Empirical Analysis of Data Acquisition Techniques: PAPI vs. CAPI

*Abstract*— Data acquisition refers to the act of collecting data on a large scale. This paper presents a qualitative and quantitative analysis of two data acquisition techniques, namely, Pen and Paper Interviewing (PAPI) and Computer-Assisted Personal Interviewing (CAPI). It cites two case studies to clearly define the difference between the two techniques. They have been compared on four factors, namely, cost, time consumed during the whole process, productivity of each interviewer and quality assessment of each techniques (average number of errors occurred per interview). It emphasizes on providing absolute numbers to clearly show the difference between the two techniques.

Keywords— PAPI, CAPI, Data Acquisition Techniques, Data mining, Data Analytics.

# Introduction

Data mining has attracted more and more attention in recent years, probably due to the popularity of the “big data” concept[1]. Data mining is the process of discovering fascinating patterns and knowledge from large amounts of data[2]. The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Commercial Data Mining, defined as “The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions”[3]. The Data Mining technique is used to predict possible future trends or to discover hidden patterns in the behavior of the data[4]. It enables us to explore the large patterns and analyze the same by means of Statistical and Artificial Intelligence in large datasets[5].

Data acquisition refers to the act of collecting data on a large scale. This data is then sent to the researcher for analysis or Data Mining. There are mainly two techniques for data acquisition, namely, Pen and Paper Interviewing (PAPI) and Computer Assisted Personal Interviewing (CAPI). PAPI is the more conventional technique, where data collection is done manually using paper and the responses by the respondent are filled by hand. On the other hand, CAPI is the more modern method where the responses of the respondent are fed to software on an android platform based device.

1. *Organization Of The Paper*

Rest of the paper is organized as follows. In section II, there are some related works that has been done on data acquisition techniques. Then, in section III gives an overview

of the two data acquisition techniques that are intended to be compared is presented. After that, in section IV there are two case studies for in-depth study of the comparison between the two techniques. The paper is concluded with comparison of the two techniques in terms of absolute data presented i.e. the four factors of comparison: cost, time consumed, productivity and quality assessment in section V.

# RELATED WORK

As a less explored area of research, there have not been many studies on Data Acquisition techniques, specifically PAPI and CAPI. Here are some of the work that has been cited by a few people:

In Literature[8] Leeuw et al., clears the dilemma related to the use of computer-assisted interviewing in social interviewing. They also state that, most professional research organizations, commercial, government and academic, are adopting CAPI with great enthusiasm.

In CAPI, interviewers visit respondent with a portable computer (generally a notebook) and conduct face-to-face interview using the computer. After the interview, data are sent to a central computer, either electronically through modem or by sending a data disk by mail[9-10]. They compare PAPI and CAPI by citing that, like PAPI in which the interviewer can hand over the paper survey to the respondent to fill sensitive responses, in CAPI as well the interviewer can hand over the device to the respondent to fill in the sensitive responses.

Literature[11] quotes that computer-assisted interviewing has rapidly become popular partly because of the expectations that it would lead to better data quality than traditional methods.

# DATA ACQUISITION TECHNIQUES

1. *PAPI*

Pen and paper interviewing (PAPI) was the first quantitative data collection method used by companies. In this method, the respondent was asked to answer the questions in a printed questionnaire and the responses were filled by the interviewer in that document. It needed very less equipment and could be carried out by less skilled task force. Therefore, it is still considered very flexible and quick. This technique is location independent, i.e., it can be successfully employed in any location. The questionnaire can be printed in the official language, i.e., English and in any regional language that is spoken in the area where the interview is to be conducted. Figure 1 illustrates the process flow of PAPI.

