

**RESEARCH METHODS AND APPROACHES IN ‘INNOVATION  
IN LIVESTOCK SYSTEMS’ RESEARCH PROJECTS AT  
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE,  
ADDIS ABABA**

**MULUGETA YITAYIH BIRHANU**

**MASTER OF SCIENCE**

**(Research Methods)**

**JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND  
TECHNOLOGY**

**2011**

Research Methods and Approaches in ‘Innovation in Livestock  
Systems’ Research Projects at International Livestock Research  
Institute, Addis Ababa

Mulugeta Yitayih Birhanu

A dissertation submitted in partial fulfillment for the Degree of  
Master of Science in Research Methods in the Jomo Kenyatta

University of Agriculture and Technology

2011

## Declaration

This dissertation is my original work and has not been presented for the award of a degree in any university.

Student Signature: ..... Date:.....

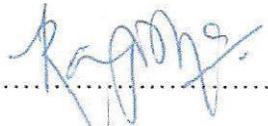
Mulugeta Yitayih Birhanu

This dissertation has been submitted for examination with our approval as University supervisors.

1. Signature:..... Date:.....

Dr. Edward George Mamati

Jomo Kenyatta University of Agriculture and Technology, Kenya

2. Signature:  ..... Date:.....

Dr. Ranjitha Puskur

International Livestock Research Institute (Addis Ababa)

## **Acknowledgments**

My utmost gratitude goes to the Regional Universities Forum for Capacity Building, (RUFORUM), for funding all my study, living and research project cost. I am also very grateful to Office of Dean, Faculty of Agriculture, JKUAT, and International Livestock Research Institute (ILRI) for arranging projects to accomplish all the attachment tasks and creating favorable conditions during my attachment period.

This research project report would not have been possible without the guidance and help of my supervisors: Dr. Edward Mamati and Dr. Ranjitha Puskur, who provided all the necessary guidance and committed their valuable time in the preparation and completion of this project. I am heartily thankful to the Innovation research team leader, Dr. Ranjitha Puskur and Mr. Antenh Girma, whose encouragement, guidance and support from the initial to the final stages enabled me to develop an understanding of the different research projects at the host institution and facilitated in the participation of different research activities.

I am also grateful to ILRI CAST unit, Jane Poole, Richard Coe and Roger Stern of SSC for their support and assistance during the initial stages of the attachment period. A great deal of appreciation is due to the members of the Faculty Agriculture JKUAT for their valuable contributions.

Finally, I would like to appreciate all graduate and post graduate students who shared their data freely and provided me the tools that enabled me to collect data to learn different research methods activities in their projects.

## Table of contents

<b>DECLARATION</b> .....	<b>III</b>
<b>ACKNOWLEDGMENTS</b> .....	<b>IV</b>
<b>LIST OF TABLES</b> .....	<b>VII</b>
<b>LIST OF APPENDICES</b> .....	<b>VIII</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>VIII</b>
<b>ABSTRACT</b> .....	<b>X</b>
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1. BACKGROUND .....	1
1.2 OBJECTIVES OF THE PROJECT .....	4
<b>2. LITERATURE REVIEW</b> .....	<b>5</b>
2.1 RESEARCH PARTICIPATION /CONSULTANCY .....	6
2.2 TRAINING IN THE RESEARCH PROCESS .....	8
2.3 WRITTEN REVIEW OF REPORTS, PROPOSAL AND PROTOCOLS .....	12
2.4 DATA MANAGEMENT .....	15
2.5 DATA ANALYSIS .....	16
<b>3. PROJECT APPROACH</b> .....	<b>20</b>
3.1 PARTICIPATION IN RESEARCH / CONSULTANCY .....	20
3.2 TRAINING IN THE RESEARCH PROCESS .....	21
3.3 REVIEW OF REPORTS AND PROPOSAL .....	22
3.4 DATA MANAGEMENT .....	23
3.5 DATA ANALYSIS .....	25
<b>4. RESULTS</b> .....	<b>28</b>
4.1 RESEARCH PARTICIPATION .....	28
4.1.1 <i>Participation in the team projects</i> .....	28
4.1.2 <i>Participation in students research projects</i> .....	31
4.2 TRAINING IN THE RESEARCH PROCESS .....	37
4.3 REVIEW OF RESEARCH PROPOSAL AND REPORT .....	40
4.3.1 <i>Review of proposal results and feedback</i> .....	40
4.3.2 <i>Review of report results and Feedback</i> .....	43
4.4 DATA MANAGEMENT .....	45
4.4.1 <i>Data management challenges</i> .....	45
4.4.2 <i>Data Management support</i> .....	48
4.5 DATA ANALYSIS .....	58
4.5.1 <i>Determining the factors which affect the intensity of forage innovation</i> .....	58

4.5.2	<i>Possible data analysis from previous studies</i>	63
4.5.3	<i>Data analysis to summarize baseline information</i>	66
5.	<b>DISCUSSION</b>	<b>67</b>
6.	<b>CONCLUSION</b>	<b>98</b>
7.	<b>RECOMMENDATIONS</b>	<b>99</b>
8.	<b>LESSONS LEARNED</b>	<b>101</b>
9.	<b>REFERENCES</b>	<b>106</b>
10.	<b>LIST OF APPENDICES</b>	<b>111</b>

## List of Tables

Table 1: Sample descriptive summary of different parameters before the outbreak, after the outbreak and now .....	30
Table 2: Sample parametric data analysis for likert scale data .....	34
Table 3: Sample Non Parametric equivalent test for the above parametric test .....	35
Table 4: Sample child Multidimensional poverty measure in Ethiopia .....	36
Table 5: Sample data analysis on the Incidence of child deprivation by region .....	37
Table 6: Training evaluation results by the trainees.....	39
Table 7: Descriptive statistics of land use without missing value codes.....	46
Table 8: Descriptive statistics of land use with missing value codes.....	47
Table 9: Sample descriptive statistics which indicates the effect of missing data handling options .....	50
Table 10: The effect of different missing data handling option during regression example.....	50
Table 11: Part of Heckman two stage model results with and without outliers .....	52
Table 12: Descriptive statistics of Intensity of Forage innovation for Successful and Unsuccessful cases .....	58
Table 13: Heckman first stage model for determinants of success in forage innovation.....	60
Table: 14 Heckman two stage outcome equations, OLS and Tobit model estimates of the factors which affect intensity of forage innovation.....	62
Table 15: Different statistical packages estimate for factors which affect the success in Trypanomosis management.....	64
Table 16: Sample summarized data on dominant land use at Ghibe valley in 1997 .....	66

## **List of appendices**

Appendix 1: Review of proposal comments to principal investigator .....	111
Appendix 2: Review of report comments to author .....	114
Appendix 3: Review of report guideline questions .....	119
Appendix 4: Review of proposal guideline questions .....	121
Appendix 5: Consultancy report summary format .....	123
Appendix 6: Data management plan .....	124
Appendix 7: Data analysis plan.....	130
Appendix 8: Household Livelihood and socio organizational impact of animal health interventions study first draft questionnaire.....	136
Appendix 9: Description of independent variable for Heckman two stage model.....	147
Appendix 10: Description of explanatory variables for binary logit model.....	150

## **List of Abbreviations**

AMMI: Additive Main effect and Multiplicative Interactions

CBTTC: Community Based Trypanosomosis & Tsetse Control

CTI: Clinical Tools Inc

FAO: Food and Agricultural Organization

GAO: USA General Accounting Office

ILRI: International Livestock Research Institute

IWMI: International Water Management Institute

JKUAT: Jomo Kenyatta University of Agriculture and Technology

ORI: Office of Research and Integrity (US department of Health and Human services)

POST: Parliamentary Office of Science and Technology

RUFORUM: Regional Universities Forum for Capacity Building

SPSS: Statistical Packages for Social Sciences

## Abstract

The project was undertaken at the International Livestock Research Institute (ILRI, Addis Ababa) in Innovation in Livestock Systems Research team. It assessed the research methods, approaches and practices followed and, challenges faced thereof by the 'Innovation in Livestock Systems' research team projects. The main purpose of the attachment was to accomplish different research methods tasks in the project and learn from the research methods activities of the project and other similar research projects. Research methods tasks comprised of research consultancy, review of research reports and proposal, introduction to the statistical software SPSS, data management and data analysis. Research consultancy to researchers and graduate students was provided based on collaborative approach. Most of the consultancy tasks focused on data management and data analysis aspects of the research projects. Challenges observed during consultancy indicated that the need of involving research methods professional starting from the inception stages of the projects. The proposal "Household livelihood, and socio-organizational impact of animal health interventions" and research report "The value of Innovation system analysis for livestock research and policy" were reviewed based on designed review guidelines. These review indicated that researchers should work more on organization of their report. Introduction to the statistical software SPSS was provided to small group of research technicians for International Water Management Institute. The training task indicated the presence of limited capacity in using different statistical tools. The data management tasks mainly focused on design of data collection instrument, data organization, data cleaning and data storage for different data sets from the team and student projects. Limited data management practice was observed among researchers and students. Data analysis to determine the factors which affect the intensity of forage innovation among positive deviants were undertaken using OLS, Tobit and Heckman Two stage models. Heckman model was found to be appropriate due to the presence of selection bias among the positive deviant samples. The result indicated that training, intensity of collaboration, dairy cooperative membership, innovation index, total livestock units and forage technologies with follow up will positively affect intensity of innovation in forage technologies. In additions, secondary data analysis on factors which affect the success in Trypanomosis management was also undertaken using logit model. The data was analyzed using four statistical software; Stata 10, SPSS16, R and GenStat 12 and all software used in this analysis gave comparable results. The result indicated that cut and carry livestock management system, intensity of access to drugs and total time for treatment will positively affect success in the disease management. Finally, baseline data for "Household livelihood, and socio-organizational impact of animal health intervention" study was consolidated from Ghibe household surveys data sets undertaken in 1997 and 1998. During the analysis challenges related to secondary data analysis such as lack of data quality were observed. For each of the above tasks the approach used, sample results, challenges faced and various lessons learned were summarized in this dissertation.

# **1. Introduction**

## **1.1. Background**

Universities in Africa have been producing professionals who serve as Scientists, Instructors, Research Assistants and Technicians in universities and research organizations. These professionals have limited research methods knowledge and skills to achieve a wide range of research objectives (Kabiru *et.al* 2008 and Teferra and Altbach 2004). This has been a challenge to researchers and graduate students in Africa to effectively accomplish their research projects and produce the required outputs which contribute to the development process of the continent. In order to tackle those challenges and improve the quality of research in Africa, different strategies have been put in place by local and international organizations. One of the main strategies has been improving the research methods knowledge and skills of professionals through research methods training or providing research methods support through research methods professionals during the course of different research projects.

Research methods professionals can be produced by integrating theoretical course works with practical activities such as by attaching students to different research projects in research institutions or universities after their course works. International Livestock research Institute (ILRI) was one of the Hosting institute for those research methods students after taking one year course works. Hence, this dissertation was produced from one of the research team in this institution, Innovation in Livestock Systems Research team. ILRI undertakes research in the following three main thematic areas: biotechnology; enhancing market opportunities; People, Livestock, and Environment (PLE). Under each theme, there are different research teams carrying out research in different fields through specific projects. Innovation in Livestock Systems Research team, which is part of the market opportunities theme, examines the processes and mechanisms by which research and knowledge generated from research conducted by ILRI and its partners are used to generate development outcomes that benefit resource poor people whose livelihoods

depend on livestock (Puskur, 2010). The research activities of the team focus on the interface between ILRI's research outputs and development outcomes. Research under this team concentrates on understanding sources of innovation in specific livestock contexts, analysis of research and non-research actors, their linkages and interactions, and analysis of policy and institutional environments (Puskur, 2010). The research activity in this team is based on qualitative and quantitative data collected from primary and secondary sources.

During this attachment period research methods tasks such as research consultancy, review of research proposal and report, introductory training on the statistical software SPSS, data management and data analysis tasks were accomplished. These research methods tasks were accomplished on the following projects: four completed studies "Positive Deviance in the Use of Improved Livestock Feed Resource in developing countries", "Ghibe Baseline Innovation Study", "Household survey for Ghibe Valley, Ethiopia" and "Impact assessment study of the Nigeria avian influenza control and human pandemic preparedness and response project"; one ongoing study "Analysis of projects/networks on land and water management in the Nile basin of Ethiopia" and one proposed study "Household livelihood, and socio-organizational impact of animal health intervention".

In addition to the above projects, research consultancy on data management and data analysis issues was provided to graduate students research projects; "Multidimensional child poverty assessment in Ethiopia", "Assessment of business reengineering implementation in Ethiopian Civil Service college", "Genotypic variance of double haploid (DH) lines under well watered and managed drought environments", "The Role of Cooperative in Triggering Natural Resource Management: the Case of Tewodros Incense production Cooperative in Amhara Region of Ethiopia" and "The role of coffee cooperatives performance in small holders livelihoods."

The task of research participation/consultancy focused on research design, data collection and data analysis in the team and student projects. Researchers and research students were consulted mainly on issues of data management and data analysis. Training on the use of statistical software SPSS was given to research technicians and Enumeration training

manual for “Household livelihood, and socio-organizational impact of animal health intervention” study was developed to use as training material and reference during data collection. Research report review on “The Value of Innovation Systems Analysis for Livestock Research and Policy” study report and proposal review on “Household livelihood and socio-organizational impact of animal health intervention” study protocol was undertaken based on the review guidelines developed for each of the manuscripts. The review comments and suggestions were given to the author and principal investigator for further improvements and to get their feedback on the review processes. Data management tasks especially development of data collection instrument, data cleaning, data organization, and preparation of data files for archiving were undertaken for different data sets in the research team. These data management tasks were accomplished based on the general guidelines and plan developed at the beginning of the attachment. Data analysis to determine the factors which affect the intensity of forage innovation was undertaken using OLS, Tobit and Heckman Two stage models and the report of the most preferable model was suggested to the team. In addition, data analyses to determine the factors which affect the success in Trypanomosis management were also undertaken using logit model. This data was analyzed using four statistical software packages (Stata 10, SPSS16, R and Gnstat 12) and all the software gave similar results. Baseline information for “Household livelihood and socio organizational impact of animal health interventions study” was consolidated from different data sets of “Household survey for Ghibe valley Studies.” Finally, from all the above tasks the different challenges of researchers, possible solutions suggested, lesson learned and their implications were documented for further reference.

## **1.2 Objectives of the project**

The overall objective of the project was to assess the research methods approach used in Innovations in Livestock Systems Team at ILRI –Addis Ababa and to accomplish a set of structured research methods tasks to provide research methods support to the team. The specific objectives of the project tasks were:

- To determine the challenge faced by researchers, research assistants and graduate students in research methods issues of planning, data collection, data management and data analysis.
- To evaluate the quality of research reports and proposals emanating from the team and develop recommendations.
- To capacitate researches and support staffs in undertaking research.
- To identify common data management challenges and develops appropriate data management practices.
- To determine factors that affects the intensity of forage technologies innovation among positive deviants in forage technologies in Ethiopia.
- To identify the factors which affect the success in Trypanomosis (Gundi disease) management in Ghibe valley Ethiopia by using data sets from previously completed study.
- To consolidate information for “Household Livelihood and socio organizational impact of animal health intervention study” from different data sets.

## **2.Literature Review**

Research process involves identification of relevant and important research problem, selection of appropriate methodology to conduct research, review of previous studies (to identify research gaps, major findings and current understanding about a given researchable problem), data management (to ease the data analysis process and retain complete, accurate and retrievable data), and data analysis (to extract the required information from the data). The quality of research projects is significantly affected by the way in which each of the above activities is accomplished in the research process. Sometimes projects fail to meet the expected objectives due to poor design, poor data quality, incorrect statistical tests used, and violation of assumptions and incorrect interpretation of results. To minimize the risk of such problems team work, involving different professionals with different knowledge and technical background is critical.

A good research team should include at least a statistician or research methods professional who can contribute to the design, data management and data analysis of the research process. In addition, a research methods professional can also support the research team by providing training, suggesting ideas in review meetings and planning, and reviewing of research proposal and reports. However, these research methods professional should have a sound grasp of ideas and knowledge on the methodological aspects of each of the steps of the research processes so as to provide appropriate advice leading to valid conclusions and inferences being drawn ( Ander and Mellenbergh, 2008).

Different scholars have elaborated their understanding and findings about the purpose, process and contribution of research participation/consultancy, training, review of research reports and proposal, data management, and data analysis tasks in the research process. The following sub-sections review the main principles and approach for various dimensions and summarises them.

## **2.1 Research participation /Consultancy**

Traditionally statisticians have provided research methods support or consultancy service to researchers during research conceptualization, study design, data management, selection and execution of data analyses and interpretation of result. Research organizations and universities have been increasingly demanding this research methods support or consultancy service from professionals or experienced researchers. Consulting is broadly defined as a process of transferring expertise, knowledge, and/or skills from one party (the consultant) to another (the client) with the aim of providing support or solving problems (Druckman 2000; Gallessich 1985). Cox (1986) described research consultancy as a craft, an art, scarcely a science, which cannot be taught but must be learned, acquired by a process similar to progressive apprenticeship under the guidance of a master. The main objective of research consultancy is to assist scientists in ascertaining the truth by providing the logic of the scientific method and contribute to the research process based on knowledge and experience gained from previous work (Clayton, 1996a).

A consultant should be a source of advice and guidance usually based on some actual, perceived or claimed specialized, knowledge, skills or expertise (Crowther & Lancaster, 2005). Therefore, a consultant should have good theoretical knowledge and practical skills in the subject areas where the consultancy service is provided. Time, patience, understanding, and empathy are all essential skills especially for collaborative consultation, but none can take the place of technical competence in pursuit of a scientific truth (Strickland, 1996). Clayton (1996b) also indicated that statistical consultation is a multifaceted operation, which requires sound logic, insightful analytical training, and client centred interpersonal communication skills. To become a successful consultant, research methods professional needs to acquire the theoretical expertise from the acquired knowledge and practical skills from addressing the problem they encounter (Hand and Everitt, 1988, Boen and Zahn, 1982).

