Data Collection in the Time of COVID – 19 Pandemic: Does the New Adopted Approach Pose an Obstacle to the Data Quality of the Produced Official Statistics? – Palestine Case

Mrs. Aya Amro, Palestinian Central Bureau of Statistics (PCBS), Ramallah, Palestine
aamro@pcbs.gov.ps

Abstract. Despite the risk and the threats posed through COVID-19 pandemic in successfully conducting censuses and surveys through delays, interruptions, diversion of funding, the Palestinian Central Bureau of Statistics (PCBS) is determined to continue collecting data on a timely basis and of a quality that is fit for purpose.

The contribution of this paper is twofold. Firstly, it introduces the adoption of CATI (Computer Assisted Telephone Interview) mode of data collection during the pandemic in the current surveys and the influence it may has on data quality. It also shed more light on the main differences between CATI and CAPI (Computer-assisted personal interviewing) modes in household surveys in particular. Secondly, it focuses on proposing a strategy regarding the samples frame of household surveys conducted using CATI, through adopting a new methodology as an additional data resource for carrying out surveys during this period. This methodology is based on the integrative role played by PCBS and other National Statistical System pillars (NSS) including a relevant private sector company; in order to establish a central geographical database (Geodatabase) of households with fixed line numbers at PCBS, where data can be used to develop a more comprehensive sample frame for CATI-based household surveys.

Key words: data collection, data quality, household surveys, CATI, CAPI. COVID-19.

1. Introduction

During the COVID-19 pandemic and the associated lockdowns, many constraints for producing reliable and timely data have tightened, as the interviewers could not be able to conduct fieldworks due to the risk associated with face-to-face interviews. In this sense, the need for an alternative method of data collection has arisen, and CATI approach seems to be the available alternative at the time. One of the household surveys this paper is presenting and was conducted in PCBS through this approach is "the Effect of COVID-19 on Socio-Economic Conditions of the Palestinian Households 2020" (COVID-19 survey).

1.1 Data Collection Using CATI

This mode of data collection in conducting surveys is used in a wide range by many countries even before COVID-19 pandemic, as in Europe and North America many government survey organizations now employ this new mode for their surveys (Leeuw, 2008). Also, a review of the literature on "Quality report of the European Union Labour Force Survey 2017" reveals that the percentage of interviews using CATI mode for this household survey reached 46.8% among the European Union countries[1]. However, this mode of data collection in PCBS is a recent use. The need for this kind of approach arises through the time of COVID-19 pandemic to be used as the main tool of data collection in many surveys. It mainly depends on using the phone to contact the respondent, where the interviewer can conduct the interview with the respondent in the form of a telephone conversation, and the interview is conducted through an application on the tablet device that reflects the survey form. The application is designed in a way that allow the researcher to move automatically through questions whenever needed. Data can be easily entered and submitted to PCBS headquarter server for the purpose of data checking and auditing.

2. Objectives

This paper discusses the effect that CATI approach have on data quality through studying some measurements related to the accuracy dimension of COVID -19 survey. Also, this paper compares between COVID -19 survey (CATI – based) with Socio-Economic Conditions survey (SEFSec) 2018 (CAPI – based) where both surveys have the same sampling units; in order to know how much the quality of data vary between CATI and CAPI adopted approaches in household surveys during the current pandemic. It also presents a proposed strategy of how to develop a central Geodatabase of households with fixed line numbers through harnessing other sources in order to serve this goal.

3. Data Quality of CATI-based Surveys

A deeper understanding of how the adopted data collection approach affects data quality is crucial. Data Quality has various Dimensions in which each of them opens doors against new challenges. These dimensions are (Relevance, Accuracy, Timeliness and Punctuality, Clarity and Accessibility, Comparability, Coherence and Completeness). One of the quality dimensions this paper focuses on is "Accuracy". A primary step for understanding data quality dimension can help us to improve it (Sidi, 2013). In this sense, the paper applies data measurement and data assessment on data of COVID-19 survey.

[1] Quality report of the European Union Labour Force Survey 2017, 2019 edition - statistical reports, Eurostat.

3.1 Data Assessment

As Part of examining the accuracy of any survey, it is recommended to conduct several comparisons against reference values. The comparison of survey results with independent and more accurate information about the population parameters is a well-known method to analyze sample quality[2].

In order to assess the relevant data quality dimension in this paper, we set multiple comparisons by the most in common comparable social indicators including the average household size, the categories of household size, and the sex of household head of the following housing surveys: SEFSec 2018 and COVID-19 survey, considering census data of 2017 as a reference.