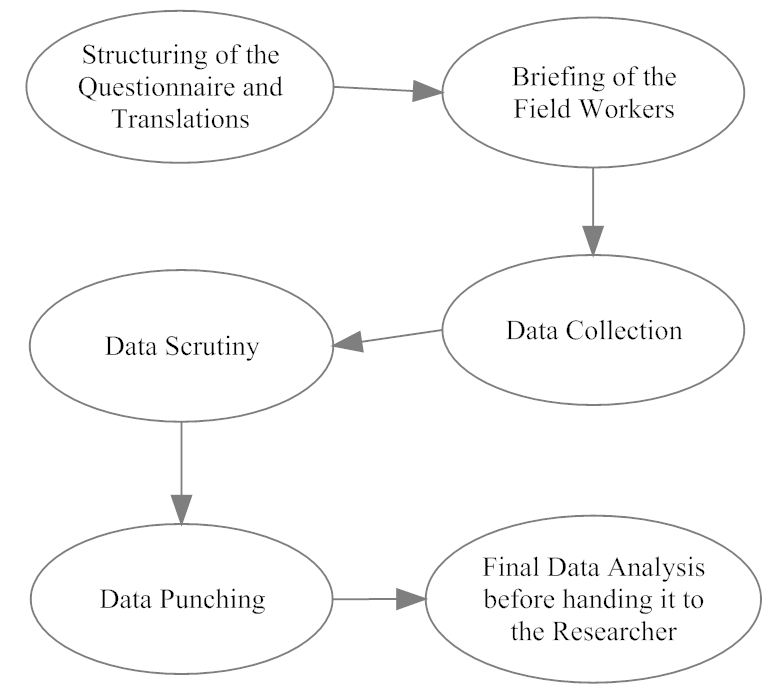


Fig.1. Process Flow Diagram of PAPI

1. *Advantages*

* For a small sample set, this method is very cost effective.
* Even if there is no network data connection in an area, this method can be easily and successfully employed to carry the desired purpose.
* The people hired to do the field work need not be highly skilled. They are required to know the regional language of the area. This helps in better interaction with the respondents and the labour employed is cheap.
* Unlike CAPI, there are no software glitches as the questionnaire is a printed.

1. *Disadvantages*

* This method is not very environment friendly as paper is used in huge numbers on which the questionnaire is printed. There is no real time data transmission of the survey to the researchers. Therefore, if there is any flaw in the way any question is to be asked and answered, that flaw cannot be rectified straight away. The quality checker has to wait till all the data is delivered after the field work. Thereafter, if a flaw is detected in the interviews, the whole or part of the field work is to be repeated.
* There is a high possibility of human error while conducting the interview, namely, spelling mistakes, wrong respondent details entered etc. This makes the process prone to inconsistent data punching and data analysis.
* Data scrutiny/entry is a cumbersome process as there is a high probability of human error in the field work, as also finger errors are unavoidable during data entry.

1. *CAPI*

Computer assisted personal interviewing (CAPI). With the technological advancements, the research industry is witnessing a shift in data surveying techniques from PAPI to CAPI. With the need for faster turnaround time and stricter quality and consistency check, many clients prefer data collection to be carried out with the help of CAPI. It is less time consuming than PAPI as no time is spent on data scrutiny and data punching. The questionnaire itself has compulsory consistency check, so the interviewer does not have to keep in mind as to which question to be asked from which group of people. Drop down menus can be employed in questionnaires for questions with a fixed set of answers, for example, gender can only have either male or female.

1. *Advantages*

* With built-in skip patterns and automated error messages, the chances of human error are very less and quality of data collection is maintained.
* With the help of real time data transactions between the field team and the researchers, the risk of repeating the field work due to errors in the script of the questionnaire is minimised,
* Since the interviewers are given their personal tabs with login and password, we can keep track of their productivity on a day-to-day basis.
* The researcher can keep track of the movements of the interviewers through GPS of the tab. This helps in planning geographical movements in field and also to monitor whether the interviewer has actually interacted with the respondents or not.
* Group interviews can be easily avoided since the tab only provides one data entry sheet at a time. This helps in eliminating any biased views by respondents.

1. *Disadvantages*

* This technique is less cost effective, when it comes to a small sample set.
* Since the ODK does not support Indian script, an image file is created and uploaded to the tab/android platform. Scripting of the questionnaire in regional language is the major problem faced while employing this technique.
* Proper training is required for the interviewers to be able to use and handle the tabs with care. This increases the cost of labour during the whole data acquisition process.
* Network connectivity, particularly in rural areas is also a major concern.

Figure 2 illustrates the process flow of CAPI.

