IARIW – CIGI 2023

IARIW – CIGI 2023

Thursday, November 2 – Friday, November 3

Data Valuation for Knowledge Sharing and Decision Making

Gabriel Kulomba Simbila
(National Bureau of Statistics of Tanzania)

Paper prepared for the Conference on The Valuation of Data November 2 – November 3, 2023
Session 9: Policy Implication of the Valuation of Data
Time: Friday, November 3, 2023 [2:30-3:30PM EST]

Data Valuation for Knowledge Sharing and Decision Making

Gabriel Kulomba Simbila (National Bureau of Statistics)

IARIW-CIGI Conference
November 2-3, 2023, Waterloo, Ontario, Canada

Digital data collection method has become popular in developing countries including Tanzania. The mandate of statistical offices is to collect data that are as accurate as possible. This paper examines the pros and cons of two methods of data collection: digital and paper-based data collection. The two methods were used national panel and integrated labour force surveys. The paper also examines the sustainability of data collected on data entry, data checks, timeliness, accuracy, omission, and errors. The statistical software will evaluate the two survey datasets and transform the data to normal distribution. Thereafter, challenges and lessons learned will be discussed in the paper.

Keywords: *Digital and Paper based data collection*

Introduction:

Data collection is an important statistical activity that facilitates informed decision-making and proper management of various tasks in an organization. It is, therefore, necessary to ensure the availability of quality data at all stages of the statistical production process.

This paper advantages and disadvantages of two approaches to data collection: computer-assisted personal interviewing (CAPI) vis-a-vis personal assisted personal interview (PAPI) is expected to improve survey data quality since the right questions are asked to the right people through the automatization of question skips and filters, potentially eliminating inconsistencies and range errors in data (Adalı et al., 2022).

In CAPI the interviewer reads questions from a handheld device, preloaded with the questionnaire, to the respondent. Thereafter, the respondent's answers are immediately entered into the device. This technique is expected to minimize the risk of committing human and instrumental errors. It also eliminates the need for manual re-keying of the data (Adalı et al., 2022).

Computer-assisted interviews save the costs of separate data entry methods, reduce processing errors, ensure that all correct questions are asked through the automatization of question skips and filters, decrease interviewer burden, and increase data quality (Adalı et al., 2022). The computer also automates the routing through the questionnaire and enables the interviewer to prepare a set of consistency checks during the interview, so that anomalies can be resolved with the respondent. These and numerous other features are believed to improve data quality, but it is unclear to what extent they actually do so and how this affects analysis (Bet Caeyer et al., 2010). Under the paper-assisted personal interview methodology, feedback or answers provided by survey respondents are recorded on paper forms or questionnaires by field enumerators (Development Bank, 2019).

CAPI can show improved quality and efficiency in the collection and management of data compared to PAPI, although initial programming investments in CAPI are costly (Mergenthaler et al., 2021).

The Problem

Several problems are associated with paper-based data collection. They include the use of quantities of paper which, besides costing a lot of money, is harmful to the environment. Moreover, paper-based data collection involves a lot of money in printing, transportation, and storage costs. There is also a problem of missing data points in the fields,

a data collector can leave a point blank, other challenges include, handwriting, data entry errors, and lack of timeliness. Thus, it is difficult to make changes in the field while using paper-based methods, and data collection may not be possible during field restrictions as was during the COVID-19 pandemic.

The duration of interviewing is one of the important considerations in designing a survey to collect quality data, but it differs in CAPI and PAPI methods. These differences, however, are not well explained in the literature (Nix1, 2014)

Another issue in data quality is missing data due to non-response or incomplete coverage; which is a real bane to researchers (Enders, 2010). It is understandable that researchers routinely employ missing data handling techniques that are objectionable to methodologists while the technical nature of the missing data literature is also a significant barrier to the widespread adoption of maximum likelihood and multiple imputation (Enders, 2010).

Data and variables

The National Panel Survey (NPS) and Integrated Labour Force Survey (ILFS) were carried out alternately by the National Bureau of Statistics in 2014/2015, 2020/21 (NPS), and 2014 and 2020 (ILFS). Both surveys are representative of the entire population. The NPS used about 2,872 and 3,352 households, the ILFS used 5120 households in each survey.

