Technical Bulletin No.2

March 2019

Use of Computer Aided Personal Interviews for Rural Household Data Collection: Tegemeo Institute's Experience

Tim Njagi, Mercy Kamau, Kevin Onyango, Jackson. Langat & Robert Toel

Summary

Use of Computer Aided Personal Interviews (CAPI) in data collection for household data in rural areas in sub-Saharan Africa is not common, and most often, rural household data in sub-Saharan Africa is collected through paper questionnaires. Often researchers are faced with the dilemma of plunging into CAPI use in spite of the scarce information available about them or stick to the tried and tested PAPI methods irrespective of the advantages in using CAPI. CAPI approaches hold much promise in SSA, mainly because most people have access to mobile phones, and many scholars have computers at their disposal. Unfortunately, only a minority use these methods in research and rarely do evaluate after use while those who do, do not share their experiences on public platforms. It is now three years since Tegemeo Institute transitioned from using PAPI to CAPI in data collection. We document our experiences in collecting rural farm household data using CAPI in Kenya, Uganda and Tanzania. We discuss the decisions that researchers willing to use CAPI to collect rural household data in sub-Saharan Africa have to make, and the trade-offs they will have to consider. We review the hardware, software, financial and logistical considerations that researchers must take into account to successfully implement a rural survey in sub-Saharan Africa using CAPI, and provide estimates of the costs, benefits and cost savings related to use of CAPI. Additionally, we highlight the challenges one should expect and alternative ways of addressing them in maintaining data quality and integrity.

Introduction

In sub-Saharan Africa (SSA), rural household data collection is mainly done through paper questionnaires. This method has been effective in reaching households living in remote areas where infrastructure is less developed. With the advancement of technology and improvement in basic infrastructure in parts of (SSA), it is now possible for researchers and research institutions to transition from use of paper questionnaires to Computer Aided Personal Interviews (CAPI) in collecting rural household data.

In Europe and USA, CAPI has been used for data collection of household data for the past two decades. As such, in developed countries, data collection methods were able to evolve from the Paper Aided Personal Interviews (PAPI) to Computer Aided Telephone Interviews (CATI) and CAPI. This advancement was made possible because the infrastructure was well developed with adequate telecommunications networks, up-to-date digital maps (Caviglia-Harris et al. 2012). These key features that support the use of CAPI are under-developed in SSA.

This paper, therefore, aims to share the experience of implementing rural household surveys using CAPI. We anticipate that our experience will benefit other research institutions and researchers expecting to transition to CAPI for rural household data collection.

About Tegemeo Institute

Tegemeo is a policy research Institute under the Division of Research and Extension of Egerton University. The Institute is one of the leading agricultural research institutions in SSA with over 20 years' experience in collecting and analysing rural household data. The mandates of the Institute include responding to contemporary policy issues/challenges and providing information to policymakers that can help in the formulation of appropriate policy strategies in agriculture, rural development, natural resources, and the environment. The Institute has developed into one of the leading authorities on agricultural policy research and analysis in Kenya and has become a reservoir of knowledge on rural livelihoods. Tegemeo Institute undertakes empirical research and analysis on topical agricultural policy issues and promotes policy dialogue and advocacy via the dissemination of various findings to a large number of stakeholders including government and the private sector.

Figure 1: CAPI Interview

The Institute was founded in 1988 as a collaborative project - Policy Analysis Matrix (PAM) between Egerton University, the University of Arizona and Stanford University. The PAM project focused on analysing the impact of policy on incentives and efficiency in agricultural production, processing and marketing in smallholder agriculture. In 1990, the United States Agency for International Development (USAID), through PAM, funded the Kenya Marketing Development Programme (KMDP) to undertake studies on improvement of marketing systems that provided analytical input that formed the basis for market liberalisation in the agricultural sector. In the same year, a report from Egerton University's planning workshop in Mombasa recommended the establishment of

Tegemeo Institute of Agricultural Policy and Development to replace PAM. Tegemeo was later formally established in 1995 as an Institute of Egerton University.

Tegemeo Institute is recognised for its strengths in providing evidence-based policy options, collection and maintenance of a reliable and up-to-date database on rural and urban livelihoods, objectivity in articulating research findings and other policy issues, a wide and strong network of affiliates, non-partisan nature and proximity to policymakers, donors and stakeholders. As a leading centre for agricultural policy research in Kenya, Tegemeo Institute has over the years been relied upon to provide an evidence-based analysis of the significant policy issues in Kenya's agricultural sector for improved agricultural productivity and poverty reduction.