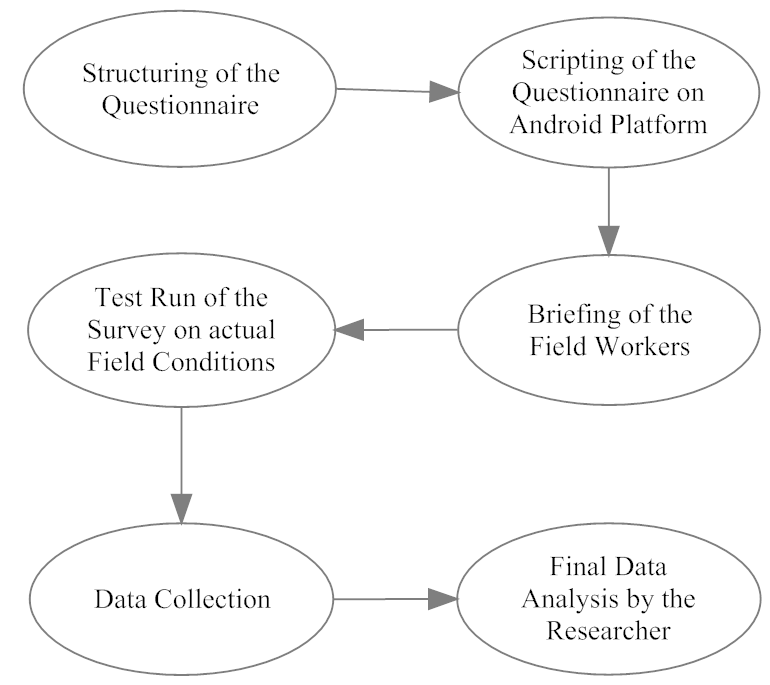


Fig.2. Process Flow Diagram of CAPI

IV. CASE STUDIES

1. *CT-DASH*

Correct and Timely Data Acquisition System for Health. This helped maternal and child health development using mobile phones.

1. *Problems witnessed*

There is a dearth of Knowledge accumulation on the Mother’s Health and Child development issues, before and after the delivery. This hampers working of various agencies in implementing various schemes, in an effective and result oriented manner. It is felt, that there is a need that the feedback on various schemes is fed correctly and timely to make them more meaningful. This will help various field agencies/workers in providing right guidance to the Mothers on their and their child’s development during pregnancy and first six to nine months of delivery. This in turn will make it easier in reaching the benefits of these schemes to the end users in a manner, as was envisaged by the policy makers.

The data collected in the surveys often suffer from human error, based on field worker’s Judgment, Data entry & Data scrutiny errors which may lead to incorrect perception and analysis.

This is more so, if the data acquisition is done manually using traditional pen and paper method. Moreover, it takes more time and effort for reporting, if it is to be collected over large samples, multiple geographies which are spread out over large areas. A need was felt to use modern technological methods to increase the speed of reporting and enhance the accuracy of data.

1. *The solution*

The Strategy adopted was Data collection through structured questionnaire and was supported by validation techniques, spot and back checks.

This whole process was supported by collection of Data using Modern Mobile/IT based systems and techniques. It was decided to collect the Data using the hand held Mobile phones. Scripting of the questionnaire was done on Android for use in appropriate mobile phones. Prior to the actual start of the data collection, the hand held mobile phones were tested in the actual field conditions in the concerned geographical area. Data transmission technology for the acquired data was also tested. Efforts were made to transfer the acquired data as soon as possible, after getting the respondent’s feedback.

1. *Inferences*

The scripted Questionnaire in the mobile phones had inbuilt logical and consistency checks. This removed the chance of capturing incorrect data and making it human error free. Data was uploaded in real time after completion of the interview, for the researcher to go through it. This helped in checking the data quality immediately and in taking timely corrective actions for changes, if any. This was much more effective in gathering error free data as the field worker collecting the data was already in the field. This technology based data acquisition helped in minimizing the data scrutiny and cleaning part, which are an integral part of data processing, thereby helping reduce time lag between Data acquisition, assimilation and analysis.

1. *Analysis*

The use of CT-DASH, Modern IT based technology/Mobile Based Data collection helps reduce the time lag between data collection & date usage, ensures error free and hence Quality data. This system aids the policy makers in taking timely and correct decisions to modify/implement various schemes in the field of Maternal and Child health. All in all, Mass usage of these methods will surely expedite the processes/efforts in improving the basic health indicators, which are so vital in overall development of our Human Resource.

Recently, it has been widely witnessed that the benefits of remote monitoring far out-weigh its drawbacks, and therefore many other techniques are emerging to address the demands of mobile health[6]. Moreover, healthcare systems with remote supervising have been shown to improve patient's quality of life and cutting on costs[7].