There are a lot of variables from these surveys, but the paper focuses on sex, age, literacy, boarding schools, and school ownership.

Methodology:

A: Data analysis in missing values

(Rubin, 1976) classified missing data problems into 3 categories. In his theory, every data point has some likelihood of missing. The process that governs these probabilities is called the missing data mechanism or response mechanism

Missing Data Assumptions

- Complete Data
- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

MCAR – Missing does not depend on data

MAR -Missing information is influenced by observed data

MNAR –Missing information is also influenced by unobserved data.

In order to identify the magnitude of missing limitations in the data and their patterns for the two applied methods, the statistical test known as Little’s Missing Completely at Random (MCAR) will be used to identify the pattern of missingness identification (Enders, 2010). Little’s procedure is a global test of MCAR that applies to the entire data set (Enders, 2010). Little’s test evaluates

mean differences across subgroups of cases that share the same missing data pattern (Enders, 2010). The test statistic is a weighted sum of the standardized differences between the subgroup means and the grand means, as follows:
Little's MCAR test Chi-Square
Assumption:

H0: The Data are MCAR
H_1: The Data are not MCAR

- Significance level (α=0.05)
- Test Statistics
- P-Value

If the P-value < Significance level: Reject H_0 and if the P-value> Significance value: No evidence to reject H_0

$$d^2 = \sum_{J=1}^{J} n_j \left(\widehat{\mu_J} - \widehat{\mu_J^{ML}}\right)^T \hat{\Sigma}_j^{-1} \left(\widehat{\mu_J} - \widehat{\mu_J^{ML}}\right)$$

Results

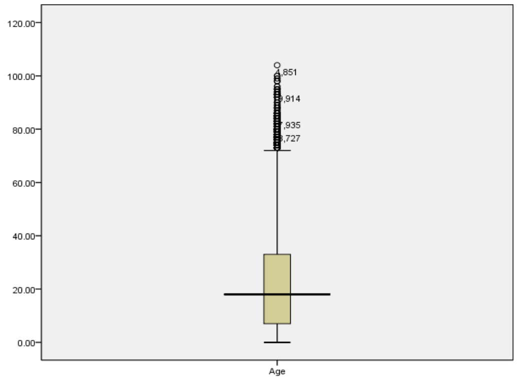
	Wave 4 (PAPI)	Wave 5 (CAPI)
Test statistic	5.013	6.704
P-value	0.025	0.01
Df	1	1
Missing Patterns	15,846 (36,3%)	17,017 (34.4%)
Total Cases	43,623	49,535

Source: *Authors analysis*

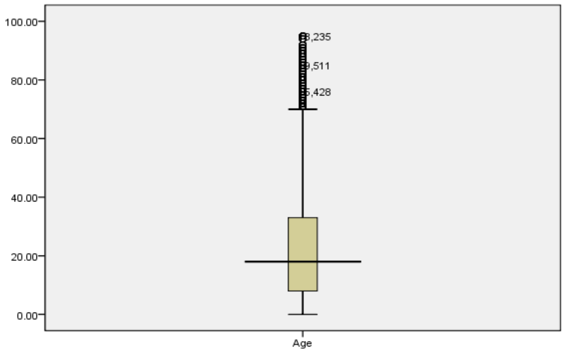
The little test statistic for MCAR shows that the p-value is below the significance P< 0.05 threshold in both two methods, therefore, the null hypothesis is rejected and concludes that the data are not missing completely at random.
The results show that there are 15,846 and 17,017 missing data representing 36.3% and 34.4% of total datasets of the selected variables. with a P-value of 0.025 in wave 4 and 0.01 in wave 5. The PAPI method is more prone to missingness than CAPI.
The research paper also explores whether these data are normally distributed and if not normally distributed the parametric statistics computation analysis can be applied.

The Boxplot

PAPI Data method



CAPI Data method



the PAPI data method as presented in a Boxplot shows the median age is 17 to 19 which represents the mid-point of the data shown by the line that divides the box into two parts, the interquartile range is about 25 ages which means that about 50% of the country has ages between 10 to 35 years. The number given by circles appears to be outside of the data representing about 25 percent of the age in the country and the value that is very far away from the remaining values indicates a greater concern and should be examined more closely to see if they appear to be reasonable values. This data is largely skewed right because the whisker and half-box are longer on the right side of the median than on the left side. The 23 percent of the age might be an outlier. There is a need to get more details about the data. The CAPI data method indicates a similar pattern, the median age is 19 to 20, and the interquartile range is about 23 ages about 50% of the country has ages between 12 to 36 years, the data is also largely skewed right representing 25 percent of the age that might be an outlier. Data needs to be double-checked again to make sure that the data are legitimate.