One of the hallmarks of Tegemeo Institute has been the evidence-based research drawing from large-scale surveys of smallholder agricultural households in 22 districts of Kenya since 1997. This household panel survey data has become one of the most comprehensive and rich sources of information on smallholder behaviour in Kenya and the broader Africa. Using this extensive database from the past and on-going studies, Tegemeo is well-placed to raise the level of debate on issues affecting Kenyan agriculture and rural livelihoods. This database has been crucial in tracking key indicators in the agricultural sector and providing policymakers with valuable insights on the impact of policies implemented since the mid-1990s on rural livelihoods.

Paper Aided Personal Interviews (PAPI) Experience at Tegemeo

Until 2014, the institute relied on paper aided personal interviews, often referred to in some quarters as paper and pencil interviews, to collect household data. Over the years, the data collection protocols for using PAPI were institutionalised. The protocols which include questionnaire development and printing; training on meaning and intent of each question in the survey as well as an accurate recording of responses; tool pre-testing and enumerator performance monitoring; field data quality assurance; questionnaire conveyance; data entry and cleaning; and data archiving. These protocols were developed to ensure that the best quality achievable through this method is attained.

In a typical process, once questionnaire development had been finalised, a key activity was printing out the questionnaires. The data that the institute collects is complex rural household data. The questionnaire is usually about 25-30 pages on average. The required number of questionnaires would be printed before the commencement of fieldwork.

PAPI requires printing, organizing transport, storage, data entry and archiving of questionnaires.

It's important to institutionalise data quality protocols and ensure compliance to guarantee quality data

For large household surveys, for example in the main surveys at the Institute, about 2000 questionnaires would be printed. Transportation of the questionnaires to the field was always a challenge. Over the course of data collection, care had to be taken to ensure the questionnaires remained clean with all pages intact. Once back at the office, the bulky nature of these questionnaires always paused challenges to safe storage and archiving. A research organisation with a number of major surveys in any given year would have to plan adequately for the volume of printing, transportation of the questionnaire both to and fro the field and storage space and archiving.

Other key considerations for the data collection protocols under PAPI included:

(i) Data integrity - The nature of rural household survey questionnaires is that they contain rather complex routings which inevitably produce errors from interviewers. Detection of inconsistencies may not be immediate. If the inconsistencies are discovered late, it may lead to complete dropping of cases from the analysis. Also, the possibility for the field staff to erroneously skip questions, section or pages in a questionnaire rise with weak supervision. There are also chances of field staff recording on wrong entry

fields. Further, the data entry process are also prone to more errors related to typing mistakes or skips. Production and Productivity

(ii) In a survey, tracking of field staff performance, as well as, quality checks are paramount. Even though PAPI allows for a first level quality check by data collection supervisors, it does not give enough room for real-time data monitoring. Completed questionnaires have to be checked and reviewed in the field, transported to data centres, entered and analysed before any meaningful feedback is given to the collection team. This process does not facilitate swift response to data quality issues. To try and mitigate this, it may be advisable to have data entry clerks in the field, especially in large surveys.

(iii) When data is captured is key in ensuring quality. For smaller surveys, data entry may begin at the end of data collection at the risk of not addressing data quality issues during collection. On the other hand, data entry in the field is expensive. Research institutions must carefully consider trade-off before making a decision on data entry. Data entry templates should also be prepared well in time.

(iv) Use of PAPI allows for verification and reference during data processing. When researchers start looking at the data, a hard copy back-up of the original questionnaire is available for verification and reference.

(v) The need to collect media information such as photos or recordings, or location data will require additional devices be purchased.