The Results of data assessments regarding indicators comparison are illustrated below:

Table 1: Percentage of Households among Selected Indicators

Average Household Size						
Region	**COVID-19 Survey**		**SEFSec 2018 Survey**		**Census 2017**	
West Bank	5.1		4.9		4.8	
Gaza Strip	6.0		5.8		5.6	
Palestine	**5.5**		**5.3**		**5.1**	
Household Size Categories						
Categories						
1-3	21.0		25.5		27.8	
4-6	47.0		44.7		45.4	
7-9	26.0		25.5		22.9	
10 +	6.0		4.3		4.0	
Total	**100.0**		**100.0**		**100.0**	
Sex Ratio of the Head of Household						
Region	Male	Female	Male	Female	Male	Female
West Bank	91.0	9.0	89.7	10.3	89.5	10.5
Gaza Strip	91.2	8.8	90.7	9.3	90.6	9.4
Palestine	**91.1**	**8.9**	**90.0**	**10.0**	**90.0**	**10.0**

Source: The Effect of COVID-19 on Socio-Economic Conditions of the Palestinian Households survey database, the Socio-Economic Conditions Survey database, 2018, census 2017 database. Palestinian Central Bureau of Statistics PCBS– Ramallah – Palestine.

As illustrated in the table above, regarding average household size and household size categories, one can notice that the overall percentage of households of COVID-19 survey in both indicators is much closer to SEFSec survey than the census 2017 with a slight difference. This may lead to many explanations, one of these explanations is that SEFSec survey is considered a panel survey; which means that each household is being visited more than once, so that every change to the household including its size is being monitored over time, considering the split households that can be tracked and included in the sample as well. This result give us a close percentage to the

[2] https://www.europeansocialsurvey.org/methodology/ess_methodology/data_quality.html

one of census 2017. On the other hand, the sample of COVID-19 survey consisted of the complete households of the latest round of SEFSec 2018 survey. Therefore, Implementing COVID-19 survey based on this sample increases the percentage of having small households with a size category of (1-3) individual, especially of those split households. The reason for this is that the response of small households can be higher than the households of a greater size. In turn, it affects the overall average household size. However, regarding the sex ratio indicator, there is no evident difference between each survey when compared to the census 2017.

3.2 Data measurement

Data quality measurement is associated to the calculation of non-sampling errors. A review of the literature on survey error and data quality "Groves, 1989" reveals four identifiable sources of error: coverage, non-response, sampling and measurement or response error (Leeuw, 2008).

As we focus in this paper on the "Accuracy" dimension, we consider using the following indicators regarding this dimension based on the National Guidelines for Data Quality in Surveys[3] in order to measure the accuracy of COVID-19 survey compared to SEFSec survey:

- Achieved Coefficient of Variations (CVs) of key variables in domains of interest
- The rate of over-coverage: The proportion of units accessible via the frame that do not belong to the target.
- Response and non-response rates.

The Results of data measurements regarding selected indicators of accuracy dimension are as follow:

Achieved CVs of key variables in domains of interest:

In order to compare the computed CV of COVID-19 survey with SEFSec 2018 survey, we first choose some of the key variables of both surveys as follow:
The key variables regarding Covid-19 survey:

1. Percentage of households in Palestine that receive assistance from one of the social protection programs.
2. Percentage of households that declaring a state of emergency because of COVID -19 pandemic is the main reason that made the main income earner to stop working during the lockdown period (March-May), 2020.
3. Percentage of households that the monthly household income decreased by the half or more during the lockdown period (March-May), 2020.
4. Percentage of households that no internet available at home is the main reason for children do not participate in educational activities.

The key variables regarding SEFSec 2018 survey:

1. Percentage of households that government wage and salary is the main source of income
2. Percentage of households that private sector wage and salary is the main source of income

[3] National Guidelines for Data Quality in Surveys: https://ndqf.in/wp-content/uploads/2021/07/National-Guidelines-for-DATA-QUALITY-in-Surveys.pdf

3. Percentage of households that wages from Israeli labor sectors is the main source of income.
4. Percentage of households in Palestine that receive assistance.

The value of computed CV's for the key indicators of COVID-19 survey ranges between 2.1% and 5.4%, while the value of computed CV's for the key indicators of SEFSec 2018 survey ranges between 2.9% and 6.3%.

Through those results, one can notice that the value of the CV's for COVID-19 survey lies within an acceptable range (Survey Data Interpretation Guide)[4], which means that data is consistent. In addition, the values of similar indicators of both surveys are quiet close.

