a larger proportion of a respondent's time compared with others. In our example we have 89 perception-based questions that take approximately 15 minutes, whereas 89 factual-based questions would only take approximately 10 minutes. This evidence can be used in the preparation of future surveys.

Timestamps were also collected when the quantity of assets was asked of respondents. This was done for each asset, meaning a large number of timestamps were collected in a relatively short space of time. This helped to shed light on interviewers' technique with roster type sections and on the reactions of the respondent to these questions, particularly when the interviewer is not under observation. The asset roster includes ownership of: radios, televisions, mobile phones, cars or trucks, computers, and bank accounts.

There are a number of insights we can draw from this data. First, the order in which questions are answered is not always consistent. From 548 interviews where asset timestamps were consistently and correctly collected, there were 18 cases where the order in which they were answered was not the same as the order in the questionnaire. One possible explanation would be that interviewers asked all of the assets and then completed the quantities afterwards in a different order to how they asked the questions. Another idiosyncrasy could be the habits that form from the feedback from previous interviews. It is impossible from this data to determine if the questions were asked in the wrong order as well as entered in the wrong order. In our survey the order in which the assets were asked is inconsequential, however, in other surveys where the question order may be of consequence, this type of paradata analysis can be important for ensuring consistency throughout data collection.

The second insight we can gain is the time difference between the different assets that were asked about. Table 3 shows the average time it took for each asset to be asked about and answered. Interviews where questions were asked in the wrong order (18 cases) and those where an asset took longer than 30 seconds (34 cases), indicating an external interruption or a consequence of questions being asked in the wrong order, were dropped (44 cases total). A *t* test of the mean length of time taken to answer questions confirms that respondents took longer to answer questions about mobile phones and bank accounts than other assets, significant at the 1% level.

There are three potential factors at play here. First, the number of times that there were significantly more households that owned, on average, one or more mobile phones or bank accounts was greater than other assets. It would therefore take respondents slightly longer to count and check the number of assets owned for these cases. Second, these assets are potentially more sensitive for

TABLE 3 Average time and quantity from assets roster

Asset	Average time taken to ask and answer question (seconds)	Median time taken to answer question (seconds)	Number of times respondent owns at least one	Average quantity of asset owners only	Number of times asset data was collected
Radio[a]	–	–	273	1.12	500
Television	6	3	83	1.06	500
Mobile phone	7	6	374	1.86	500
Car, truck or motorbike	4	3	57	1.14	500
Computer	4	3	27	1.26	500
Bank account	6	4	102	1.82	500
Total	5	3	916	1.50	3,000

Note. [a]The assets were displayed in a list within a questionnaire roster, with radio being the first in the list. The timestamp for each asset was triggered when the respective quantity was entered. The timestamp immediately before the first one in the asset list was at the end of the previous section. The timestamp for the first asset therefore includes the time during which the interviewer was introducing and explaining the asset module to the respondent. The data for radios are therefore not comparable with the other assets in the list.

respondents to talk about, particularly for the bank account questions, and respondents therefore took more time in answering. Third, mobile phone and bank accounts are more likely to be individual items, whereas in most households, radios, TVs, cars, and computers would typically be shared between the members of the household. It could therefore take respondents a longer time to answer questions about mobile phones and bank accounts if they have to think or try to ask about the ownership of other household members.

5 | USING GPS COORDINATES

The collection of GPS data is becoming a requirement for all serious CAPI fieldwork projects (Gibson & McKenzie, 2007)[13] . Large public datasets such as demographic and health surveys now include geocoded data at the cluster level for all surveys. Here we present an additional use to those outlined by Gibson and McKenzie (2007) that are realized while data collection is taking place for assessing adherence to protocols, ensuring there is no falsified data and overall data quality checks.

5.1 | Monitoring random walks using geographic information systems

In continuation of our example case study, we implemented a random walk protocol to find households. Figure 2 presents an example of the GPS data collected at the location of each interview and the order of these interviews to assess interviewer adherence to the protocol and to assess its effectiveness. This map was created using QGIS and Google satellite photographs for the village outlines. Satellite photographs and identifiers of the village have not been included to protect the identities of respondents. Unfortunately, the quality of satellite images and the definition of our village and subvillage boundaries means a more detailed and robust analysis was impossible once fieldwork had been completed. Figure 2 presents the best example available to demonstrate the potential of GIS data in fieldwork monitoring because the village has a clear and up to date satellite photo, two clear subvillages, no additional unused subvillages and was in a rural setting with clear village boundaries.