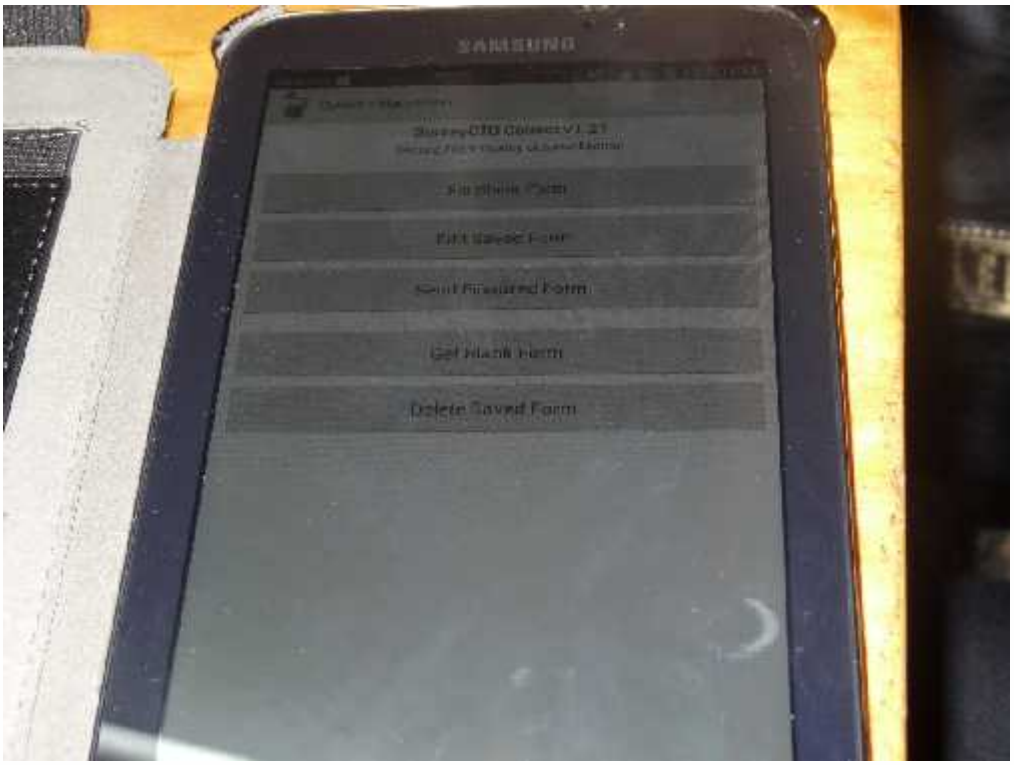

Figure 2: CAPI interface

Computer Aided Personal Interviews (CAPI) Experience

The Institute migrated to CAPI in 2014. In the beginning, we piloted on a smaller survey before rolling out to a larger study later in the year. The key lessons learnt from the smaller study were necessary for, the more extensive survey. By 2014, a limited number of platforms were available for use primarily in collecting complex data from rural households. We have documented our experience to help organisations or individual researchers minimise

errors and omissions as they transition to CAPI. We discuss the key considerations that they ought to take before making the transition.

3.1 Hardware & Software choices

The selection of the hardware is usually informed by the intended software to be used. At the moment, the available range of software for data collection is increasing and improving by the day. Similarly, the cost of equipment has been coming down especially with the growth in the number of low-cost Windows and Android devices being released to the market growing each year. For any researcher or institution that is about to embark on CAPI for the first time, the choices at the moments are much greater than two years ago.

In deciding on the choice of software, key points for consideration include:

(i) Reliability, is this software reliable and able to handle the type and amount and complexity of data intended to be collected?

(ii) Support, how easy and frequently one can get technical support in case of bugs, system failure, challenges in system design and general support during data collection;

(iii) Data transmission, whether the software allows for customisation of data transfer

(iv) Reviews from past users, reviews are great feedback to any software developer in developing and improving the user experience for any given software. A negative review should not imply failure, but one should be able to see whether future version releases have addressed the reviews. Usually, each vendor will have documentation of the released version detailing improvements.

(v) Complexity, some softwares are easy to learn, while others require an expert programmer to be able to use the software.

(vi) Cost, we have two categories: freeware, most of which require an expert user/developer to customise the software to suit specific data needs and subscribed software which is very simplified for user experience but one has to pay a period license fee

(vii) Updates - reliable software programmes keep improving the version of the program from time to time to improve the user experience. While, in some cases, the upgrades are minor - aimed at fixing serious problems with the version, at times, the upgrades are significant and require re-orientation of users to fully utilise and appreciate features that have been added or changed.

(viii) Add-on features – paid for software platforms will usually add value to open source software. Some of these features may include integrating the open source software with applications that may improve the data collection experience. Also, some of the add-on features may be about aggregating and visualising, and not necessarily in data collection. Depending on the type of data that is expected to be collected, this could be an essential factor

For hardware considerations, key points include:

(i) Warranty period and validity, although the standard warranty for many electronic gadgets is usually one year, this may differ depending on the manufacturer and product. Also, some manufacturer offer warranty that is limited to certain locations, e.g. a device with a one-year European warranty and a three month worldwide warranty. In developing countries, most of these devices are usually imported. Therefore the warranty conditions become critical.

(ii) Support, the availability of a local dealer for any equipment is a significant advantage.