According to Batanero (2000), the role of consultant in empirical research is complicated by three facts; (a) researchers use different research methods to achieve the same goals, (b) the theoretical, practical, and statistical constraints on researchers differ when the researchers have different objectives, and (c) the relationship between the substantive and statistical hypotheses is often oversimplified by researchers. This makes the task of consultants very difficult and challenging. However, the involvement of the consultant in the research teams has the following two advantages. Firstly, it helps to ensure that complex data are correctly analyzed and interpreted and secondly, it enables the researcher to learn new developments in statistical procedures and software tools (Batanero, 2000). Nevertheless, to get the job done well it is important for the consultant and researcher to understand the limits and extent of the data, the statistical tools, and the amount of time available for the consultant to analyze the problem.

Scientific consultation can have different approaches. Blake and Mouton (1976) describe five approaches that can be used to consult: prescriptive, theories and principles, confrontation, acceptance and collaborative consultation. In prescriptive consultancy, the researcher will be told what to do in line with his challenge. It is the responsibility of the consultant to make the decision. In theories and principles, the consultant helps the researcher to internalize theories so as to deal with the challenge of using new theories. Confrontation approach challenges the researcher's assumptions to view the challenge in a new perspective. In acceptance consultancy, the consultant reassures and supports the researcher's approach to deal with the situation.

Boen and Zahn (1982) also discussed straight and collaborative approaches. In the straight approach, the researcher asks the consultant for a solution to a problem and the consultant takes up the problem, works on it, and then provides the researcher with a written or verbal presentation of the analyses and the findings. In the collaborative approach, the researcher is looking for ways to improve his or her research methodology and statistical operations and the emphasis is on the consultant working with the researcher to clarify goals and achieve desirable research objective which can contribute to improving the livelihood of

the community or contribute to scientific knowledge; therefore, the researcher also provides the consultant an opportunity to learn more about the scientific problem and possible recommended solutions which also helps the research methods fellow to learn from the reach experience of researchers.

Generally the role of research methods professional is enormous starting from project conception to report writing and data archiving. Ander and Mellenberg (2008) summarized consultancy questions pertaining to research design, measurement, data analysis, and report writings which almost cover the entire research process. When researchers use complicated design and approaches, the need for statistical consultancy becomes much more acute and important.

As a research methods professional and trainee student the collaborative approach is the best and most preferable approach which should be followed in research institutions like ILRI which has a lot of expertise and professionals in different research methods issues. This approach helps both the consultant and researcher to share their experiences and ideas each other in order to solve problems. In it will empower the researchers and make them competent as the research methods fellow mainly assists them to know the methodological principles and idea behind each of the challenges and let them to solve similar challenges by themselves in the future. So they will not depend on the consultant and the consultant will also learn from those experienced researchers to improve his skills and knowledge on different subject while they are working together.

## **2.2 Training in the research process**

Training is a process in which a trainer and trainees work together to transfer information from the trainer to the trainees to develop the latter's knowledge, attitudes or skills so that they can perform tasks better (Health link Worldwide, 2003). It can also be described as an educational process in which people can acquire new information, re-learn and reinforce existing knowledge and skills, and most importantly have time to think and consider what new options can help them improve their effectiveness at work (Amherst College, 2010).

Training involves the development or strengthening of knowledge, skills, and attitudes of an individual to help to accomplish the expected tasks effectively and efficiently.

There are many factors which need to be taken into consideration when designing training processes to support learning and change in behaviour. The most important and essential element is the specification of learning goals in line with the broader capacity development program and priorities of a given organization. Hackett (2003) also mentioned that well formulated training programme has four key stages: defining training needs; designing and planning training; providing the training; and, evaluating the outcome of training. The World Bank (2008) evaluation study also found the issue of training design to be of critical importance, and within that targeting of training content was found to be the most important training design factor driving training success. In addition, this evaluation also found that for well-targeted training, organizational and institutional capacity gaps need to be correctly diagnosed, specific training needs must be assessed, and participants should be selected in a strategic manner.

Most experts agree that human learning, training, and performance improvement initiatives should begin with a needs assessment. Needs assessment is a process for examining and framing people related problems and performance improvement opportunities. It might be initiated in response to a problem or opportunity, or it might be used in ongoing learning or performance improvement efforts (Gupta, 2007). According to Gupta, there are four approaches for training needs assessment: first, knowledge and skills assessment; second, job and task analysis; third, competency-based needs assessment and the last are strategic needs. Training needs can be assessed using survey, interview, or other data gathering instruments.

The impact of training for research and development activities is enormous. Researchers should give training and/or should take training about new research methods approaches and new statistical software. According to the finding of training need assessment survey at National Centre for Research Methods, PhD students, junior researchers and academic

employers of research staff need training in some specific issues such as in interviewing; qualitative analysis ; statistics/quantitative methods (at all levels); use of statistical software; and, longitudinal data analysis (Wiles *et.al*, 2005). This survey also indicated that researchers, especially at more junior levels, recognize the need for training in a range of methods.

For successful training all the necessary training materials such as instructional manual, reference materials, power point presentations, software, data sets and other relevant information should be prepared in advance based on the objective and content of the study. Each of the training materials should be designed in a way to support the training objectives and goals. For instance, instructional manual for enumeration training can serve as training guide during the training and as a reference for answering questions, which may arise during the interviews, because content of the manuals should refers to all the questions on the questionnaire and includes concepts and definitions, techniques of interviewing and duties and obligations of enumerators and supervisors (FAO, 2010).

A team which includes researchers and other research support staff with the right skill and knowledge is one of the main components of a successful research project. Team members need training in basic job-related research skills as well as specialized training relevant to the unique clinical, theoretical, and social complexities of the research being conducted (Nelson and Beddy, 2008). According Nelson and Beddy training can be conducted either at individual level or at group level. Individual training involves the research team members completing a training exercise in order to gain knowledge or skills, but doing so independently of other team members. Group level training is analogous to “classroom style” teaching, in which a trainer provides explicit instruction via lecture format and then answers questions from the audience.

Training can be given at any time and point in the research process depending on the objectives of the training. For instance, in research institutions, training can be given to researchers, research assistants, laboratory technicians, enumerators and supervisors and or

management personnel. Enumeration training is one of the most common practices before survey, census, and other similar studies to help them properly undertake the data collection process. Training of supervisors and enumerators before and during survey significantly affects the quality of the data collected and eventually the final output of the study. An important aim of enumerator training is to develop enumerators' capacity to motivate respondents to give complete and accurate answers and to ensure the quality of data collected. This training should also equip the enumerators with the knowledge and skills for doing their job well, since survey requires a special technical skill and experience. The exact amount of time devoted to enumerators training, the design of training sessions, and the content of the training program depend on particular organizational setting and what interviewers are going to be doing (Fowlers, 1984). Researchers and research assistants also need training on issues related to new research methods and approaches and statistical software or tools in order to cope with emerging and new challenges and approaches in the field of enquiry.

Training should be an interactive process in which both the trainer and trainees participate in the training process actively. According to Etling (1997) successful training can be achieved through infusing flexibility in presenting the topics based on the trainees need, providing different options at each step of the training, choice and sequence of topics based on needs of participant , group exercise or practical work. A trainer's primary role is to help participants learn and a good trainer encourages participants to discover things and learn for themselves (Health link worldwide, 2003). According to this paper, three things can help stimulate participants' curiosity in training: involving people as active participants in the learning process, rather than passive recipients of information; ensuring the training is relevant to the participants' day-to-day work and; using a variety of media and methods. These points are very important especially for training related to research methods. Training sessions can include mainly lectures or presentation, discussion and practical sessions. The practical sessions helps the trainees to relate training to their jobs or help them practice what has been covered in the presentation session. The practical session can

be scheduled after the presentation session, so trainees can apply the theory they have learned into practice.

Evaluation is one of the most important aspects of any training events. The trainer should explain the main objectives of the training at the beginning of the training. At the end of the training, the trainees should be asked to assess the effectiveness of the training programme, the main purpose being to provide feedback to devise better training approaches in the future. Evaluating the training which includes monitoring addresses how one determines whether the goals or objectives were met and what impact the training had on actual performance on the trainees. According to FAO (2011) evaluation on the basis of the time dimension, evaluation may be classified as formative evaluation and summative evaluation. Formative evaluation involves the collection of relevant and useful data while the training programme is being conducted to identify the drawbacks and unintended outcomes and it is helpful in revising the plan and structure of training programmes to suit the needs of the trainees and Summative evaluation is done at the end of the programme and makes an overall assessment of its effectiveness in relation to achieving the objectives and goals. Generally good training should be designed on the basis of the trainees' needs and delivered with the active participation of the trainees. Training evaluation is extremely important to assess the result of any training programme based on the pre specified goals and objectives of the study. Training should be evaluated by the trainers and trainees during and at the end of the training.

### **2.3 Written review of reports, proposal and protocols**

Review of research reports and proposals has become one of the most common practices in the scientific community to support scientific quality, maintain ethical standards, avoid inappropriate research, train and nurture researchers, improve the research process, and to further advance knowledge (Harvey, 2009). The most common method of reviewing research proposal and reports is peer review. Peer review is a system whereby independent experts assess the quality of the manuscript to ensure that the science is sound (POST, 2002).

A global survey conducted by Mark Ware Consulting (2008) found that peer review is widely supported and it improves the quality of published paper. Peer review can be used for allocation of research funding, publication of research findings in scientific journals and to assess the quality of research conducted by different units (Godlee & Jefferson, 2003). The ability to develop fundamentally sound, objective critiques of scientific manuscripts is an essential element of the peer review process and it is an important professional skill for investigators (Seals and Tanka, 2009). Peer review is mostly applicable in scientific journal publication or research reports than in research protocol. However, the quality of scientific papers cannot just be maintained by reviewing only the final report of the research but also by the review of research protocols and proposals which can also add more value during the beginning of the study before investing any resources.

Eysenbach (2004) described the advantages of the research protocols review as: highlighting good-quality studies at an early stage, contributing to a register of selected trials, reducing publication bias against negative (neutral) or incompetent findings, promoting recruitment of cooperating centre and trial participants, helps researchers in funding applications, preventing the implementation of poor research, preventing data dredging by documentation of intended analyses and it establishes priority of an important idea. Peer review of protocol will encompass suggestions for improvement and an expert opinion on the value of the research plan, which helps the author to incorporate the suggested changes and resubmit the revised version.

According to Godlee and Jefferson (2003) peer reviewer gives useful comments on the importance, originality, presentation, validity, ethical aspects, and other features of the study. Therefore it improves the quality of research report. Published scientific results should also be scrutinized by experts in the field and given scores for quality and quantity according to established rules (Seglen, 1997). Hames (2007) states that peer review helps to prevent publication of bad work, to check that the research reported has been carried out well and there is no flaw in the design and methodology. He also added that peer review

ensures that the work is reported correctly and unambiguously, the results presented are interpreted correctly and helps select work that will be of the greatest interest to the readership.

Peer review can be slow and expensive, inconsistent, biased and abused (Smith, 2006). However, Smith has also mentioned that peer review is likely to remain central to science and journals because there are no obvious alternatives that can substitute peer review and, scientists and editors have a continuing belief in peer review. He also added that peer review can be improved through standardizing procedures; opening up the process; blinding reviewers to the identity of authors; training reviewers; being more rigorous in selecting and deselecting reviewers; using electronic review; rewarding reviewers; providing detailed feedback to reviewers; using more checklists; or creating professional review agencies.

Before the review of any manuscript it is critical to develop good reviewing criteria. Ramos *et.al* (2008) classified review criteria into those focusing on structural and validity issues. The structural component includes theoretical components (introduction and discussion), experimental design (the methods section), and the results sections. The validity component includes internal validity (why the inferences on the effect of a given independent variable can be incorrect?), and external validity (how the inferences can be generalized across populations, contexts, etc.?).

Even though the role of peer reviews in the research process is still controversial, it is very important to the integrity of the research process. Peer review should be conducted based on clear guidelines and criteria in order to minimize its shortcomings by standardizing the process. Researchers and research methods professionals who are participating in the scholarly arena have a professional obligation to participate in the review process in order to support researchers, maintain the quality of research projects and contribute to the production of knowledge.

## **2.4 Data management**

Data management can be defined as the process of designing data collection instruments, maintaining data sheets, entering data into computer files, checking for accuracy, maintaining records of the processing steps, and archiving it for future access (Patel *et.al* ,2004). It is a process that ensures the data obtained as part of the study protocol is accurate, of high quality to verify the results obtained, can be reproduced, and is stored according to the legal requirements. Effective data management practice based on clear understanding of the objectives and purpose of the study should be in place starting from the conception stage of the research project. Data management has multifaceted and complex effects on every research activity and should be addressed before starting any data collection activity. But in order to avoid unexpected problems, unnecessary corrective steps, and delays in data verification and analysis, it is important to consider data management issues during all phases of the research project (Scheleicher &Saito, 2005).

Data management practice ensures that the variability in the data derived from the phenomena under study and not from the data collection process, and to facilitate accurate, appropriate, and defensible analysis and interpretation of the data (Schoenbatch and Rosamond, 2000). Managing data well enhances the quality of the scientific process, ensures that high quality data is obtained, and increases the longevity of data and opportunities for data to be shared and re-used.

According to Tavakoli (2007), data management has three stages: data preparatory, data organization, and data analysis and dissemination. Data preparatory stage includes instrument construction and refinement, development of data collection procedures, personnel training, instrument coding, software programming for data entry, planning for data set creation, development of data security procedures, and documentation. Data organization includes data entry, data conversion, data editing, data cleaning, data manipulation, merging, backup, preliminary data analysis, and documentation. The last stage: data analysis and dissemination, includes data analyses to meet objectives of the

study, data access procedures, manuscript preparation; documentation, and archiving data. Muraya, Garlic and Coe (2002) described the steps to be followed in data management as planning, designing field data recording sheets, collection of data, checking of raw data, data entry and organization, backup of data files, processing of data for analysis, checking of processed data, maintenance of a data processing log and archiving data for future use.

The Principal investigator should clearly plan the main component of data management's tasks such as data ownership, data collection, data storage, data protection, data retention, data processing and analysis, data sharing and archiving at the conception stage of the research project. This data management plan should be incorporated into the protocol of the study. The plan should include description of procedures for data collection, types of data collected, database software used, details on methodology implemented to ensure data validity and quality, and reference to the delegation of those responsible for data collection.

Generally data management is one of the essential components of a good research project to which researchers should pay due attention from the inception stage of the project. Without a responsible data management system it is not possible to support individuals or group research, optimize investment in research, reuse and recombination of different data sources, disseminate research results and recommendation and, maintain scholarly records. Therefore research organizations should have in place a data management policy starting from the inception stage of research projects.

## **2.5 Data analysis**

Data analysis is a process, which can be used to establish what the research findings are and how they answer the research questions. Concern with analysis should (1) begin during the design of a study, (2) continue as detailed plans are made to collect data in different forms, (3) become the focus of attention after data are collected, and (4) be completed during the report writing and reviewing stages (GAO,1992). Data analysis helps to makes sense from large amount of data, demonstrate how data support or do not support

a theory and communicate our findings quickly and convincingly. Before starting any data analysis, one needs to understand the data properly. The main goal of data understanding is to gain general insights about the data that will potentially be helpful for further steps in the data analysis process (Berthold, Borglet and Klawonn, 2010). Data understanding generally helps to understand the nature and type of attribute, the presence of outliers and extent of missing values in the data.

Choice of appropriate analysis methods depends on the objectives of the analysis, the design used and, type of measurements taken. The objectives of the data analysis are primarily determined by the objectives of the trial (Coe, Stern & Allen, 2002). The design of the study has great role in determining the type of analysis that is used. Study design can be experimental, quasi-experimental, or observational studies. The heart of data analysis, answering the major research question, is inextricably linked to the research design and the research design frames the type of analysis that you can and cannot do (Trochim, 2005). The choice of appropriate data analysis technique is a complex issue because real-life data often contain mixtures of different types of data. This makes the choice of analysis technique somewhat arbitrary. It is quite possible that two statisticians confronted with the same data set will select different methods of data analysis, depending upon what assumptions they take into account while interpreting the results of analysis.

Data analysis can be categorised into qualitative and quantitative data analysis based on the nature of data collected, the major assumptions made during analysis and, the type of information extracted from the analysis. Qualitative data analysis is an inductive process of developing a set of themes or a conceptual framework that captures the main ideas in the data while quantitative data analysis can be descriptive or confirmatory analysis, depending on the research questions (Barker, Pistrang and Elliot, 2002). Qualitative data analysis involves three separate sets of processes, which are identifying meaning, categorizing the information, and integrating. Qualitative data can be analyzed as the research progresses continually by refining and reorganizing in light of the emerging results but for quantitative data the analysis will be left until the end of the data collection

process, and if it is a large data set, statistical software is the easiest and most efficient method to analyze the data properly (Dawson 2009). The descriptive part of quantitative data analysis provides first estimates and summaries of information about the variability or uncertainty in the data and it indicates unexpected patterns and observations that need to be considered when doing formal analyses (Coe, Stern and Allen, 2002). Trochim (2005) also says descriptive analysis helps to describe what is going on in the data and inferential analysis helps to draw conclusion from the data to a case that is more general.