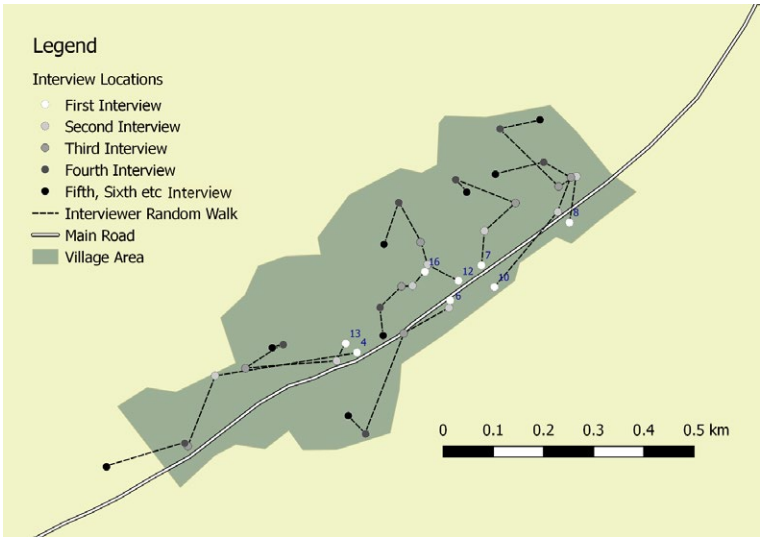

FIGURE 2 Random walk map

Figure 2 shows how GPS data can be used to review random walks and identify any idiosyncrasies that are taking hold in interviewers' random walk or household selection. It shows the paths taken by interviewers through a village in the sample area. The darker green area represents the main residential area of the village; the points, the location of interviews; the dotted line, the route interviewers took through the village; and, the numbers, arbitrary identifiers for interviewers. The order of interviews can be determined from the colour of the dots with white being the first moving through grey, then to black which is the final interview(s) of the day. The white and black line represent the main road that runs through the middle of the village. This map can be used to assess protocols for ensuring households are randomly selected, as well as to investigate patterns in interviewers' behavior in their household selection or adherence to protocols.

Without the appropriate supplementary information about household density and electronic village maps, we only feasibly assess the random walks visually. With household density, spatial analysis of the distribution of households with GPS of interviews can be performed to assess whether a suitably random sample of households was selected using the protocol. With electronic maps, adherence to random walk protocols can be formally assessed.

Selecting starting locations for our random walk was designed to give a random spread of starting locations across the village. In our example, there is a collection of starting points around the middle of the village that is unexpected, however the white points represent the first interview, not the starting location. This pattern is unexpected and if it were to re-occur in other villages would be cause for further investigation into protocols and adherence.

In addition, there are some interviews that appear to take place outside of the village. The outline of the village comes from aerial photographs from an unknown date, therefore, it is likely that the village has expanded since the photo was taken and that these interviews took place at these new households. Alternatively, it is also possible that the interview took place away from the physical household, such as in the land surrounding the household, or by taking shade under a nearby tree. This behavior was observed by field teams.

By looking at the GPS coordinates of consecutive interviews, it is also possible for researchers to calculate the distance traveled by interviewers between interviews. This is visually described in Figure 2 and numerically in Table 2. Again, this information can be used for checking effectiveness and adherence to protocols. By analyzing GPS coordinates of consecutive interviews, taking place on the same day and in the same cluster, we found that on average interviewers moved 134 meters between each interview, which is reasonable based on the random walk protocol and size of clusters. However, as noted previously, GPS points only represent successful interviews and do not take account of the precise route taken, therefore, distance between points should be considered an approximation of the route taken and distance covered by interviewers.

Future developments in mapping in developing countries should allow researchers to refine such spatial analysis of paradata. There is currently a lack of funds to create and update maps on the African continent for instance. Such maps require high-quality information and data, which is costly and demands specific skills. Currently, the African continent has a poor mapping coverage at the scale of 1:25,000. According to the African Development Bank (AfDB, 2017), this only represents 2.9% of the area of the continent, while it reaches 86.9% for Europe.

