A review of Census and Survey Processing System for Generating Reliable and Quality Statistics in Nigeria

I. Ahmed[1]; T. D. Okpe[2]; M. U. Adehi[3]

[1,3]Department of Statistics,
Nasarawa State University Keffi, Nigeria.

[2]Department of Mathematics and Statistics,
Federal Polytechnic, Nasarawa, Nigeria.
E-mail: ibrahimloko@gmail.com[1]; thankgodokpe@gmail.com[2]

Abstract — **This paper presents how Census and Survey Processing System (CSPro) coped with data processing system in a complex large scale survey. A recent large scale surveys used the CSPro package and SPSS for data entry fact sheet and tabulation generation for the survey. The outcomes of using CSPro and data processing system methods in a large scale survey was discussed. It is suggested that the use of CSPro and SPSS has achieved a better data quality than some other data processing packages. Some of the advantages for using CSPro in Nigeria such as improvements in data quality and turnaround times are discussed..**

Keywords - *Census Survey Processing (CSpro), SPSS, Data Quality, large surveys.*

I. Introduction

The purpose of this paper is to outline the experiences of the large scale surveys using Census and Survey Processing System (CSPro) and SPSS package for data processing in Nigeria. The survey team used the CSPro package for data processing during the conduct of the large scale surveys in recent times. This paper provides an overview of each phase of the data processing for survey from preparation to performance. It is hoped the information shared in this paper provides key insight for those national and state level survey agencies, researchers and statistical offices that may be considering using CSPro as an alternative method for data processing for surveys. [9].

CSPro is a package for entry, editing, tabulation and dissemination of census and survey data. CSPro was developed jointly by the U.S. Census Bureau, Macro International and Serpro S.A., with major funding from the U.S. Agency for International Development (USAID). CSPro includes a data entry application, a batch editing application and a tabulation application. Within the data entry application there are facilities for defining the structure of your data (the data dictionary), and creating data entry forms. During the data entry process there has the facility to verify data by retyping values and comparing them to the values entered previously. It also has the facility to compare two data files created during double data entry, and produce a report on the discrepancies found.

II. CSPRO AND SPSS FOR DATA COLLECTION AND ANALYSIS

CSPro is the most frequently used software for data entry and initial analyses of data from general surveys and population censuses. Current DHS surveys also use CSPro. However, every data file must have a Data Dictionary, even if it is only being used for simple data analysis such as constructing frequency tables for selected variables CSpro has powerful component which is used to export data files to different application files for further analysis [3]. *Statistical Package For Social Science (SPSS) is seen as one of the most powerful and widely used statistical package for both researchers and individuals.*

Main Phases of Data Collection
1. Questionnaire design
2. Pilot testing
3. Data Collection
4. Data Editing
5. Development of Electronic questionnaire
6. Data Entry
7. Data validation
8. Data processing

Main Phases of Data Processing
a. Preparation
   - Software and hardware selection
   - Develop data process cycle
   - Training in use of survey software

b. Processes
- Data processing in large scale surveys
- Software and Hardware Selection

Software Selection

An initial concern was the size and complexity of the large scale questionnaire and software could cope with a questionnaire of that magnitude. The survey data collection instrument contained 5 modules namely Household, Ever married women, unmarred women, Village and Health facilities were used. The ability to include smooth data process with national level to field agency and skip instructions, define values and incorporate rules were important criteria. The most important requirement however, was to obtain easy-to-use software for use by the questionnaire designers and ultimately the field data mangers and data entry operators. [1]

With the tabulation application in CSpro you can produce cross-tabulations and frequency tables which are useful for exploratory data analysis and error checking. When you need to transfer to a statistics package to analyze your data further, CSPro provides a useful Export feature that transfers the data to formats readable by Excel and a variety of statistical packages. If requested this feature will generate syntax files for STATA, SPSS and SAS that contain the instructions for reading the data and for labelling the variables. The above mentioned distinguished advantages, such as improvements in data quality and turnaround times it has used for recent large scale surveys in Nigeria [4].

Hardware Selection

Machine should have at least P4 processor with 256 RAM and windows XP with higher version operating system. Better to format Hard disk before start of project. Install latest virus protection on the machine.

Develop Data Process Cycle

Total process in survey system can be divided in two parts Data Entry Machine and Supervisor machine the following is the process diagram will give clear picture.

The above listed instruction manuals is helpful guide field agency data processing in charge staff to smooth data process and generate reliable data from surveys and experiments.

Survey Data Process

Main activities of survey data process is listed below:

Receiving a Questionnaire from Field
Questionnaire Editing
Data Entry
Verification
Comparison
Validation
Coverage Report
Field Check and Table Generation
Sending Data

Once questionnaire arrives in the central office, it is kept intact. One office editor will be responsible for editing and coding of questionnaire of a complete EA, one data entry operator will enter all the questionnaires for initial keying and a second keyer will enter the questionnaires for verification, and so on. Each survey data will be saved in a separate data file for that state, rather than into one large data file for the entire country. This is to protect against loss of data due to hardware or software failure.