The rate of over-coverage:

Coverage errors result from inadequate representation of the target population based on the units in the sampling frame. Over-coverage occurs due to the inclusion of units that do not belong to the target population. Factors contributing to over-coverage regarding COVID-19 survey are:

1. The phone number of the household is out of service.
2. The phone number of the household is incorrect.

The calculated over-coverage rate through those factors of the total sample is considered an estimate of the whole sample frame. The over-coverage rate for both Covid-19 survey and SEFSec 2015 survey is 6.26%, 5.64% respectively. The over-coverage rate of COVID-19 survey indicates that the sampling frame is adequately representative of the target population.

Response and non-response rates:

The term response usually refers to the level of participation in survey or interview research. Nonresponse error represents the gap between the sample and the respondents (Liu, 2012). In the case of COVID-19 survey, the cases of completed or partially completed households are considered response cases, whereas, non-response cases were attributed to different factors, which are:

1. The household refuses to cooperate.
2. The phone of the household is switched off.
3. No one of the households' member answered the call.
4. The respondent is an unqualified member to give answers.

The results are as follow:

Table 2: Response and Non-Response Rates

	COVID-19 Survey	SEFSec 2018 Survey
Response Rate	93.6	90.2
Non-Response Rate	6.4	9.8
Total	**100.0**	**100.0**

[4] Survey Data Interpretation Guide: https://www.toronto.ca/wp-content/uploads/2017/12/93c0-tph-survey-data-interpretation-guide-aoda.pdf

The results show high values in response rate in favor of COVID-19 survey. This can be explained by the period of which this survey was conducted in. Due to the lockdowns across Palestine in that period, many households were forced to stay home. In this sense, we can come up with a theoretical conclusion that the household was more likely to be a respondent than a non-respondent. The response rate of COVID-19 indicates that the sample itself is representative and the possibility of a bias is very slight when comparing the non-response rate of COVID-19 survey with some of the European Union's countries such as Sweden (43.4), Denmark (45.0), and Netherlands (48.4)[5].

In general, we can conclude that data quality is not significantly affected by using CATI mode. On the contrary, results of both data assessment and data measurement of the accuracy dimension for survey output data has showed that this mode is of a good quality as CAPI mode. However, several constraints can affect data quality of CATI-based surveys other than the one we have discussed in this paper. One of the important indicators must be taken into account regarding data quality is the response burden indicator[6], which is used to measure and compare the average length of completion of the questionnaire. This indicator is a crucial one in assessing CATI-based surveys as the length of the questionnaire plays a major role in the interview made by telephones; this we may discuss in other studies or papers.

4. Sample Frame of CATI-Based Household Surveys

Survey samples should reflect the underlying target population adequately. Samples of conducted household surveys by PCBS are selected either from a master sample of the total household's frame of census 2017, or from other large-scaled household survey sample frame. In our case, the sample of COVID-19 survey is selected from a master sample of SEFSec 2018 sample frame, with a total sample size of 9,926 households. As we mentioned earlier in the section of Data Quality of CATI-based Surveys about the effectiveness of CATI mode of data collection, PCBS is more likely to adopting this mode of data collection in conducting surveys even after the COVID-19 pandemic. In this sense, the main challenge in this phase is the lack of a comprehensive household sample frame for fixed line numbers or phone numbers. This is somehow a crucial issue facing the progress of conducting surveys with high quality during this crisis. In the section below, we illustrate a strategy of how to generate an expanded sample frame of CATI-based household surveys.

4.1 A Proposed Strategy on How to Extend the Sample Frame of CATI-Based Household Surveys

Each household survey has a different sample size; those samples are selected from a representative sample frame of each administrative level in Palestine. In order to build an integrated central geodatabase of households with fixed line numbers, we need to get benefit of other data sources, including mutual cooperation between PCBS and other NSS's and private sectors associations that provide relative data of the fixed line or mobile phone networks systems across Palestine. NSOs need to build a broad coalition of all segments of society and make sure all producers and users of data are counted and benefit from the systematic implementation of open data principles across the NSS ((UNSD), 2019). In this section, we discuss a proposed strategy of how to extend the sample frame of CATI-based household surveys as an initial step

[5] Quality report of the European Union Labour Force Survey 2017, 2019 edition, statistical reports - Eurostat

[6] Statistics Code of Practice for the European Neighborhood South countries (based on the European Statistics Code of Practice), principle (9), Eurostat, April, 2016

to create a central geodatabase that includes all the available fixed line numbers of Palestinian households along with spatial identification data of these households.