5.2 | Using start and end GPS

GPS coordinates were taken at both the start of an interview, and at the end. Table 4 displays the key statistics for our start and end coordinates. The mean difference refers to the mean absolute difference between the start and end latitude or longitude. *T* tests were conducted for the mean difference

TABLE 4 GPS coordinates at the start and end of interviews

Coordinate	Observations	Mean difference	Standard deviation	Two-sample *t* test with unequal variances (average of start vs end)
Longitude	798	0.0002317	0.0002995	$t = 0.0281$
Latitude	798	0.0003025	0.0008793	$t = -0.0560$

in the latitude and longitude to check their significance. In each case we cannot reject the null hypothesis that the difference between the start and end coordinates is zero. For latitude and longitude these minor differences are therefore explainable by measurement error, or by the interviewer moving around a respondent's home during the interview.

6 | INTERVIEWERS' EFFECTS

Interviewers' characteristics are often neglected by researchers, although they can significantly impact response rates, actual responses, and therefore the overall data quality. Personal characteristics of interviewers such as their gender, education, and survey experience (which are observable variables) can be used to check if specific traits affect survey outcomes. Personality and behaviors (such as voice, speech characteristics, social skills, and visual contact) are also likely to play an even more important role in face-to-face interviews but are more difficult to capture except via in-field direct observations by a field supervisor. Several researchers provide an analysis of interview length looking at various interviewers' characteristics (e.g., Böhme & Stöhr, 2014; Couper & Kreuter, 2013). Various methodologies can be used to analyze these effects, including cross-tabulations and multilevel or random effects models. In some research fields, such as research on willingness-to-pay for environmental goods (e.g., Bateman & Mawby, 2004), or research on sensitive topics (e.g., Anglewicz et al., 2013), interviewer effects must be seriously considered by researchers. These could impact interview length, responses to certain types of questions, and rates of refusal to participate to the survey.

As part of this study we collected a range of variables relating to the interviewers' personal background and experience. This included their age, gender, education level, and the number of research projects they had previously worked on. A summary of these characteristics is shown in Table 5.

We do not find any correlation between the interview length and interviewer characteristics[14] . Prior to this survey, all 16 interviewers had worked on a range of two to 20 surveys, with a mean of 5.5. We do not find a correlation between past survey experience and interview length. This could be explained by the fact that all interviewers took part in a 6-day training course prior to the commencement of fieldwork. More precisely, the interviewer training was attended by a total of 21 participants, made up of three supervisors and 18 interviewers. All interviewers were introduced to the project and given an initial overview of the survey questionnaire and protocols. Of the 18 interviewers trained, 16 were selected to be a part of the field teams, based on their performance in training. The training concluded with an outdoor

TABLE 5 Summary of interviewer characteristics

Age	27.6	Mean	**Education level**	12.3%	Certificate
	28	Median		6.2%	Diploma
				81.4%	University degree
Gender	61.4%	Male	**Previous surveys**	5.5	Mean
	38.6%	Female		3	Median

field practice with real respondents in order to increase interviewers' capacity to navigate throughout the questionnaire. This could suggest that interviewer training and attitude is more important than experience, which would have implications for recruitment and training procedures of future projects.

During fieldwork, researchers should also be mindful to conduct interviewer-specific checks to ensure that there are no biases in the data owing to which interviewer conducted the interview. Often, these can be very simple checks, such as for the number or rate of refusals, the frequency of answers that disable other questions or sections, and the values of key variables.

One of the key sections of this survey was the listing of household members at the start of the survey. Interviewers were asked to record the names and ages of all household members, as well as their knowledge of the household and natural gas. These questions were used to decide the eligibility of each of the household members, and for the selection of the main respondent. It was therefore vital that this section be completed accurately, so as to ensure there were no biases introduced to the data by the selection of the respondent. Table 6 shows a summary of these key variables.