1. *PAPI vs. CAPI : Industrial application*

This case study shows the comparison between PAPI and CAPI on 4 factors, namely, cost, time consumed, productivity of each interviewer and quality assessment (average number of errors).

Costing factor includes cost of structuring the questionnaire and translations, printing of the questionnaire, data entry and data scrutiny charges in case of PAPI and scripting of the questionnaire on android platforms in case of CAPI. While costing for final data analysis, labour costs remain same for both. Time consumed factor takes into account the days spent on structuring the questionnaire and translations, printing of the questionnaire, data entry and data scrutiny in case of PAPI and scripting of the questionnaire on android platforms in case of CAPI. While days spent on final data analysis and data collection in the field remains the same for both. Productivity of each interviewer is measured on the number of interviews conducted by him in a day. Quality assessment factor takes into account the number of errors that can occur and can halt the data acquisition process in each of the two techniques.

1. *Overview*

In a study, interviews were conducted over a specific sample set using both techniques, i.e. PAPI and CAPI, of collecting data, the following was observed. This helped in comparing the two techniques on the four factors, i.e., costing of the whole process, days spend during the whole process, the productivity of each interviewer in a single day and the average number of errors occurred during each interview.

# RESULTS

This section presents the comparison of the two techniques i.e. PAPI and CAPI on the basis of costing of the whole process, days spend during the whole process, the productivity of each interviewer in a single day and the average number of errors occurred during each interview.

Fig.3a compares the two techniques on the basis of cost. CAPI was found out to be more cost effective over the same sample set than PAPI. While using PAPI, Rs.34800 was spent. On the other hand, while using CAPI only Rs.25000 was spent while conducting the survey for the same sample set.

Fig.3b compares the two techniques on the basis of days spent for the entire process to be conducted successfully. CAPI was found to be less time consuming completing the process in as less as half the time consumed as in comparison to PAPI.

Fig.3c compares the productivity of interviewers employed to complete the process using the two techniques. An interviewer employed with PAPI completed 5 interviews in a day, while an interviewer employed with CAPI was able to finish 7 interviews in a day’s time.

Fig.3d acts as a quality assessment for the two techniques. It was witnessed that on an average 2 errors per interview were found while using CAPI. On the other hand, on an average 5 errors per interview were found while using PAPI.

After assessing the two techniques over the four factors, it was found that, CAPI has an edge over PAPI. CAPI is approximately 40% more cost effective than PAPI. CAPI takes 50% less time to complete the same process with the same sample set than PAPI.