Researchers make a wide range of mistakes during data analysis. Kuzon *et al* (1997) described the most common mistakes during data analysis as “the seven deadly sins of statistical analysis” which includes using parametric analysis for ordinal data, inappropriate use of parametric analysis (overlooking the normality assumption and the size of sample), failure to consider type II statistical errors; using unmodified t test for multiple comparisons; reporting standard error instead of standard deviation; failure to rely on a statistician or rely too much on a statistician and underutilization of ANOVA, multivariate regression, logistic regression and non linear regression. Miller (2009) also described these mistakes as “The Seven deadly sins” which includes; over-reliance on null hypothesis testing, failure to perform prior to conducting the study, using asymptotic statistical computation with small sample, ignoring missing data, failing to consider the multiple problems when using multiple comparison analysis, using stepwise procedure to select variables in multiple regression, failing to perform or report model diagnostics. Therefore before data analysis researcher should consider whether the statistical assumptions for that specific analysis are met or not and whether the research questions are answered properly. Statistical assumptions are very important to get good reflection from the data, the study design and important issues in the data. For instance using mean and standard deviation for non-normally distributed data will give misleading information and instead the use of median or inter quartile range will give a better estimate. Repeated measures study design requires a repeated measures analysis while binary dependent

variable requires a categorical analysis which does not strictly need the normality assumption.

There are wide ranges of statistical software ranging from spreadsheet, which can do limited analysis through software made for specific analysis to all-purpose software, which can do all statistical analysis. It is however not possible to identify the best statistical software to analyze a given data set because all the packages have their own range of tools to support in the analytic process (Liwins & Silver, 2004). According to Altman & McDonald (1997) choosing regularly updated software that uses well-documented algorithms, checking for accuracy, substantive and statistical plausibility of the software and testing with complex or problematic models for accuracy are the most important points that should be considered when choosing statistical software. Researchers should avoid using software that have not been tested adequately for accuracy or that do not perform well.

The task of data analysis should be considered starting from the inception stage of the project and continue as focus of attention during data collection, data management and completed only after the review of the report. Both qualitative and quantitative data analysis should be undertaken based on properly designed data analysis plan which considers a variety of data analysis options. But researchers should understand which data analysis options best answer the research question or the objectives of the study and finally they should clearly describe the strength and weakness of the specific data analysis option used in the body of their report.

### **3. Project approach**

#### **3.1 Participation in research / consultancy**

Research participation or consultancy task was accomplished for different research projects including graduate and postgraduate student projects mainly focusing on data management and data analysis. Collaborative consultancy approach which mainly involved two interdependent stages was used to assist both researchers and research students. In this approach firstly, researchers explained their challenges and then the research methods student asked different questions to clarify the problems to be solved and got an in depth understanding about the nature of the problem. The questions mainly included the background of the study, the main question to be addressed from researcher's perspective, the design of the study, the measurement taken and any preliminary analysis done. Secondly, the trainee research methods student explained different methodological approaches and statistical techniques that could be applied to address the challenges of the researcher and then the researcher could ask questions to clarify about the different methodological approaches suggested. The student had tried to answer most questions presented by the researchers mainly from his previous research methods knowledge acquired from the course work and by reading how other researchers have approached similar challenges. Moreover for cases that required specialist input needs, more time and effort, relevant sources of references and information were suggested to the researchers.

After each of the consulting activity, a consultancy report was prepared and documented for further reference. The goal of this report is to summarize the relevant issues and lesson learned from each of the consultancy /research participation activity. The report contains the nature of the problem to be addressed, the design of the study, variables collected, and preliminary activities to solve the problem, suggestions and conclusion.

### **3.2 Training in the research process**

To accomplish the training task the following four different stages were used: identification of the training needs, development of training materials, delivering the training, and evaluation of the training. The training needs were identified by consulting the research team leader and research assistants. The needs included enumeration training for “Household livelihood, and socio-organizational impact of animal health intervention study” and ‘Introduction to different statistical packages’ for research technicians and other researchers from different projects.

Training materials such as power point presentations, reference materials and datasets which considered the level of the trainees’ statistical knowledge, were used during the training. Instructional manual was also developed by integrating information from FAO, Statistics South Africa and other research institutes’ enumeration training manuals.

Introduction to the statistical package SPSS’ training was given to research technicians from IWMI for four consecutive days. Each of the training sessions had a presentation and practical exercise in which the trainees were practicing the use of software using simulated and real research data sets from the project. The trial version of the software was used for the practical session.

At the end of the training, the entire training activity was evaluated by both the trainees and the trainer using questionnaire designed for that purpose and summary of these evaluations was documented for further reference.

### **3.3 Review of reports and proposal**

The written review of research reports and proposal was accomplished based on review guidelines developed before starting the review process (Appendix 3 and 4). These guidelines, which helped to standardize the review process, were developed based on common and applicable scientific manuscript peer review principles. The review process mainly focused on the general structure of the manuscript, the main contents to be included, and the way the contents appear in the text.

The main intention of the proposal review was not editing the contents but to examine the overall strengths and weaknesses of the proposed study and suggest ideas for improving the content and coherence of the article. Similarly, before reviewing the proposal, available literature relevant to the proposed study was consulted to understand the logic and scientific principles behind the proposed study. Then all the sections of the proposal were read without making any notations in order to understand the author's intent, style, and arguments. Subsequently each section of the proposal was critically assessed to provide appropriate recommendation whether the content provides robust scientific evidence to address the problem being assessed or not. Finally, constructive comments and list of possible suggestions were given to the principal investigator for feedback and possible arguments on some ideas which were not clear during the review. From the principal investigator's feedback, different lessons were learned. The review process and results of the review were documented for further reference.

Before starting the report review, other related articles published from the project and similar studies were assessed in order to understand the line of thought of the author. Then the appropriateness of the title, abstract, background, methods, results, discussion, conclusion and reference parts of the report were critically assessed based on the review guidelines developed. After the review process constructive comments and list of possible suggestion about each section of the report were given to the research team leader in order to get feedback and possible suggestions on the review comments and suggestions. The

review comments contained a brief summary of the contents of the article, list of major comment and possible options for improvement.

### **3.4 Data management**

The data management task mainly included three major components: Designing the data management plan, identification of the data management challenges and providing data management support to the team. The data management plan was developed by consulting the project team leader and other researchers (Appendix 6). Then data management challenges were identified during the different data management activities of the team and student projects and by assessing the quality of different databases the team research. Since most of the data sets of the project are stored in SPSS files, the majority of data management tasks were accomplished by using this statistical package.

Assessment of the team data management practice was undertaken using different data sets from previous studies together with their data collection instruments. Variables definition, data cleanness, data storage system, data description and Meta data issues were critically evaluated to identify possible data management challenges and take corrective measures on data management errors before analyzing the data. Definition of the most important variables was checked using the data collection instrument and SPSS data management guidelines to take appropriate corrective measures.

Data management support was given based on the plan and assessment results on data cleaning, data collection, and data organization, data, suggesting possible solutions for data management challenges related to missing values and outliers and design of data collection instruments. The data cleaning tasks for discrete or continuous variables were undertaken by using the descriptive statistics options of the statistical package. In additions, after identifying the extent of missing values and outliers in different data sets, possible data management options to minimize their effects and obtain unbiased estimates were suggested. For the different data management tasks data management log book was prepared to record each of the data management tasks and corrective measures taken. In

the log book variable definitions errors such as variable coding, missing value labeling, data entry errors and the correction made were recorded.

Qualitative data was collected from different organizations working on land and water management activities throughout Ethiopia. The data was collected together with IWMI research technicians comprising of three people. After the data collection activity, the information was entered in excel to organize the information.

Draft data collection instrument for “Household livelihood and socio-organizational impact of animal health intervention” study was developed using the following seven general steps: identifying the core objectives of the study, defining research questions under each core objective, listing the most important variables under each research question, anticipating how the data will be analyzed (according to the stated objectives and research question), developing tentative questions which can answer each of the research questions, ordering the questions properly and checking the inclusion of all the variables and, pre-testing the questionnaire and pilot study (both the pretest and pilot test have not been accomplished yet).

Finally to prepare some of the team data files for archival Meta data and data file description were prepared and included in the data sets. The Meta data includes information about the Title of the project, Authors, Data set overview, Brief description of data capture instrument, Data collection and processing and Data formats. The Meta data were prepared by using the questionnaires, project protocol and by consulting the project research team members. The data file description includes format of the data, number of records, number of variables and type and number of derived variables.

### 3.5 Data analysis

The data analysis task comprised three different activities: developing the data analysis plan, accomplishing data analysis task for different projects and preparing data analysis report. The data analysis plan was developed by consulting the research team leaders and researchers (Appendix 7). The task accomplished from different research projects data sets mainly from “Positive deviance in the use of improved livestock feed resource in developing countries”, “Ghibe baseline innovation study” and “Household survey for Ghibe valley.”

The main data analysis tasks included computation of variables, multicollinearity check, descriptive statistics and model building. The dependent and independent variables for each of the data analysis activity were identified by considering the objectives of the data analysis.

#### A. Computing the dependant and independent variables:

A composite indicator (response variable) which was referred to as index of innovation related activity (INNOVA) was created using a range of proxy indicators on forage related innovation activities and their values were normalized for aggregation. The proxy indicators take value 1 if the household is using the technology at the moment and 0 otherwise.

Different explanatory variables were also computed from proxy indicator variables and their values were also normalized for aggregation. If the values of all proxy variables take dummy value, one and zero, the mean response of the variables was taken as an indicator. But if there are any numeric variables among the binary indicators their values were normalized using the following formula to make the measurements comparable (Kiggundu, 2006).

$$\frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

Where  $X_i$  refers the value of the variable,  $x_{\min}$  is the minimum value and  $x_{\max}$  the maximum value.

## **B. Multicollinearity checks and descriptive statistics**

After selecting possible explanatory variables, multicollinearity check was done to eliminate variables with strong correlation and improve the general model fit. Therefore, variables which had strong correlation with one or more of the variables were excluded from the analysis. The descriptive statistics of most variables were summarized in consultation with the research technician which was useful in assessing the nature of the variables before including them in the model.

## **C. Model building**

### **Determining Factors which affect the intensity of forage innovation**

Data analysis to identify factors which affect the intensity of forage innovation activities was undertaken using three statistical models: OLS, Tobit and Heckman Two stages. These models were constructed using Stata 10 and R statistical packages. The results of the analysis were compared based on the statistical principles and nature of study to suggest the best model which could explain the data well. But only the result of Stata 10 was reported to the innovation team.

### **Identifying factors which affect the success in Trips management**

After identifying different possible data analysis options it was decided to undertake analysis to identify the factors which affect the success in Trypanomosis management based on its relation to previous study reports and other study objectives. The main purpose of the analysis was to explore the possibility of using secondary data to extract information other than the main study objectives. The response variable, observing improvement in tsetse management, was a binary response variable which takes the value 1 if the farming households are observing improvement in tsetse management and zero otherwise and regressed with different independent variables. This analysis was done using the following four different statistical software; Stata10, SPSS 16, GenStat 12, and R.1.1.11. The main purpose of using different statistical packages was to assess the difference in results obtained using different statistical packages for a given analysis.

### **Summarizing baseline data**

Baseline information for “Household livelihood and socio-organizational impact of animal health intervention study” was summarised from different data sets of “Household survey for Ghibe valley, Ethiopia” studies. Data sets stored in SPSS format were used for this analysis. These data sets had different challenges such as absence of label, errors in value coding which needs to be corrected before the analysis. After correcting some of the data sets using the questionnaire the data was summarised using the descriptive statistics option of SPSS. However, before starting the data summary process, the variables which are very important for the impact assessment were identified and collected together in one data files.

## **4.Results**

### **4.1 Research participation**

Using the designed approaches, research participation/ consultancy task was accomplished in different research projects for the team and student projects. The major purpose of this activity was to learn from the scientists and help them in solving the research methods challenges they encounter during the course of their research project. The main research methods challenge of researchers and students, lessons learned and support given for different research projects are presented in the following sections.

#### **4.1.1 Participation in the team projects**

Research methods support for “Impact assessment study of the Nigeria avian influenza control and human pandemic preparedness and response” and “Analysis of projects /networks working on land and water management in the Nile basin of Ethiopia” studies was provided in order to solve different challenges the scientist encountered and to contribute to the project research activities.

#### **Participation in impact of avian influenza study**

The “impact assessment study of Nigeria avian influenza control and human pandemic preparedness and response project” was undertaken to assess the response to the avian influenza outbreak in 2006 (ILRI, 2011). The impact assessment team was comprised of scientists from different technical backgrounds working on different issues. During the attachment period there was an opportunity to assist one of the team members working on the economic impact analysis component in solving data management challenges, descriptive analysis and qualitative data summary.

## **Challenges**

The scientist received data sets in Microsoft Access format to summaries the economic impact of the outbreak and report the findings. After running the summary statistics in Microsoft Access, the scientist found that the result of the analysis was not in appropriate format, as it was giving codes in the form of a1, b1, c1 instead of variable labels. Consequently, it was necessary to go back to the questionnaire to gain an understanding of the data. This was very challenging for questions which had multiple responses. To sort out this problem, discussions with the research technician were initiated and checking of data started to check the data entry format with reference to the questionnaire. The following problems were then detected in the data sets: the variables did not have description, the field names were not properly defined and some of the definitions were not also explanatory. After entering the description of the variables using the questionnaire the summary statistics was run. In addition, there were also some data entry challenges in some of the data sets. Then the scientist alerted the researchers responsible for the data management task regarding those data entry errors and another set of data files in SPSS format was received. However, since the scientist was not using the SPSS package, that proved to be another challenge.

## **Proposed solutions**

The following actions were taken to support the scientist to overcome this problem; producing the descriptive statistics of all the variables and submitting the summary result and transferring the whole data sets from SPSS data file to SAS data file, which the scientist was using at that time. Since the variables contain both categorical and numeric variables the descriptive statistics options of SPSS were used to get the frequency table and summary statistics. After transferring all the 12 data sets using Stat transfer both the summary statistics and the transferred data were submitted to the scientist. All the above activities were done together with the research technician in the innovation team.

***Table 1: Sample descriptive summary of different parameters before the outbreak, after the outbreak and now***

Observations	Minimum	Maximum	Mean	Std. Deviation
Number of birds processed per month_ Before	60.0	1800.0	645.000	783.5603
Number of birds processed per month_ During	.0	300.0	95.000	141.7745
Number of birds processed per month_ Now	30.0	3000.0	977.500	1362.5069
Price of processed bird_ Before	.0	1300.0	402.500	607.7486
Price of processed bird_ During	.0	400.0	125.000	189.2969
Price of processed bird_ Now	.0	1600.0	525.000	736.5460
Price of unprocessed bird_ Before	.0	1200.0	360.000	566.6274
Price of unprocessed bird_ During	.0	300.0	95.000	141.7745
Price of unprocessed bird Now	.0	1500.0	487.500	693.4635

After finishing this support the scientist shared a presentation about the model that was being constructed using simulation techniques. The presentation was informative about how and when to use simulation using real data sets and construct models from survey or other similar studies.

### **Participation in Analysis of project/Network study**

Scientists and research technicians from innovation in livestock systems team and IWMI, who are working on “Analysis of projects / networks working on land and water management in the Nile Basin of Ethiopia” study held different meetings on the data collection status of the study. The study was being carried out at national level with the main objective of exploring opportunities for facilitating an institutional structure for joint learning and exchange between different key players working on land and water management in Ethiopia. For the member of the data collection team, there was an opportunity to participate in those meetings and share experience and lessons.

The first meeting was on the general assessment of the data collection activities that we were doing together with the research technician from IWMI. During this meeting, after providing a brief description of the data collection status, two important issues related to the data collection instrument and process were raised for discussion. The first issue was the length of the questionnaire which was highlighted by some of the respondents. The questionnaire had seven main sections and 28 subsections. This issue was very important as some of key informants may be busy and not have adequate time to respond to all the questions. The second important issue was about the total sample size of the study as it was included only six national projects working on water and land management. It was suggested to include other additional samples based on specific criteria to get sufficient information from an adequate number and diverse set of projects or networks working on water and land management and setting criteria to select those sample projects or networks. Considering the scope and objectives of the study, the team decided to include additional projects and networks that have different focus and provided diversity in interventions in order to improve the quality and scope of data which was to be collected.

#### **4.1.2 Participation in students research projects**

Data management and data analysis support was provided to the following five graduate student projects: “Coffee cooperative performance on smallholder livelihoods”, “Genotypic variance of double haploid (DH) lines under well watered and managed drought environments”, “Assessment of BPR (business reengineering process) implementation in Ethiopian civil service college”, “The Role of Cooperative in Triggering Natural Resource Management: the Case of Tewodros Incense production Cooperative in Amharic Region of Ethiopia” and “Determinants of multiple dimension of child poverty in Ethiopia.” Research students approached the trainee graduate fellow with different data management and data analysis challenges and possible suggestions and support were provided based on the nature of their study. In the following section, summary of data management and data analysis challenges and possible suggestion given are presented.

#### **4.1.2.1. Consultancy on data management**

##### **Challenges**

Except one, the four students were using the SPSS statistical package for data management and data analysis. The main data management challenge of these students included; defining the variable in the variable view, design of questionnaire and; lack of knowledge and skill about data cleaning. The most common difficulties associated with variable definition were correctly setting the value column and the measure columns of the variable view. Even though these issues look very simple to most researchers or research methods professionals, they were not correctly defined in most students data file. For example, in some data files either there was no coding for categorical variables or the coding was incorrect.

Almost in all of the data files the measurement column was not properly defined. Missing value definition was also one of value setting challenge for students. Even though SPSS has good facility to manage different types of missing values, most of the students had no information about how to use this feature. Some of them were entering zero for missing values without specifying zero as missing value in the column.