The average household size is 4.45 people, which is in line with official statistics in Tanzania (Tanzania Demographic and Health Survey—TDHS, 2016). Turning to each individual interviewer, the average household size for each interviewer was within 0.8 standard deviations of the overall mean. Additionally, the average number of eligible household members for each interviewer was also within one standard deviation of the mean. This suggests that all interviewers followed the correct protocols for the listing of household members, and selection of respondents. Significant

TABLE 6 Average household size per interviewer

Interviewer ID	Total household interviews	Total household members recorded	Mean household size	Mean number of eligible household members
630617	50	213	4.26	2.06
631329	51	238	4.67	2.16
631405	47	214	4.55	1.68
631422	51	214	4.20	2.08
631515	47	204	4.34	1.77
631521	49	259	5.29	2.41
631525	51	216	4.24	2.49
631529	50	240	4.80	2.22
631532	52	231	4.44	1.92
631538	50	214	4.28	2.32
631579	48	216	4.50	1.94
631581	50	265	5.30	2.58
631583	49	209	4.27	1.90
631591	56	238	4.25	1.70
631606	49	211	4.31	1.59
631615	50	181	3.62	2.00
Mean	50	223	4.45	2.05
Std dev	2.1	21.2	2.14	0.92
Min	47	181	3.62	2.58
Max	56	265	5.30	1.59

deviations from the mean could indicate that interviewers are not following the correct protocols for listing and selecting household members, which can lead to significant respondent selection bias.

None of our interviewer-specific checks revealed any causes for concern. However, this may not always be the case. For example, Himelein (2015) tests the existence of interviewer effects for subjective and objective questions for a household survey in Timor-Leste, and find they exist in both with a stronger magnitude for subjective questions. So, for future research it is important that researchers do monitor interviewer trends during fieldwork. This can help to identify potential issues with data quality and prevent interviewer biases from affecting the data. Where such issues are evident, interviewers may require additional supervision or training, or in extreme cases be removed from data collection activities.

7 | DISCUSSION AND CONCLUSION

Data and statistics shape realities, and so having reliable data is key to informing and supporting decision-making (Desiere, Staelens, & D'Haese, 2016; Jerven, 2017). Survey paradata are a powerful tool for researchers in any field. However, within development economics there is still much work to be done to raise awareness of the uses and benefits of paradata. In this paper we have aimed to address this by presenting an overview of the types of paradata available to researchers and demonstrating their uses and potential through our survey of households in southern Tanzania. In particular, we present useful lessons relating to three key types of paradata: (i) Timestamps; (ii) GPS coordinates; and (iii) Interviewer characteristics.

Our discussion of timestamps showed how they can be useful for the planning and preparation of survey fieldwork, how they can help researchers monitor in-field activities, and how they can be used to evaluate data quality in the post-field phase. Our analysis of timestamps helped us to bring the survey length in line with our fieldwork and budgetary parameters, while preserving the quality of the overall research. It also suggested that interviewers were generally following the questionnaire correctly. While our timestamps did not uncover any particular issues, researchers should be conscious of the potential issues that may be uncovered through such analysis, such as interviewers not following survey protocols precisely, which could lead to a wide range of bias and measurement errors. All surveys should include as a minimum start and end timestamps, however we strongly recommend that future surveys make use of multiple timestamps spread throughout all sections of the survey. This can help to identify sections that need to be reduced in length or cases where interviewers are not following survey protocols or instructions.

Analysis of GPS coordinates can similarly be used to ensure that sampling protocols are followed correctly by enabling researchers to track interviewers' movements in the field. This could, for example, highlight cases where the stated random walk protocol is not being followed correctly, or where clusters have been mis-identified. In our survey, analysis of GPS coordinates showed some unexpected results, however, these were isolated cases and not of concern. With the growth of geographic information systems (GIS) and supplementary data in the developing world, the potential for geocoded data to be used during fieldwork and for analysis can soon be realized.

Finally, analysis of interviewer effects can help to uncover unwanted biases or issues in the data. While we did not uncover any individual interviewer effects, future research should ensure that such effects are monitored throughout fieldwork to prevent bias in the data collected. This is particularly important with regard to the interview length, selection of respondents, and nonresponse rates.

