UGANDA MARTYRS UNIVERSITY

SCHOOL OF POST GRADUATE STUDIES

Master's dissertation

STAFF CAPACITY AND DATA QUALITY IN VI AGROFORESTRY

AUTHOR: NABUKEERA VIVIAN

SUPERVISOR: MR. BWANIKA GODFREY

A RESEARCH PROPOSAL SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF A MASTER OF SCIENCE DEGREE IN MONITORING AND EVALUATION OF UGANDA MATYRS UNIVERSITY

DEDICATION

I dedicate this work to my Mother, the unsung hero who has stood by me and seen me evolve into a useful being.

AKNOWLEDGEMENTS

This research is a product of many contributions from individuals and organizations. First and foremost, I would like to express my deepest gratitude to my Supervisor Mr. Bwanika Godfrey for his wholehearted scientific guidance, encouragement and competent supervision. His unwavering interest in my study gave me intrinsic motivation to be persistent.

I extend my sincere gratitude to the management of Vi Agroforestry in Kampala Uganda for their interest in my research and logistical support offered to me during the period of data collection.

I wish to thank my family and friends for their encouragements and support throughout my study. A vote of thanks extends to my colleagues and friends at Uganda Martyrs University Monitoring and Evaluation class 2016 who gave valuable comments during formal and informal discussions.

I owe distinguished appreciation to all the staff of farmer organisations. They willingly spared their scarce time especially during interviews, to share with me their knowledge and experiences on a number of issues concerning data quality and staff capacity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study was conducted amongst the Vi Agroforestry partners. It is entitled “Staff capacity and data quality” the study was guide by three major objectives that is: To assess how capacity of staff in designing data collection tools affect project data quality.

To evaluate the capacity of staff in data collection and its relationship to project data quality and to assess how the ability of staff to analyze data affect project data quality

The main objective of this study can be summarized as the relationship between staff capacity and the quality of data among Vi Agroforestry partner staff. understanding the different capacities of staff who handle data and how the presence or the lack of these capacities have affected data quality.

A correlational research design was applied as a research design in this study and correlational data was majorly used in the analysis. This non-experimental type of research was helpful in identifying the relationship between variables and seeing the frequency of occurrence in the variables. Purposive sampling was used to identify the sample and staff who are knowledgeable or who have handled data were interviewed. A total of 61 staff were interviewed out of a population of 70 staff and this sample was determined using a 95% confidence level. Other primary data collection methods used were key informant interviews and observations. Statistical data analysis involved the use of descriptive and inferential analytical techniques.

Staff capacity in designing data management was found to have an impact on the data quality and this is backed by the analysis that on the different staff capacities that are discussed in the later chapters.

Capacity of staff in designing data collection tools was assessed by analyzing the staff's capacity to interpret indicators, their capacity to construct statements and their capacity to pretest developed tools and the results show that much as there are some staff with capacities in the design of data collection tools, there is still a lot of work to do to effect data quality positively. This is evidenced in only 38.4% of the respondents who agreed to having capacity to construct statements. There is still need to ensure good quality data collection through the selection, training and supervision of data collectors.

The capacity of staff in data collection was looked at in different dimensions and these included their capacities in the use of different data collection methodologies, the results showed that only 48% of the respondents strongly agree using knowledge gained in interview skills trainings compared to 78% of the respondents who have had trainings on interview skills, the results reflect a compromise in the data quality produced more on the data quality dimension of accuracy and validity. It is advisable to use standardized data collection tools, which have already been tried and tested in real life situations, and improve these if necessary to maximize data quality.

The capacity of staff in data analysis was assessed in their knowledge in data imputation, visualization, data editing and data integration. the results show that all the respondents who have done data analysis have not done data deduplication. And this is evidenced with the result of a coefficient of 0.000. The officers should be as critical of the methodological approaches to using qualitative data analysis software as they are about the fit between research question, methods and research design.

CHAPTER ONE

1.0 INTRODUCTION

This research set out to investigate the relationship between staff capacity and the quality of data among Vi Agroforestry partner staff. This particular topic was chosen due to the need to investigate the reasons behind the poor quality data in Vi Agroforestry partner farmer organisations and find possible ways to make it better. staff capacity in data management was looked at in 3 dimensions and those are capacity of staff in planning and designing of data collection tools, capacity of staff in data collection and the capacity of staff in data analysis. The quality of data was looked at from 6 different dimensions and those were timeliness, completeness, Accuracy, Validity, consistency and uniqueness.

chapter one therefore presents the background of the study, statement the problem, hypothesis statements, conceptual analysis, objectives of the study significance of the study and the justification of the study,

1.1. Back ground to the problem

Data quality has become a critical concern in the management information in organisations. Studies have shown that data quality is both a multi-dimensional and a hierarchical concept and it falls into one of four general ways: as excellence, value, conformance to specifications or meeting or exceeding consumer expectations Liu et al 2012. Although research efforts on data quality presented in the existing literature have addressed a significant advance in its short history, a generally accepted data quality model has not appeared yet. In their paper on data quality, Liu et al 2012 define data as an evolutional construct that evolves from collection quality, to organizational quality, to presentation quality and finally to application quality.

However, we live in an era of unprecedented data abundance and aggregation. The sheer variety of new information available on the internet. One serious problem we need to address is that of "dirty data"., missing or inaccurate information that resides in the abundance and aggregation of data in lives today. Dirty data can have several pernicious effects: increase costs and inefficiencies, creates liability risks and undermines the reliability and benefits of information technology, including the potential to streamline service delivery.

M&E systems produce data that is used to document progress towards the programs' goals and objectives. Often these systems produce data that is incomplete, tardy, owing to insufficient capacity of staff in monitoring and evaluation.

Data quality is viewed using different dimensions and those are; completeness, accuracy, timeliness, consistency and validity. Data quality is majorly affected by the way data is managed. Data must be of high quality if it is to be relied upon to inform decisions. Accurate, complete and timely data show what is happening on the ground, bad data call the system itself into question.

Why is it important to have quality data?

Data quality is important because we need accurate and timely information to manage projects and accountability, good information to manage project effectiveness, to prioritize and ensure the best use of resources.

The quality of data can be improved and the first step for improving data quality is to uncover data defects through data profiling or data archeology which is the process of analyzing the data for completeness, correctness and reasonability.

Many organizations still sidestep long-term data quality improvement practices in favor of achieving short-term goals. However, an increasing number of organizations realize that the consequences of not addressing the poor quality of data may result in adverse effects

On the other hand, data management as one of the staff capacities, concerns the dealing with data in a scientific context. Often more importance is given to results, analysis and derived conclusions than to the data themselves. However data are a product of the science enterprise and are more and more understood as a valuable research output themselves (dataONE 2012b; Ludwig and Enke 2013; Data service 2012 – 2015a). data can be qualitative or quantitative and comprises also of photos, videos or audio file resulting from different source as field experiments, model outputs or satellite data.

Ideally data management is considered as an integral part of the organization implementation. It includes the collection phase, the processing and the analysis of data, the documentation and preservation. Different data management activities are associated with each step. Well managed data will further facilitate reuse by oneself or others over time, replicate or validate research results and in organizations facilitate decision making.

Without data management, data could be subjected to data loss more easily and this could happen due to technical problems, due to missing information and due to not storing data in an appropriate way.

This study will have a specific focus on how data is captured in different organisations partnering with Vi Agroforestry, how staff enter data and how they prevent and fix duplicate records.

The quantity of information is a serious issue in evaluating data quality. A study on the use of graphs to aid decisions and a phenomenon called information overload was once conducted by Chan (2001). The scholar assumed that processing too much information can lead to making poor decisions. An experiment was conducted to show whether business managers would perform differently when treated with different loads of data. One group of subjects was given information with high load, whereas the other group of subjects was given information with nominal load. The results demonstrated that business managers under nominal information load could make higher quality decisions than those under high information load. This demonstrates that having more information is not necessarily better, or, in other words, does not necessarily lead to higher decision making performance. The phenomenon of information overload could be proven in this study.

Eppler & Muenzenmayer (2002) came up with a conceptual framework for information quality in the website context. They generally distinguish between content quality and media quality. For content quality, they further distinguish between relevant information and sound information. Attributes that can be associated with relevant information are as follows: comprehensive, accurate, clear, and applicable. Concise, consistent, correct, and current are attributes that make information sound. Media quality can be divided into the categories optimized process, with attributes like convenient, timely, traceable, and interactive, as well as reliable infrastructure, with attributes like accessible, secure, maintainable, and fast. Difficult navigation paths on a website are deemed an example of the convenience attribute.

1.2 Problem statement

Staff capacity in data management is key to the quality of project data. Quality data is a source of informed decision making and it also helps to gain deep insights into beneficiaries behavior, trends and creating extra ordinary experiences. quality data is a sure result of good data management. **(** *Mantra et al 2014)* if well managed data can be in position to answer a number of questions. Incorporating data management as a routine part of project implementation can save organizational time and resources in the long run. The M&E department at vi-agroforestry has time and again trained data collectors and staff of partner organizations on the how to design and interpret data collection tools, however not much success has been achieved in the quality of data produced by the partner staff.

One serious problem that needs to be addressed is that of “dirty data”—missing or inaccurate information that resides in the abundance and aggregation amongst the Vi partners. This results from the inefficient capacities to collect data using different data collection techniques like focus group, interviews and observations. This is accompanied by the inefficient skills in data analysis resulting into poor interpretation if data

Without good quality data, poor decisions will be taken, traceability of evidence will be lost, the organization will lose the insights they need to make the data useful. This study will therefore seek to scrutinize the different capacities of partner staff with an effort to improve data quality

Overall Objective

The purpose of this study was to find the relationship between staff capacity and the quality of data among Vi Agroforestry partner staff and to understand the different capacities of staff who handle data and how the presence or the lack of these capacities have affected data quality

1.3. Study objectives

I. To assess how capacity of staff in designing data collection tools affect project data quality.

II. To evaluate the capacity of staff in data collection and its relationship to project data quality.

III. To assess how the ability of staff to analyze data affect project data quality

1.4.Research questions:

I. Does staff capacity in designing data collection tools affect data quality?

II. Is there a relationship between staff capacity in data collection and project data quality?

III. Does the ability of staff to analyze data affect project data quality?

1.5. Scope of the study

The study will focused on the different capacities in data management staff of partner organizations in Vi Agroforestry have and how they use them to create good quality data.

1.6. Geographical Scope

The study was carried out amongst the partners of Vi Agroforestry, staff of Mubende District farmers Association, Mpigi Farmers Association, Wakiso Farmers Association, Sembabule District Farmers Association, Masaka District farmers association, Bugiri district farmers association, Uganda cooperative alliance and Uganda National Farmers Federation. These are the partners that are part of the consortium setting of Vi Agroforestry and are partially implementing the activities of Vi Agroforestry with in their members. They are therefore mandated to report to Vi Agroforestry, and it is the mandate of Vi Agroforestry to develop the capacities of farmer organisations to produce quality data.

1.7. Time scope

This study focused on the period between 2012 and 2016 since this is the period within which Vi Agroforestry has collaborated with the partners in Question

1..8. Content Scope

This study investigated the capacities the partner staff of Vi Agroforestry have in designing data collection tools, collecting data and data analysis, it went further to relate the effect these capacities have on the quality of data focusing specifically on the dimensions of data quality and those are consistency, accuracy, timeliness, validity, uniqueness and completeness.

1.9**. Significance of the study**

As a monitoring and evaluation officer in Vi agroforestry, this research gave an insight of the capacities of the different staff in monitoring and evaluation in Vi agroforestry, the M&E officers in the partner organizations and the capacities of the data collectors,

As an outcome of this research, areas of weakness were identified and recommendations made to improve on the capacity of staff in M&E so as to influence good quality results.

The secretariat (Vi agroforestry) will benefit through an improvement in the data received from the partners and thereby improving the quality of reports they write to the donors

The study was published to offer ways on how staff capacity can be improved, on what staff capacities need to be developed if a project is to realize quality results

As a result of improved data quality, the donors will be in position to get reliable information from the different partners that represents the impact of their funds.

The partners will be able to improve on the quality of data produced by their staff which will be used as a reliable reference for resource mobilization.

Reliable and accurate data is a basis of decision making by management, the decisions made in most cases do have an impact on the direct beneficiaries who are the communities, therefore with this study an effort to improve on the quality of data will be made by the different partners.

1.10 Justification of the study

Several studies have been conducted on data quality. However none of the studies has related data quality to staff capacity. This study therefore seeks to bridge the gap that has not been researched about by the different authors

This study seeks to analyze the staff capacities and their effect on data quality

A study on improving data quality by Yoshiko and Devesh 2007 examine the construction and use of data sets in political science, the core of this paper analyses the incentives and capabilities facing the four key actors in the data supply chain i.e the respondents, data collection agencies, international organizations and finally academic scholars.