The other data management challenge the students faced was related to inappropriate design of questionnaire. Since some of the questions in some research projects were not coded correctly, data analysis with such coding gave results which could not be interpreted. This challenge was very common especially with Likert scale type questions.

The concept of data cleaning was not clear for some students and others had limited practice of how to do it. Even though there are different techniques that researchers can use to check errors from their data sets, we used two general approaches specifically for categorical and continuous variables. Errors in categorical variables are mostly related to value records out of the pre defined value ranges, this task was accomplished by producing the frequency table report from the descriptive statistics sub menu of Analyze menu. The

analysis report would indicate the minimum and maximum value records for a given variable and if the value did not give sense or fell out of the expected range, that value was checked and corrected by referring to the appropriate questionnaire. For example in one situation, ethnicity was coded as 1= Amhara 2 = Oromo 3 = Gurage, and 4 =Tigray but the frequency report gave minimum value of 1 and maximum value 5. The value 5 was an error in this particular case because it was out of the range 1 and 4. Then this value was checked from the questionnaire and corrected accordingly. Errors in continuous variables are related to the expected minimum, maximum and mean values of the observation and any unusual value can be checked from the questionnaire and corrected accordingly.

### **Proposed solutions**

Different possible solutions and recommendations were provided based on the nature of their challenges. The most important solutions includes providing clear descriptions about the nature of data management problem and their possible solutions such as defining the variables and measurement columns, recoding different responses using the recode sub menu of Transform menu in SPSS and analyze them using the appropriate methods, increasing their awareness with appropriate explanation about the advantage and disadvantage of setting missing values in SPSS data base , providing, description of the data cleaning activities which mainly includes two major procedures; checking for errors and, finding and correcting the errors using their own data sets

#### **4.1.2.2 Consultancy on data analysis (Challenges and proposed solutions)**

Even though most students had defined research objectives and questions very well, there was no data analysis plan in at least one of the research projects. Data analysis support was provided to various research students and the summary of the three most important challenges and possible solutions suggested are presented below.

The first challenge was related to Likert scale data analysis, where respondents rate their preference or perception about a given question. Although this is one of the most common

data type collected in surveys, there was very little understanding about the possible ways of analysis. One of the students during our discussion indicated that since most of the students in his faculty are using t test or other parametric tests he wanted to use the same. However, using parametric test on non parametric data such as Likert scale data is not appropriate.

Since likert scale data does not fulfil the parametric assumptions such as normal distribution and constant variance the use non parametric test was suggested by explaining the possible bias created using different examples. For instance in order to test the presence statistically significant difference in perception about different forage technologies adoption between successful and unsuccessful innovative groups the following parametric and non parametric equivalent tests were used. But the result indicates different P values for the presence of significant difference in perception between successful and unsuccessful positive deviants.

***Table 2: Sample parametric data analysis for likert scale data***

<b>Group Statistics –Independent t test</b>						
	Status	N	Mean	Std. Devi	Std. Error Mean	Sig. (2-tailed)
Availability of fodder	1	527	2.3340	1.49092	.06495	
	2	76	2.7105	1.31469	.15081	.024
Availability of crop residue	1	527	3.0930	1.31274	.05718	
	2	76	2.7895	1.22546	.14057	.048
Availability of improved grazing land	1	527	2.9962	1.39117	.06060	
	2	76	2.8421	1.28637	.14756	.363
Availability of green forage purchase	1	527	3.3321	1.28081	.05579	.453
	2	76	3.2105	1.31976	.15139	

**Table 3: Sample Non Parametric equivalent test for the above parametric test**

	Ranks- Nonparametric test		Mann-Whitney test		
	Status	N	Mean Rank	Sum of Ranks	Sig. (2tailed)
Availability of fodder	1	527	293.30	154567.50	0.001
	2	76	362.35	27538.50	
Availability of crop residue	1	527	307.01	161796.50	0.053
	2	76	267.23	20309.50	
Availability of improved grazing and	1	527	303.94	160175.50	.429
	2	76	288.56	21930.50	
Availability of green forage purchase	1	527	304.07	160247.50	.415
	2	76	287.61	21858.50	

Where 1= freely accessible, 2=Do not know, 3=slightly expensive 4=Expensive and 5=Very expensive  
 Status, 1=Successful 2=Unsuccessful

Table 2 and 3 indicated that the impact of using parametric and non parametric test for likert scale data. The parametric test indicated the presence of statistically significant difference on availability of fodder and crop residue perception at 5% percent between successful and unsuccessful households while the non parametric equivalent indicated us the presence statistically significant difference fodder availability perception at 1% and at 10% for availability of crop residue. From this result any one can reach on two different conclusions especially if the researcher fixed its P value at 1%, 5% or 10 %.

The second challenge was related to in the use of statistical tools: the student wanted to get Biplot from the data collected in two replications and five locations of two different countries. However, since the statistical tools that she was using was not providing the Biplot, she requested support on this issue. After a good discussion about the objectives and nature of the study, the data was analyzed using GenStat 12 and the AMMI, Additive main effect and multiplicative interactions analysis and the Biplot report was submitted to the researcher. This data was also analyzed with the statistical package R and the result of the analysis was very comparable to GenStat.

The third challenge was related to designing research projects using secondary data. The student wanted to conduct his research based on secondary data on child poverty in Ethiopia. The research student has tried to assess the available literature related to his topic of interest and the possible data analysis options. But he was not sure how to choose good research topic based on the available data, the type of analysis that could be done and whether it was possible or not to use secondary data for his research project. After we discussed about the necessary of collecting primary data, different poverty assessment techniques was identified based on previous research and the available data. Then the most recent poverty analysis technique, multidimensional poverty analysis was preferred. The data was analyzed using SPSS statistical package. The results from the analysis were impressive and informative. Samples of the analysis result are given in table 4 and table 5.

**Table 4: Sample child Multidimensional poverty measure in Ethiopia**

Measures	K1=union	K=3	K=4	K5	K11=intersection
A	0.4123	0.459	0.496	0.531	0
H	0.990	0.836	0.698	0.532	0
Mo	0.408	0.383	0.346	0.282	0
M1	0.167	0.147	0.119	0.079	0
M2	0.253	0.214	0.169	0.108	0

Where K= Different cut of points

A = Average Poverty Gap A is the average number of deprivations a poor person suffers.

H =Headcount H= number of poor people/ the total population

Mo = Adjusted Headcount, is calculated as H times A.

M1 = the adjusted poverty gap M1 is given by HAG or the M0 measure multiplied by the average poverty gap.

M2 = the squared poverty gap (HAS)

**Table 5: Sample data analysis on the Incidence of child deprivation by region**

Dimensions	Deprivation percentage by Regions				
	ADDIS ABAB	AMHA	OROMI A	SNNP	TIGR AY
Health	15.2%	33.3	10.6	21.2	19.7
Care	27.6	21.6	17.2	15.5	18.1
Education	3.6	17.3	25.8	30.0	23.3
Freedom	2.5	14.4	26.7	35.4	21.0
Shelter and environment	11.6	21.6	20.9	24.7	21.1
Sanitation facilities	5.0	25.4	18.4	27.7	23.5
Safe drinking water	.3	17.6	21.7	34.0	26.4
Asset	14.5	20.3	20.1	24.3	20.8
Nutrition	3.6	17.0	13.5	46.2	19.7

Table 5 tells us that SNNP (southern nations and national people) state is the region where the most deprived children are found. On six dimensions the region was most deprived than other regions. Child deprivation in Addis Ababa is relatively better than the other regions.

#### **4.2 Training in the research process**

Generally two types of training needs were identified during the discussions with the research team leader and research technicians. These were enumeration training for “Household livelihood and socio-organizational impact of animal health intervention study” and “Introduction to different statistical packages” for research technicians and researchers from different projects.

## **Enumeration training**

Enumeration training has not been given yet as the starting date of the project is extended to September. However training manual and other necessary materials for the training have been developed to serve as a guideline during the training and as a reference during enumeration. The manual contains the following seven main sections:

- Context of the research and enumeration
- Survey personnel related to data collection
- Materials and survey instruments
- Enumerators task and conduct
- Enumeration process
- Handling the questionnaire
- Post interview procedures

## **Introduction to SPSS training**

An introduction to the statistical package SPSS was given to a small group of research technicians from IWMI. PowerPoint presentation and reference books were used to facilitate this 4-day training course. Each training session had two sub sessions; brief description of statistical principles and their application in SPSS and practical session in which the trainees were practicing different data management or analysis activities using either simulated or real data sets. The main topics covered during the training included; General Introduction to SPSS statistical package, data entry, data cleaning, descriptive statistics and inferential Statistics. The trainees had the opportunity to practice each of the above topics using their own personal computers. During the training great emphasis was given to the data management aspects, especially variable definition, data entry and data cleaning, because these issues were the most important challenges observed from different project data sets I got to work with.

At the end of the training the trainees evaluated the general training approach and outcome based on questionnaire developed for this purpose. The result the evaluations is given below

**Table 6: Training evaluation results by the trainees**

**Impact of the training**

How useful did you find the course		How demanding did you find the course?	
Rate	Frequency %	Rate	Frequency %
Very useful	100.0%	Very easy	.0%
Useful	.0%	Easy	.0%
Not sure	.0%	Medium	100.0%
Not useful	.0%	Difficult	.0%
		Very Difficult	.0%

**Level of understanding**

Theoretical session understanding		Data cleaning understanding	
Level	Frequency %	Level	Frequency %
Very good	.0%	Very good	.0%
Good	.0%	Good	.0%
Medium	100.0%	Medium	100.0%
Poor	.0%	Poor	.0%
Very poor	.0%	Very poor	.0%

Data entry understanding		Data analysis understanding	
Level	Frequency %	Level	Frequency %
Very good	.0%	Very good	.0%
Good	.0%	Good	.0%
Medium	100.0%	Medium	100.0%
Poor	.0%	Poor	.0%
Very poor	.0%	Very poor	.0%

### General rating and recommendations

How did you rate the overall standard of teaching?		How did you rate the quality of the course notes?		Would you recommend the course to other people?	
Rate	Frequency %	Rate	Frequency %	Rate	Frequency %
V. good	100.0%	V. good	100.0%	Yes	100.0%
Good	.0%	Good	.0%	Perhaps	.0%
Medium	.0%	Medium	.0%	No	.0%
Poor	.0%	Poor	.0%		
Fair	.0%	Fair	.0%		

The trainees rated the training as very important but it was neither easy nor difficult for them. But they indicated that their level of understanding on descriptive statistics, data entry, data cleaning and data analysis sessions is still medium. They also rated the material covered as medium in scope, high in practical content, considered statistical and computing knowledge of the trainees. In additions they rated the overall standard of teaching and quality of the training as very good. Finally, the trainees concluded that the training was very interesting and useful, with excellent teaching style and continuous practical work. However, the time allocated to cover the training content was not sufficient and this needs to be revisited.

According to the trainer self assessment, the learning outcomes of the training were fully achieved because after the training, the trainees are able to define variables, enter data, clean data and analyze data using the statistical package.

## 4.3 Review of research proposal and report

### 4.3.1 Review of proposal results and feedback

The “Household livelihood and socio-organizational impact of animal health intervention” study protocol was reviewed based on the guidelines developed using different scientific sources. This guideline had six main parts which include different criteria to review the major sections of the research proposal. Since some aspects of the proposal were reviewed

by other scientists, my review focused on the aspects which were not observed by the reviewer. After completing the review activity general comments and suggestions were given to the principal investigator to indicate opportunities for improvements. Comments were given under each of the four main sections: title, introduction, methods, and reference and general comments about the organization of proposal.

### **Brief summary of comments and suggestion**

- The term “*household livelihood*” in the title is broad and complex which can be measured by different indicators, so suggestion was given to substitute this term with the term *food security* to simplify it and make it measurable.
- The term “*socio-organizational*” in the title indicates the impact of the intervention on the social and organizational aspects of the households. But this issue was not clearly spelt out in the objective and methodology part of the study. And how this will be addressed?
- The rationale of the proposed study should be clearly indicated including more explanation why it is worth doing. Even though it is briefly mentioned at the end of the 4th paragraph, suggestion was given to add more clarification on these issues for readers who are not familiar with the background and the intervention.
- Although there are different known factors which affect the quality of the research to meet the desired objectives, the limitation and scope of the study was not clearly described. So it is better to mention the limitation and scope of the study by considering other confounding factors which can affect food security status of households other than the project intervention.
- The review of previous research on the subject is a critical component of a good research proposal, but this is not reflected in this proposal. Was no similar work carried out on the subject previously?

- Objective three can be modified to make it more informative, possibly to read, to identify the factors that explain the effectiveness of community based Trypanosomosis & Tsetse control (CBTTC) intervention designed to enhance household livelihoods at Ghibe.
- Attributing land use change only to the project intervention may not be acceptable as there are multiple socio-economic, political and environmental factors which affect land use change. So it is better to consider this issue and modify the objective.
- The term ‘other intervention’ in the research question one is too broad and general while a research question should be very specific that could be answered specifically with the study. Therefore it is better to mention the specific intervention.
- In the methods section the source of imaginary data for land use change analysis was not described. So it will be more informative if it can be described briefly in this section.
- The data analysis part only talks about the inferential and qualitative data analysis and it does not say anything about descriptive analysis which helps to describe different variables. So it is advisable including the descriptive part in the data analysis section.
- The statistical package that will be used for data analysis was not mentioned and it is better to mention it.
- There are some references cited in the text but not in the reference section and vice versa. This section needs revision.
- Since the proposal did not include executive summary and data management plan it is better include this section in the protocol.

The principal investigator responded to these comments and suggestions and expressed that they were very useful to enrich the proposal, while described the logic behind the choice of title and the objectives of the study.

#### **4.3.2 Review of report results and Feedback**

The research report “The value of innovation system analysis to livestock research and policy” was reviewed based on the guidelines developed for this purpose. Generally the following comments were given on the contents of the manuscript.

- The title was catchy but it does not fully reflect the content of the study if looked at in isolation from the complete report.
- The abstract briefly summarizes the content of the study and is well written
- The introduction part was written very well. However specific research hypothesis was not formulated or stated in this paper because the study was conducted mainly to explore the experience and lessons in relation to innovation in the area and to understand the circumstances surrounding the positive change processes. The approaches used to understand the process of change and the lessons learned are briefly stated.
- The methodology section clearly described how the study was conducted and the specific procedures or approaches used. However sampling method for key informant interviews was not clearly described and the sample size for key informant interviews and PRA was not clearly indicated. In addition, even though the definition of variable (INNOVA) was provided in part three of the there is no clear description about the type of proxy indicators used to estimate INNOVA and there is no clear description why the researcher used Tobit model.
- The results section of the report includes both the descriptive and inferential statistics and appears to be comprehensive and complete. However the

organization of this section can be improved to make it more readable and keep the reader captive.

- The discussion part of the study was not clearly and separately indicated in the manuscript and part of it was included under the results section. Even though the interpretation for the regression output was presented, the link to the results and the objectives of the study was not clear in the discussion part. The interpretation of the result did not consider the sources of potential bias and other limitations due to precision of measurement.
- The conclusion part of the report was presented under the fourth section named “What can this kind of diagnosis are used for and how?” and it provides a summary of the main findings and their implication for use. However the limitation of the study and potential further research in the same or similar areas was not clearly indicated in this section.
- The reference system followed APA style which includes both the author and date of publication. However, all the references cited in the text, (such as ILRI, 2007; Mytelka, 2000 and Kiggundu, 2010) were not listed in the reference section and; there are also references listed (Mulatu. W 2005; ILRI, 2006 and World Bank 2006b) which were not mentioned anywhere in the report.
- General comments on the use of Tobit model and improving the organization of the report were provided in detail.

The above comments on the different parts of the report were given to the research team leader, as the author has left the team, for feedback and suggestion. The team leader gave general feedback that the comments and suggestion given were very valuable and very thoroughly done.

## **4.4 Data management**

This task focused on four studies; “Household survey for Ghibe valley, Ethiopia”, “Positive Deviance in the use of improved livestock Feed Resources in Developing countries”, “Household livelihood and socio-organizational impact of animal health intervention study”(data collection instrument development),“Analysis of projects/networks working on land and water management in the Nile basin of Ethiopia.”

### **4.4.1 Data management challenges**

The current data management practices of the team were assessed to identify challenges, to learn from experiences and to prepare the data for further analysis. Generally there is good data management practice in the Innovation in Livestock Systems Research projects. Most of the data sets are stored in SPSS formats and some of them contain project protocol, summary of analysis and analysis reports and variables description. However none of the data files have Meta data and data file description.

As it is described above, even though the general data management practice of the team is very good, there are still some challenges in the data files of different studies. The most common data management challenges identified included value labelling, specifying the measurement scale, inconsistent variable naming, absence of label and, challenges related to data collection instruments. Some of these challenges were also observed in the data files of student projects during the consultancy tasks.

#### **A. Challenges in value labeling of categorical variables**

Value labelling was one of the major challenges observed from different data sets of the project which mainly included absence of missing value codes, using missing value codes without specifying them in the missing value column, using incorrect codes for categorical variables and mixing the codes of successive variables. For example there were two missing value codes which were used in one of the variables such as “NA” and “Not

applicable” where “NA” may refer to ‘not answered’. But in the missing value column only one of the codes was specified as a discreet missing value and SPSS considers the other missing value code as one of the response values. Some categorical response questions had also repeated response codes for the same questions such as 0=” No” and 2=”No”. This creates a big challenge during data analysis and data entry.

The other error related to value labelling was using zero in missed or not applicable responses. Sometimes when the question was not applicable, zero was entered as the value of the response for that case. For instance in the question aimed at assessing the resource level of households between current and 10 years ago, the respondents were asked to state the amount of land they used for different crops. For this question zero was entered as a response value for all households who responded that they were not using their land for a given crop or did not answer the question and there were no missing value codes in the missing columns. In such type of situation zero was not the correct value of the response because the question was either not applicable to the household or not answered due to different reasons. SPSS considers zero as a real value and analyses accordingly. The following simple descriptive analysis indicates the impact this error in the final output of data analysis.