Beyond our own investigations there is a plethora of paradata that can be collected and analyzed from surveys, particularly those conducted using CAPI methods. For example, the field and interview

conditions can affect the efficiency of the interview and even data quality. Adverse weather conditions, particularly in areas with underdeveloped infrastructure, can prevent interviewers from reaching their samples, or cause severe delays. Even seemingly minor issues such as the comfort level of interviewers and respondents during the interview could potentially have an effect on the quality of the data collected. During our survey we did track weather conditions, however there was very little variation meaning that we were unable to detect the impact of the weather on our survey. This is therefore one area in which future surveys, taking place under more variable weather conductions, could shed more light on the potential impact of weather conditions on interview length and data quality.

As a result of our research, we make a number of recommendations for researchers using CAPI surveys. First, researchers should understand the different types of paradata and feel comfortable in their ability to collect, analyze, and use them. Second, researchers should collect a wide range of paradata in their surveys. The effort required to do so is minimal, as much paradata can be collected automatically as part of the wider data collection, yet the opportunity cost of uncollected paradata can be significant. Third, paradata should be monitored and analyzed from the very first day of field, rather than waiting until the end of data collection. This will enable the early identification of potential issues and help smooth the data collection process. Fourth, when issues are discovered from the analysis of paradata, solutions should be implemented to help improve the quality of data being collected, such as additional training on interviewing techniques or survey protocols.

While we have not been able to discuss and analyze them here, other researchers may find other types of paradata such as keystrokes, respondent contact attempts, and audio recordings useful for their own work. Future research in the field of development economics should certainly aim to take greater account of paradata and develop ways in which it can be used to improve data quality. This in turn can help to ensure that development related policies and decisions are based on the most accurate and precise data.

ACKNOWLEDGMENTS

We would like to thank the respondents who have agreed to participate in the survey carried out for this study. This survey benefited from the financial support of the UONGOZI Institute. We thank the editor and two anonymous reviewers. The usual disclaimers apply.

ENDNOTES

[1] *The Journal of Development Studies* devoted a Special Issue on the topic, entitled "Statistical Tragedy in Africa? Evaluating the Database for African Economic Development" (*The Journal of Development Studies*, Vol. 51, No. 2, 2015).

[2] Paradata are not new, but the advent of CAPI has helped collect more systematic paradata and formalize their use.

[3] It must be noted that the timestamp draws its information from the date and time settings of the actual hardware, and so it is important that these are set correctly prior to field launch and not altered during fieldwork.

[4] Questionnaires, sampling strategy and detailed field protocols are available from the author upon request.

[5] An urban ward is an administrative structure for one single town or portion of a larger town. A rural ward is composed of several villages.

[6] An mtaa is a Swahili word, best translated as an urban neighborhood.

[7] By household, we mean people who generally sleep and eat in the dwelling, who pool their resources to buy food and other necessities, and who have a common head of household who makes major decisions concerning the household's budget.

[8] We used the survey CAPI software, "surveybe" to conduct this selection. Details about the procedure and SQL codes developed are available in Choumert-Nkolo, Cust, Mallet, Taylor, and Terenzi (2018).

[9] A total of 803 households were contacted; however, three households had no available respondent, 12 had no eligible respondent, and five refused to take the survey.

[10] The sections selected were those that had an average length of over 5 minutes and were present in the tool throughout the main fieldwork.

[11] There was a small uptick in survey length at the very end (3 December 2016). There are a couple of potential explanations. The most likely is due to only eight of the 16 interviewers conducting interviews on this day. Across all completed interviews, the survey length of the eight interviewers working on this day was 5 minutes longer than the other eight interviewers. There is also a much smaller sample size on this day, with only 39 interviews being completed by eight interviewers, compared with an average of 75 interviews by 16 interviewers on other days. Additionally, this was a Saturday, so interviewers and respondents may have felt more relaxed. Fieldwork was also originally planned to have finished the day before and so some interviewers may have experienced some fatigue on this day.

[12] Questions in this section included, for example:

- Overall, today in your community, how does the gas industry impact access to water? (1) Very negative, (2) Negative, (3) Neither negative or positive, (4) Positive, (5) Very positive, (–99) Don't know.
- Overall, today in your community, what impact does the gas industry have on employment opportunities? (1) Very negative, (2) Negative, (3) Neither negative or positive, (4) Positive, (5) Very positive, (–99) Don't know.