Chongliang and Wang on his study on data quality management and data quality control analyzes several current problems in the field of scientific data management i.e the data quality problem still not solved, the absence of data living period design in data quality management and data quality management not always that clear, the study however neglects the issue of staff capacity the more reason to carry out this research.

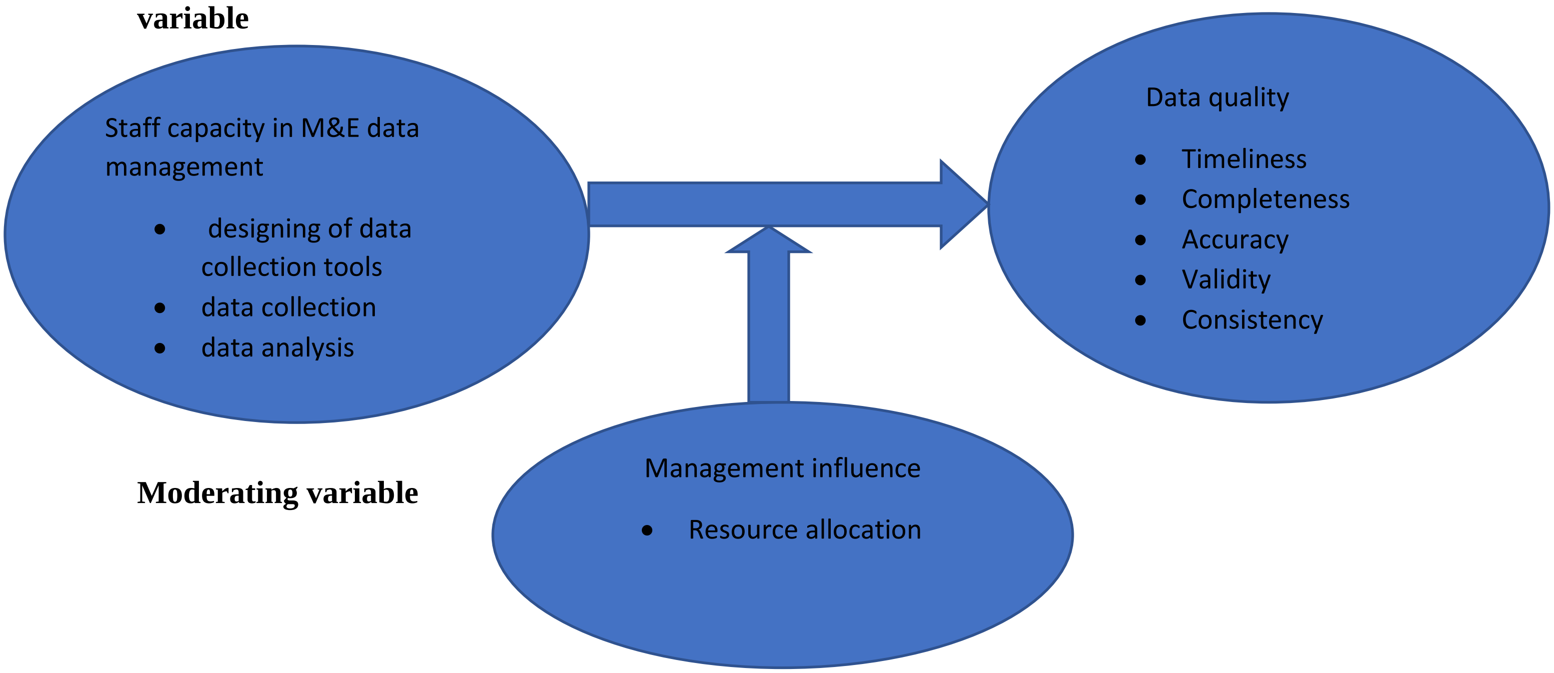
(Mantra et al 2014) who looks at the importance of data management in the dimension of answering research questions, this study also looked at the importance of data management to creating quality results. Scientific literature is available on the effect data management has on quality data, this thesis tried to fill the gap that relates how staff capacity in different dimensions especially in data management and how it affects data quality an area that has not been researched on.

The practical relevancy can be found in the fact that good data management is a sure way to producing quality data, therefore if the hypothesis is supported, it will benefit the partner organizations of Vi agroforestry in defending their results with good quality data

Figure 1.1: Conceptual Framework:

Source: Wang et al.2001;Lee et al 2002. Frame work.

independent variable **dependent variable**



Moderating variable

This study will focus on the quality of project data that is seen to be mainly determined by staff capacity in data management. Good quality data is the dependent variable that is majorly influenced by Staff capacity data management which will be looked at in different dimensions and that is: staff capacity in designing and the interpretation of data collection tools, Their ability to analyze data, their capacity collect data. Data quality on the other hand will be assessed using the five dimensions of data quality and that is timeliness, accuracy, validity, consistency and completeness. The moderating variable will be the management influence which will be assessed in terms of resource allocation.

1.12 Definitions of terms

Accuracy

Data should provide a clear representation of the activity/interaction

Data should be in sufficient detail

Data should be captured once only as close to the point of activity as possible

Validity

Data should be recorded and used in accordance with agreed requirements, rules and definitions to ensure integrity and consistency

Reliability

Data collection processes must be clearly defined and stable to ensure consistency over time, so that data accurately and reliably reflects any changes in performance

Timeliness

Data should be collected and recorded as quickly as possible after the event or activity

Data should remain available for the intended use within a reasonable or agreed time period

Relevance

Data should be relevant for the purposes for which it is used

Data requirements should be clearly specified and regularly reviewed to reflect any change in needs

The amount of data collected should be proportionate to the value gained from it

Completeness

Data should be complete

Data should not contain redundant records

Compliance

Data must comply with regulations on data protection and data security

Data management:

NCHRP report 666(Cambridge systematics, inc. et.al 2010) defines data management as " the development, execution and oversight of architectures, practices and policies to manage the information life cycle needs of an enterprise in an effective manner as it pertains to data collection, storage, security, data inventory, analysis, quality control, reporting and visualization. The international organization for standardization (2003) offers a more concise definition of data management as the activities of defining, creating, storing, maintaining and providing access to data and associated processes in one or more information systems. This synthesis focused on data governance, sharing, ware housing and quality.

Data governance:

Data governance deals with ensuring that the data are managed properly. It is the establishment, execution and enforcement of authority over the management of data assets (Cambridge systematics Inc. et al 2010; Ladley 2012). The terms “ data governance” and “ data business planning” are often used interchangeably or as components of one another(Stickel and Vandervalk 2014). Ladley (2012) suggested there should be a distinction between managing data (data management) and ensuring data are managed properly (i.e, data governance)

Chapter 2

2.0 LITERATURE REVIEW

2.1 Introduction

Conducting a literature review is a means of demonstrating the author's knowledge about a particular field of study, including vocabulary, theories, key variables and phenomena, and its methods and history. Conducting a literature review also informs the student of the influential researchers and research groups in the field (Randolph, 2009).

This chapter therefore, presents review of related literature, definitions of key study concepts as presented by other researchers. In addition, there will be an overview of research related to each of these topics. There are various factors that can influence the quality of data produced. One main assumption of this study is that capacity of staff in designing data collection tools, data collection and data analysis has a great impact on the quality of data produced. At the end of the chapter, a summary of factors that can have an effect on data quality will be listed. These factors were extracted from existing literature.

2.2 Theoretical Framework of the Study

There are a number of theoretical frameworks for understanding data quality. A systems theoretical approach influenced by American pragmatism expands the definition of data quality to include information quality and emphasizes the inclusiveness of the fundamental dimensions of accuracy and precision on the basis of the theory of science(Ivanov,1992), this framework is what underpins the study in question. the theory says that One certain way to improve the quality of data is to improve its use. If an organization wants to improve data quality, it needs to ensure that there is stringent use of each data element, (Orr 1998) There are a large number of studies on data quality and data management, a few researches have been done on the factors that influence data quality

Connecting for health data framework defines data quality as "the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data." Data accuracy is one of the "foundational features" that contribute to data quality (along with other attributes such as timeliness, relevancy, representation, and accessibility). In addition, data quality has two essential components: content (i.e., the information must be accurate), and form (i.e., the data must be stored and presented in a manner that makes it usable). These definitions are important to keep in mind when considering ways to minimize data inaccuracies, as they illustrate why the task of fixing dirty data requires more than merely providing "right" information.

When people think about data quality, they usually only refer to accuracy. Indeed, data are normally considered of poor quality if typos are present or wrong values are associated to a concept instance, such as a person's erroneous birth date or age. However, data quality is more than simply data accuracy. Other significant dimensions such as completeness, consistency and currency are necessary in order to more fully characterize the quality of data.

The definition of capacity adopted in the European Centre for Development Policy Management (ECDPM) study is the emergent combination of attributes, capabilities and relationships that enables a system to exist, adopt and perform, the attributes in this study are focused on data management which impacts on data quality, Data quality is a multi-dimensional concept (Wang et al.2001;Lee et al.2002; Batini and Scannapieca 2006). Accuracy, timeliness, consistency and completeness are examples of these dimensions. The literature consistently organizes these quality dimensions in four categories: intrinsic, contextual, accessibility and representational (Wang and Strong 1996;Pipino et al.2002; Hazen et al 2014). Intrinsic dimensions (e.g accuracy) describes the quality of objective and native data. Contextual dimensions (e.g relevancy) are dependent on the context in which the data are used. Representational dimensions refer to data understandability and conciseness. Accessibility refers to data sharing and security.

There are very few examples in the literature of monitoring of “capacity” itself. However monitoring of performance is being adopted as one way of formulating conclusions as to the capacities that are being developed and which need further development. Lavergne(2005) summarized the distinction between capacity and performance in the context of learning network on program based approaches

Morgan(2005) discussed the idea and practice of systems thinking and their relevancy for capacity and capacity development with particular reference to organizations as learning entities. Monitoring and evaluation of experience is therefore central to systems thinking in so far as feedback to stakeholders on the practical results of the organization’s work contributes to the emergence of analytical capacities.

Lara El Mouallem BE Eng., MSc Farhad Analoui PhD University of Bradford, In their study on the need for capacity building in human resource management issues, defined capacity as the ability of humans, institutions and societies to perform successfully, to identify and reach their goals, and to change when necessary for sustainability, development and advancement purposes (Ubels et al., 2010). Capacity development is considered an endogenous dynamic process that relies on one’s motivation, effort, and perseverance to learn and progress (Lopes and Theisohn, 2003) which enables organizations to change, flourish and grow. Some of the major capacities that enhance growth include leadership development and knowledge networking (Lopes and Theisohn, 2003).

2.3 Review of related literature according to objectives

2.3.1. Staff capacity in Data analysis skills and data quality

Davenport and DJ Patil noted in their October 2012 *Harvard Business Review* article on the rise of the data scientist that the advent of the big data era means that analyzing large, messy, unstructured data is going to increasingly form part of everyone's work. Managers and business analysts will often be called upon to conduct data-driven experiments, to interpret data, and to create innovative data-based products and services. To thrive in this world, many will require additional skills. Davenport and Patil recognize the need for skills to analyze large data sets but they don't relate these skills to data quality which directly influences the differences within the quality of data produced. This study examined the minimum analysis skills by data managers to be able to produce quality data.