Methodology

The methodology is based on the mutual relations between the pillars of the national statistical system, and the telecommunications company in Palestine in particular. This company owns a database that includes the entire fixed line numbers available to Palestinian households provided with some identification data of each number including the name of relative customers on the national level. There are main distributors of the wired phone numbers set by the company in residential areas depending on the population density of each area. Those distributers are either small boxes containing up to 60 landlines in areas with low-density population per box for different housing units, or they are on a form of large boxes (Lockers) containing up to 800 fixed line numbers in areas with high density population, especially in city centers. In addition, each distributor is associated with spatial coordinates x and y based on the Palestinian coordinates system used in PCBS.

In order to achieve the proposed strategy, we need to work on a random sample of around 1000 households to be selected from the households' database of census 2017 for around one or two Enumeration Areas (EA). This strategy relied mainly on the process of linking the identification data of these households including the name of household head or any relative names, along with the spatial coordinates x and y of each building in the selected areas. Data is disaggregated at each administrative level (Region, Governorate, Locality, EA) with the registered landline numbers of the households in the database of the telecommunication company.

The first step of the linking process is based on matching item (1), through locating the housing units connected to each distributor within its enumeration area by matching each building x and y coordinates in the selected enumeration areas from PCBS with x and y coordinates of the distributors. Then, each household is being identified from the located housing units or buildings. The second step is based on matching item (2), which represents the name of household head or a relevant name according to the relation of head of household data in the census database, and the subscriber name of which the fixed line number is registered by at the company database. Eventually, a list of households with their available fixed line numbers and relative x and y coordinates of the building at each administrative level is generated (see fig. 1).

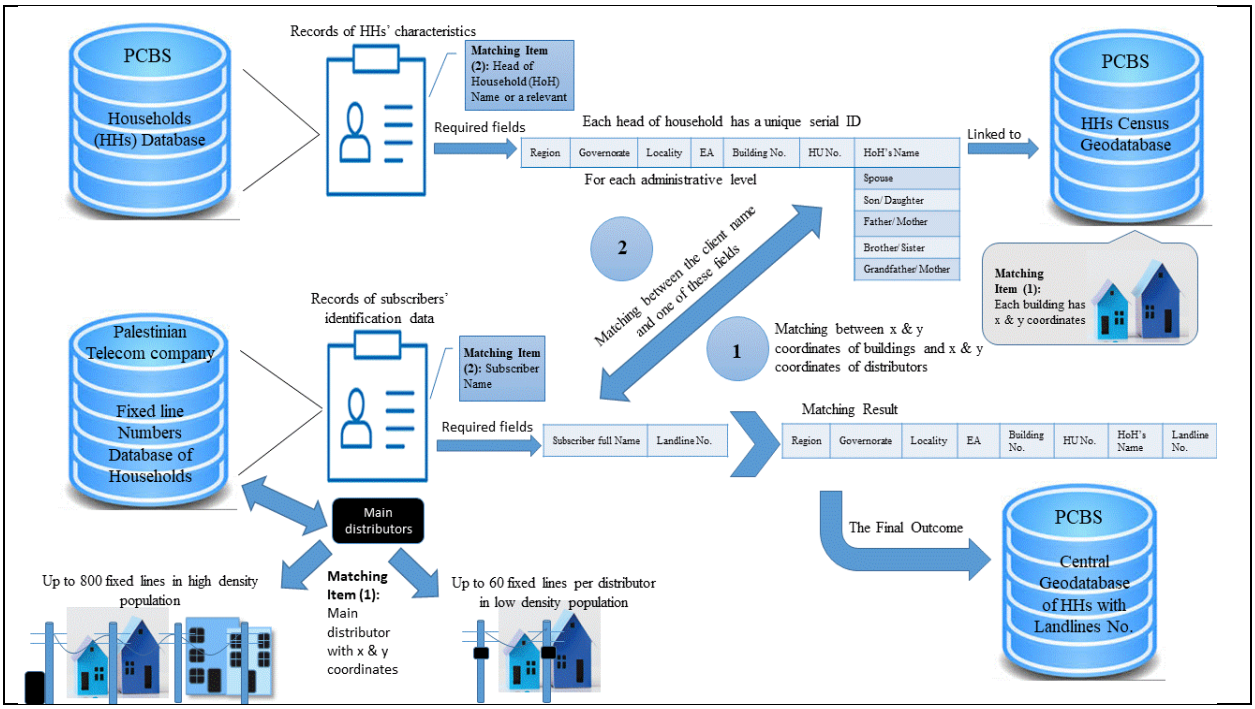

Fig. 1. Linking Process

Source: Designed by the author.

4.2 Associated limitations

The proposed linking process is subject to many limitations that must be considered such as:

- The extent of the company's cooperation in the process of linking and providing data at the individual level.
- Defining the ownership and the reference of data between the two parties after the completion of the linking process, and establishing a geographic database of households and how it may influence data confidentiality.
- The subscriber may have more than one fixed line number registered by his/ her name, which can affect the matching process between the subscriber name and household head name.
- The possibility that the utilization of the housing unit at PCBS database is not for habitation only but for habitation and work, and the fixed line number could be registered by the name of the work place.