[13] Collecting GPS coordinates has been a fairly established part of household surveys, and CAPI makes it easier to collect accurate GPS.

[14] Interviewers aged 28 or above had an average interview length of 59.7 minutes, compared with 57.6 minutes for those aged 27 or younger. Mean comparison tests between these two agegroups do not indicate significant differences. Similarly, mean comparison tests for gender and education level did not reveal any significant differences in interview length.

REFERENCES

AdDB. (2017). *Economic benefits of open data in Africa*. Abidjan, Côte d'Ivoire: Statistics Department (ECST), African Development Bank Group. Retrieved from: https://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/Economic_Benefits_of_Open_Data_in_Africa_March_2017.pdf

Anglewicz, P., Gourvenec, D., Halldorsdottir, I., O'Kane, C., Koketso, O., Gorgens, M., & Kasper, T. (2013). The effect of interview method on self-reported sexual behavior and perceptions of community norms in Botswana. *AIDS and Behavior*, *17*(2), 674–687. https://doi.org/10.1007/s10461-012-0224-z.

Arthi, V., Beegle, K., De Weerdt, J., & Palacios-López, A. (2018). Not your average job: Measuring farm labor in Tanzania. *Journal of Development Economics*, *130*, 160–172. https://doi.org/10.1016/j.jdeveco.2017.10.005.

Banks, R., & Laurie, H. (2000). From PAPI to CAPI: The case of the British Household Panel Survey. *Social Science Computer Review*, *18*(4), 397–406.

Bateman, I. J., & Mawby, J. (2004). First impressions count: Interviewer appearance and information effects in stated preference studies. *Ecological Economics*, *49*(1), 47–55. https://doi.org/10.1016/j.ecolecon.2003.12.006.

Beegle, K., Carletto, C., & Himelein, K. (2012). Reliability of recall in agricultural data. *Journal of Development Economics*, *98*(1), 34–41. https://doi.org/10.1016/j.jdeveco.2011.09.005.

Beegle, K., Christiaensen, L., Dabalen, A., & Gaddis, I. (2016). *Poverty in a rising Africa*. Washington, DC: World Bank. Retrieved from https://openknowledge.worldbank.org/handle/10986/22575

Böhme, M., & Stöhr, T. (2014). Household interview duration analysis in CAPI survey management. *Field Methods*, *26*(4), 390–405. https://doi.org/10.1177/1525822X14528450.

Caeyers, B., Chalmers, N., & De Weerdt, J. (2012). Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment. *Journal of Development Economics*, *98*(1), 19–33. https://doi.org/10.1016/j.jdeveco.2011.12.001.

Carletto, C., Jolliffe, D., & Banerjee, R. (2015). From tragedy to renaissance: Improving agricultural data for better policies. *The Journal of Development Studies*, *51*(2), 133–148. https://doi.org/10.1080/00220388.2014.968140.

Choumert-Nkolo, J. C. (2018). Developing a socially inclusive and sustainable natural gas sector in Tanzania. *Energy Policy*, *118*, 356–371.

Choumert-Nkolo, J., Cust, H., Mallet, M., Taylor, C., & Terenzi, L. (2018). *Randomising within-household respondent selection* (EDI Working Paper 2018:3). High Wycombe, U.K.: Economic Development Initiatives (EDI) Ltd.

Christiaensen, L. (2017). Agriculture in Africa—telling myths from facts: A synthesis. *Food Policy*, *67*, 1–11. https://doi.org/10.1016/j.foodpol.2017.02.002.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment, in *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 41–49). Alexandria, VA: American Statistical Association.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A*, *176*(1), 271–286.

De Weerdt, J., Beegle, K., Friedman, J., & Gibson, J. (2016). The challenge of measuring hunger through survey. *Economic Development and Cultural Change*, *64*(4), 727–758. https://doi.org/10.1086/686669.

Demombynes, G., Gubbins, P., & Romeo, A. (2013). *Challenges and opportunities of mobile phone-based data collection: Evidence from South Sudan* (World Bank Policy Research Working Paper No. 6321). Washington, D.C.: World Bank.