To facilitate evidence-based decision-making, organizations need efficient methods to process large volumes of assorted data into meaningful comprehensions (Gandomi & Haider, 2015). The potentials of using BD(Big Data) are endless but restricted by the availability of technologies, tools and skills available for BDA. According to Labrinidis and Jagadish (2012), BDA (big data analysis) refers to methods used to examine and attain intellect from the large datasets. Thus, BDA can be regarded as a sub-process in the whole process of *insight extraction* from BD. It is certain that for BD to realize its objectives and progress services in business environment, it requires the correct tools and approaches to be analyzed and classified effectively and proficiently (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 2015)

Furthermore, to sort through data, so that valuable information can be constructed, human analysis is often required. While the computing technologies required to facilitate these data are keeping pace, the human expertise and talents business leaders require to leverage BD are lagging behind, this proves to be another big challenge. As reported by Akerkar (2014) and Zicari (2014), the broad challenges of BD can be grouped into three main categories, based on the data life cycle: data, process and management challenges:

The challenges are significant such as data integration complexities (Gandomi & Haider, 2015), lack of skilled personal and sufficient resources (Kim, Trimi, & Chung, 2014), data security and privacy issues (Barnaghi, Sheth, & Henson, 2013), inadequate infrastructure and insignificant data warehouse architecture (Barbierato, Gribaudo, & Iacono, 2014), and synchronising large data (Jiang, Chen, Qiao, Weng, & Li, 2015). Advocates such as Sandhu and Sood (2014)perceive that the potential value of BD cannot be unearthed by simple statistical analysis. Zhang, Liu et al. (2015) support this perspective and state that to tackle the BD challenges, advanced BDA requires extremely efficient, scalable and flexible technologies to efficiently manage substantial amounts of data – regardless of the type of data format

In a new Avanade survey, more than 60 percent of respondents said their employees need to develop new skills to translate big data into insights and business value. Anders Reinhardt, head of Global Business Intelligence for the VELUX Group — an international manufacturer of skylights, solar panels and other roof products based in Denmark — is convinced that “the standard way of training, where we simply explain to business users how to access data and reports, is not enough anymore. Big data is much more demanding on the user.” Executives in many industries are putting plans into place to beef up their workforce’s skills. They tell me that employees need to become:

Ready and willing to experiment: Managers and business analysts must be able to apply the principles of scientific experimentation to their business. They must know how to construct intelligent hypotheses. They also need to understand the principles of experimental testing and design, including population selection and sampling, in order to evaluate the validity of data analyses.

Adept at mathematical reasoning: How many of your managers today are really "numerate" — competent in the interpretation and use of numeric data? It's a skill that's going to become increasingly critical. VELUX's Reinhardt explains that "Business users don't need to be statisticians, but they need to understand the proper usage of statistical methods. We want our business users to understand how to interpret data, metrics, and the results of statistical models."

Some companies, out of necessity, make sure that their employees are already highly adept at mathematical reasoning when they are hired. Capital One's hiring practices are geared toward hiring highly analytical and numerate employees into every aspect of the business. Prospective employees, including senior executives, go through a rigorous interview process, including tests of their mathematical reasoning, logic and problem solving abilities.

Able to see the big (data) picture: You might call this "data literacy": competence in finding, manipulating, managing, and interpreting data, including not just numbers but also text and images. Data literacy skills must spread far beyond their usual home, the IT function, and become an integral aspect of every business function and activity.

These and many other qualities must be possessed by company or organizational staff in order for quality data to be realized

Procter & Gamble's CEO, Bob McDonald, is convinced that "data modeling, simulation, and other digital tools are reshaping how we innovate." And that has changed the skills needed by his employees. To meet this challenge, P&G created "a baseline digital-skills inventory that's tailored to every level of advancement in the organization." At VELUX, data literacy training for business users is a priority. Managers need to understand what data is available, and to use data visualization techniques to process and interpret it. "Perhaps most importantly, we need to help them to imagine how new types of data can lead to new insights," notes Reinhardt.

Tomorrow's leaders need to ensure that their people have these skills, along with the culture, support and accountability to go with it. In addition, they must be comfortable leading organizations in which many employees, not just a handful of IT professionals and PhDs in statistics, are up to their necks in the complexities of analyzing large, unstructured and messy data.

"Capacity development resource guide on data analysis and use" suggests that Data analysis and use can help policy and governance stakeholders make strategic, informed decisions related to the identification and prioritization of health issues and to policy and program development, implementation, and monitoring and evaluation (M&E). The broad topic of data analysis and use includes a series of linked but discrete actions, including the assessment of data needs, collection and analysis of data, synthesis and interpretation of data, and translation and targeted communication of data to decisionmakers.

While some individuals and organizations maintain involvement throughout each step—from data collection to data use—more often, these functions and competencies are separated within an organization and performed by professionals with different skill sets. Data producers design and implement research and information systems and, often in partnership with communication experts, will also synthesize and communicate the data collected through these efforts. Data users employ data to answer a specific question or inform a decision in the policy or program process.

Data are essential to supporting effective policy and program development, implementation, and M&E. Data are analyzed throughout these processes to help stakeholders understand health issues, advocate for change, design appropriate strategies, prioritize interventions, and develop and amend action plans. It is important that data be continuously fed back into the policy development process to ensure that decisions are being made based on the most current evidence.

(Foreit et al., 2006). Successful use of data in health policy is achieved when evidence-based information is considered and applied to the process of policy and program design, advocacy, policy dialogue, planning, resource allocation, and program review or improvement. Communication and dissemination (printing and distributing reports) of information is often necessary to facilitate data use, but it is not sufficient to ensure the information will actually be used to inform decisions or actions. Individuals, institutions, and the health system must have the capacity and supporting structures to regularly gather, analyze, interpret, share, store, and use data. At the individual level, while the level of skill required depends on the person's role, all policy stakeholders benefit from the ability to analyze and interpret data and translate and distill complex data into useful evidence to support the efforts of policymakers, program managers, and advocates.

Such competencies help (1) advocates understand the underlying data and promote action, (2) policymakers to lead the discourse on health issues and guide implementation, (3) planners and managers to make and implement program and service delivery decisions, and (4) stakeholders such as journalists and citizen watchdog organizations to monitor accountability. Institutions must have mechanisms to regularly produce high-quality data (e.g., on health trends, service delivery, expenditures, impact) and policies and procedures that support the synthesis and use of data and ensure the flow of information throughout the health system

2.3.2 Staff capacity in data collection and data quality

The capacity to collect data is often presented by clear data collection protocols that identify who is collecting what data, when, from whom and for what purposes (Grant makers for effective organizations, 2016b). Data collection involves the use of different data collection methodologies like interviews, focus group discussions, Observations among others and it’s in the capacities to use these methodologies that determines the quality of data produced in relation to timeliness, accuracy, consistency, validity and completeness.

There are many activities directly related to data collection. Apart from data entry and file naming (which are discussed below), the choice of an appropriate software format for the collected data has to be taken into consideration. Many different programs are used for data collection, ranging from spreadsheets and statistical software to relational database management systems and geographic information systems (Michener and Jones 2012). Every software and format has advantages and disadvantages, depending on what kind of analyses are planned, software availability and costs, the hardware which is used to capture the data and discipline specific standards and customs (UK Data Service 2014b). Also, the file format for working with the data can differ from the formats used for storing and long-term preservation.

There are a number of methods of data collection and interviews is one method of collecting survey data. There are a number of techniques that can be employed when carrying out interviews so as to produce quality data and some of them are: Rather than asking respondents to fill out surveys, interviewers ask questions orally and record respondents' answers. This type of survey generally decreases the number of ―do not know‖ and ―no answer‖ responses, compared with self-administered surveys. Interviewers also provide a guard against confusing items. If a respondent has misunderstood a question, the interviewer can clarify, thereby obtaining relevant responses (Babbie, 1992). As noted previously, personal interviews are a good way to gather information from community leaders, particularly those who might be unwilling or too busy to complete a written survey

Another method of collecting information is the focus group. Focus groups are useful in obtaining a particular kind of information that would be difficult to obtain using other methodologies. A focus group typically can be defined as a group of people who possess certain characteristics and provide information of a qualitative nature in a focused discussion. Focus groups generally are composed of six to twelve people. Size is conditioned by two factors: the group must be small enough for everyone to participate, yet large enough to provide diversity. This group is special in terms of purpose, size, composition, and procedures. Participants are selected because they have certain characteristics in common that relate to the topic at hand, such as parents of gang members, and, generally, the participants are unfamiliar with each other. Typically, more than one focus group should be convened, since a group of seven to twelve people could be too atypical to offer any general insights on the gang problem. A trained moderator probes for different perceptions and points of view, without pressure to reach consensus. Focus groups have been found helpful in assessing needs, developing plans, testing new ideas, or improving existing programs (Krueger, 1988; Babbie, 1992.There are a number of standards to put in mind when conducting focus group discussions and those are going to act as a basis for this study, the standards mentioned are but not limited to ability to develop a discussion guide, providing an incentive for participation, selection of focus group participants, moderating discussions and analysis of results. Much as different authors write and have researched about the different data collection methods, a lot has to be done to relate the capacity of the users of these methods to the standards that these methods carry. There are a number of techniques a data collector may employ so as to get good quality data and these are: the moderator needs to remain neutral to ensure that everyone feels comfortable expressing their opinion, eliciting further information from shy participants, dealing with shy participants by acknowledging their opinion and soliciting other opinions, paraphrasing long unclear comments by participants,these and more techniques can be taken into account for quality data collection.

According to the guidelines to strengthen data quality in mobile CBHIS, National programs and donor-funded projects increasingly rely on decentralized models of care to expand coverage of health services, ensure linkages to health facilities, and reach the most vulnerable populations. New emphasis has been placed on community-based models in which frontline health workers are expected to provide services and collect and report data. For example, the global “90-90-90”1 targets recognize that achieving equity in HIV prevention and care will require an emphasis on community-based approaches and systems (Joint United Nations Programme on HIV/AIDS, 2014).

2.3.3 Staff capacity in designing data collection tools and data quality

Data quality is affected by the design of data collection applications and platforms. Designing data collection tools involves interpretation of indicators, construction of statements, pretesting of tools before their use and involvement of stakeholders in the design of data collection among others and the absence or presence of these capacities has a great impact on the accuracy, timeliness, validity, completeness and consistency of data. Built-in components in the data collection workflow (such as skip logic, automated calculations, data validations, and instructional prompts) are used to increase the accuracy and completeness in data collection. data quality is also affected by the feedback and supervisory structures in place for encouraging data collectors and their supervisors to collect and report high quality data. Data quality is affected by behavioral factors of the people collecting data, such as their attitudes, values, and motivation.

Researches often use surveys, interviews and focus groups to collect new data. There may be instances where questions have been designed and researched and shown to measure what it intends to measure (valid) and does so consistently (reliable). When these questions exist, it is usually worth contacting the author and asking permission to use them.(Newman 1997) Offers guidance on how to design questionnaires. This also provides flexibility in the analysis of the responses. On the other hand, standardized questionnaire items often represent the least common denominator in assessing people's attitudes, orientations, circumstances, and experiences. By designing questions that will be appropriate for all respondents, it is possible to miss what is most appropriate to many of the respondents (Babbie, 1992). This is gaged on the capacity involved in designing data collection tools. This study will therefore bring to light how capacity in designing data collection tools affects data quality.

A variety of media and methods might be used to deliver, disseminate and facilitate information sharing each having associated advantages and disadvantages. Literature, for example, can reach the largest number of people, but does not provide the opportunity for feedback. Radio provides for speedy communication to a wide audience and often encourages feedback, but is not appropriate for communicating detailed or complex information (Muthiah, 1991). Face to face contact can lead to greater understanding and more frank discussion and feedback. A matrix based approach involving discussions with the key stakeholders may provide a useful means of agreeing which approach might be most appropriate (see for example Table 13 below). Further practical guidelines are given in Maine *et al.* (1996). The authors show the different data collection methodologies but don't relate how these different methodologies and how their ways of being designed affect data quality which is brings us to the rationale of this research.

2.4 Summary of the literature

In summary there are a number of identified factors that influence data quality as suggested by different scholars. However as suggested by Anders Reinhardt, head of Global Business Intelligence for the VELUX Group, there's a need to develop the capacities of company staff in data management if quality data is to be realized, above others company and organizational staff need to be adept at mathematical reasoning at least possess minimal statistical skills to enable them perform basic analyses, data management as pertains data collection, storage, security reporting and visualization all make up the qualities that have a direst influence towards data quality.

(Gandomi and Haider, 2015) show the challenges in data analysis like the integration complexities, they also mention insufficient personnel skills which is in a way related to this study. However this study in particular seeks to go in detail to analyze the specific insufficient personnel skills as regards data analysis, data collection and design of data collection tools.

(Newman 1997) defines the different standards when designing data collection tools. These standards will form a basis for research during this study however he doesn't relate how the lack of these qualities affects data quality.

3.CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter discusses the methodology that was employed in the study. Research methodology a comprehensive set of strategies which is used to gather evidence and to analyze specific data (Amin 2005). This section consists of the following sections; research design, study population, sampling techniques, data collection methodologies, reliability and validity, data management and ethical values

3.1 Research design.

A correlational research design was applied in this study and correlational data was majorly used in the analysis. This non experimental type of research was helpful in identifying the relationship between variables and seeing the frequency of occurrence in the variables.

A reconnaissance survey in which the researcher was introduced to the local leaders and the timeframe for data collection was arranged. The importance of such informal meetings is emphasized especially in acquiring support and erasing natural suspicion (Freudenthal and Narrowe, 1991).

3.2. Study population

The population of this study comprised of staff from Vi Agroforestry partner organizations, some of the Vi Agroforestry staff who were mainly the key informants and farmer trainers.