(iii) Add-ons, while functions such as cameras, SIM and GPS are a must carry for high-end equipment such as tablets, they may not be available in some low-cost equipment. Depending on the data needs, the

specifications of the equipment may also change. For example, for data collection that involves capturing images, one may want to have a device that has high specs for the camera. If one is using the gadget to collect GPS data, then the accuracy required and the time it takes to pinpoint a location inform the choice of device.

(iv) Cost is probably the most important factor as it guides the equipment that will be purchased

(v) Reliability and reviews – experience from other users is also an important consideration in assessing the quality of equipment. Usually, forums on the manufacturer's or dealers' website or other online forums allow users to give feedback on the experience such as satisfaction with the product, challenges in use, what they are happy with or unhappy about particular equipment.

(vi) Battery longevity – There is a lot of emphases that is now placed on the battery capacity and how long it is expected to last. For example, newer tablets have batteries that are now projected to last 7-8 hours. However, the longevity of battery life varies with use and tablets with higher capacity batteries are likely to be more expensive. The decision point would be whether to buy battery packs or go for equipment that has high capacity. In the case of tablets, the current new tablets and phablets have a capacity of between 3000 and 5000 mAh, while the versions released two years ago had 2000 mAh.

(vii) Other factors are the version of the operating system (android or windows) version that the gadget runs on. The version of the operating system is an important consideration due to the compatibility with the intended software to be used, support, and compatibility with third-party applications.

3.2 Design and quality assurance

The design of the questionnaire is mainly influenced by the complexity of the data that is being collected and the ability of the software to accommodate the complex data requirements. However, the design of the questionnaire plays a crucial part in data quality assurance. Simple design easily allows for some quality checks on the data being corrected and can provide immediate feedback to the enumerator on the data being collected. Take the example of a household roster. To gather information on household demography, the section requires that the same set of questions be asked to all members of a particular family. A data manager will also want to control for the responses such that errors in data capture are minimised. For instance, the household should only have one household head. A check needs to in build to runs in the background such that it compares the responses of each member in the family and will only become visible if the conditions on the tests are violated. In this way, a simple design becomes complex quite fast.

There is a trade-off between the simplicity of the design and the efficiency while the questionnaire is being administered. For the check just explained earlier, if the questionnaire has very many checks, some that become quite complex, the questionnaire may move very slowly depending on the hardware option being used. Also, depending on the software being used, it may result in crashes, especially if the software is not well equipped to handle numerous complex checks. Also, there is also a penalty of the battery life in that the battery is likely to run out fast when handling complex questionnaire designs. If the quality of the hardware is low, it will be easy to notice that they tend to heat up while administering the tools.

Building these checks may require good command in programming for some software platforms. In many cases, it is not possible to find a sufficient intersection of programming and technical skills, e.g. agronomist or health within individuals or institutions. Typical agriculture sector agencies may not require these services often to hire one. On the other hand, programmers may not fully understand the needs of the technical data to build in these checks independently. As such, while a developer can quickly code a request, he is unlikely to understand the logical flow or responses to add these changes to the system.

3.3 Logistics

How the survey is organised is essential when using CAPI. First, the enumerators depending on the battery capacity of the equipment may need to be stationed where they can access electricity on a daily basis. Secondly, the need/decision to upload data on a real-time basis also affects how the survey is organised. While appreciating that mobile phone technology has penetrated many rural areas in SSA, these areas may not have good internet connectivity. Usually, we find that they may have a 2G connection. Some pockets of areas especially near markets, along highways and in rural towns will have 3G connectivity while 4G connectivity is usually found in the major cities.

Experience on transitioning from PAPI to CAPI

Transitioning from PAPI to CAPI is a great experience. How smooth the transition depends on the preparedness of an organisation regarding researching and getting as much information as possible, how they can build their skill base to handle CAPI, and support that they can rely on either from software vendors or other skilled users.

For the Institute, much of the learning was self-taught. Researchers first learnt about the tools through software provider website and interaction with the support provided by software vendors. Additionally, online forums provided an excellent and varied source of knowledge from other users' experience. Learning by doing has worked well for the institute but also takes time. Over time, the Institute has invested in online courses, and short courses, the latter which can be quite expensive especially if they are offered internationally.

However, there are common mistakes that can happen in CAPI. The first is logical errors in the code. Lack of errors in the questionnaire code does not imply that there may not be a mistake in the logical flow of the questionnaire. For example, a skip rule that becomes active if the value is less than a certain threshold, and is mistakenly captured as if the value is greater than the threshold is not easy to trace. While compiling/debugging, there will be no error to be reported. However, this code will result in missing data or capture of the wrong data. Having more than one reviewer can help trace out such errors.