After the data entry operator completes the verification, a report with discrepancies in both data file will be generated. The can be done after comparing both the files and correction must be made in verified data, which will be used for further process cycle.

The data entry application has been designed in such a way that it will be able to identify any inconsistency at the time of entering the data. However, there are some inconsistencies that require the attention of subject matter specialists or senior staff to resolve them. The result of executing this option is a report listing these types of inconsistencies. The report should be printed out and passed along with the package the complete questionnaires to an editor. The second level editor will analyze the messages and decides whether data need to be changed or not. [6] The second level editor will use the "Error Manual" to identify and properly resolve the inconsistencies. If the editor decides that no more changes are required or if after running the secondary editing no more messages are displayed. The data will kept in final folder, as soon as all the processes on a state data get over. He should update the log after uploading the data on file transfer protocol (FTP) site or any storage media for purpose of storage and security of information [1].

Data processing is, generally, "the collection and manipulation of items of data to produce meaningful information." In this sense it can be considered a subset of information processing, "the change (processing) of information in any manner detectable by an observer.

Data processing may involve various processes, including:

I. Validation – Ensuring that supplied data is correct and relevant.
II. Sorting – "arranging items in some sequence and/or in different sets."
III. Summarization – reducing detail data to its main points.

IV. Aggregation – combining multiple pieces of data.
V. Analysis – the "collection, organization, analysis, interpretation and presentation of data."
VI. Reporting – list detail or summary data or computed information.
VII. Classification – separation of data into various categories.

Data analysis uses specialized algorithms, software and other statistical calculations that are less often observed in a typical general business environment. For data analysis, software suites like **SPSS** or **SAS**, or their free counterparts such as **DAP** or gretl are often used. However, CSpro and SPSS have been found to be most widely used in the data production process as it allows for flexibly data validation and Analysis [7].

CSpro Support Mobile Data Collection
Mobile data collection or mobile surveys is an increasingly popular method of data collection. Over 50% of surveys today are opened on mobile devices. The survey, form, app or collection tool is on a mobile device such as a smart phone or a tablet often times called Computer Assisted Personal Interview (CAPI) which is designed at the development stage using Cspro with all skip instructions and response option embed in the application.

These devices offer innovative ways to gather data, and eliminate the laborious "data entry" (of paper form data into a computer), which delays data analysis and understanding. By eliminating paper, mobile data collection can also dramatically reduce costs: one World Bank study in Guatemala a 71% decrease in cost while using mobile data collection, compared to the previous paper-based approach. With application like Cspro researcher can be performing different task at the same time which include: Data collection, data validation, data entry and editing which will have come as different phase of data production before analysis [4].

III. USING CSPRO

CAPI surveys using CSpro are faster, simpler, and cheaper However, lower costs are not so straightforward in practice, as they are strongly interconnected to errors. Because response rate comparisons to other survey modes are usually not favorable for CAPI surveys, efforts to achieve a higher response rate (e.g., with traditional solicitation methods) may substantially increase costs [10].

1. The entire data collection period is significantly shortened, as all data can be collected and processed in little more than a month
2. Interaction between the respondent and the questionnaire is more dynamic compared to e-mail or paper surveys. Online surveys are also less intrusive, and they suffer less from social desirability effects'
3. Complex skip patterns can be implemented in ways that are mostly invisible to the respondent.
4. Pop-up instructions can be provided for individual questions to provide help with questions exactly where assistance is required.
5. Questions with long lists of answer choices can be used to provide immediate coding of answers to certain questions that are usually asked in an open-ended fashion in paper questionnaires
6. Online surveys can be tailored to the situation (e.g., respondents may be allowed save a partially completed form, the questionnaire may be preloaded with already available information, etc.).
7. Online questionnaires may be improved by applying usability testing, where usability is measured with reference to the speed with which a task can be performed, the frequency of errors and user satisfaction with the interface

Fig. 1. CAPI questionnaire used for data collection (Holding Questionnaire)
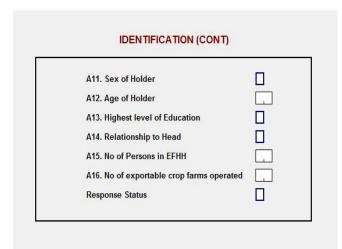
Fig. 2. CAPI questionnaire used for data collection (Identification Session)

SPSS (Statistical Package for Social Sciences)

SPSS is one of the most popular data analysis software packages. It supports various statistical methods and procedures. SPSS was first developed in 1968 at the Stanford University for internal use only. Starting from March 2009, the name SPSS had been changed to **P**redictive **A**nalytics **S**oft**w**are (**PASW**) Statistics.

In July 2009, the company which owned PASW announced that it was being acquired by IBM. As of January 2010, it became "SPSS: An IBM Company". By October 2010, IBM SPSS was fully integrated into the IBM Corporation. Recent versions of SPSS Statistics can handle multiple data sets with an almost unlimited number of variables and cases. It allows data and outputs to be imported and exported using a variety of formats including Microsoft Excel and various text formats. Users can operate SPSS through a menu (and dialog box) driven graphical interface as well as command line (syntax) interface.