The entire population of Vi Agroforestry partner staff is 70 and a total of 61 interviews were carried out and among those interviewed were farmer trainers, partner field staff, M&E partner staff and coordinators. The respondents were chosen depending on their vast knowledge and understanding as far as data quality is concerned.

3.3. Sampling Techniques

Purposive sampling technique was used since the sample selection was based on the researcher's own judgement on the understanding and the application of the data quality procedures. Furthermore purposive sampling was favored due to its time and cost effectiveness.

3.4.Sample size

The sample size was determined using a Qualtrics sample size calculator and a confidence level of 95% with a margin error of 5% and a population size of 70 was considered during the calculation. Which provided a sample size of 60 respondents. The different categories however were determined using a 99% confidence level and a 5% margin error.

Table 3.1: showing the sample size

Category	Sample size
Partner M&E officers	**9**
Field officers	**35**
Farmer trainers	**15**
Key informants	**2**
Total	**61**

The sample size was used because the study was historical in design which involved studying the past events that could as well determine the future data quality. It also involved synthesizing data from many other different sources and getting information different people as well.

3.5. Data collection methods and instruments

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories: secondary methods of data collection and primary methods of data collection. For this study, both methods were used to collect information. For primary data collection methods this study stuck to individual interviews and focus group discussions. And for secondary data collection, partner reports over the years were reviewed

3.5.1. Secondary data

De Zeeuw and Wilbers (2004) highlights the importance of secondary data in getting a general picture and examining possible contradictions with earlier thoughts of the researcher. Accordingly, both published and unpublished literature related to the study area and the topic were reviewed

3.5.2. Primary data collection methods

De Vaus (1996) points out that the process of obtaining qualitative data especially through opinions can be complex. Because of this complexity, a multidisciplinary approach is often required. In this study, primary data sources were basically data collectors and the officers who make meaning out of the data. This data was gathered through a diversity of tools to enable data to be triangulated in order to increase its validity and reliability. Among the tools used was the questionnaire which was used in the study to acquire information from the staff of partner organizations. Some of the questions adopted in the questionnaire were open ended to enable respondents express their opinions innately.

3.6. Key informant interviews.

These are interviews with specially selected individuals who have a long period of experience in a certain community or specialized knowledge or skills in a certain topic (de Zeeuw and Wilbers, 2004). The interviews were done using a semi structured questionnaire.

Chambers (1992: 519) emphasizes the importance of key informant interviews in unearthing "invisible crucial social facts" and getting a better understanding of complex situations. The objective of using key informants is to collect information and gain more understanding of issues in a short period of time. Such information is used to develop a checklist for further investigation and to supplement the data collected using other tools (Jackson and Ingles, 1998).

Jackson and Ingles (1998) add that researchers should take note that this tool should be supplemented with other tools in order to collect sufficiently reliable data. In that sense therefore, Interviews with a few selected individuals were carried out accompanied with focus group discussions for the farmer trainers. The figure below shows a data collection process that was followed.

3.7. Reliability and Validity.

Validity of research can be explained as an extent at which requirements of scientific research method have been followed during the process of generating research findings. Oliver (2010) considers validity to be a compulsory requirement for all types of studies. Indeed, issues of validity and reliability of research instruments are of great significance to the findings of any scientific research. Moreover, as Dörnyei (2007) adds, validity and reliability issues serve as guarantees of the results of the participants' performances. In its broader context, validity refers to the degree to which a study reflects the specific concepts it aims to investigate. Two types of validity are discussed in social science literature: internal and external (Berg, 2007). Internal validity refers to the extent to which an investigation is actually measuring what it is supposed to measure. This type of validity answers the question: Are the differences found related to the measurement?; while external validity answers the question: validity in this research was ensured by:

I. drawing a sample from a larger population that was used to draw conclusions from the larger group from which the sample was taken.

II. Discussing the project design with my supervisor or a group of colleagues to help ensure that validity is preserved at every stage of the process

III. Internal validity was ensured by establishing a causal relationship between two variables and that is staff capacity and data quality. The effects observed in the study were due to the manipulation of the internal factor and not due to another factor.

On the other hand, reliability refers to the extent to which a research instrument yields the same results on repeated trials. Reliability is a concern every time a single observer is the source of data, because we have no certain guard against the impact of that observer's subjectivity" (Babbie, 2010, p.158). According to Wilson (2010) reliability issues are most of the time closely associated with subjectivity and once a researcher adopts a subjective approach towards the study, then the level of reliability of the work is going to be compromised. With this particular study reliability was ensured by using a retest method where the questionnaire was given to some respondents and the same set of questionnaire was given to the same respondents after two weeks and the equations for the two tests was as follows: $X_1 = X_1 + \varepsilon$ and $X_2 = X_1 + \varepsilon_2$ with

Assumptions: $V(\varepsilon_1) = V(\varepsilon_2)$ and $\rho(\varepsilon_1, \varepsilon_2) = 0$ thus $\rho(x) = \rho X_1 \rho X_2$

3.8. Measurement of variables

The study used the Likert scale. A *Likert scale* is the sum of responses on several *Likert item*s. A Likert item is simply a statement that the respondent is asked to evaluate by giving it a quantitative value on any kind of subjective or objective dimension, with level of agreement/disagreement being the dimension most commonly used. Well-designed Likert items exhibit both "symmetry" and "balance". The format of a typical five-level Likert item, was used:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Both qualitative data and quantitative data were collected during this study, a sample of 60 respondents was interviewed. The data included a range of opinions about data quality. All the qualitative data collected was rated in frequencies using tables and items recorded in percentages.

3.8.1. Data analysis

Data from the interviews was compiled, sorted, edited and coded to ensure that required accuracy, completeness, validity, consistency and uniqueness. The data was cleaned, coded and then entered into Epinfo software where it was analysed. For qualitative data analysis, the researcher created themes, categories and patterns, relationships among these themes were established and in-depth explanations and interpretations made.

3.8.2 Univariate Analysis

This was the first stage in data analysis and it involved descriptive and inferential statistics which led to production of frequencies and corresponding percentages for the study variables, both explanatory and outcome variables

3.8.4 Multivariate analysis

Under multivariate analysis, the joint effect of the independent variable on data quality was investigated. Only variables that exhibited statistical significance during cross tabulation were included in the multivariate model. This was done using classic and logistic regression

Qualitative data analysis involved the identification, examination, and interpretation of patterns and themes in textual data and determining how these patterns and themes helped answer the research questions at hand. In order for qualitative data to be analyzable it was first grouped into the meaningful patterns and/or themes that were observed.

3.9. Ethical considerations.

Permission was sought from the respondents during data collection and verbal consent to participate to participate in the study out of one's volition was obtained from all the respondents of different levels.

Special care was taken when delivering sensitive or difficult questions during the interview

This study stood on the assumption to keep all the findings anonymous. All interviews were done under the condition of anonymity.

4. CHAPTER FOUR

4.0 PRESENTATION AND DISCUSSION OF RESEARCH FINDINGS

4.1. Introduction

This section presents the results from the analysis and the discussions of the findings. The sections brings to detail the responses to questions on whether staff capacity in designing data collection tools affects data quality, whether there is a relationship between capacity of staff in data collection and project data quality and whether the ability of staff to analyze data affect project data quality? The Data analysis involves univariate, bivariate and multivariate analysis.

4.2. Response Rate

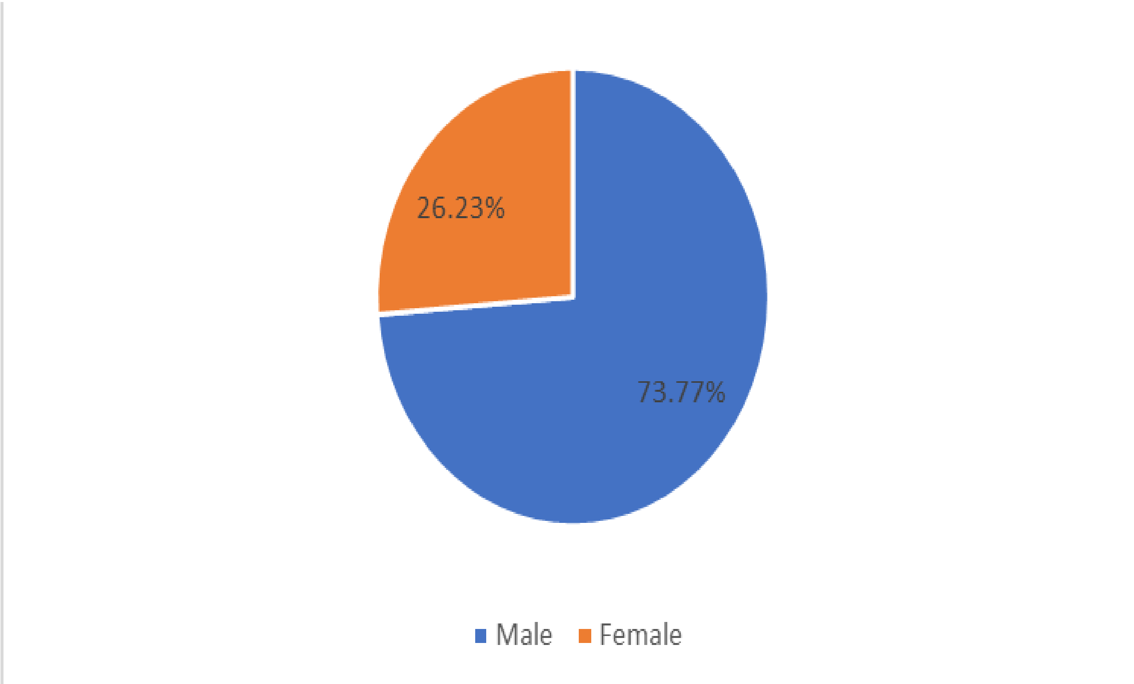
There is an overall total of 86 staff in Vi farmer organizations and out of 86 staff only 61 staff were interviewed which is 70% the sample size was determined using the formula $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ this formula provided a 95% confidence representation of the entire population of study.

4.3 Description of the respondents

This looks at the response rate, sex, age and educational background of the respondents.

4.3.1 Sex distribution

Figure 4.1 Sex distribution of respondents



A total of 61 respondents were interviewed 73% were male and 26% female. The lower percentage for the females is explained by the higher recruitment of males compared to the females. Out Of the 61 respondents, 35 were field officers who interact with data through data collection during surveys, 9 were Monitoring and Evaluation officers of partner organizations who are responsible in designing data collection tools and carrying out data analysis, 15 were farmer trainers who are responsible in the training of fellow farmers and who directly interact with data through data collection and then 2 key informants who were the M&E staff of Vi Agroforestry and they responsible for quality assurance.

4.3.2 Age

Table 4.2 Age of the respondents

Age Group	Frequency	Percent
27 - 36	39	63.93 %
18 - 27	11	18.03 %
36 - 45	7	11.48 %
45 - 54	3	4.92 %
54 - 70	1	1.64 %
TOTAL	**61**	**100.00 %**

Source: ***study findings 2018***

The majority of the respondents were between the age of 27 and 36 making a percentage of 63.93% , 18.03% were between the age of 18 and 27, 11.48% were respondents between the age 36 and 45, 4.92 were between the age 45 and 54 and the least represented age group was that of respondents between 54 and 70 years of age. Respondents were people who have dealt with data in one way or another. The highest percentage of respondents between the age 27 and 36 can be explained by the employees mainly being in the youthful age due to the dynamic nature of the field work accompanied by riding of motor cycles in difficult terrains which is easily manage by the youths

4.3.3 Education levels of the respondents

Table 4.3 education level of the respondents

Level of Education	Frequency	Percent
Secondary	1	1.64 %
Post Secondary	59	96.72 %
Missing	1	1.64 %
TOTAL	**61**	**100.00 %**

Source: study findings 2018

The findings show that over 96% of the respondents have attained a post-secondary level education. The post level education is inclusive of tertiary and university education. The biggest percentage of the respondents can read and write and therefore capable of interpreting and administering a questionnaire which has a big impact on the quality of data produced amongst farmer organizations. The 96% is inclusive of the respondents who have trainings in data management a prerequisite to good operational and analytical systems which results into reduced costs, improving existing organizational processes as well as improving decision making, all this is a resultant of good quality data. The educational level trend more so the 96.7% can be explained by the recruitment process in the farmer organizations which prefers and is strict on recruitment of diploma/bachelor graduates rather than secondary school graduates. The findings show that Partner organizations with staff that have staff with some skills and trainings in data management have a high level of good quality data compared to their counterparts with no skilled staff.