It is advisable to develop the questionnaire on paper first before coding. The paper version can then be cross-checked by a number of persons to ensure that the questions in the code are consistent with what was expected to be captured.

Secondly, a very new program can have bugs that may result in wastage of time in the field, or lead to loss of data. It is, therefore, ideal to have a program that has existed for a while or pilot a very new software with a smaller survey to reduce the risk.

Cost is the primary factor for the main comparison between PAPI and CAPI.

A crucial advantage of CAPI is the efficiency in administering the questionnaire. Pre-coded skip patterns make the flow easier and faster as it filters questions based on previous responses — for example, the use of demographic questions to filter questions on schooling. Children below school-going age are automatically skipped — similarly, the use of location information to ask a site-specific question and so on. In a panel data set, it is easier to identify households as data from the previous wave can be pre-loaded with the identifying information for the family. Under PAPI, the likelihood of wrong identification of households increases.

While CAPI requires a significant capital outlay, the running costs after that are likely to be lower. In the long term, it becomes much cheaper than PAPI. A key misconception is that under CAPI, an enumerator will take less time and therefore additional savings. In our experience, there was not much significant difference in the time required

to administer the questionnaire. However, there were key cost components that will be completely gotten rid of under CAPI. During preparations, there is no heavy duty printing of questionnaires. After data collection, there is no data entry, and the costs associated with the shipping back of questionnaires to the data centre is avoided. The time taken to process the data is also shorter as data is available immediately after it is collected. Additionally, under CAPI, purchase of smartphone, tablets or phablets allows for the collection of different types of data reducing the need to buy extra equipment while spreading their cost to more than one survey.

However, the review of questionnaires under CAPI is not as assured as under PAPI. This weakness emphasises the need to have data quality checks while data is being collected.

Under both CAPI and CAPI, training of enumerators and supervisors is critical. Regarding training requirements, we may not conclude that the training is shorter under CAPI. If the recruitment process is sound, then high ability enumerators will be brought on board. They should easily understand how to operate tablets or phablets. As such, more emphasis is still on the content of the questionnaire. On the same length, the process of data is also similar under the two systems.

Conclusion

The objective of this paper is to share the institute's experience while migrating from PAPI to CAPI to encourage the use of CAPI in data collection of agriculture and rural household data. Each method has their advantages and disadvantages. PAPI allows for greater field personnel supervision and review of questionnaires during data collection. Also, maintaining hard copies of the questionnaire allows for reference and verification during data processing. However, it can suffer from data integrity, for example, if the writing is not legible, if pages of the questionnaire get lost, or if the questionnaire gets defaced. Also, data cleaning is also prone to errors and require an iteration of checks. This is corrected for by in building data checks that will return error messages during collection alerting an enumerator to the mistake in data capture. Further, collected data can be immediately accessed and check for consistency and feedback to data collection teams provided much faster as when compared to PAPI.

Use of CAPI significantly reduces the storage need as CAPI is less bulky. The costs associated with printing and courier of the questionnaires to and from the field are avoided. On the other hand, use of CAPI may require some periodic purchase of licenses and procurement of cloud storage. In our experience, training, cleaning of the data and data processing protocols are more or less the same. However, Heather (2003), notes that consistent and correct enforcement of the routeings as made possible in CAPI brings benefits in data quality and a reduction in data cleaning and editing post-fieldwork.

Data accuracy and consistency are vital in any research process, and collection methods that would guarantee these will always be preferable. Each organisation should assess their data needs to inform their transition. However, regardless of the approach used, having good enumerators and supervisors goes a long way in ensuring that the data collected is of good quality.

References

Belcher Martin, Chris Wilson, Sara Gwynn 2014 Monitoring, Learning and Evaluation ICT Capacity Review and Recommendations, Tegemeo Institute Nairobi

Caviglia-Harris, Jill, Simon Hall, Katrina Mullan, Charlie Macintyre, Simone Carolina Bauch, Daniel Harris, Erin Sills, Dar Roberts, Michael Toomey, and Hoon Cha. 2012. "Improving Household Surveys Through Computer-Assisted Data Collection: Use of Touch-Screen Laptops in Challenging Environments."

Heather Laurie 2003 "From PAPI to CAPI: consequences for data quality on the British Household Panel Study."

Smith, Tom W., and Jibum Kim. 2015 "A Review of Survey Data-Collection Modes: With a Focus on Computerizations."