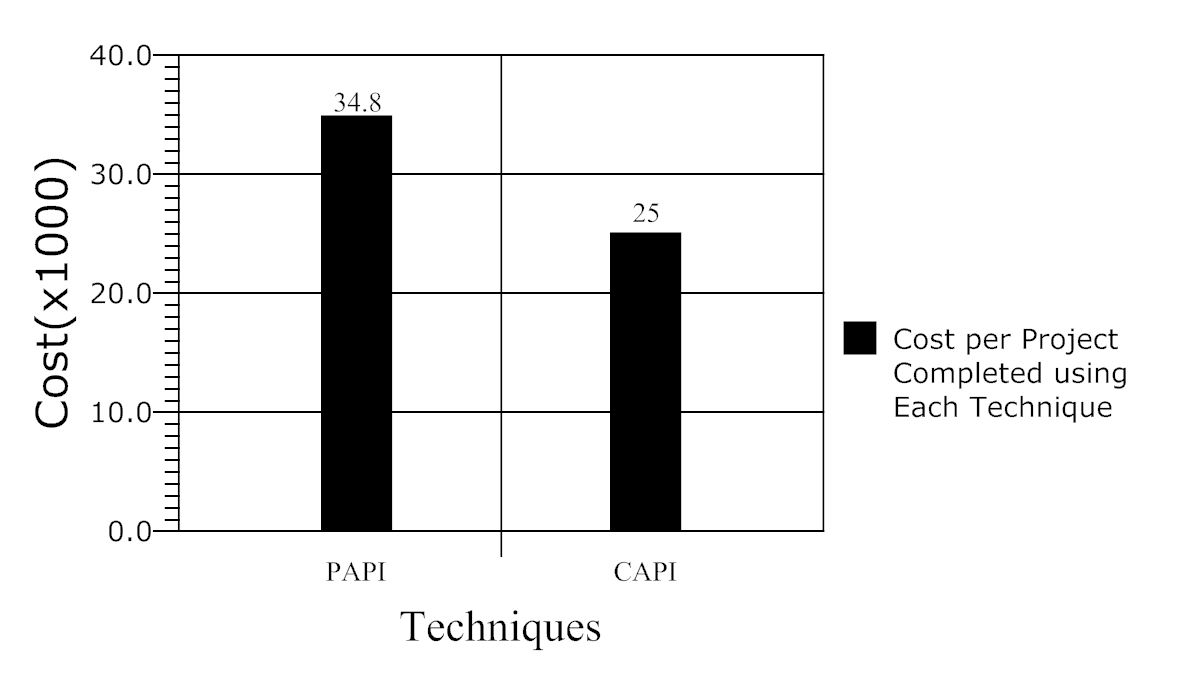


Fig.3 (a) Cost vs. Technique graph

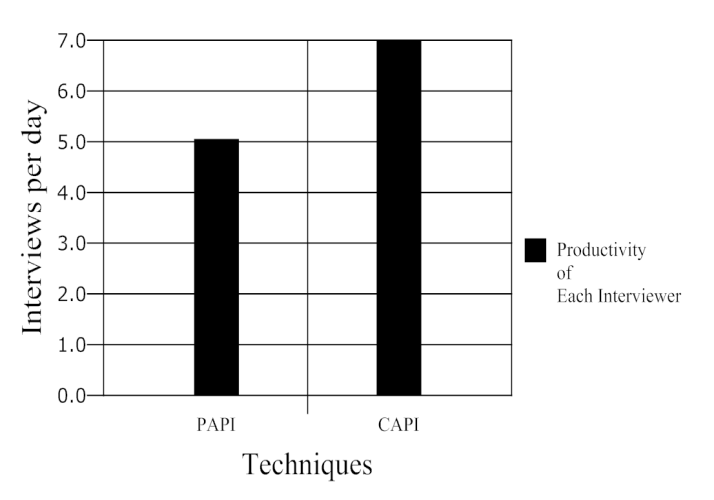


Fig.3 (c) Productivity per Interviewer vs. Technique Graph

The productivity of interviewers using CAPI is 40% more than the interviewers using PAPI. On an average the number of errors occurring in PAPI is 60% more than in CAPI.

Table 1 summarises the comparison between the two techniques. It clearly shows that PAPI is overrun by CAPI by a significant margin in costing, total time consumed in the process, productivity of each interviewer and average number of errors per interview. The commercial use of CAPI is more beneficial in the industry than PAPI. However, the only instance when PAPI can be considered as an equivalent technique to employ is when the sample set is small. When the sample set is small, the efficiency of PAPI and CAPI is practically equal.

TABLE 1. COMPARISON RESULTS

|  |  |  |
| --- | --- | --- |
| **Factors** | **Techniques** | |
| **PAPI** | **CAPI** |
| **Costing (Rs.)** | 348000 | 25000 |
| **Time consumed (days)** | 26 | 13 |
| **Productivity per Interviewer** | 5 | 7 |
| **Avg number of errors per interview** | 5 | 2 |