**Table 7: Descriptive statistics of land use without missing value codes** (when value zero was used in place of missing value codes)

<b>Descriptive Statistics</b>					
	N	Minimum	Maximum	Mean	Std. Deviation
Maize (ha) presently	603	.000	3.000	.20894	.325605
Wheat (ha) presently	603	.000	3.500	.34904	.500594
Sorghum (ha) presently	603	.000	4.000	.14749	.354542
Maize (ha) 10 yrs ago	602	.000	3.000	.19268	.339755
Wheat (ha) 10 yrs ago	602	.000	3.250	.29273	.450443
Sorghum (ha) 10 yrs ago	602	.000	2.000	.12757	.321203

**Table 8: Descriptive statistics of land use with missing value codes** (when missing value codes was used instead of using zero)

<b>Descriptive Statistics</b>					
	N	Minimum	Maximu	Mean	Std.
Maize (ha) presently	350	.001	3.000	.35997	.358259
Wheat (ha) presently	340	.025	3.500	.61903	.526672
Sorghum (ha) presently	158	.001	4.000	.56289	.496656
Maize (ha) 10 yrs ago	303	.001	3.000	.38282	.395842
Wheat (ha) 10 yrs ago	303	.025	3.250	.58159	.485001
Sorghum (ha) 10 yrs ago	137	.010	2.000	.56057	.459803

Tables 7 and 8 demonstrate the change in mean land allocation for different crops when using the value zero for missing information.

The other value labelling problem was using incorrect codes for different variables. For instance for the question “How important are the following as objectives of producing livestock?” Such as “For oxen draught” had five possible Likert scale responses in the questionnaire, but in the variable view definition of SPSS it had eight nominal response codes which were not related to the question. Mixing the value code of successive variables was also another problem observed in some variables. For the question “importance of oxen for drought” the possible response codes given in the questionnaire was a Likert scale data ranging from ‘Not important’ to ‘Very important’. However in the data entry format three different responses of the previous question which said “for home used milk, for breeding and for market” were included as a value code for this variable.

## **B. Challenges in specifying the measurement scale**

The second most important challenge was errors in specifying the measurement scale. The measurement columns of most variables were not properly set or left as the default value. For instance some of the nominal scale variables were set as scale or ordinal variables and some of the ordinal variables were set as nominal or scale variables. This error was very common in both the research team and student projects.

## **C. Inconsistent variable naming and absence of label**

The third most important challenge identified was inconsistent variable naming across the data base. For example in one of the data sets C7.2 and C7\_2 was given as variable names for two different variables in the same data set. For some variables, the question numbers were given as the variable name and there were no labels to describe the type of information. This posed a big challenge when the baseline data were being summarized, as it was difficult to get part of the data collection instrument for some of the studies, it was not possible to identify the variables for appropriate analysis.

### **4.4.2 Data Management support**

The main data management support which was given to the research team included correcting the value labeling and measure types of the variable definition, data cleaning, suggesting missing data and outliers handling techniques, data collection instrument development, summarizing qualitative data and, data organization for archival. In the following sections the steps used to accomplish each of the above tasks and the results obtained are presented.

#### **A.Data cleaning and handling missing values and outliers**

This task was done for positive deviance study data sets, using the descriptive sub menu of the Analyze menu in the SPSS window. This procedure helped to identify errors and

correct them by cross checking with the original questionnaires. The task was accomplished in two ways specifically for categorical and continuous variables.

Errors in categorical data were checked using the menu Analyze - descriptive statistics - frequency options and by selecting the minimum and maximum values of the statistics options. Then the maximum or minimum values were observed and compared with the true values in the questionnaire. Using this procedure it was found that some errors were created during coding of some categorical variables.

Possible errors in continuous variables were checked by using the menu Analyze - Descriptive statistics – Explore options and by setting the statistics options to get the mean, standard deviation, minimum and the maximum value and outliers. The descriptive statistics of some variables indicated the presence of some outliers and unlikely values in some of the variables. For example, the observation to the variable Household education for case 432122 had value 17, which is unlikely among the farming households of rural Ethiopia but after checking the value from the questionnaire it was found that the value is 14. Similarly the value of experience in cropping for case 432109 was recorded as 1981 while the true value is 27. Using this procedure data cleaning was performed on most Positive Deviance study data set variables.

### **B. Handling missing value**

Missing value was one of the challenges in some of the data sets in different studies. Missing values reduce the sample size and result in loss of data and thwart the purpose of the research. SPSS and other statistical software packages have four possible missing data handling options during data analysis. These are either leaving the missing value as it is, using list wise deletion, using pair wise deletion or mean imputation. For example as indicated in table 3, with 603 cases of the positive deviance study the percentage of missing data when simple descriptive analysis was done with list-wise deletion was almost 20% while the same analysis was done using pair wise deletion was 2%.

**Table 9: Sample descriptive statistics which indicates the effect of missing data handling options**

Variables	Type	Sample size	Mean	St.dev	CI lower	CI upper	Difference in CI
ALTCAT	List wise	80.8%	1.653	0.054	1.547	1.760	0.213
HHEDUCA	List wise	80.8%	4.41	1.168	4.08	4.74	0.66
ALTCAT	Pair wise	100%	1.558	0.049	1.453	1.649	0.196
HHEDUCA	Pair wise	98.2%	4.23	0.149	3.93	4.52	0.59

Table 9 depicts that the mean value and standard error of the variables in List wise deletion is greater than that of pair wise deletion and the confidence interval is wider in list wise deletion than pair wise deletion. List wise deletion decreases the number of sample size drastically than pair wise deletion

**Table 10: The effect of different missing data handling option during regression example**

Variables	Parameters	Type of deletion		
		List wise	Pair wise	Mean imputation
DISTANCEMKT	% included	50	100	100
	Coefficient	0.001	-0.002	-0.001
CREDIT	% included	50	99	100
	Coefficient	-0.010	-0.008	-0.005
AVAILFREXSERVI	% included	50	100	1000
	Coefficient	.118	0.107	0.106
ALTCAT	% included	50	100	1000
	Coefficient	0.000084	0.00007	0.000067
TLU	% included	50	65	100
	Coefficient	0.18	0.022	0.025
TYPEHOUSE	% included	50	99	100
	Coefficient	-0.23		-0.015

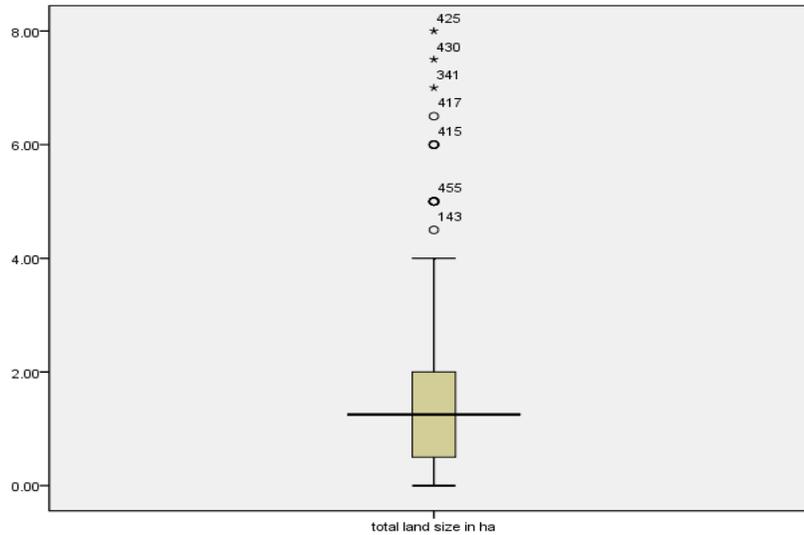
Table 10 demonstrates the effect of missing data handling option in the parameter estimates when intensity of forage innovation was regressed with different explanatory

variables. List wise deletion decreases the total samples size by almost 50%, while pair wise deletion keeps the total sample size almost close to 100%. The effect of missing data handling on the estimated coefficient was also serious. For instance the estimated coefficient sign of the variable “distance from market centre” was changed from positive to negative when we were changing from list wise deletion to pair wise deletion. Almost for all variables the coefficient estimate in pair wise deletion and mean imputation are different from that of the list wise deletion. Based on the above observation pair wise deletion was suggested during data analysis of the positive deviance study.

### **C. Dealing with outliers**

The first step in dealing with outliers was checking their value from the data collection instrument to confirm whether the value is a true observation or data entry error. Accordingly, the variables land size, total training, intensity of collaboration, altitude, and distance to nearest district, age, and total livestock unit had outliers. There are different options to deal with the outliers such as deleting the value, deleting the variable, transforming the value or transforming the variable. However before taking any measure on the outlier it is recommended to assess the impact of the outliers by including and excluding them in the model and observe the effect of the outlier on the parameter estimate or checking the trimmed mean, which removes the 5% extreme values on both sides. For instance the variable “total land” had 7 outliers which affected the estimate of mean value and standard deviation and eventually affects the entire results of any analysis. The score of the outliers were checked and found to be 7, 7.5 6, 6.5 and 8. The estimated mean values with the outlier was 1.55 hectare while the 5% trimmed mean was 1.43 hectare which gives a big difference of 0.11 hectare .

**Figure: Sample outliers for total land size data**



In additions to the descriptive statistics option, estimate of Heckman two stage models was used to assess the impact of outliers in different variables coefficients estimate by including and excluding the outliers in the data. In this model the dependent variable was intensity of forage innovation and different continuous and categorical parameters were used as independent variables. Sample of the analysis result which indicates the effect of outliers on the parameter estimates of different variables is presented in Table 10.

**Table 11: Part of Heckman two stage model results with and without outliers**

Variables	Analysis with outlier		Analysis without outlier	
	Coefficient estimate	Significance	Coefficient estimate	Significance
LANDSIZE	0.0076	0.365	0.004	0.340
TLU	0.0025	0.173	0.0092	0.003
TOTTRAIN	0.0832	0.014	0.062	0.082
INTENSCOLLABO	0.175	0.031	0.147	0.072
FSIZE	0.0032	0.472	0.0027	0.570

The outliers in total livestock unit changed the coefficient estimate in the outcome model from highly significant to highly insignificant, the outliers in intensity of collaboration and total training changed the coefficient estimate from significant at 10% to significant at 5%, the outliers in family size changed the coefficient estimate slightly and the outliers in land size changed the coefficient estimate by large amount (Table 11). Hence the outliers were handled based on the above assessment results.

#### **D. Qualitative data collection and summarizing**

Data for analysis of project/networks working on land and water management in the Nile basin of Ethiopia was collected from different projects or networks working on land and water management. Since the study was qualitative in nature the interview was led by an experienced research technician and the response of the key informants were recorded under each question. Some of the challenges faced during data collection include:

- Inappropriate responses - The interviewer asked information about major changes in the project but the key informant was answering about change in government structure which actually could affect the project activities. Even though the response was related to the question it was not major information needed.
- Inconsistency in providing information - The respondent was asked to list the main types of land and water management interventions and they gave list of interventions but when they were asked to highlight the most adopted intervention from the list some respondents were indicating new ideas which was not mentioned earlier.
- Some respondents found it difficult to respond to some questions as they could not understand them. For example since the question “Did certain organizational habits and practices restrict interaction, knowledge sharing and learning” was not easy to answer; a list of indicators had to be provided to the respondents to get the responses.
- One of the respondents pointed out that the questionnaire was long and result in tiring the respondent and lead to give incorrect or low quality responses. For instance, looking at the questionnaire one of the respondents said he had no time to go through the interview and answer all the questions.

- Some of the respondents were not able to provide information on some questions as they had not started implementing interventions. But the questions were designed assuming that most of the interventions in the projects have been already implemented. Therefore there was a need to identify the correct respondent and project or network that could provide the required information before the data collection activity. This highlights the importance of consulting the respondents in case of qualitative studies before selecting the sample.

This qualitative data collected from different projects was summarized in excel format in seven different sections based on the questionnaire for further analysis and data manipulation. The main purpose of summarizing the data was to facilitate the data analysis process and understand the nature of the data better. Before starting to summarize the data, the best possible options to do it were discussed with the scientist. It was organized in such a way that the response of all projects/ networks for each question were put together separately to observe the pattern and type of responses.

The other data management support was summarising qualitative data collected for “impact assessment study of Nigeria avian influenza control and human pandemic preparedness and response project.” Data files together with sample qualitative data summary were received from the scientist. After reviewing the nature of the data sets and variables, a thorough discussion was held with the research technician on how to summarise the data files and the procedures which can be followed to produce a good summary. The following five steps were followed to summarise the data: knowing the data, understanding the objectives of the analysis, categorizing the information, identifying patterns and connection between and within categories and, bringing all the major dimensions together. The final summary was done for three different groups separately; feed millers, breeders and processors based on the nature of the variables and sampling techniques used.

## **E.Design of data collection instruments**

It is not unusual to see researchers struggling when they start designing data collection instruments. Even though several procedures have been proposed for designing questionnaire, there are no uniform and agreed principles which help to design good survey data collection instruments. This resulted most researchers to include unnecessary and poorly formulated questions which do not contribute to meeting the objectives of the study. The problems of a poorly designed data collection instrument do not end with data collection and entry but it affects the analysis as well. Based on a review of relevant literature, the following 8 steps were defined and used to draft the data collection instrument for “Household livelihood and socioeconomic impact of animal health intervention study.”

### **Step1. Identifying the core objectives of the study**

Before writing the questions, the core objectives of the study were taken from the study protocol developed by the principal investigator. Since a questionnaire that is written without a clear understanding of the objectives of the study is inevitably going to overlook the important issues and waste participants' time by asking unnecessary questions, great care was taken to consider all the objectives of the study. Therefore the *three core objectives* were considered from the study proposal.

### **Step 2 Define research questions under each core objectives**

After identifying the core objectives of the study, three *research questions* were also taken from the study protocol. For each of the research questions, relevant questions were developed to cover the important issues. Each of the core objectives had at least one research question which would help to effectively achieve the stated objective of the study.

### **Step 3: Make a list of variables under each research questions**

For each of the research questions, a potential list of variables that would answer the question was identified. These variables formed the basis for formulating the questions. All the variables included fall the following four major categories: Background variables, Dependent variable(s), Independent variables and others: intervening, antecedent, spurious etc. For example, under research question two the following variables were listed:

#### **Research question two - sample variables list:**

- Total area cultivated
- Conversion from cultivated land to permanent grassland
- Cultivation of permanent grassland
- Forest planting
- Forest removal
- Digging ponds
- Hedgerow planting
- Hedgerow removal etc.....

### **Step 4: Anticipate how the data will be analyzed according to the stated objectives and research question**

This is an important step which most researchers do not consider during questionnaire design. After identifying the variables the possible ways in which each of the variables is going to be analyzed was identified by considering the research question and objectives of the study. This helped to consider the nature of the question and expected responses during the data collection activity.

**Step 5: Develop tentative questions which can answer each of the research questions and in turn the objectives of the study.**

After anticipating the ways how the data is going to be analyzed, tentative questions were developed under 4 main sections using the variables under each of the research questions.

**Step 6: Order the questions properly and check the inclusion of all the variables**

After the tentative questions were developed they were ordered in such a way that answering one question does not influence or bias response to the following questions, with those belonging to the same category of the household characteristics put under the same section and subsection of the questionnaire. After ordering the questions, a check to ensure inclusion of all the variables was made and now the questionnaire is ready for comments and suggestions by the research team.

**7: Pre-test the questionnaire and Pilot study**

The pre-test helps to check whether the questions and responses are valid, reliable, appropriate, necessary and sufficient. The questionnaire will be administered to a small number of people who resemble or are drawn from the population of interest without selecting them randomly.

**E. Preparing data for storage**

The data files from different studies in the team are stored properly together with the most common and necessary information. But there are also other information which should have been included in the data files such as Meta data, data file description, data collection instrument. Meta data and data file description were included in the data sets to ease the data storage and data access processes. The data sets were organized with appropriate format for archival and one copy of the original datasheets, which was made and filed with an appropriate organization or entity information, was submitted to the research team.

## 4.5 Data analysis

Three data analysis tasks were accomplished during the attachment. The result of the analysis and the specific procedures used are presented below.

### 4.5.1 Determining the factors which affect the intensity of forage innovation

Data set from 603 positive deviant households in forage innovation project collected from three regions of Ethiopia was considered. From 603 cases 76 of them have zero intensity of forage innovation and 527 have intensity of forage innovation between 0 and 1. A model which explains the probability of the existence of intense innovation activities was constructed using intensity of forage innovation as dependent variable and different demographic and socio economic factors as explanatory variables. The main interest here was to estimate the level of innovation activities by their intensity and to examine whether there was any variability among positive deviant farming households or not. However some of the farming households who were unsuccessful in forage innovation have zero forage innovation intensity (Table 12), which indicates the presence of strong relation between the intensity of households' forage innovation and their success which was defined as the current use of any forage technologies for animal feed. Therefore it was important to consider the probability of household success before determining their intensity in forage innovation. In other words farming households who are also not successful in forage innovation are not expected to have intense innovative activities as they are not currently using the technologies due to various reasons.