4.3. Designing data collection tools

Table 4.4. Frequency statistics on ability to develop a reporting template before data collection

Responses	Percent	Cum. Percent
Disagree	5.26 %	5.26 %
Agree	63.16 %	68.42 %
Strongly agree	31.58 %	100.00 %
TOTAL	**100.00 %**	**100.00 %**

Source: study findings 2018

Data presented in table 4.3 show that a large number of respondents 63.16% agree to having ability to develop reporting templates before data collection, there is a percentage of respondents who disagree on having ability to develop a reporting template before data collection. This could be explained by the lack of knowledge in the development of reporting tools as one of the respondents during an interview clearly stated that she has knowledge and what it takes to develop reporting tools.

In an open response some of the respondents indicated a clear lack of knowledge to even interpret indicators or worse still use a logframe. Questions concerning development of data collection tools or reporting formats were addressed to partner monitoring and evaluation officer who are supposedly technical in the field of development of tools and analysis, however a percentage 5.26% as seen in the table of results have no idea on the development of reporting templates before data collection. The cycle of data quality begins from being able to at least interpret a log frame and its indicators and data quality is compromised right from the start if there's failure to interpret indicators. poorly interpreted indicators will directly lead to poorly drafted reporting tools which is a sure way to poor quality data. This therefore means that staff need to be capacitated in the development of reporting tools and the interpretation of logframes.

Table 4.5. Frequency statistics on the ability to construct statements

Responses	Percent	Cum. Percent
Disagree	23.08 %	23.08 %
Agree	38.46 %	61.54 %
Strongly agree	38.46 %	100.00 %
TOTAL	**100.00 %**	**100.00 %**

Data source: study findings

Perhaps the most important part of the survey process is the creation of questions that accurately measure the opinions, experiences and behaviors of the public. Accurate random sampling and high response rates will be wasted if the information gathered is built on a shaky foundation of ambiguous or biased questions. Creating good measures involves both writing good questions and organizing them to form the questionnaire. A good percentage of respondents (38.46%) strongly agreed to having the ability to construct statements, this is basically because a few of these have had a few trainings on questionnaire development and have also had experience in the development of the same.

Questionnaire design is a multistage process that requires attention to many details at once. Designing the questionnaire is complicated because surveys can ask about topics in varying degrees of detail, questions can be asked in different ways, and questions asked earlier in a survey may influence how people respond to later questions. Researchers also are often interested in measuring change over time and therefore must be attentive to how opinions or behaviors have been measured in prior surveys. This study revealed a significant percentage of respondents 23.08% that disagreed to the ability to construct statements during the design of data collection tools, the reason behind this is their lack of knowledge and experience in designing questionnaires more so on the part of constructing statements

One of the techniques of designing data collection tools is designing closed-ended questions, the choice of options provides, how each option is described, the number of response options offered and the order in which options are read can all influence how people respond.

Further still, The choice of words and phrases in a question is critical in expressing the meaning and intent of the question to the respondent and ensuring that all respondents interpret the question the same way. Even small wording differences can substantially affect the answers people provide.

Table 4.6: Frequency statistics on respondents discussing developed tools before data collection with the project stake holders

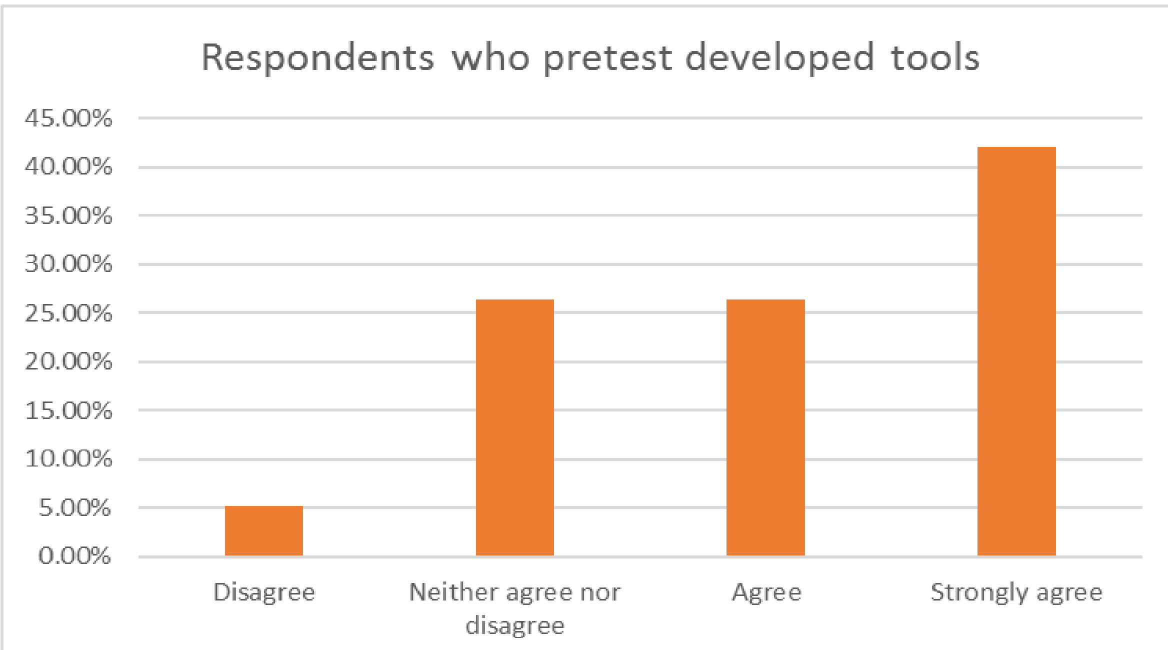
Response	Percent	Cum. Percent
Agree	50.00 %	60.00 %
Disagree	5.00 %	5.00 %
Neither agree nor disagree	5.00 %	10.00 %
Strongly agree	40.00 %	100.00 %
TOTAL	**100.00 %**	**100.00 %**

Source: primary data

The study results reveal that 40% of the respondents strongly agree to discussing developed tools with the stakeholders and incorporating their views with in the design, however there is a 10% proportion of respondents who don't discuss developed tools with the stake holders. The respondents during the interviewed revealed that when he develops these tools, they are just adopted as they are, he farther revealed that the stakeholders would in any case give abstract information due to their lack of knowledge in the subject area.

Involvement of stakeholders in the development of tools is one sure way of improving their quality and validity. It is crucial to understand and manage the six core dimensions. However, there are additional factors which can have an impact on the effective use of data. Even when all six dimensions are deemed to be satisfactory, the data can still fail to achieve the objective. Data may be perfectly complete, unique, timely, valid, accurate and timely. However if data items are in English and the users don't understand English then it will be useless. It may be useful to ask these additional questions about your data. Usability of the data - Is it understandable, simple, relevant, accessible, maintainable and at the right level of precision? This is what is looked at during the involvement of stakeholders at the design of data collection tools. One of the key benefits of this approach is that stakeholder participation is not only used to gather data, but also draw conclusions. Deciding what data will be created and how - this should be communicated to the whole research team.

Figure 4.2 Frequency statistics on the respondents who pre-test developed tools before data collection



Data source: primary data

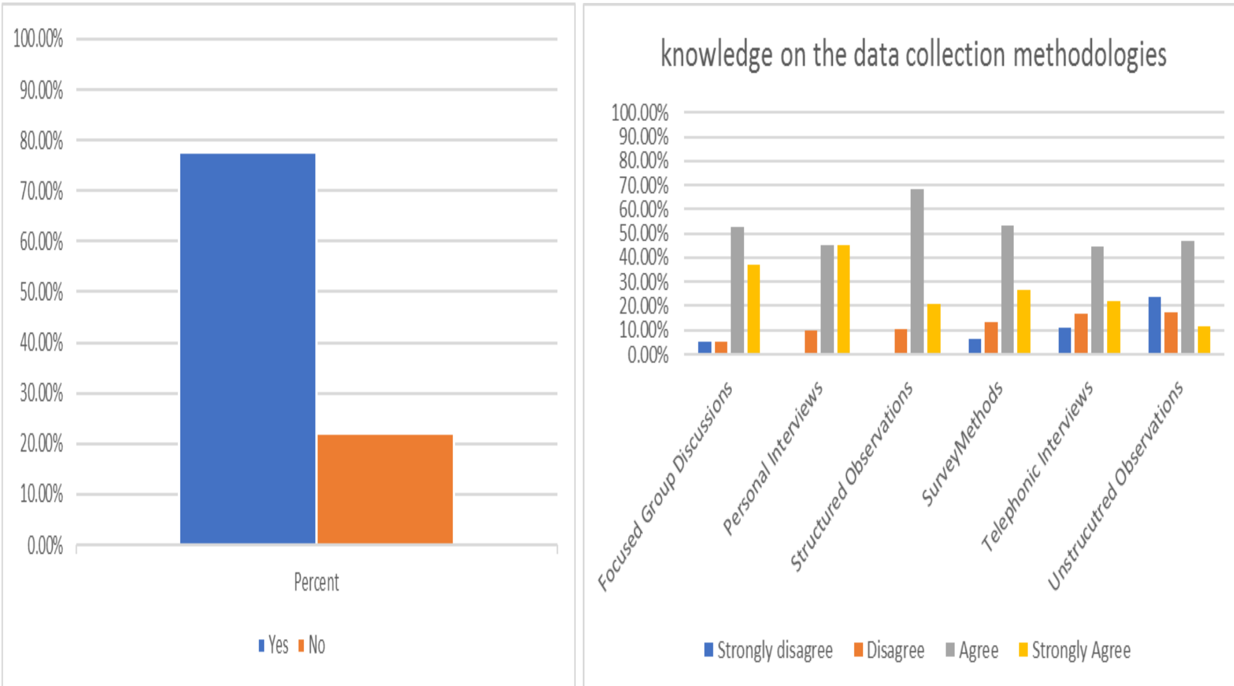
Results from the study show that over 25% of the respondents neither agree nor disagree to pretesting data collection tools before data collection and 5% disagree to pretesting the tools before data collection. This is because from the responses got during the study, the respondents don't have enough resources to pretest the developed tools and most of the respondents did not know of the existance of pretesting and have never had of it.

A very important part of the questionnaire construction process is its piloting, known as pretesting. This involves testing the research instrument in conditions as similar as possible to the research, but not in order to report results but rather to check for glitches in wording of questions, lack of clarity of instructions etc. - in fact, anything that could impede the instrument's ability to collect data in an economical and systematic fashion. Pretests should be conducted systematically, with potential respondents and using the same method of administration. The temptation to hurry over them, using just a convenience sample, should be avoided.

It is also beneficial to pretest the questionnaire with specialists in question construction, who may be able to pick up potential difficulties which might not be revealed in a pretest with respondents.

The lack of knowledge to do questionnaire pretesting poses a very great risk to the kind of data produced and this was evidenced by some of the responses got from the respondents as one of them clearly stated it as a waste of time and resources to carry out a pretest of the tools to be used, with such kind of mentality and attitude towards pretesting, data quality is surely compromised. Nevertheless, cognitive pretesting is considered an important part of the questionnaire design research process –the only way to determine in advance whether a questionnaire causes problems for interviewers or respondents (Presser et al., 2004) and also as a valuable addition to psychometric techniques when validating complex tools (De Silva et al., 2006).