The research paper also examines both data methods using one of the three approaches of normal distribution tests to check the skewness and kurtosis in order to ensure data quality. Skewness is the measure of the symmetry of the data distribution, and kurtosis is the measure of the peak of the distribution.

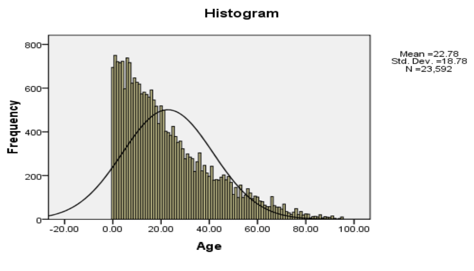
Descriptive Statistics

CAPI DATA		PAPI DATA	
Skewness	Kurtosis	Skewness	Kurtosis
Statistics	Statistics	Statistics	Statistics
1.038	0.583	1.070	0.690

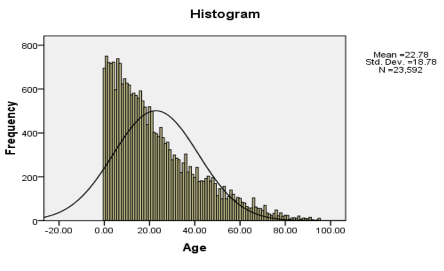
Source*: Authors Analysis*

The descriptive table above shows that the skewness for both method data are 1.038 and 1.070 which are much higher than the normal skewness of 0 value close to 0 between -1 and 1 close to 0. Similarly, the kurtosis values for both method data are 0.583 and 0.690 which are far above 0. These results the data is not normally distributed.

CAPI DATA



PAPI DATA



The two originated datasets show a non-normal distribution graph, all are skewed and require conduct of a parametric and non-parametric statistic.

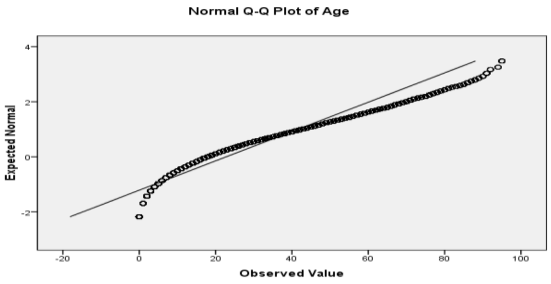
Tests of Normality

	CAPI DATA			PAPI DATA		
Age	Kolmogorov-Smirnov[a]			Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.	Statistic	Df	Sig.
	0.114	23592	0.000	0.118	21027	0.000

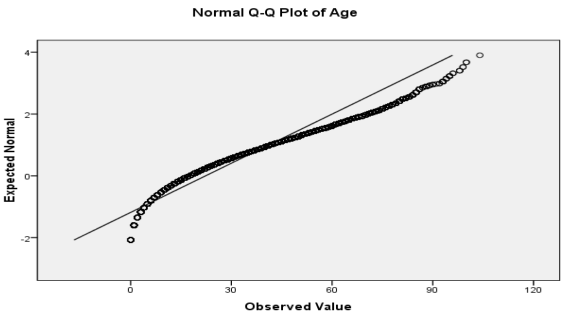
Source: Authors Analysis

The Kolmogorov-Smirnov[a] tests show that in both method data, the significance value is 0.000 which is less than the Significance value of 0.05. When the significance value is less than 0.05, we reject the null hypothesis and conclude that data are not normally distributed, in the Q-Q plot reveals the same results

PAPI DATA METHOD



CAPI DATA METHOD



For both methods, the expected value and observed value do not fit the line, there are a few dots away from the lines, and the value of kurtosis also indicates this. We conclude that the data is not normally distributed as is often the case with survey data. Therefore, therefore, it is important to do parametric statistics tests and ensure that the data does not have outliers. A data transformation is a mathematical procedure that can be used to modify or adjust variables that violate the statistical assumptions of normality.