***Table 12: Descriptive statistics of Intensity of Forage innovation for Successful and Unsuccessful cases***

Status	Mean	Std. Deviation	Minimum	Maximum
Unsuccessful	.0000	.00000	.00	.00
Successful	.4614	.22362	.17	1.00
Total	.4032	.25918	.00	1.00

Before running the model, multicollinearity check was done and variables which had strong correlation were removed from the model. For instance, variables (years of experience in crop and livestock farming and age of household head), (distance to nearest district center and distance to nearest market center) had strong correlation. Therefore, years of experience and distance to nearest market center were excluded from the analysis. This was done by considering the correlation level of the variables with other variables and their theoretical importance. Three different models which can be used to determine the factors which affect intensity of innovation were constructed using Stata 10. The main purpose of running the three types of models was to assess the impact of selecting models without considering the theoretical and practical realities of the data and to indicate the effect of sample selection bias in survey where the respondents are selected with specific criteria or self select themselves.

As it is indicated in Table 12, the OLS and Tobit model gave comparable results because the variables have almost the same coefficient estimate and significance level.

**Table 13: Heckman first stage model for determinants of success in forage innovation**

Variables	Coef	Std. Err	z	P> z	Dy/dx
X <sub>1</sub>	.4897522	1.083134	0.45	0.651	2.7%
X <sub>2</sub>	1.028668	.5452139	1.89	0.059	5.7%*
X <sub>4</sub>	-.310838	.5770197	-0.54	0.590	-1.75%
X <sub>6</sub>	.1297083	.0521236	2.49	0.013	0.73%**
X <sub>8</sub>	-.171518	.2631656	-0.65	0.515	-0.88%
X <sub>9</sub>	.0886317	.2779905	0.32	0.750	0.53%
X <sub>13</sub>	.1268114	.0480504	2.64	0.008	0.7%**
X <sub>15</sub>	.5533921	.2883892	1.92	0.055	3.41%*
X <sub>16</sub>	.023014	.0352179	0.65	0.513	0.12%
X <sub>18</sub>	-.271823	.6415706	-0.42	0.672	-1.53%
X <sub>19</sub>	.8146351	.3722572	2.19	0.029	9.11%
X <sub>20</sub>	.2069918	.2809682	0.74	0.461	1.16%
X <sub>21</sub>	.7019241	.2468874	2.84	0.004	4.80%**
X <sub>23</sub>	-.005087	.3702884	-0.01	0.989	-0.02%
CONS	-1.46148	.6321981	-2.31	0.021	
lambda	-.131290	.0643616	-2.04	0.041	
Rho	-0.73478				
sigma	.1786794				

\*significant at 10%, \*\* significant at 5%

The Lambda coefficient from the selection model has negative value and it is significantly different from zero at 5%, which indicates the presence of selectivity bias in the sample. This suggests that the error terms of the selection model, success in forage innovation and

intensity of forage innovation model (outcome) are negatively correlated. This means unobserved factors that make the success of forage innovation are more likely tend to be associated with lower intensity in forage innovation. Therefore, using OLS method to determine factors which affect the intensity of forage innovation will give us biased result and hence the use of Heckman two stage models is appropriate in this case.

Different variables significantly affect the success of forage innovation. For example total training of the household head affects positively the probability of success in forage innovation at a significant level of 10%. A unit increases in training increases the probability of success in forage innovation by 5.7%. Living in the same part of the country will also have positive effects on the probability of success in forage innovations. Households who are always living in that part of the country will have 9.11% more probability of success than those migrant households. As expected total livestock unit will positively and significantly affect the success of forage innovation. A unit increase in total livestock unit will increase the probability of success in forage innovation by 0.7%.

**Key to variables names (Appendix 9: definitions of each variable)**

X <sub>1</sub>	Intensity of Collaboration	X <sub>9</sub>	Sex of household head	X <sub>16</sub>	Education level of Household Head
X <sub>2</sub>	Total Training	X <sub>10</sub>	Intensity of forage use Diversity	X <sub>17</sub>	Altitude category
X <sub>3</sub>	Age of household head	X <sub>11_1</sub>	Open grazing	X <sub>18</sub>	Age of household head
X <sub>4</sub>	Distance from nearest district center	X <sub>11_2</sub>	Mixed system	X <sub>19</sub>	Migration
X <sub>5</sub>	Farm enterprise index	X <sub>12</sub>	Dairy cooperative membership	X <sub>20</sub>	Intensity of Access to forage technology
X <sub>6</sub>	Family size	X <sub>13</sub>	Total Livestock unit	X <sub>21</sub>	Consultancy about forage
X <sub>7</sub>	Total land owned privately	X <sub>14</sub>	Innovation index	X <sub>22</sub>	Forage technology with traders connection
X <sub>8</sub>	House of the respondent	X <sub>15</sub>	Forage technology with follow up	X <sub>23</sub>	Intensity of Adverse Effect

**Table: 14 Heckman two stage outcome equations, OLS and Tobit model estimates of the factors which affect intensity of forage innovation**

Variables	Heckman/ outcome		Heckman dy/dx Coeff	Tobit		OLS	
	Coef	P> z		Coeff	P> z	Coeff	P> z
X <sub>1</sub>	0.146	0.057	0.154	0.215	0.006	0.201	0.004
X <sub>2</sub>	0.070	0.041	0.086	0.127	0.000	0.117	0.000
X <sub>3</sub>	-0.078	0.215	-0.082	-0.005	0.936	-0.011	0.852
X <sub>4</sub>	0.014	0.775	0.009	-0.043	0.388	-0.028	0.530
X <sub>5</sub>	0.125	0.000	0.125	0.157	0.000	0.145	0.000
X <sub>6</sub>	0.004	0.441	0.006	0.009	0.045	0.008	0.061
X <sub>7</sub>	0.007	0.412	0.007	0.004	0.710	0.003	0.763
X <sub>8</sub> <sup>*</sup>	-0.009	0.711	-0.011	-0.051	0.031	-0.040	0.062
X <sub>9</sub> <sup>*</sup>	-0.002	0.929	-0.001	-0.007	0.799	-0.005	0.841
X <sub>10</sub>	0.045	0.124	0.045	0.238	0.000	0.202	0.000
X <sub>11,1</sub> <sup>*</sup>	-0.079	0.054	-0.079	-0.059	0.168	-0.060	0.120
X <sub>11,2</sub> <sup>*</sup>	-0.030	0.308	-0.030	-0.027	0.417	-0.030	0.321
X <sub>12</sub> <sup>*</sup>	0.060	0.007	0.060	0.106	0.000	0.100	0.000
X <sub>13</sub>	0.010	0.001	0.012	0.012	0.000	0.012	0.000
X <sub>14</sub>	0.137	0.014	0.137	0.134	0.027	0.121	0.028
X <sub>15</sub> <sup>*</sup>	0.107	0.001	0.107	0.142	0.000	0.139	0.000
X <sub>16</sub>	-0.0009	0.775	0.000	0.003	0.243	0.002	0.398
<b>Lambda</b>	-0.13	0.041					

\*Categorical variables

A unit increase in total training will increase the intensity of forage innovation by 8.6% and a unit increase in intensity of collaboration will increase the intensity of forage innovation by 15.4%. A unit increase in farm enterprise index will increase the intensity of forage innovation by 12.5%. This means farming households who are participating in different farming enterprises are more innovative than others. Livestock management system, open grazing system negatively and significantly affect the intensity of innovation than the reference cut and carry system at 10%. Which means households who are using

open grazing system 7.9% less intensive than others using cut and carry system. Even though it is not statistically significant mixed systems users are less intensive than cut and carry system. Dairy cooperative memberships, which was coded as 0 = No and 1 Yes, will positively affect the intensity of forage innovation. Farming households who are member of the dairy cooperative are 6.0% more intensive than non members. Introduction of forage innovation with traders' connection has positive and statistically significant impact than the other way round. Farming households who gets forage technologies with follow up and training have 10.7% more intensity of forage innovation than others. Total livestock unit and intensity of farmers innovation related activities have also positive and significant effect on the intensity of forage technology innovation.

#### **4.5.2 Possible data analysis from previous studies**

Sometimes it is possible to extract information from previous studies to answer a research question not originally intended during the project design stage. This type of analysis is called secondary analysis. In the following section, the secondary analysis which was done on the Ghibe Innovation Baseline study data set is presented. The main objective of the analysis was to determine the factors which affect the success in observing improvement in Trypanomosis management.

The success in Trypanomosis management around Ghibe area has been recorded by different researchers working in the area. However, the data sets of innovation baseline study reveal that almost 38% of the farming households responded that they are not observing improvement in Trypanomosis management. This can be attributed to different socioeconomic and demographic factors. Therefore, it is very important to identify those factors which affect the success in the management of the disease in order to provide appropriate support to the farmers and design appropriate interventions. Since the response variable observing improvement in disease management, which takes 1 if the household observed improvement and zero otherwise, is a binary variable, logit model was used for

this analysis. The model was constructed with four different statistical packages Stata 10, R .1.11, SPSS 16 and GenStat 12. The result of the analysis is showed in table 14

**Table 15: Different statistical packages estimate for factors which affect the success in Trypanomosis management**

	SPSS			Genstat		Stata		R	
	B	Sig.	Exp(B)	B	Sig.	B	Sig.	B	Sig.
Z <sub>1</sub>	-0.1630	0.7270	0.8500	-0.1610	0.7310	-0.1631	0.7270	-0.1631	0.7273
Z <sub>2</sub>	0.0060	0.6530	1.0060	0.0057	0.6500	0.0057	0.6530	0.0057	0.6526
Z <sub>4_1</sub>	0.9570	0.0040	2.6040	0.9520	0.0040	0.9569	0.0040	0.9569	0.00382
Z <sub>4_2</sub>	-0.2510	0.6590	0.7780	-0.2510	0.6580	-0.2506	0.6590	-0.2506	0.6589
Z <sub>5</sub>	0.6050	0.1930	1.8310	0.5990	0.1970	0.6050	0.1930	0.6050	0.1926
Z <sub>6</sub>	0.1790	0.0470	1.1960	0.1780	0.0480	0.1792	0.0470	0.1792	0.04678
Z <sub>7</sub>	-0.3110	0.0530	0.7320	-0.3110	0.0540	-0.3113	0.0530	-0.3113	0.05348
Z <sub>8</sub>	0.0020	0.8690	1.0020	0.0026	0.8620	0.0025	0.8690	0.0025	0.8693
Z <sub>9</sub>	-0.0990	0.8340	0.9060	-0.0990	0.8340	-0.0991	0.8340	-0.0991	0.8337
Z <sub>10</sub>	-0.4600	0.1480	0.6310	-0.4620	0.1450	-0.4600	0.1480	-0.4600	0.1477
Z <sub>11</sub>	0.6270	0.0460	1.8720	0.6240	0.0470	0.6270	0.0460	0.6270	0.04602
Z <sub>12</sub>	-0.4630	0.2240	0.6300	-0.4610	0.2250	-0.4626	0.2240	-0.4626	0.2241
Z <sub>13</sub>	0.0300	0.6760	1.0310	0.0306	0.6730	0.0303	0.6760	0.0303	0.6761
Z <sub>14</sub>	0.0250	0.2380	1.0260	0.0253	0.2400	0.0254	0.2380	0.0254	0.2385
Z <sub>15</sub>	0.0640	0.0430	1.0660	0.0632	0.0450	0.0638	0.0430	0.0638	0.04335
Cons	-50.73	0.2390	0.0000	-49.80	0.2490	-49.97	0.2470	-49.97	0.2474

Key to Variable names (Appendix 10: definitions for each variable)

Z <sub>1</sub>	Sex of household head	Z <sub>9</sub>	Combined use treatments
Z <sub>2</sub>	Age	Z <sub>10</sub>	Injection practice
Z <sub>3</sub>	Distance from nearest Town	Z <sub>11</sub>	Crops for market
Z <sub>4</sub>	Cut and carry system = Z <sub>4_1</sub> Confinement =Z <sub>4_2</sub>	Z <sub>12</sub>	House of the respondent
Z <sub>5</sub>	Gundi cooperative memberships	Z <sub>13</sub>	Total land owned privately
Z <sub>6</sub>	Intensity of Access to drug	Z <sub>14</sub>	Treatment year
Z <sub>7</sub>	Intensity of familiarity with technologies	Z <sub>15</sub>	Time for treatment
Z <sub>8</sub>	Cost for treatment		

The four statistical software packages gave almost equivalent coefficient estimate for each parameter and almost the same P value. These results indicate that success in the management of Trypanomosis disease would positively affected by intensity of access to drugs and total time spent for treatment at 5% significant level. A unit increase in intensity of access to drugs will increase the probability of success in the disease management by 119.6% and a unit increase in time for treatment will increase the probability of success in the disease management by 106.6%. However, intensity of familiarity with different treatments will negatively and significantly affect success in Trypanomosis management at 10% significant level. Even though the effect of Gundi cooperative membership on success of Trypanomosis management is positive, this effect is not statistically significant. Households who use cut and carry management system are 260% more successful than the reference communal grazing system users. Households who are market oriented in crop production are more successful than the others by 187% at 5% significant level.

#### 4.5.3 Data analysis to summarize baseline information

Baseline data was summarized from “Ghibe household survey study” which was conducted in 1997 and 1998. Since most of the variable definitions were not included in the label columns of the SPSS data files, summarizing some of the variables was very challenging.

The summary of indicators mainly included the following main category;

- Land use and land cover
- Household income
- Change in production
- Social human welfare
- Soil characteristics

***Table 16: Sample summarized data on dominant land use at Ghibe valley in 1997***

	Dominant land use			
	Frequency	Percent	Valid Percent	Cumulative Percent
Grow Crops	499	38.5	87.2	87.2
Grazing	45	3.5	7.9	95.1
Fallow	22	1.7	3.8	99.0
Forest	5	0.4	0.9	99.8
Homestead	1	0.1	0.2	100.0
Total	572	44.1	100.0	

## **5. Discussion**

### **5.1 Research participation**

The research participation task indicated the need for research methods support to researchers and students especially in data management and analysis. Research methods support should start from inception stage of research projects, as some problems created during the design stage of the project lasts up to the end of the study. For instance, since researchers and students do not consider the role of sample size and power calculation in the overall quality of the research, some research projects had very small sample size leading to biased estimates or conclusions. The contribution of research methods professional during data management or analysis is often constrained by the design of the study, and therefore researchers should consult research methods professional or statistician at the beginning of their project especially at the design stage.

The collaborative consultancy approaches which were used to provide support are among the best approaches to address research methods challenges or problems of researchers or students. The approach can help researchers to improve their research methodology and statistical techniques and at the same time it gives the consultant an opportunity to know more about the scientific background of the research problems and possible solutions recommended (Boen and Zahn, 1982).

During the attachment it was observed that most researchers and students appreciate the importance of data management after they have faced some challenge during data analysis or when they are not able get the expected result after the data analysis. But researchers should design their data management plan at the inception stage of the project (Monash University, 2011), which should be assessed by experienced data managers or researchers to assess whether it meets the standard data management practice. Data management plan has increasingly been one of the important research funding criteria and most donors need

this plan together with the research protocol. A good data management plan has several advantages for researchers and other secondary users of the data.

One of the most important data management challenge observed during consultancy was challenges in using data sets in different format. If data files are not in a format which can be used by all potential users, some researchers in the team may face problems in manipulating those data sets. Especially senior scientists may not have the expertise to use different statistical software as they may usually use single software which is convenient or suitable for their field of specialization. But generally researchers should at least know the most common ways data managements such as data transfer which help to convert data sets from one format to another. For instance some data transfer software such as *stat transfer* has good facility to transfer data sets without changing the special characteristics of the variables. There are also other general options which can be used to store data sets in a way that can help most researchers use in different statistical packages, such as text or CSV forms. However, to provide successful data management support to researchers, research methods professional should have at least the basic data management knowledge in different statistical packages.

Almost in all of the research projects supported, there were common types of data management challenges which include questionnaire design, variable definition, data cleaning and data storage. Mostly researchers, research assistants and students working on data management activities were challenged to define the variables view of the SPSS package especially correctly setting the value of categorical variables, missing values codes and the measure column. Before rushing to enter the data, the variables should correctly be defined based on the information in the data collection instrument. The variable view of SPSS has 10 main components which include Name, type width, decimal, label, value, missing, and columns, align and measure. But the most common errors observed were coding of the value of the observation and setting the measure column. This is mainly because of either limited knowledge about the statistical principles or lack of understanding of the data management aspects of the software. Some students did not have

any value coding for categorical data and others had incorrect codes or codes which do not give the required information after analysis. But researchers such as Pallant (2007) suggested that each response must be assigned numerical code before it can be entered into the SPSS package. These eases data entry, data cleaning, data manipulation, data analysis and interpretation of the results.

Missing data handling and data cleaning activities are not given due consideration among researchers. There is very little knowledge about the impact of missing values handling on the overall result of the project especially among research students, may be due to their limited experience and exposure to data management issues related to missing observations. Data cleaning is also one of the important data management activities which should be undertaken before any data analysis but there was very little understanding among researchers and students about this activity.

The other research methods challenge for research students observed during the consultancy was data analysis. For instance most of the students had research objectives and research questions but there was no data analysis plan which clearly indicates the type of analysis and expected output from the analysis. This is mainly due to lack of awareness and experience. Some of the data analysis challenges are highly related to the design of data collection instruments. This is mainly because researchers do not consider the data management and data analysis aspect when they are designing research instruments. For instance some students were not able to enter data of some questions after they completed their survey as some of the questions could not easily be coded. Therefore after developing the data collection instrument it is good to design the data entry format and test with at least with simulated data in order to identify any challenges during data entry or manipulation and make the possible adjustment before the data are collected.