Figure 4.3 Percentage of Respondents' awareness of the data collection methods

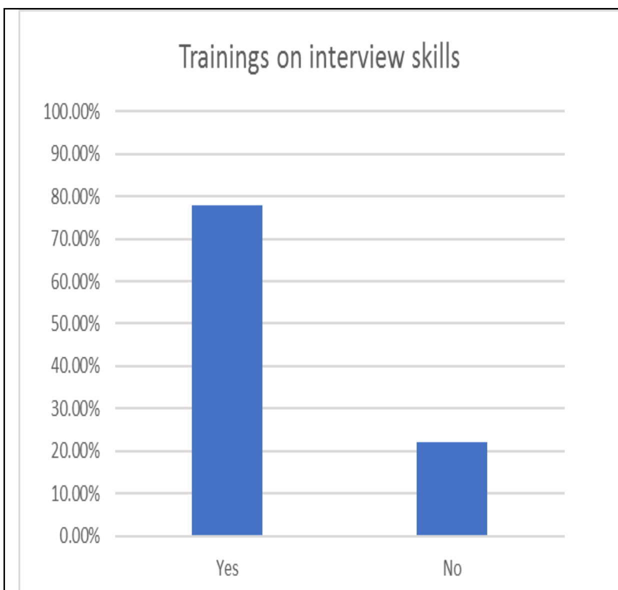


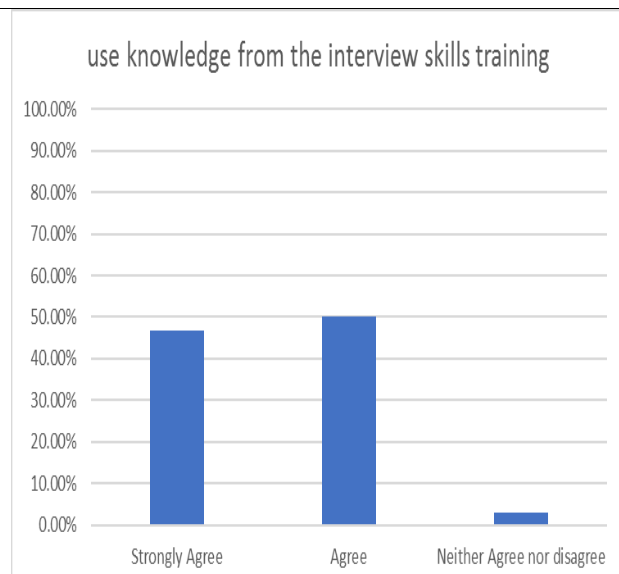
Source: primary data

78% of the respondents were found to have knowledge of the existing data collection methodologies, among the methodologies in question were focused group discussions, personal interviews, structured observations, telephonic interviews and unstructured observations. The results show that most of the respondents had knowledge in structured observations with a percentage of 68, the least popular methodology was unstructured observations with 25% of the respondents strongly disagreeing to having knowledge on the use of the methodology. The respondents revealed to not having knowledge on the use of the unstructured observations.

The study also revealed that the commonest data collection methodology used was the personal interviews and these are used with the help of unstructured questionnaires much as the assessment of knowledge in personal interviewed revealed that there are respondents who use personal interviews but with necessarily not enough knowledge on its use. Like all other data gathering methods, they are most effective through triangulation with other methods. Furthermore, a disadvantage of one data collection method is most likely the advantage of the other. O'Leary (2004) further remarks that it is worth remembering that one method of data collection is not inherently better than another. However the research point out that personal interviews are more preferred than the rest of the methodologies as one of the respondents noted that the personal nature of interviews can also aid in soliciting for higher response rates. Respondents are more likely to be committed to providing meaningful information in an interview than they would otherwise do with questionnaires (Abawi; 2013). The refusal rate for personal interviews is typically very much smaller than non-response rate for questionnaires and other methods (Akbayrak: 2000).

figure 4.4 distribution of frequency statistics on the trainings on interview skills and the use of knowledge gained from the interview skills





Data source: primary data

The research shows that 78% of the respondents have had trainings on interview skills and 22% have not had any trainings on interview skills. However for those who have had trainings in interview skills only 48% strongly agreed to the use of knowledge gained from the interview skills trainings got. Some of the respondents revealed that they could hardly adopt to new techniques taught to them during the trainings and that by the time the surveys take place they have already forgotten about the techniques taught to them during the interview skills trainings.

The value of interviewing is not only because it builds a holistic snapshot, analyses words, reports detailed views of informants; but also because it enables interviewees to "speak in their own voice and express their own thoughts and feelings" (Berg, 2007: 96).

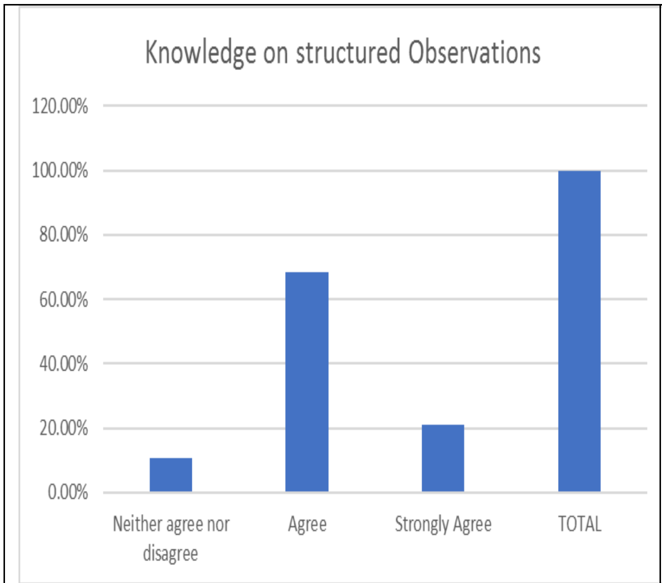
From the fact that interviews are only as good as their interviewer comes the downside. Their being highly dependent on the skills and abilities of the interviewer can have a negative impact on the effectiveness of the interview as well as the quality of data collected itself. Like all qualitative methods, the heavy reliance on the interviewer becomes problematic as the outcomes may not be perceived as reliable largely due to the fact that they mostly come from researcher's personal judgments and interpretations. For the reason stated above, interviewing, like all qualitative data collection methodology has often been considered as being more subjective, or prone to individual interpretation (Miles & Huberman, 1994).

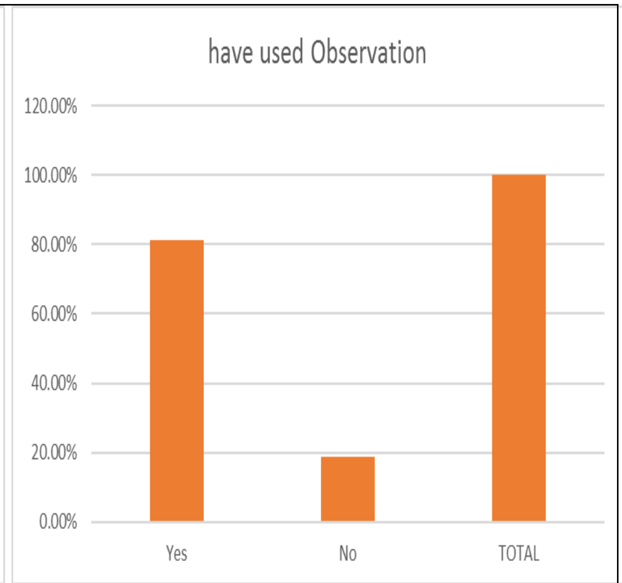
It is worth noting that interviewing was found to be very susceptible to a certain level of bias. In unstructured in depth interviews, in one of the interactions with the respondents it was observed that interviewers directed an interview in a direction best suited to their world view, which in turn affected the validity of results produced. On the same note, the interviewer's own perceptions may hinder the free flow of information as they begin interpret responses, albeit to suit their own expectations, regardless of the message the interviewee meant to put across. Risk of bias can also be heightened due to fatigue, as well as becoming too involved with interviewees (Abawi: 2013).

In terms of involvement with interviewees, the researcher may develop some form of empathy for the interviewer which may block their objective analysis of the issue at hand. It is however worth noting that bias can also be subject several factors such as skill level of the interviewer, world view, and other socio political factors. A certain level of bias can thus be reduced through training and adherence to research ethics. Hoyle, Harris and Judd (2002) go on to state that proper training and proper interviewer behavior helps greatly in achieving the goals of an interview.

The fact that interviews require a great amount of time to collect information is another disadvantage of the method. They require careful preparation of which, ideally this requires a lot of time and effort. Typical planning will involve arrangements for visits, seeking necessary permission, confirming arrangements, and at times rescheduling appointments to cover absences and crises which need more time (Akbayrak, 2000). The same author is also of the belief that any interview under half an hour is unlikely to valuable, whilst an interview that lasts more than an hour may be unreasonable on busy interviewees. Also to be considered is that, the analysing and transcribing requires time. Mouton, Hawkins, McPherson, and Copley (1987) calculate that one hour recorded interview material requires up to ten hours in order to conduct transcription from dictation. In addition to the above, interviewing can also be tiresome, especially if there is a large number of participants.

Figure 4.5 Relationship on the knowledge of structured observations and the use of observations





Source: primary data

The research findings show that 80% of the respondents have used observation as a methodology of data collection however there is a percentage of 30% of those who have used the methodology without necessarily having good knowledge or a clear understanding of observation a method of data collection. Validity is a related dimension of data quality because, in order to be accurate, values must be valid, the right value and in the correct representation, therefore this statistic implies that a lack of knowledge in the use of observations ideally means getting invalid data. when inaccurate - data may not be fit for use.

In a visit to one of the organizations it was evident through checks on the different data gathered through observations that not clear instruments were used during the collection of the said data. It was also observed that there was a lot of incompleteness with the data, there was not enough information to draw conclusions and the reported data did not contain enough information to represent performance measure of activities. Completeness is a data quality element that according to the research seems to have been adulterated.

Staff Capacity In Data Analysis

Table 4.7 Frequencies of respondents who have carried out data analysis

Response	Frequency	Percent
No	9	52.94%
Yes	8	47.06%
Total	17	100.00%

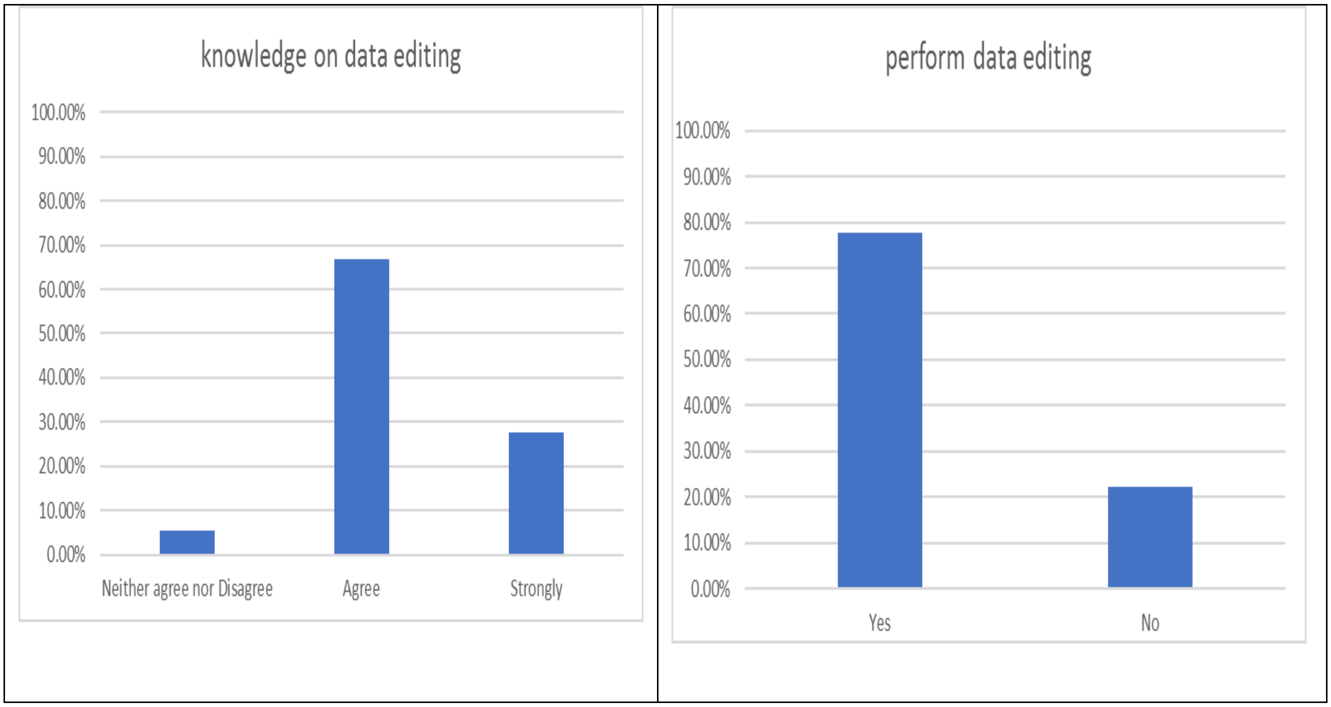
Source: primary data

A total of 17 respondents were interviewed on data analysis. The respondents included the monitoring and evaluation officers, key informants and some field officers of farmer organizations who deal with data on the level of data analysis. Out of the 17 respondents 52% of them have not carried out data analysis. The respondents were asked of the different statistical soft wares they use in analysis, their knowledge in the use of the common analysis soft wares like SPSS and excel, how often they carry out analysis, their trainings in data analysis.

Some of the respondents who had knowledge on the existence of the data analysis softwares however didn't have the expertise to use them. This was unveiled by the simple questions asked to them on the few functionalities of the soft wares and much as they had agreed to having used them they were not well conversant with the packages. Analysis looks at validity and acceptance. In scientific contexts it is tempting to think about validity. Here, a data analysis is successful if the claims made are true. This definition has the advantage that it removes the subjective element of acceptance, which depends on the audience to which an analysis is presented. But validity is an awfully high bar to meet for any given analysis.