SPSS is user-friendly. Even beginners can do basic statistical analysis with the software. It offers excellent on-line help, complete users' manuals and self-learning tutorials. The package supports almost all statistical methods, which allows users to perform basic to advanced analysis on data sets. SPSS has good support for data management and data documentation. Data files from CSpro after being exported are ready for analysis using SPSS as all the variables have been properly defined and need not to be defined or redefine except on cases where new data needs to be generated from the original data file to create new variables. The majority of household surveys are analyzed with SPSS, and/or many final survey data sets are available in SPSS (*.sav) format. Microsoft Excel can also be used to

format and arrange the outputs from SPSS analysis and make presentations and reports [8].

Table 1. SPSS output after exporting the data from CSpro.

Cross tabulation Mathematics is a challenging subject * Student Performance			Student Performance		Total
			Fail	Pass	
Mathematics is a challenging subject	Strongly Agree	Count	22	42	64
		% of Total	9.2%	17.5%	26.7%
	Agree	Count	38	57	95
		% of Total	15.8%	23.8%	39.6%
	Neutral	Count	7	9	16
		% of Total	2.9%	3.8%	6.7%
	Disagree	Count	26	21	47
		% of Total	10.8%	8.8%	19.6%
	Strongly Disagree	Count	6	12	18
		% of Total	2.5%	5.0%	7.5%
Total		Count	99	141	240
		% of Total	41.2%	58.8%	100.0%

Source: *Project on Student Believe and their performance in mathematics, 2018*

IV. Conclusion

The complexity of some recent Survey made data collection a serious challenge, In-house staff and field workers including researcher faced with innovation and determination. Upon reflection, it was quite evident that strict survey controls were needed for the data process operations, such as weekly reporting for complex surveys, constant follow-up and increased observation of the recent survey data processing which can be done using CSpro as one of the instrument of data collection. However, the following outlines some of the benefits that can be assumed by using the CSPro package.

Skip instructions, ranges and validation are built into the program in the data entry software, provides greater security for data, improved data quality and large amount of survey data in limited time of delivery after which Statistical Package for Social Science (SPSS) can then be employed to analyze complex statistical models for better decisions. The recent survey are optimistic using CSPro as an efficient method of data process in large scale survey. As experience with the technology grows, its benefit hopefully can be exploited over a wide range of projects. National Bureau of Statistics (NBS) today uses CSpro and SPSS for complex surveys which could have taken years and time to plan and implement but was done within a short period of time and little human resources. It is therefore recommending that scholars, academia, researchers, individuals, companies and organizations, schools, NGOs to put more effort in research and development while sensitizing on the need to use CSpro and SPSS for data collection and analysis.

Acknowledgment

References

[1] Biffignandi, S. (2012). Handbook of Web Surveys. Wiley Handbooks in Survey Methodology. **567**. New Jersey: John Wiley & Sons. ISBN 978-1-118-12172-6.

[2] Bohme, Frederick; Wyatt, J. Paul; Curry, James P. (1991). 100 Years of Data Processing: The Punchcard Century. United States Bureau of the Census.

[3] Dillman, D.A. (2006). Mail and Internet Surveys: The Tailored Design Method (2nd ed.). New Jersey: John Wiley & Sons. ISBN 978-0-470-03856-7.

[4] Groves, R. M. (2001). Survey errors and survey costs. New York: John Wiley & Sons. U.S. Census Bureau and Macro International http://www.census.gov/ipc/www/cspro/doc.html.

[5] Mavletova, Aigul; Couper, Mick P. (2013). "Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference?". Survey Research Methods. **7** (3): 191–205. doi:10.18148/srm/2013.v7i3.5458 –via ojs.ub.uni-konstanz.de.

[6] Mellenbergh, G.J. (2008). "Surveys". Advising on Research Methods: A consultant's companion. Huizen, The Netherlands: Johannes van Kessel Publishing. pp. 183–209. ISBN 978-90-79418-01-5.

[7] Quantitative research methods in educational planning Web:http://www.sacmeq.org and http://www.unesco.org/iiep.

[8] Statistical Services Centre, University of Reading web: http://www.reading.ac.uk/ssc

[9] Truesdell, Leon E. (1965). The development of punch card tabulation in the Bureau of the Census, 1890. United States Department of Commerce.

[10] Vehovar, V.; Lozar Manfreda, K. (2008). "Overview: Online Surveys". In Fielding, N.; Lee, R. M.; Blank, G. The SAGE Handbook of Online Research Methods. London: SAGE. pp. 177–194. ISBN 978-1-4129-2293-7.

[11] Okpe T.D. (2018). Statistical Analysis of Student Believe and performance in mathematics in Polytechnic Academy Nasarawa and Government Secondary School (GSS). A paper presented at the 5th School of Applied Science National Conference, Federal Polytechnic Nasarawa.