The main data analysis challenge of students includes identifying the appropriate model, testing the model assumptions and analysis of survey Likert scale data. Even though Likert scale data is one of the most common data collected during surveys, in which respondents

rate their preference or perception about a given question, there were different challenges related to this data analysis. These challenges can be categorised as inappropriate design of the likert scale questions and lack of knowledge about the nature of these data and the associated analysis techniques. For example the response for one of the question about the perception of farmers was coded as 1 = not important, 2 = slightly important ,3 = important , 4 = very important and 5 = neutral . Such type of coding has basically two important problems. Firstly, it is unbalanced scale or forced choice scale, which has 3 positive and one negative response which forces the respondent to give positive response than considering both types of response. Secondly, the weighting given to neutral response is not correct. The neutral response should have been between the negative and positive responses to indicate the respondent is indifferent or unable to provide positive or negative responses. Analyzing such type of data using non parametric analysis usually provides a bias in the result and interpretation of the output. Because the highest weight is given for the neutral while it should have been for either of the extreme negative or positive values.

Different researchers have various assumptions about the analysis of Likert scale data and some of them have very serious argument against the use of parametric tests for liker scale data. Since many statistical tests rely heavily on distributional assumptions, such as normality and constant variance in the error terms, when these assumptions are not satisfied, the commonly used statistical tests often perform poorly, resulting biased and incorrect estimates. But on the other hand non-parametric tests are designed to have desirable statistical properties when few assumptions can be made about the underlying distribution of the data and can give better results than parametric. In other words, when the data are obtained from a non-normal distribution and other parametric criteria, a non-parametric test is often a more powerful statistical tool than its parametric ‘normal theory’ equivalent. Since the Likert scale data frequently used in households’ survey typically violates the assumption of normality necessary for parametric tests, most research methods professionals encourage the use of non parametric test for such type of data. In additions the ordinal scale nature of the data also violates the frequent assumption that data are from a continuous distribution. Therefore using parametric test for such type of data is not

recommended and researchers should check whether the assumptions of the parametric test are met or not before using them. However if the assumptions are not met the use of non parametric test is the best option.

The other challenge related with Likert scale data was interpreting the results after using nonparametric test, as it is not straightforward like other data analysis results. This challenge will be further complicated if the researcher is not using consistent values coding of the possible responses. For instance the interpretation of mean rank score obtained after doing Mann-Whitney U test for response coded from Very Important to Not very important is very different from Not very important to Very Important. This challenge would be worse if the value coding was not following the standard Likert scale coding. However if the coding error is not serious, it is possible to recode the values again and do the analysis.

Generally, the potential role of research methods professional in the research process is enormous and they should be involved from the inception stage of the research projects. By doing so they will have a room to contribute their inputs and avoid potential pitfalls in the research projects. If researchers seek research methods support only when they face challenges during the course of the project, the research methods professional may not have the possibility to add more value to that project and their role will be limited to only solving problems and not supporting the research process.

## **5.2 Training in the research process**

Research methods professional can also provide training to researchers or research assistants to transfer information and develop knowledge and skills. Most training events involve training needs assessment, preparation of training materials, delivering training and evaluation. Training after needs assessment helps to achieve the expected training outcomes effectively as it will engage the trainees actively during the course of the training. Training needs assessment can be done by questionnaire, group discussion, individual discussion with participant, self assessment, and discussion with managers and

by observation (Healthlink Worldwide, 2003). However training needs assessment should not only focus only on the training needs of the participant but it should also assess the background knowledge and skills of the participant. This helps to design the appropriate training strategy, the necessary content of the training and other preparatory work.

The statistical tool training task indicated different important points that should be considered for any statistical software training. Firstly, before starting the training it is better to confirm that the trainees have at least the basic statistical and research methods knowledge in order to accomplish the training outcomes properly. But if the trainees do not have sufficient statistical knowledge, it is better to help them first to gain the basics of research methods and statistical principles before starting the training. Secondly, the background knowledge of the trainees on statistical theories and methods will also determine the required time for practical and theoretical sessions. If there is limited time and the trainees have good statistical background then it is better to design the training session in a way that has more practical session than descriptive or theoretical sessions. But in case if the trainees do not have sufficient statistical knowledge there is a need to allocate balanced sessions between the theoretical and practical sessions in order to achieve the desired training objectives.

During the design stage of SPSS introduction training it was assumed that the trainees' statistical background to be above average and the training was designed to have more practical sessions than theoretical ones. But this assumption turned out to be wrong during the training and the trainees were found it difficult to understand well the data analysis and interpretation parts, which was due to their limited statistical background knowledge and data analysis experience. This was clearly reflected by their assessments that even though the training had more practical session than theoretical, their understanding on descriptive statistics, data entry, data cleaning and data analysis sessions is still medium.

The other most important lessons observed during the training task was training session's time management. The time allocated to a given training session should be sufficient in

order to deliver the required knowledge and skills. In their final evaluation the trainees indicated that the time allocated for the amount of material covered in the different sessions of the training was not sufficient and they suggested that enough time should be allocated to cover the most important concept and undertake practical exercises using the statistical package in other similar training. This was really one of the very important suggestions given from the trainees which should be considered in the future. However, the background knowledge and readiness of the trainees can also affect the amount of time used for each session of the training. If the trainees have good background knowledge, practical sessions are more effective than the theoretical sessions. If insufficient, there should be a balance between time spent on introducing statistical principles and practical sessions. During this training the participation of trainees was very high and this encouraged the trainees to facilitate the training session actively.

Training evaluation was one of the important activities accomplished to finalize the training task. Training can be evaluated by senior management, trainers or trainees. For instance, in their evaluations the trainees described about the importance of the training, its scope, content and standard of teaching. This would help to improve the training content and approach used in the future. It also helped to determine what the participants have learned and gave the opportunity for the trainees to reflect useful feedback in an organized manner, to develop future training plan, to ensure trainees and learners follow-up their training with relevant actions to apply. Training evaluation can be on daily assessment based or at the end of the training. In this training, while the trainer used daily evaluation, the trainees evaluated the training process only at the end of the training. On reflection, it was felt that a daily assessment would have been better for this training as it could have brought out the trainees' perceptions on the content of training earlier and could have been improved immediately.

Even though training can be evaluated by using different tools such as Observation, Questionnaire, Interview, diaries and Self recording of specific incidents, questionnaire was used to assess this training. A well designed questionnaire could be the best approach

in statistical software training, as it can give both qualitative and quantitative feedback to measure achievement of the objectives. But lessons from this training evaluation indicated that the questions should not be extremely specific and penetrating and mostly it should be general which mainly focuses on the training and learners personal objectives. Because specific and penetrating question may lead to biased feedbacks as the respondents may be tempted to tell their personal feeling freely.

### **5.3 Review of research protocol and report**

Review of research reports and proposal is necessary to improve research quality and ensure the relevance of research outcomes. Research methods professionals could be peer reviewers of reports and proposal and peer review is one of the most common approaches used to ensure the quality of the research proposal or reports. Participation in peer review processes is encouraged because it provides expert scrutiny of a project, and helps to attain high standards and encourage accurate, thorough and credible research reporting (Australia government, 2007). The result of proposal and report review has clearly indicated the role of research methods professional in peer review and the importance of peer review in improving the quality of research projects. However, peer review of research report or proposal needs sound and adequate background knowledge on the subject under review and the common approaches used to solve similar problems. This will enable the reviewer to provide assessment on both impacts and results of the study and validity or methodology used.

Research methods professional can also work as an interdisciplinary reviewer as it is defined by Pautasso (2010), where the reviewer is not knowledgeable about some areas of the paper's research questions, data or methodology. This approach was partially used for both review activities. In this case the research methods professional's have limited contribution to the research methods/ validity aspects of the study than on the impacts or results. But during interdisciplinary review, unlike standard review, research methods

professional can increase contributions by improving their understanding of the subject before undertaking the review process through reading or discussion with other experts. For instance, during the review task although the subject under review and the background knowledge of the reviewer were similar, there was no sufficient knowledge about the model described both in the research report and protocol. However, after referring to others work on the subject, sufficient understanding was gained to undertake the review both on the research methods aspect and on the validity of the research.

In additions to its advantage by standardizing the review process, use of review guidelines can ease the process. Before reviewing any manuscript it is important to have a minimum set of criteria and guidelines to undertake the review process. These guidelines should not focus exclusively on what is wrong and what needs to be improved, as this can be discouraging for many authors, but it should be approached in a supportive ways to identify strengths of the writing and/or the research ideas and highlights concepts that need consideration for further improvement. Peer reviewers should always approach someone else's paper as if it were their own and should be considerate and thorough, and always be respectful of the author's ideas.

The proposal review provided various suggestions that could help the principal investigator to further improve the protocol. The recommendations given to the principal investigator to improve the title, objectives, methods and data analysis section and the feedback from the principal investigator indicated important lesson what peer reviewers and researchers would consider in the review process. For instance the principal investigator gave feedback on each recommendation by acknowledging some of the issues and describing the logic behind on other part of the proposal content. This was mainly because the principal investigator considered the review process as an assessment to further improve the proposal and not evaluation of the manuscript. This was really very interesting approaches which gave an opportunity to the reviewer to learn from his comments and suggestion and develop his critical thinking ability too. Because from the principal investigator feedbacks for the comments and suggestions, it was also possible to understand the line of thought of the principal investigator which was not very clear previously.

The review result of “**The value of innovation system analysis for livestock research and policy**” indicates that the abstract was brief and explanatory which included the main points of each section in the report; the background information is adequate; the methodology used were properly described; the result section is comprehensive and; the referencing techniques appropriate. However, some minor improvements in the title, methods and model specification part were suggested. Although the title creates positive impression and stimulates readers’ interest it can be improved to indicate the subject and scope of the study with more accuracy, to help the reader identify key variables, both dependent and independent, and suggest a relationship between variables which supports the major finding of the study.

The methods section of research paper is one of the most important sections which provide information to judge the validity of the study. Hence in this section, the author must provide a clear and precise description of how the study was conducted, and the rationale for the specific procedures chosen. It must be written with adequate information in a way that the study could be repeated by others to evaluate whether the results are reproducible and the audience can judge whether the results and conclusions are valid (Fathalla, 2004). Even though the methods part of the report was good, the variables used to estimate the dependant variable were not clearly described in detail and it is not possible to repeat the analysis using different statistical software and models. Therefore, there should have been clear description about the methodology used and the variables included in the computation of the dependant and other independent variables.

Evaluating the quality of survey research reports using standardized criteria has also been the common practices among the scientific community and it was found that this research report is able to indicate the quality of the survey as it has included the most important aspects of the study. Poor quality survey research report can be attributed to ineffective reporting of sufficiently rigorous surveys or poorly designed and/or executed survey, regardless of the reporting quality (Draugalis *et.al*, 2008). Good research report provides opportunity for the reader to judge the quality of the research by clearly describing the

most important aspect of the research and the reporting process. The report can satisfy the standard criteria of most survey research reports. Because it indicates the research questions which were to be answered by the survey, the sample and sampling procedure, the design used, the research instruments used, the method used to test the instrument and the total response rate. However there were also some important points which should have been included in the report such as the power of the sample size (total population in order to help the reader to judge the power) and the variables included in the estimation of the dependent variables for the Tobit model. In additions if a reader looks at this report separately from the general survey report, the reporting techniques may not look sufficient. The report should have also indicated the limitation of the study and research gaps in this area which could be addressed in future.

Generally the result of the proposal and report review clearly indicated that the need of adding more effort to improve the organization research manuscript in order to make the science more sound and interesting.

#### **5.4 Data management**

Even though it has been overlooked by many researchers and research institutions, data management is one of the most important tasks in the research process which should be given great attention before and after the data collection task. Before starting any scientific research project, the principal investigator and the research team must address issues related to data management such as; data collection, data ownership, data storage, data protection, data storage, data analysis and data access (CTI, n.d.), because the quality of properly designed and implemented research project can be easily flawed by poor data management practice. Therefore in order to improve the quality of research projects due attention should be given to the designing of appropriate data management plan for the project if possible at the planning stage of the project and including it in the project proposal for further review and reference.

To achieve the desired research objectives and store the data for further reference or give access to secondary users, through standardize procedures, the presence of appropriate data

management strategy has great role. In this aspect there are few research organizations that have very strict regulation on the data management task which can be looked as a good sample for other research organizations and universities throughout the world. For instance the government of Australia has a policy that addresses the ownership of research materials and data including their storage, retention beyond the end of the project, and appropriate access to them by the research community (Australian Government, 2007). In additions to helping to achieve the research objectives clearly, these types of policy and regulation can facilitate the research process to move smoothly by setting transparency and accountability in the research processes.

#### **5.4.1 Data management challenges**

The data management task of the attachment mainly included designing of the data management plan, identification of the data management challenge and finally providing data management support to the team projects.

The first step in improving the data management practices of a given institutions is identifying the possible data management challenges which can be observed during the course of different research project and designing appropriate strategy to minimize those challenges and manage data properly. Therefore during the data management task of this attachment different challenges and lessons which can help the team and other projects to improve their data management practices were identified. Accordingly, the most common data management challenges identified from the SPSS data files of the project were; errors in defining the variables such as absence of variable label, incorrect specification of the values for categorical variables, and the measure column.

Before entering the data in SPSS it is important to tell SPSS about the name and type of variables and the coding instructions which are mostly called defining the variables. To define the variables we need to have information about the name, type, width, decimal, labels, values, missing, and columns, align and measure section of the variable view options. This is the most important part of the data management aspect in SPSS which can

also facilitate to archive the full information of the data and to setup the data for better analysis. Generally the coding of the variables in the values label should be similar to the coding in the questionnaire. This makes easy the data entry task and refereeing back the questionnaire during data cleaning and analysis for unexpected or unusual observation. However if there is a need to change the code of the questionnaire, this should be done by an experienced researcher after the data entry is completed based on the questionnaire or before starting the data entry, if it will be entered by well experienced data entry clerk because inappropriate or incorrect value coding can create a big problem in data management and data analysis activities.

SPSS has a very good facility to handle missing values, which involved coding of the missing values in the value column and setting them in the missing column as discrete or range values. Labelling missing value codes without defining the missing value column will let SPSS consider the missing value as real observation and perform data analysis using those values as other observation in the data. But sometimes it possible to see this error in different research projects databases due to lack of awareness on the impact of this error on the data management and analysis. SPSS also allows the user to assign different missing value codes for different observations missed due to different reasons. This helps to identify the extent and source of missing observations and take the appropriate correction measure. Therefore researchers should always use this facility properly in order to avoid bias resulted from data management errors and achieve their research objectives properly.

Using zero for observations which are missed due to one or another reasons changes the parameter estimates and total sample size included for different analysis. As it is clearly indicated in table 1 and 2, the use of value zero instead of missing value codes or using missing value coded for observations which have zero value leads to biased estimate of different statistics. For example when zero was considered as real observation, instead of missing value code, the mean land allocated for maize indicated an increase from 0.192 to 0.208 but when zero was considered as missing value code the change in the mean

value of land allocated to maize indicates a decrease from 0.382 to 0.359. This is because of the wrong understanding that all households who did not allocate their land to a given crop due to different reasons were taken as they allocated zero value irrespective of their different reasons why they did not allocate during that specific period. For instance, if there were households in which their area was not good for producing that crop or households who could not produce the crop due to agro ecological factors, the amount of land allocated for a crop not produced could not be zero but it is a 'not applicable' missing values case. In a real situation household may not also allocate their land for a given crop in a given year because of different reasons such as due to change in weather or other reasons and if value zero is entered instead of missing value codes the estimate will be biased. The descriptive statistics in table 1 also depict that the number of farming households involved in maize production is the same now and ten years ago, while table 2 shows that the number of farming households cultivating the three crops increased over the 10 years period time which could be expected. Therefore this is one of the most potential data management errors which can be resulted from such types of coding problems. For instance by using the above results it is possible to reach on two different conclusions. Based on the result from table 1 it possible to conclude that the number of farmers growing a given crop are the same now and ten years ago while by using table 2, it is possible to conclude the number of farmers have increased than before ten years ago. Hence researchers should clearly identify the cause of missing value and code their values properly both in the value labelling and missing value column of SPSS. This will avoid errors in the estimates of parameter and any inferences made based on them.

The measure column of the variable view of SPSS is also one of the important information which should be defined before data entry. This column refers the level of measurement in which the response variable under consideration was recorded. It distinguishes three different types of scales: nominal, ordinal and scaled variables. Nominal variables are also known as categorical variables when the numerical codes given are simply used to differentiate one response from the other. Ordinal refers to variables where the numerical

codes reflect an ordering of some sort, but where the distance between the categories can vary. Scale variables includes interval and ratio levels of measurement, where any numeric codes have meaning in terms of number relations that go beyond mere category and order. In most research data files this column did not set properly or simply set as the default value. Setting the level of measurements properly in SPSS has different advantages. It helps to determine the correct statistical test to use and whether the numeric data value can be treated as representative of a numeric measure of a characteristic, representing the rank or order within the distribution of data values, or is simply a short-hand substitute for a category. The measurement scale can be easily identified by looking the data, or the value column. In additions the statistical package sometimes may not activate the data for selection during analysis. For example if a categorical variable is set as scale data, during defining of variables in multiple response options this variable will not be activated for selection. Therefore, this column should be set correctly in order to maintain the quality of the data and facilitate the data management and data analysis activities.

Inconsistent variable naming which was observed in some of the data sets was also one of the data management issues that should be considered very well. Although it is not easy to give consistent and uniform variable names in long questionnaires, which have large number of variables, the naming system should not be the one that would be confusing and irrelevant during the data analysis process. This problem will be worse especially if the data is to be transferred to other statistical package for analysis. This will creates a big challenge in identifying the variables during data analysis and even during data management activities during the attachment. Therefore based on the variables naming rules of the statistical packages researchers should define the variable names in more consistent and descriptive ways in which other researchers and data entry clerks could understand very well.

### **5.4.2 Data management support**

The data management support provided on data cleaning, missing data handling and outliers handling indicated that the importance of considering these issues before any data analysis task.