It was revealed that some of the organizations are still using the manual analysis, this was unveiled by observation accompanied by personal interviews during the researchers' visits to the organizations. The purpose of the visits was to explore the use of computer-based qualitative and quantitative data analysis software packages. The advantages of using qualitative data analysis software include being freed from manual and clerical tasks, saving time, being able to deal with large amounts of qualitative data, having increased flexibility, and having improved validity and auditability of qualitative research. The partner organisations however who are not yet enjoying this privilege miss out on the above. Concerns include increasingly deterministic and rigid processes, privileging of coding, and retrieval methods; reification of data, increased pressure on researchers to focus on volume and breadth rather than on depth and meaning, time and energy spent learning to use computer packages, increased commercialism, and distraction from the real work of analysis.

Figure 4.6 The relationship between staff who have knowledge on data editing and those that can perform data editing



65% of the respondents have knowledge in data editing and they agree on having knowledge in data editing. However the same research shows that over 75% of the respondents perform data editing. There's a very big risk therefore that there are staff who carry out data editing without necessarily having knowledge in the same. This is evidenced in the graphs above. There is therefore a difference of 10% of the respondents who carry out data editing by default. Data editing is defined as the process involving the review and adjustment of collected survey data. The purpose is to control the quality of the collected data. Data editing can be performed manually, with the assistance of a computer or a combination of both.

When the researcher collects the data it is in raw form and it needs to be edited, organized and analyzed. The raw data needs to be transformed into a comprehensible form of data. The first steps in this process are to edit the data. The edited data is then coded and inferences are drawn.

The editing of the data is not a complex task but it requires an experienced, talented and knowledgeable person to do so.

With editing the data the researcher makes sure that all responses are now very clear to understand. Bringing clarity is important otherwise the researcher can draw wrong inferences from the data. Sometimes the respondents make some spelling and grammatical mistakes the editor needs to correct them. The respondents might not be able to express their opinion in proper wording. The editor can rephrase the response, but he needs to be very careful in doing so. Any bias can be introduced by taking the wrong meanings of the respondents point of view.

During this research data editing was looked at through different dimensions and these were rotating about the purpose of editing; clarity of responses, making omissions, avoiding biased omissions, logical adjustments and checking handwriting.

The editor has a great responsibility to edit the surveyed data or other form of responses. The editor needs to be very objective and should not try to hide or remove any information. He should not add anything in the responses without any sound reason. He should have to be confident in making any changes or corrections in the data. In short, he should make least changes and only logical changes. He should not add anything that shows his opinion on the issue.

There a number of qualities that were assessed during data collection and these helped us to understand the importance partner staff attach to data editing, the qualities assessed were, The data editor should have three qualities; he should have to be Intelligent, objective and experienced in editing the data. He should know how important is the handling of data to the researcher. He should try to avoid the slightest chances of bias, which means that he should also be honest with his work. His data editing will play a major role on the final inferences that the researcher will draw from the data.

Figure 4.7 Relationship between staff that have done data analysis and their knowledge on data analysis dimensions

Variable	Coefficient	95% Confidence	Limits	Std Error	F-test	P-value
Knowledge On DataImputation	-0.071	-0.765	0.622	0.293	0.0593	0.814555
Knowledge On Data Intergration (Yes/No)	0.786	-0.658	2.229	0.610	1.6565	0.239004
Knowledge On Data Visualisation (Yes/No)	0.310	-0.685	1.304	0.420	0.5419	0.485572
CONSTANT	0.357	-0.810	1.524	0.494	0.5234	0.492844

Correlation Coefficient: r^2 = 0.52

The coefficient on knowledge on data imputation of -0.071 implies that some of the respondents have not done data analysis and they have no knowledge on data imputation. Data imputation is the process of replacing missing data with substituted values. There are three main problems that missing data causes: missing data can introduce a substantial amount of bias, make the handling and analysis of the data more arduous, and create reductions in efficiency. Because missing data can create problems for analyzing data, imputation is a way to avoid pitfalls involved with list wise deletion of cases that have missing values. This research therefore revealed that there was no knowledge on data imputation for the respondent who have carried out data analysis. One of the aspects of data quality is the completeness of a dataset which involves the degree to which the data relevant to a particular application domain is included in the dataset.

An incomplete dataset contains missing records, missing attributes, missing metadata, or missing schema information.

The relationship between respondents who have done data analysis before and those with knowledge in data integration show a coefficient of 0.78 which is high above 0.5 implying that for those who have done data analysis certainly had knowledge on data integration. Data integration means combining information from various sources into something useful. It’s about efficiently managing data and making it available to those who need it therefore. Knowledge on data integration is a justified attribute for data quality because it increases the value of data through unified systems which includes combining data of varying natures, applying spatial information to non-spatial data, Easily accessible data means easily transformed data. People will be more likely to integrate the data into their projects, share the results, and keep the data up to date. This cycle of available data is key for innovation and knowledge-sharing. Therefore, the absence of data integration affects the quality of data that is produced and shared.

The coefficient on data visualization 0.3 imply that some of the respondents who responded to having done data analysis have some knowledge on data visualization. Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Data visualization tools have been important in democratizing data and analytics and making data-driven insights available to workers throughout an organization. Data visualization further simplifies data and makes it easier to interpret and utilized by the other people of the organization. One aspect of data quality is relevancy and the relevance of a dataset involves the degree to which it meets the needs of the data consumer.

A comparison is made still of some respondents who have done data analysis and have knowledge in data intergrattion(0.78) but don’t have knowledge in data visualisation.(0.3).

The correlation coefficient which is 0.72(the square root of 0.52 the coefficient of determination) reinforces the fact that some of the respondents who have done data analysis have good knowledge on data.

The coefficient determination 0.52 implies that only 52% of the independent variables actually explain the dependent variable.

Figure 4.8. Relationship between staff that have done data analysis and their knowledge on data editing, imputation,validation and deduplication

Variable	Coefficient	95% Confidence	Limits	Std Error	F-test	P-value
IperformDataEditing (Yes/No)	0.500	-0.737	1.737	0.523	0.9130	0.371135
IPerformDataImputation (Yes/No)	0.417	-0.655	1.488	0.453	0.8454	0.388440
IPerformDataValidation (Yes/No)	-0.167	-1.404	1.071	0.523	0.1014	0.759390
IPerformDataDeduplication (Yes/No)	0.000	-1.010	1.010	0.427	0.0000	1.000000
CONSTANT	0.000	-0.714	0.714	0.302	0.0000	0.000000

Correlation Coefficient: r^2 = 0.34

A relationship between the respondents that have carried out data analysis and those that have performed the data quality aspects in data analysis show a coefficient of 0.5 under data editing and this implies that almost a half of the respondents that have carried out data analysis have also performed data editing.

The coefficient 0.4 that is under data imputation implies that just a small fraction of respondents that have carried out data analysis have also done data imputation. However, those that have done data editing have also done data imputation, not all who have done data imputation have done data editing.

Further still the results show that all the respondents who have done data analysis have not done data deduplication. And this is evidenced with the result of a coefficient of 0.000

The coefficient correlation of 0.58 (square root of the coefficient of determination) shows that there is a moderate correlation meaning that for anybody for example who has done data analysis has also done data editing.

5. CHAPTER 5

5.0 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the summary of findings, conclusion and recommendation that is offered on the topic of study entitled staff capacity and data quality in Vi Agroforestry partner organizations.

5.2 Summary of key findings

Satisfying data quality and data accuracy requirements is a key step in the implementation of real-time information programs / projects. Data quality is a key factor in the effectiveness of congestion management that relies on data from various sources. The study initially provided an overview of previous studies followed by a discussion of data quality issues and staff capacities associated with partner organizations

There is a relationship between data quality and staff capacity and this is evidenced in the results of the study that have been discussed in the previous chapters, some of the Vi Agroforestry partners have capacities that affect data quality however some still lack the same capacities something that has compromised the quality of data amongst the partner organizations. The capacities assessed in this study were: staff capacities in designing data collection tools, staff capacities in data collection and staff capacities data analysis.

The results revealed that partner staff have some abilities in designing data collection tools and this was evidenced by their knowledge on some aspects like their ability to develop reporting tools before data collection, capacity to construct statements during the designing of data collection tools and 31.5% of the respondents strongly agree to having knowledge in the design of data collection tools. There is therefore a percentage of 68 of the respondents who still cannot design data collection tools

The research on the capacities of respondents in data collection show that the respondents use the data collection methodologies much as a high percentage of them don't have knowledge and expertise in the use of the methodologies, this can be seen in the results comparing the use of observations as a data collection tool which show that 80% have used observation as a data collection methodology however only 20% have the knowledge in data collection.

47% of the respondents have done data analysis however a few have knowledge in data analysis. This was analyzed using different dimensions and some of them where knowledge in data editing, data deduplication, data imputation and their effects on data quality.

5.3. Conclusions

5.2.1 Capacity in designing data collection tools:

Good data management includes developing effective processes for: consistently collecting and recording data, storing data securely, cleaning data, transferring data (e.g., between different types of software used for analysis), effectively presenting data and making data accessible for verification and use by others. Commonly referred to aspects of data quality are:

I. Validity: Data measure what they are intended to measure.

II. Reliability: Data are measured and collected consistently according to standard definitions and methodologies; the results are the same when measurements are repeated.

III. Completeness: All data elements are included (as per the definitions and methodologies specified).

IV. Precision: Data have sufficient detail.

V. Integrity: Data are protected from deliberate bias or manipulation for political or personal reasons.

VI. Timeliness: Data are up to date (current) and information is available on time

Designing data collection tools is not necessarily the sole responsibility of the monitoring and evaluation officers as the studies revealed however the benefits of ‘involving other stake holders are clear and well spelt out and this process can go ahead to involve programme staff, participants and/or other stakeholders in giving their input into the ; identifying key results and determining what constitutes ‘success’; contributing to collecting the data; and analyzing and interpreting the results. E.g developing data collection tools requires one to have skills in interpreting the log frame or clear understanding of the log frame is a prerequisite for developing data instruments for quality data

5.2.2 Capacity in data collection and data quality

As it is clear from this study, there is a range of data collection methods available for use in research. The study reveals that there is still lack of knowledge in the data collection methodologies and some of the data collection methodologies discussed in the study were the use of interviews and the use of observations, where for example 80% of the respondents agree to the use on observations much as their knowledge in the use of the same was contradicted. BERNARD (1994) adds to this understanding, indicating that participant observation requires a certain amount of deception and impression management. Most anthropologists, he notes, need to maintain a sense of objectivity through distance.

The most appropriate combination of methods depends on the nature of what is being researched and the context of the research. In other words, what are the research questions that need to be answered; by when must the answers be available to support decision making by different stakeholders; what resources are available for the research; and what other conditions may constrain the research? Forexample, Heavy reliance on the interviewer signifies a great need for skills, and thus proper training. In cases where interviews are being held at a larger scale, these need to be organized and trained which may require a lot of financial resources (Akbayrak: 2000). The cost factor thus begins to show up as a critical success factor in interviews. This becomes a major disadvantage, especially face to face interviews. They typically require one to be available to conduct the interviews. At the of the day as we have seen in our study results it all affects the quality of data produced. One of the data collection methodologies used is observations

5.2.2. capacity in data analysis and data quality

The findings show that there is still a gap in the data analysis amongst partner staff this is evidenced in the relationship that was drawn using regression where all the respondents who had done data analysis had not done data deduplication. And this is evidenced with the result of a coefficient of 0.000.

Even when data have been collected using well defined procedures and standardized tools, they need to be checked for any inaccurate or missing data. This is known as data cleaning, and it also involves finding and dealing with any errors that occur during the writing, reading, storage, transmission or processing of computerized data. Ensuring data quality also extends to ensuring appropriate data analysis and presentation of the data in the reports so that the findings are clear and conclusions can be substantiated. This often also involves making the data accessible so that they can be verified by others and/or used for additional purposes such as for synthesizing results. Eising 2010 adds to this by saying that "The main benefits of data analysis are rather self-evident. How can someone improve their processes and identify problematic issues if they are not willing to look at the data? The answer, of course, is that they cannot make reliable improvements without data analysis. The key word here is "reliable!" Most people have a general idea about possible changes that "should" or "could" improve their processes. However, when it comes to these sorts of changes there is the inherent risk that the change does not have the desired result. There can also be unexpected consequences that impact some other aspect of that organization in a negative manner".

which will mean logistical costs to get to the interviewee. Even when held over the telephone, the call costs have to be factored in. In addition to the expensive nature of interviews, they can also be time consuming as compared to other data collection methods**.**

5.3. Recommendations

It is advisable to use standardized data collection tools, which have already been tried and tested in real life situations, and improve these if necessary to maximize data quality. Where adaptations to the local context are necessary, or when data collection tools need to be developed, it is important to conduct a pilot test first (and improve the tool) before using it more generally. Using experienced data collectors, providing training for data collectors on a specific task or tool and/or supervising data collection across multiple data collectors can also help to reduce bias (e.g., inappropriate prompting for answers during interviews) or errors (e.g., misunderstanding which project elements need to be observed) in the data obtained.