Desiere, S., Staelens, L., & D'Haese, M. (2016). When the data source writes the conclusion: Evaluating agricultural policies. *The Journal of Development Studies*, *52*(9), 1372–1387. https://doi.org/10.1080/00220388.2016.1146703.

Gibson, J., & McKenzie, D. (2007). Using global positioning systems in household surveys for better economics and better policy. *The World Bank Research Observer*, *22*(2), 217–241.

Grosh, M., & Glewwe, P. (2000). *Designing household survey questionnaires for developing countries*. Washington, D.C.: World Bank.

Himelein, K. (2015). Interviewer effects in subjective survey questions: Evidence from Timor-Leste. *International Journal of Public Opinion Research*, *28*(4), 511–533.

Jerven, M. (2016). Africa by numbers: Reviewing the database approach to studying African economies. *African Affairs*, *115*(459), 342–358. https://doi.org/10.1093/afraf/adw006.

Jerven, M. (2017). How much will a data revolution in development cost? *Forum for Development Studies*, *44*(1), 31–50. https://doi.org/10.1080/08039410.2016.1260050.

Jerven, M., & Johnston, D. (2015). Statistical tragedy in Africa? Evaluating the database for African economic development. *The Journal of Development Studies*, *51*(2), 111–115. https://doi.org/10.1080/00220388.2014.968141.

King, J. D., Buolamwini, J., Cromwell, E. A., Panfel, A., Teferi, T., Zerihun, M., & Emerson, P. M. (2013). A novel electronic data collection system for large-scale surveys of neglected tropical diseases. *PLOS ONE*, *8*, e74570. https://doi.org/10.1371/journal.pone.0074570.

Landry, P. F., & Shen, M. (2005). Reaching migrants in survey research: The use of the global positioning system to reduce coverage bias in China. *Political Analysis*, *13*(1), 1–22. https://doi.org/10.1093/pan/mpi001.

Leeuw, E. D. de (2008). *The effect of computer-assisted interviewing on data quality: A review of the evidence* (WWW Document). Utrecht, The Netherlands: University of Utrecht. Retrieved from http://dspace.library.uu.nl/handle/1874/44502

Leisher, C. (2014). A comparison of tablet-based and paper-based survey data collection in conservation projects. *Social Sciences*, *3*(2), 264–271. https://doi.org/10.3390/socsci3020264.

MacDonald, M. C., Elliott, M., Chan, T., Kearton, A., Shields, K. F., Bartram, J., & Hadwen, W. L. (2016). Investigating multiple household water sources and uses with a computer-assisted personal interviewing (CAPI) survey. *Water*, *8*(12), 574. https://doi.org/10.3390/w8120574.

Nicolaas, G. (2011). *Survey paradata: A review* (National Centre for Research Methods Review Papers No. 17). Swindon, U.K.: Economic and Social Research Council.

Oya, C. (2015). Who counts? Challenges and biases in defining "households" in research on poverty. *Journal of Development Effectiveness*, *7*(3), 336–345. https://doi.org/10.1080/19439342.2015.1068358.

Randall, S., & Coast, E. (2015). Poverty in African households: The limits of survey and census representations. *The Journal of Development Studies*, *51*(2), 162–177. https://doi.org/10.1080/00220388.2014.968135.

Rizzo, M., Kilama, B., & Wuyts, M. (2015). The invisibility of wage employment in statistics on the informal economy in Africa: Causes and consequences. *The Journal of Development Studies*, *51*(2), 149–161. https://doi.org/10.1080/00220388.2014.968136.

Rosero-Bixby, L., Hidalgo-Céspedes, J., Antich-Montero, D., & Seligson, M. A. (2005). *Improving the quality and lowering costs of household survey data using personal digital assistants (PDAs). An application for Costa Rica*. Paper presented at the meeting of the Population Association of America, March 31–April 2, 2005, Philadelphia, PA.

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: Evidence from African survey and administrative statistics. *The Journal of Development Studies*, *51*(2), 116–132. https://doi.org/10.1080/00220388.2014.968138.

Tasciotti, L., & Wagner, N. (2017). How much should we trust micro-data? A comparison of the socio-demographic profile of Malawian households using census, LSMS and DHS data. *The European Journal of Development Research*, *30*(4), 588–612. https://doi.org/10.1057/s41287-017-0083-6.