The data transformation of the non-normal distribution of a variable is based on the variable of age series, using one of the three approaches: log difference, sqrt method, and inverse method.

The sqrt method:

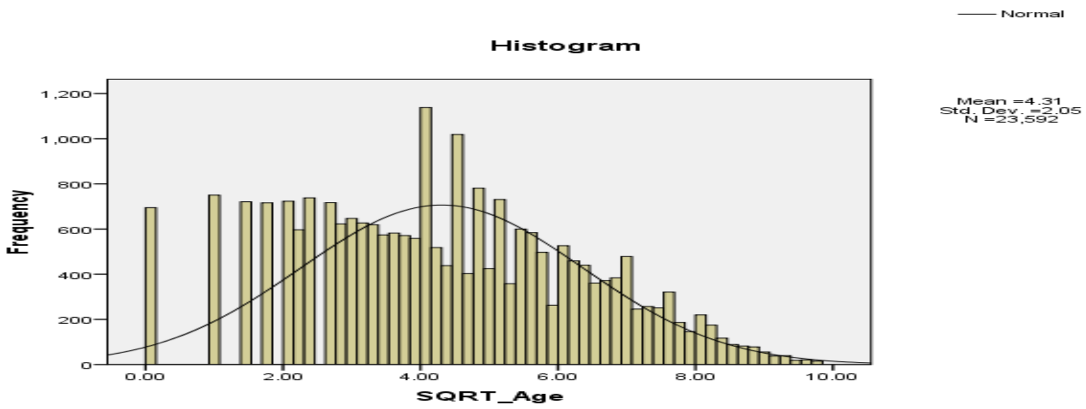
Descriptive Statistics

SQRT_Age		
N	Valid	23592
Skewness		.098
Std. Error of Skewness		.016
Kurtosis		-.533
Std. Error of Kurtosis		.032

Source: *Authors Analysis*

The Skewness value is 0.098 which is closer to 0 and the kurtosis is -.533 which is also close to 0, we conclude that the data is normally distributed.

In the histogram, it appears to be perfect distributed we also say that, the data is normally distributed and can be used in the analysis



Duration of interview for selected variables of ILFS by methods PAPI and CAPI

	PAPI	CAPI
Variable	**2014**	**2020**
The average duration of the interview (seconds)	111	112
Number of questions	197	220
Mean duration per question (in second)	33.8	30.5
Median duration per question (in second)	32.4	19.2

Source: ILFS 2014 & 2020

The average duration of the interview (in seconds) divided by the number of questions, gives the mean duration of 33.8 seconds per question for PAPI (2014) and 30.5 for CAPI (2020).

Conclusion remarks:

Data valuation provides a better understanding of the source data and facilitates its best use. It helps ensure accurate planning and forecasting. It also enables policymakers and planners to use data effectively and adapt their strategies accordingly. A shift in data collection methodology from paper-based to computer-assisted personal interviews provides a potential for improved accuracy, timeliness, costs, and consistency due to the programming of consistency checks found in an electronic questionnaire and impossible with paper questionnaires. Despite the use of technology in digital data collection, there is still a need for data validation to ensure the data quality and accuracy before its analysis and processing.

Acknowledgment:

Lastly, I would like to acknowledge with gratitude, the financial support from the International Association of Research in Income and Wealth (IARIW) and the Center of Innovation Government Innovation (CIGI). This paper would not have been possible without their support.

References

Adalı et al., 2022 Evaluating the Demographic and Health Surveys Mode Switch from PAPI to CAPI: An Experiment from Turkey. Social Science Computer Review, 40(6), 1393–1415. https://doi.org/10.1177/08944393211009566

Bet Caeyer et al., 2010. A Comparison of CAPI and PAPI through a Randomized Field Experiment 1.

Development Bank, A. (2019). The CAPI Effect: Boosting Survey Data through Mobile Technology. The CAPI effect boosting survey data through mobile technology. www.adb.org

Enders, C. K. (2010). Applied Missing Data Analysis (Vol. 1).

Mergenthaler et al., 2021. Going digital: added value of electronic data collection in 2018 Afghanistan Health Survey. Emerging Themes in Epidemiology, 18(1). https://doi.org/10.1186/s12982-021-00106-3.