I recommend that organizations consider the capabilities of the statistical packages, their own computer literacy and knowledge of the package, or the time required to gain these skills, and the suitability of the packages for their work. The intelligence and integrity that an organization brings to the work process must also be brought to the choice and use of tools and analytical processes. The officers should be as critical of the methodological approaches to using qualitative data analysis software as they are about the fit between research question, methods, and research design.

A particularly important issue that needs to be addressed and fully described in the data collection plan is the sampling strategy used. The designer of the data collection plan needs to Start by defining the 'population of interest' from which to draw the sample. This is the group or 'units' of interest in the geographical area of interest during the time of interest. Sampling is the process of selecting units from the population of interest to study these units in detail with the aim of drawing conclusions about the larger population. Sampling error exists within any sample. In other words, no sample will yield exactly the same information as if all people in the population of interest were included in the data collection. Hence, information collected from a sample is used to make estimates about the population of interest. Different ways of sampling will introduce different types of bias when assessing the results of a project. As such, it is important to clearly describe the strengths and weaknesses of the sampling strategy used and to address these, where possible, in the data analysis and/or describe how they affect the conclusions drawn about the project tested.

During designing data collection tools a few things have to be put in consideration and such are: poorly constructed questionnaires; failing to give people the opportunity to really talk during interviews; lack of good translators; and failure to record the data accurately. Ways to avoid or minimize these often common mistakes include: obtaining expert input into the design or adaptation of data collection instruments; ensuring good quality data collection through the selection, training and supervision of data collectors (including ensuring gender balance in interview teams); providing private spaces for the interviews so that people feel more able to express their views, or using technology to provide more ‘privacy’ in responses and to avoid obtaining ‘socially desirable’ responses; and using certified and experienced translators

Ways to support a good quality data analysis may include: ensuring ‘initial’ conclusions are challenged by holding feedback sessions with key stakeholders or intentionally searching for disconfirming evidence; looking for patterns; and ensuring ‘cross-cutting’ issues such as equity are addressed.

There is still some work to do amongst the partners as regards analysis and reporting, focusing reporting according to the different data collection instruments rather than answering the KEQs, or using complicated graphs that need a great deal of explanation in order to be understood. Good data visualization is an important tool for communicating key messages, for example, by creating a visual ‘map’ of key ideas for textual data and adding specific quotes in boxes to illustrate summary findings, or by showcasing key findings through carefully selected spreadsheets.

References:

[1] Akbayrak, B. (2000) “A Comparison of Two Data Collecting Methods: Interviews and Questionnaires”

[2] Daniel F., Casati F., Palpanas T., Chayka O., Cappiello C. (2008) "Enabling Better Decisions through Quality-aware Reports", International Conference on Information Quality (ICIQ)

[3] G.E. Liepins, 1989 “Sound data are a sound investment,” Quality Progress, vol. 22, no. 9, pp. 61-64

[4] Hansen, M. (1991) Zero Defect Data, MIT. Masters thesis

[5] Jack E. Olson (2003), "Data Quality: The Accuracy dimension", Morgan Kaufmann Publishers

[6] Kahn, B., Strong, D., Wang, R. (2002) "Information Quality Benchmarks: Product and Service Performance," Communications of the ACM, April 2002. pp. 184–192.

[7] Price, R. and Shanks, G. (2004) A Semiotic Information Quality Framework, Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato

[8] MEASURE Evaluation. (2017) Improving data quality in mobile community-based health information systems: Guidelines for design and implementation (tr-17-182). Chapel Hill, NC: MEASURE Evaluation, University of North Carolina.

[9] T.C. Redman, 1995, “Improve data quality for competitive advantage,” Sloan Management Rev., vol. 36, no. 2, pp. 99-109

[10] R.Y. Wang, H.B. Kon, and S.E. Madnick,1993 “Data quality requirements analysis and modeling,” Proc. Ninth Int’l Con& on Data Engineering, pp. 670-677, Vienna

[11] R.Y. Wang, M.P. Reddy,1992 and H.B. Kon, ‘ Toward quality data: An attribute- based approach,”

[12] R.Y. Wang, D.M. Strong, and L.M. Guarascio, 1994, Beyond Accuracy: What Data Quality Means to Data Consumers, (No. TDQM-94-10). Cam- bridge, Mass.: Total Data Quality Management Research Program, MIT Sloan School of Management

[13] Wambugu, S. & Villella, C. 2016. Health for health information systems in low- and middle-income countries: Challenges and opportunities in data quality, privacy, and security(tr-16-140). Chapel Hill, NC: MEASURE Evaluation, University of North Carolina.

[14] Woodall, P., Borek, A., and Parlikad, A. (2013), "Data Quality Assessment: The Hybrid Approach." Information & Management **50** (7), 369–382.

[15] S. B. (2009). Qualitative research: A guide to design and implementation. San Francisco, CA: Jossey-Bass Saldaña, J. (2011). Fundamentals of qualitative research.

[16] Maxwell, J. A. (2013). Qualitative research design: An interactive approach. (3rd ed.). Thousand Oaks, CA: Sage Publishing Merriam

[17] Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1): 1-1

[18] Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modelling. *Structural Equation Modeling*, *2*(3): 255-273.

ANNEX

Dear Respondent,

I am VIVIAN NABUKEERA a Master of Science student in Monitoring and Evaluation (MSc. M&E) at Uganda Martyrs University.

I am undertaking a research study entitled, "data quality and staff capacity" with reference to Vi Agroforestry. This research is a partial requirement for the award of a master degree of science in monitoring and evaluation of Uganda Martyrs University.

You are randomly selected to respond to the following interview questions to the best of your knowledge for the success of this research. The interview will take only five to ten minutes of your time. Your responses and information provided shall be treated confidently for academic purposes strictly.

Thanks

1. Respondents Background information

a) Organization :…………………………………………………………………….

b) Designation **………………………………………………………………………..**

c) SEX

Male ☐ Female ☐

D) Age group

18-27 Years	27-36 Years	36-45 Years	45-54 Years	Above 54 Years

E) Level of Education

Primary	
Secondary	
Post Secondary	

2. Trainings on data

a) Have you ever had any training on data management?

YES ☐ NO ☐

b) IF YES What kind of trainings did you get?

c) When was the training done (year)…………………………………….

3. Designing of Data collection tools(tick)

a) What criteria do you follow when designing a data collection tool?

……………………………………………………………………………………

b) Rate your Knowledge on the following data collection instruments

	Very good 5	Good 4	Fair 3	Poor 2	No knowledge at all 1
questionnaire					
checklists					

Survey approach					

c) Do you carry out data quality assessment in your organization?

YES ☐ ☐ NO

d) Which of the following measures do use when testing the quality of the data collection tools? (**tick**)

I. Self-evident measures (refers to the fact that the instrument appears what it is supposed to measure) ☐

II. Pragmatic Measures (procedure tests the practical value of the research tool) ☐

e) As a designer of data collection tools, rate your practice on the following

	Strongly disagree 5	**Disagree 4**	**Neither agree nor disagree 3**	**Agree 2**	**Strongly agree 1**	**comments**
Must know intervention results before designing tools						
I am in position to define indicators						
I have ability in Constructing statements which statements						
develop a reporting format before data collection						
I discuss developed tools before data						

collection with the project stakeholders						
I pretest developed tools put under quality						

4. Data collection

a) What kind of data do you collect in your organization? May be specify

……
………..

……
……………

b) Are you aware of the existing data collection methods?

YES ☐ NO ☐

Rate your Knowledge on the following data collection methods

	Strongly disagree 5	**Disagree 4**	**Neither agree or disagree 3**	**Agree 2**	**Strongly Agree 1**
Structured observations					
Unstructured observations					
Personal interviews					
Telephonic Interviews					
Focused group discussions					
Case study method					
Survey method					

5. Which of the following data collection methods do you use when collecting data?(tick)

a) Observations may be they could give reasons each used

b) Personal interviews

c) Focused group discussions

d) Case studies

e) Survey methods

6. Do you know about data cleaning?

YES ☐ ☐

7. **I Do have knowledge on these data cleaning processes (**Data editing, Data validation, Imputation)

	Strongly disagree	**Disagree**	**Neither agree or disagree**	**Agree**	**Strongly Agree**
Data editing					
Data validation					
Data imputation					

a) How do you handle data cleaning?

……

……

………………………………………………………………………………..

8. Do you perform any of the data cleaning processes below(tick)

	YES	**NO**
Data editing		

Data validation		
Data imputation		
De duplication		

What are the challenges associated with data cleaning?

……

……

…………………………………………………………………………………………….

9. **Do you have knowledge about data coding?**

YES ☐ **NO** ☐

10. **Have you ever carried out data coding?**

YES ☐ **NO** ☐

11. **How do you carry out data coding?**

……

……

……………………………………………………………………………….

12. **Where do you store your data in your organization?**

……

…………………………………………

……

…………………………………….

13. What do you require when beginning data entry?(tick)

a) Photocopy of the report you are entering

b) Highlighter and pencil

c) Data dictionary

14. During data entry I carry out the following

	Strongly disagree 5	Disagree 4	Neither agree nor disagree 3	Strongly agree 2	Agree 1
Use of working dictionary/ glossary					
Allocating Sample identification(unique identifiers)					

15. Data validation

a) I have knowledge on the following data validation procedures

	Strongly disagree	**disagree**	**Neither agree nor disagree**	**Strongly agree**	**Agree**
Test for missing data					
Test for correct field length					
Test for class or composition use one					
Test for range or reasonableness use one					
Test for invalid values					
Cross reference checks					

Test for comparison with stored data					

16. Actual Data Analysis

b) Do you store data in your organisation

c) Which software do you use to store data in your organization?

d) Have you ever carried out data analysis before?

YES ☐ NO ☐

e) How is data analysis in your organization done?

……

……………………………………………………………………

f) what are the associated challenges

……

…………………………………………

17. Rate your Knowledge on the following

	YES	NO	Comment
Descriptive statistics(average, mode, median)			
Data visualization			
Data integration(combining data residing in different sources and providing			

Interview tool for data collectors

18. Respondents Background information

a) Organization :……………………………………………………………………….

b) Designation **………………………………………………………………………..**

c) SEX

Male ☐ Female ☐

D) Age group

18-27 Years	**27-36 Years**	**36-45 Years**	**45-54 Years**	**Above 54 Years**

E) Level of Education

Primary	
Secondary	
Post Secondary	

19. Have you had any trainings on data collection?

YES ☐ ☐ NO

20. If yes how has the trainings impacted you as far as data collection is concerned?

……

……

……………………………………………………………………………………………..

21. Do you enjoy your experience of data collection?

YES ☐ NO ☐

22. What are the challenges you face during data collection?

………

………

………

………

……………………………………

23. Which of the following data collection methods do you use when collecting data?(tick)

f) Observations

g) Personal interviews

h) Focused group discussions

i) Case studies

j) Survey methods

24. Have you ever used observation as a data collection methodology?

Yes ☐ No ☐

25. Which observation types do you employ?

- Direct observation (observation of an event personally by the observer when it takes place)

- Indirect observation

26. Have you carried out interviews before?

YES ☐ NO ☐

27. What kind of interviews have you carried out? (tick)

    Face to face interviews

    Focused group interviews

28. Have you had trainings on interview skills?

    YES ☐ NO ☐

29. I use the knowledge got from the interview skills training

1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree
5. Neither agree nor disagree

30. I am objective when collecting data

    1. Strongly agree
    2. Agree

3. Disagree
4. Strongly disagree
5. Neither agree nor disagree

31. Rate your knowledge on the use of the following data collection instruments.

	Very good 5	Good 4	Fair 3	Poor 2	No knowledge at all 1
questionnaires					
checklists					
Survey approach					

32. Have you used the mobile data methodology before (MSALM platform)?

YES ☐ NO ☐

33. What’s your experience in the use of mobile data collection methodology?

……

……

……

……

……………………………….

34. What challenges have you faced with using the mobile data collection methodology?

……

……

……

……

………………………….