TDHS. (2016). *Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS-MIS) 2015–16*. Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF.

United Nations. (2008). *Designing household survey samples: Practical guidelines* (Studies in Methods, Series F No. 98). New York: Statistical Division, United Nations.

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–95). Hoboken, NJ: Wiley.

How to cite this article: Choumert-Nkolo J, Cust H, Taylor C. Using paradata to collect better survey data: Evidence from a household survey in Tanzania. *Rev Dev Econ*. 2019;00:1–21. https://doi.org/10.1111/rode.12583

APPENDIX

TABLE A1 Average time per interview section

Section	Date	Pilot (preparation phase)	Fieldwork (fieldwork phase)					Final sample (21 November to 3 December—prevailing number of questions)
		19 November	21 November	22 November	23 November	24 November	25 November to end	
All	Number of questions	348	308	249	222	210	210	210
	Average interview length (mins)	113[a]	97.02	85.79	74.39	64.49	51.26	58.84
	Average time per question (seconds)	19.53[a]	18.90	20.67	20.11	18.43	14.65	15.84
1. Selection of the respondent	Number of questions	12	12	13	13	13	13	13
	Average time taken (mins)	9.18	8.21	8.75	7.23	5.64	5.63	6.10
	Average time per question (seconds)	45.90	41.05	40.38	33.37	26.03	25.98	28.34
2. Household roster	Number of questions	20	13	15	17	20	20	19
	Average time taken (mins)	9.77	4.89	4.52	6.03	4.33	4.07	4.33
	Average time per question (seconds)	29.31	22.57	18.08	21.28	12.99	12.21	13.95
3. Household characteristics	Number of questions	105	97	84	80	65	65	69
	Average time taken (mins)	28.4	21.68	18.6	15.18	14.27	8.42	10.85
	Average time per question (seconds)	16.23	13.41	13.29	11.39	13.17	7.77	9.22
4. Community life	Number of questions	12	12	10	9	7	7	8
	Average time taken (mins)	5.51	5.11	4.83	4.17	3.85	2.77	3.23
	Average time per question (seconds)	27.55	25.55	28.98	27.80	33.00	23.74	25.37

(*Continues*)

TABLE A1 (Continued)

Section	Date	Pilot (preparation phase)	Fieldwork (fieldwork phase)					Final sample (21 November to 3 December—prevailing number of questions)
		19 November	21 November	22 November	23 November	24 November	25 November to end	
5. Perceptions of natural gas activities	Number of questions	75	74	55	48	48	48	50
	Average time taken (mins)	20.53	23.96	15.69	12.25	11.25	7.05	9.33
	Average time per question (seconds)	16.42	19.43	17.12	15.31	14.06	8.81	10.91
6. Social licence to operate	Number of questions	15	15	11	11	7	7	8
	Average time taken (mins)	4.64	3.93	3.49	3.24	3.2	1.74	2.50
	Average time per question (seconds)	18.56	15.72	19.04	17.67	27.43	14.91	19.00
7. Energy use	Number of questions	15	12	9	6	4	4	5
	Average time taken (mins)	7.43	.	4.3	3.38	3.2	2.06	2.41[b]
	Average time per question (seconds)	29.72	.	28.67	33.80	48.00	30.90	32.93[b]
8. Use of fiscal revenues	Number of questions	13	13	13	11	11	11	11
	Average time taken (mins)	5.54	.	4.87	4.28	3.8	3.06	3.33[b]
	Average time per question (seconds)	25.57	.	22.48	23.35	20.73	16.69	17.99[b]
9. Knowledge of natural gas, environment and networks	Number of questions	81	60	52	44	35	35	38
	Average time taken (mins)	21.99[a]	.	20.4	18.15	14.58	15.61	16.23[b]
	Average time per question (seconds)	17.65[a]	.	23.54	24.75	24.99	26.76	26.66[b]
	Number of interviews completed	16	48	32	71	75	557	783

Note. [a]Of 16, 8 values had to be imputed with weighted averages for those sections because of technical issues with timestamp triggers. [b]Weighted average calculated without value from the 21 November. Unable to impute values for 21 November because all relevant timestamps were missing.