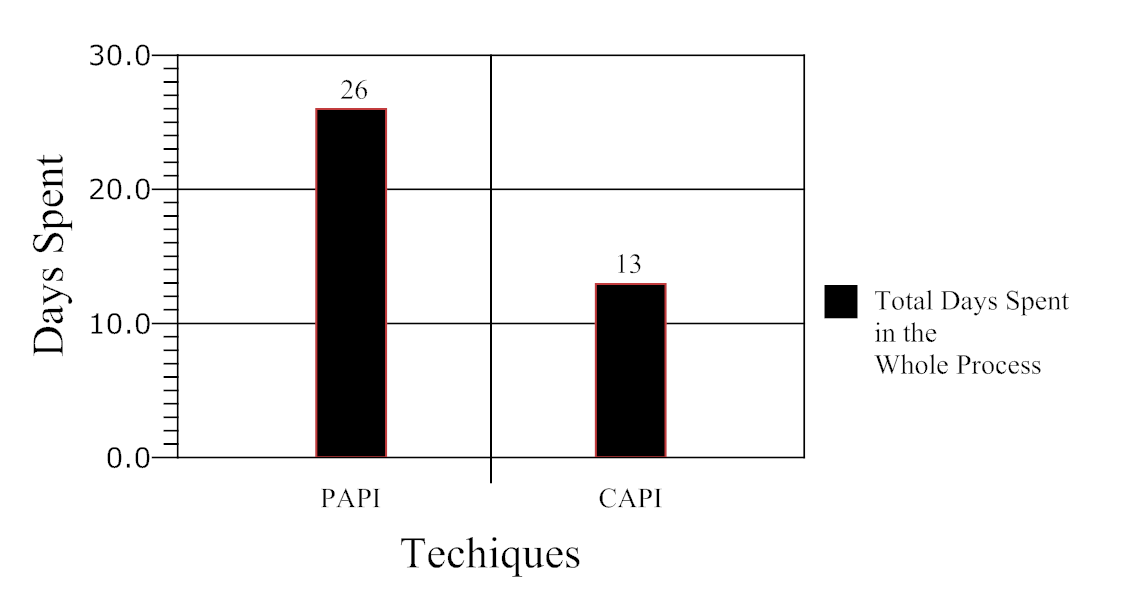


Fig.3 (b) Days Spent vs. Technique graph

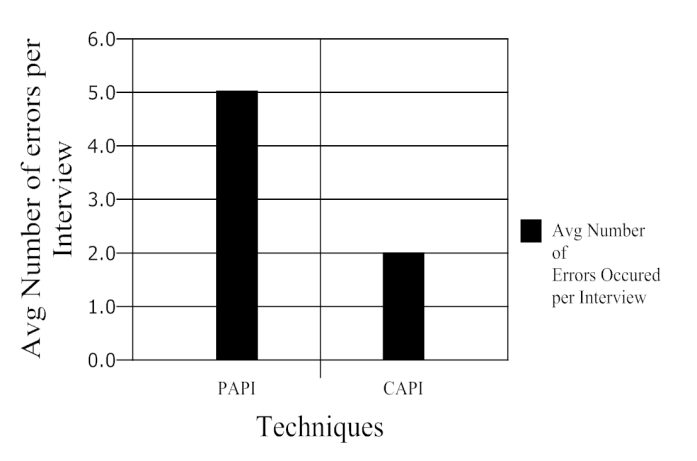


Fig.3 (d) Avg Number of Error per Interview vs. Technique Graph

VI. REFERENCES

[1] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan and Yong Ren, “Information Security in Big Data: Privacy and Data Mining,”vol. 2, 2014.

[2] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques,”San Mateo, CA, USA: Morgan Kaufmann, 2006.

[3] Lect. Shital P. Bora, “DATA MINING AND WARE HOUSING,”2011.

[4] Dr. B. L. Shivakumar and S. Alby, “A Survey on Data- mining Technologies for Prediction and Diagnosis of Diabetes,” International Conference on Intelligent Computing Applications, 2014.

[5] E. Papageorgiou, I. Kotsioni, A. Linos, “Data Mining: A New Technique In Medical Research,” HORMONES, 4(4):pp 189-191,2005.

[6] Majid A. Al-Taee, Nadine A. Jaradat, Dima M. Abu Ali, “Mobile Phone-Based Health Data Acquisition System Using Bluetooth Technology,” IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2011.

[7] G. Malkar, “Healthcare without bounds: trends in remote patient monitoring 2009,” SpyGlass Consulting Group, March 2009.

[8] Edith D. de Leeuw, Joop J. Hox and Ger Snijkers, “The Effect Of Computer-Assisted Interviewing On Data Quality: A Review,” Journal Of The Market Research Society, vol. 37 number 4, 1998.

[9] W. E. Saris, “Computer Assisted Interviewing,” SAGE Quantitative Applications in Social Sciences, no. 80, 1991.

[10] R. P. Baker, “New Technology in Survey Research: Computer-Assisted Personal Interviewing (CAPI),” social science computer review, vol.10, pp 145-157, 1992.

[11] M. Denny, L. Galvin, “Improved Quality At The Touch Of A Button: Use Of Computers For Data Collection,” *Proceeding At The Market Research Society Conference*, 1993.