5. Conclusion and Future Work

Due to COVID-19 pandemic that invade the whole world in early 2020, many constraints faced the interviewers in fully achieving their field works. In this sense, PCBS showed a high level of risk management through switching the initially adopted mode of data collection in conducting many household surveys. The use of CATI mode in household's surveys during this pandemic was a successful experience; as the results of data quality showed a high accuracy in COVID-19 survey data. This may lead to the conclusion that this mode of data collection can be adopted not only when personal interviews are difficult to be carried out, but it also can be considered for the long term of conducting surveys in PCBS, either as a secondary tool or as the main tool of data collection, or by moving to mixed-mode data collection.

The current crisis is somehow serving to steer our thinking to keep looking for the alternatives, through seeking for a better change in the adopted strategies regarding data collection even after the age of COVID-19 pandemic. Some of the recommendations to take in consideration in the foresee future are:

- Consider adopting CATI mode of data collection in carrying out future surveys at PCBS even after COVID-19 pandemic.
- In order to improve the proposed strategy through generating a comprehensive sample frame for CATI-based surveys, it is recommended to take advantage of surveys and censuses that will be implemented during the coming period, such as the agricultural census 2021, by adding a question related to fixed line or mobile phone number in the relative data collection questionnaire.
- Reconsidering the strategic plans regarding the use of administrative records as an alternative resource of some surveys, as well as, moving toward a Register-Based Census.
- Establishing data sharing platforms between PCBS, civil society and private sectors associations in order to allow statisticians to share a large number of indicators derived from big data sources that can be linked as much as possible.
- Embracing open data principles and practices, in order for NSOs to raise their standing as the trusted institution that ensures all users have ready access to high-quality data and statistics that meet national and international demand for information, while protecting privacy and confidentiality in line with the Fundamental Principles of Official Statistics ((UNSD), 2019).

References

1. Biemer, P. P., & Lyberg, L. E. (2003). Introduction to Survey Quality.
2. De Leeuw, E. D., & Collins, M. (1997). Data Collection Methods and Survey Quality: An Overview.
3. De Leeuw, E. D., Hox, J. J., & Snijkers, G. (1995). The Effect of Computer-assisted Interviewing on Data Quality. A Review. *Market Research Society. Journal. 37*(4).
4. De Leeuw, E. D.(2008). The Effect of Computer-Assisted Interviewing on Data Quality: A Review of the Evidence.
5. Eurostat Quality (2019), Report of the European Union Labour Force Survey 2017, 2019 edition, statistical reports.
6. Eurostat (2016), Statistics Code of Practice for the European Neighborhood South countries (based on the European Statistics Code of Practice).
7. Giuliani, G., Grassia, M. G., Quattrociocchi, L., & Ranaldi, R. (2004). New methods for measuring quality indicators of ISTAT's new CAPI/CATI Labour Force Survey.
8. Hubrich, S., & Wittwer, R. (2017). Effects of improvements to survey methods on data quality and precision - Methodological insights into the 10th wave of the cross-sectional household survey "mobility in Cities - SrV." In *Transportation Research Procedia* (Vol. 25).
9. Palestinian Central Bureau of Statistics (2019), Population Final Results - Detailed Report Palestine - Population, Housing and Establishments Census 2017, Ramallah – Palestine.
10. Palestinian Central Bureau of Statistics (2019), Socio-Economic Conditions Survey, 2018 - Main Findings, Ramallah – Palestine.
11. Pullum, T. W., Juan, C., Khan, N., & Staveteig, S. (2018). The effect of interviewer characteristics on data quality in DHS surveys. *DHS Methodological Reports No. 24*, (September).
12. Robert M. Groves ، Floyd J. Fowler, Jr. ،Mick P. Couper ،James M. Lepkowski ،Eleanor Singer & Roger Tourangeau (2004). Survey Methodolgy.
13. Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*.
14. United Nations Statistics Division (UNSD) (2019), Open Data Practices in Official Statistics and their Correspondence to the Fundamental Principles of Official Statistics.
15. Survey Data Interpretation Guide: https://www.toronto.ca/wp-content/uploads/2017/12/93c0-tph-survey-data-interpretation-guide-aoda.pdf
16. National Guidelines for Data Quality in Surveys: https://ndqf.in/wp content/uploads/2021/07/National-Guidelines-for-DATA-QUALITY-in-Surveys.pdf