Data cleaning which helps to check the data for any errors, which could be created during data entry or data manipulation, is one of the most important task which needs to be accomplished before doing descriptive and inferential analysis. Data cleaning helps to check the data from potential errors especially from outliers and missing values and to test the normality of the data which should be assessed before determining any type of analysis. Data cleaning involves in defining and determining the type of errors, searching and identifying the errors, correcting the errors, documenting the errors types and modifying the data entry format for future use. In SPSS the data cleaning task can be performed for categorical and continuous variables separately. Data cleaning for categorical variables mainly involves identifying value codes or an observation out of the range while for numeric or continues data it can also involves some descriptive and exploratory statistics such as minimum, maximum, mean values or outliers. Since the quality of any analysis result is only as good as the quality the data it is always good to give the necessary time for data cleaning.

Handling of missing data could be a potential source of bias during survey data analysis and the interpretation of the results is always challenging when the number of missing values is large. Researchers and research methods professional should have good understanding about missing data handling techniques in order to use the available information, especially if there are challenges in data collection, to avoid biased results in the presence of missing data and the commonness of missing data in most studies.

There is growing recognition that failure to address issues of missing data can lead to biased parameter estimates and incorrect standard errors (Davey and Savla 2010). But there are no also common applicable procedures to handle missing values, and different approaches may lead to different results. However to avoid concerns over data-driven

selection methods, it is essential to pre-specify the best missing data handling methods in the methodology section of the study protocol or analysis plan. But it is very important to note that the strategy employed to handle missing values might in itself provide a source of bias.

According to Davey and Savla (2010) data can be missed at random or non random. Random missing occurs when the probability that an observation is missing depends only on the values of the observed data and non random missing occurs in situations where the probability that an observation is missing depends on the values of the unobserved variables. This can be identified by looking the number of missing values whether it is large or small, the wording and design of the question and finally using SPSS add-on module called “Missing Values Analysis” that will statistically test whether missing values are random or non-random. If the missing values are non-random, then the study is not accurately measuring the intended constructs because results of the study may have been different if the missing data was not missing. Therefore there should be clear decision about how the missing data will be handled in SPSS before the data analysis.

Table 3 and 4 indicates the three different methods that can be applied to handle missing data: List wise deletion, pair wise deletion or mean imputation. Each of the above methods has their own advantages and disadvantages. For instance when there are few missing values (very roughly, less than 5% of the total number of cases) and if those values can be considered to be missing at random; that is, whether a value is missing does not depend upon other values, then the typical method of list wise deletion is relatively “safe” but if the missing values are non random on different variables and large in number using list wise deletion may cause losing most of the variables and cause reduction in sample size. This was clearly indicated on table 3 and 4, when list wise deletion was used the sample size is 81% and 50% for most variables in the descriptive and regression analysis while this percentage rose to 100% in pair wise deletion and mean imputation. This indicated that while list wise deletion decreases the sample size up to 50%, pair wise deletion or mean imputation can maintain the required sample size in spite of their limitations. Therefore,

most statisticians prescribe caution regarding the use of list wise deletion, even though it is the default method of missing data handling in various statistical software (Wayman 2003).

In pair wise deletion, all estimates are calculated on the basis of all cases that are available for a pair of variables. Pair wise deletion is better in some cases than list wise deletion when sample size is small or missing values are large, because there are not many values to begin with, and it helps not to omit even more like list wise deletion. For instance the difference in confidence interval column of table 3 indicated that pair wise deletion resulted in narrow confidence interval than list wise deletion and the standard deviation is lower for both variables in pair wise deletion than list wise deletion, which is desirable in most data analysis activities. But Haitovsky (1968) states that pair wise deletion is less commonly used than list wise deletion mainly because of the potential for covariance matrices that are less than full-rank and other statistical problems. Davey and Savla (2010) said that it is unlikely that a pair wise approach will correctly adjust parameter estimates and standard errors and therefore without stronger statistical justification, this approach is probably best avoided in statistical analysis.

The third option, Mean imputation, replaces missing data with the average of valid data for the variable in question. Mean imputation helps to use the full sample of the study as indicated in table 4. Table 4 indicates that the total sample size in mean imputation is 100% for all the variables while there are variations in Pair wise deletion. But mean imputation could potentially lead to very serious bias as it estimate parameters by using the same value of other observation and hence it artificially reduces the variance of the variable in question, in addition to diminishing the relationships with other variables (Wayman 2003). Therefore, this method is not mostly recommended by statistician and researchers.

Currently there are also other approaches which can also be used to deal with missing values. One of the options is using multiple imputations which involve predicting the missing values by using the existing values from other variables. Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data,

but also by incorporating the uncertainty caused by estimating missing data (Wayman,2003). According to Schafer (1997), multiple imputation methods demonstrated superior properties when compared to simple imputation methods and list wise deletion. By imputing more than one value for each missing observation, uncertainties due to imputation is removed from the analysis and since imputed values are estimated by using information from other variables in the data the estimates are more acceptable than mean imputation.

Generally from the missing data handling activities it is possible to understand the absence of universally applicable and accepted techniques for all types' data sets and hence researchers should consider the various missing data handling option based of the nature of their data sets before undertaking any data analysis in SPSS and other statistical packages.

Outliers handling is also one of the very important data management aspect which should be considered before any data analysis. Observations which are considered to be extreme or unusual based on the distribution of the data are considered as outliers. Outliers can occur due to errors in measurement, observation, sampling, modelling or data transformation, and natural variability in the data and can have advantages or disadvantages (Clark 1989). For instance they can be used as diagnostic tools in model building, can be of intrinsic interest and can help detect specific observations. The statistical methods which can be used to handle outliers vary depending on the context of the study and the source of outliers. Therefore, following discussion tries to show researchers an overview of how to cope up with untraceable sources or natural values of the observation.

Outliers can have different effects on the parameter estimates of any analysis such as resulting bias or distortion of estimates, inflated sums of squares (which make it unlikely you'll be able to partition sources of variation in the data into meaningful components), distortion of p-values (statistical significance or lack thereof can be due to the presence of a fewer even one unusual data value), faulty conclusions (it's quite possible to draw false conclusions if you haven't looked for indications that there was anything unusual in the

data). Choi (2009) explained that outliers can dramatically change the magnitude of regression coefficients and even the direction of coefficient signs (i.e., from positive to negative or vice versa). These issues were clearly observed from the positive deviance study when the factors which affect the intensity of forage innovation were determined using Heckman two stage models. Table 5 indicates the effect outliers have on the magnitude of parameter estimates and the significance level. Therefore any hypothesis tests which specify positive or negative causal relationships between dependent and independent variables, reporting estimated coefficients without outlier diagnostics can lead to incorrect inferences and compromise otherwise scientifically valuable findings. If researchers ignore outliers especially in the dependent variables the estimate of parameter and other inferences based on that result can be misleading.

There are various options to handle outliers in a given study. For example as indicated in table 5 after checking that the value of the outlier is the true observation, it is possible to assess the effect of outlier in the parameter estimate by including and excluding the outlier value in the analysis and make decision based on the result of analysis, theoretical background and judgment of the researcher. If the inclusion of the outlier caused bias in the parameter estimate or gave distorted P value, it is mostly recommended to exclude the outlier from the analysis and report the analysis results with and without outlier in order to provide information for the reader. But if the inclusion of outlier does not have effect on the parameter estimate and P value it is recommended including the outlier in the analysis. However this method is not the only solution and there are other options which could be considered such as transforming the variable or substituting the value with other reasonable value. Hence researchers should identify and design the best approach to handle outliers based on the cause of the outlying and the nature of investigation under consideration.

### **Qualitative data collection and summarising**

Qualitative data which is mostly obtained in the form of spoken or written language rather than numbers can be collected with interview, observation, documents or artifacts.

Production of interview data requires awareness of the complexity of self-reports and the relation between experience and language expression. Different challenges associated with qualitative data collections were identified during the data collection task such as unrelated responses, inconsistent responses, unable to understand the question and problems associated with the length of the question. However, to handle all the above challenges, it needs more practice, skill and time than similar challenges in quantitative data collection. Sometimes it was observed that respondents could not give the appropriate response due to lack of understanding the question or lack of the required knowledge on the subject.

The facilitator during qualitative data collection should be well experienced in order to guide the respondents in the appropriate direction and get them on track to obtain the required information and achieve the research objectives easily. Sometimes respondents give inconsistent information due to lack of understanding the main information required or lack of readiness on some specific issues. This can be avoided by providing appropriate information in a very clear and simple words and enough time to think over it. The length of the questions has also important role in qualitative data collection. The questionnaire should not be very long in order to get the required information from different respondents, because some respondents may not have time and willingness to provide responses for long questions. Unlike quantitative data collection, the questions are held to be a constant stimulus so that the only variation in answers can be assumed to originate from the respondent, during qualitative data collection it was possible to adjust the question to the individual being interviewed in order to get sufficient information. But it needs good skill to explore information from each participant and therefore interviewing skill needs to be practiced and mastered before beginning the data collection task.

Summarizing qualitative data was also one of the data management support provided to the team. The data was summarized in a word processing document before undertaking the final analysis by using different procedure depending on the nature of the data collected and the type of information needed. Generally the first most important step in qualitative and quantitative data analysis is to understand the data very well. But unlike quantitative

data analysis, in qualitative data it is not the marks on the paper the analyst consider but on the meanings represented in the texts recorded. In qualitative data analysis it is not possible to analyze the printed words by counting how many times a particular word appears in the text, but the information, the ideas and thoughts that have been expressed by the respondents. Therefore, the qualitative data analysis task indicated that researchers should clearly understand the general ideas and thoughts of the collected information before starting the data analysis task. Then based on the objectives of the study the data can be categorized to identify patterns and relationships between different categories and finally all major categories should come together to get the findings behind the data collected. But this task indicated that summarizing qualitative data is very challenging than analyzing quantitative data.

### **Design of data collection instrument**

Design of the data collection instrument has also its own impact on the data management and data analysis activities of research projects. Careful and precise description of data collection instruments and procedures are vital for obtaining consistent and reliable data. The most common data collection instrument in survey is questionnaire, designed well; it is a good scientific measuring instrument. It is very important to carefully think through the purpose of the research before rushing into the field with a questionnaire. Questionnaire should be more than a mere inventory of what would like to be known, but consideration must be given to some of the problems involved in getting what is hoped for and the best method of gathering that information. Different lessons were learned during the data collection instrument design tasks. The different steps used to design the data collection instrument resulted good data collection instrument which can help to achieve the desired research objectives. The data collection instrument should be designed very carefully in order to get the required information without wasting resources. Development of data collection instrument involved varied procedures which may include; identifying the core objectives of the study, defining the research questions, listing the variables under

each of research questions, anticipating how the data will be analyzed, developing tentative questions, ordering the tentative questions, pre test the questions and pilot study.

Before starting to write the questions it is very important to critically think about the main research objectives of the study and break down this main objective in to other core objectives. Then under each core objective the research questions and the variables which will be measured to answer each of the research questions should be clearly stated. This will help to avoid writing unnecessary questions and consider only the most important variables which help to answer the research questions. After identifying the most important variables, it is always good to think about the possible data analysis options which will be undertaken using each of the variables. This will help in deciding the measurement scale which will be used to collect information using each of the variables. After carefully accomplishing the above tasks the researcher can now start writing tentative questions and undertake the pre test using non-randomly selected samples from the study population. This step will help us see the importance of the variables identified in field condition and to make modification on the variables and other parts of the data collection instrument. In addition it also helps to assess the amount of time required to complete the questionnaire and to check whether the information will be analyzed, to clarify directions, question wording, or response categories where necessary. Finally there should be pilot study to assess the questionnaire, the interviewers, and all other aspects. During pilot study it is possible select up a sample of up to 50 respondents and collects the data. The data will be coded and analyzed to assess the relevance of information to the objectives of the study. Questions that are not providing useful data should be discarded, and final revisions of the questionnaire will be made.

The steps mentioned here are very important in the design of survey data collection instrument, but they are not exhaustive. Researchers or statistician should design every step carefully based on the nature of their study as the final results of any research are only as good as the quality of the data collection instrument and the data collected.

## **Preparing the data for storage**

Data storage has become increasingly the most common data management practice in the scientific community. Research data which is defined as the material, originally recorded by or for the investigator, commonly accepted in the scientific community as necessary to validate research findings should be stored in appropriate format after the study is completed. It includes the raw data collected, data collection instrument, study protocol, notebooks, as well as any other records that are necessary for the reconstruction and evaluation of reported results of research and the events and processes leading to those results, regardless of the form or the media on which they are recorded. Data storage is crucial to a research project for the following reasons: Properly stored data has the following different advantages; it is a way to safeguard your research investment, data may need to be accessed in the future to explain or augment subsequent research, other researchers might wish to evaluate or use the results of your research, stored data can establish precedence in the event that similar research is published. Storing data can protect research subjects and researchers in the event of legal allegations (ORI, 2011).

Even though it is not common to see in most data files, stored data should at least contain the study protocol, metadata, data file description and data sets in different format. During my attachment period there was an opportunity to see different completed projects data files and most data files do not contain Meta data and data file description. Therefore both Meta data and data file description were prepared by consulting the research technicians and using the study protocols for most data sets. However it was really very hard to create for some old data sets due to absence of both data collection instrument and study protocol. Meta data and data file description are among the most important information which helps to understand any archived data. Meta data is data about the data and is generally descriptive information or documentation about statistical data. It facilitates sharing, querying, and understanding of statistical data over the lifetime of the data. Data file description gives information on the data file format, number of observations and description about each of the variables included in the data sets. Especially if there are computed variables, it can give us the variables included in the computation and the

procedures used to compute the variables. Therefore any archived data should contain at least study protocol, Meta data, and other data descriptions during archival in order to understand the nature of the study without any help from the research team who have good knowledge about the data or study.

## **5.5 Data analysis**

Data analysis was one of the tasks accomplished during the attachment in order to support the research team and learn from the team experiences. Data analysis generally involves different techniques to describe, explore, understand, prove and predict information based on sample data sets collected from population using some sampling strategy. Data analysis is more than manipulating numbers and it is an activity that should be considered in all stages of a study. Both qualitative and quantitative data analysis requires understanding a variety of data analysis methods such as planning data analysis early in a project and making revisions in the plan as the work develops, understanding which methods will best answer the study questions given the data that have been collected and, once the analysis is finished, recognizing how weaknesses in the data or the analysis affect the conclusions that can properly be drawn (GAO 1992). Data analysis generally involves two major components: descriptive and inferential analysis. The descriptive statistics help to describe and summaries the data and allow presentation of the data in a meaningful ways to allow simpler interpretation. Descriptive analysis cannot help in making conclusions beyond the data that has been analyzed or reach conclusions regarding any hypotheses we might have made. But inferential statistics allows using the sample to make generalization about the population from where the sample is drawn. In the data analysis tasks described in this attachment both the descriptive and inferential analysis were done for different studies. But the report focuses more on the inferential analysis used and the methodology applied to choose the appropriate model.

In the first data analysis task, determining the factors which affect the intensity of forage innovation, the dependant variable in the analysis which was a continuous variable between zero and one and the independent variables contain both categorical and numeric

variables. Using these variables three different statistical methods were applied to estimate the level farming households' forage innovation intensity and to identify the factors which affect the intensity of innovation.

The first option was to use OLS methods; however, this method would be misleading as there are observations which have zero values that could destroy the linearity assumption of the model. As indicated by the descriptive statistics summary in table 4, there are 76 unsuccessful cases which have zero intensity of forage innovation. This even indicated that the presence of strong relation between being unsuccessful and zero intensity of innovation which should be considered even before determining the intensity of innovation. But in situations where the dependent variable to be modelled is limited in its range, using OLS estimation would result in biased and inconsistent parameter estimates (Heckman, 1979).

The second option was to use Tobit model. However, the main limitation in Tobit model is that the same set of variables and coefficients determine both the probability that an observation will be censored and the value of the dependent variable, intensity of innovation. In additions Tobit model does not also allow a full theoretical explanation of why some observations are censored. However, sample selection model can easily demonstrate these and address the shortcomings of both OLS and Tobit model by modifying the likelihood function.

According to Heckman (1979) sample selection bias may arise in practice for two reasons. Firstly, there may be self selection by the individuals or data units being investigated and secondly, sample selection decisions by analysts or data processors to operate in much the same fashion as self selection. Sample selection model uses different set of variables and coefficients which determine the probability of success and the value of intensity forage innovation. But these variables may overlap, to a point, or may be completely different and then it allows for greater theoretical development because the observations are said to be censored by some other variable. This allowed taking account of the censoring process because selection and outcome are not independent.

Heckman two stage models helped to explain intensity of forage innovation as a function of several explanatory variables in a situation where we only observed innovation index for those households that are successful. But there could be a very clear theoretical reason why some households which are not successful did not have innovation index. In this case self selection problem might arise because unobservable factors that affect the success in the adoption and use on the one hand and, the intensity of adoption and use on the other hand might exist and affect both situations. In this case the farmer may thus self select himself / herself for adoption and use decision because of different unobservable factors. To overcome this problem, the Heckman two stage econometric estimation procedures which have selection and outcome equation were preferred. Success in the adoption and use of the technology is a dummy variable that takes a value 1 if the household is using the technology now and 0 otherwise. The dependent variable in the outcome model, the maximum intensity in the use of the technologies by the household heads now, is a continuous variable measured by innovation index of the household head which was computed from proxy indicators of forage innovation.

Since there are two models, the selection and outcome model, it is possible to have a theory that specifies why some households are successful and others are not. This could mainly be due to different socioeconomic and demographic characteristics, lack of interest to continue their innovative activity, lack of the required knowledge and skills or others. Therefore in the framework of the sample selection model, it is possible to specify one equation for whether or not a household is a successful innovator or not and a different equation for the intensity forage innovation.

One of the most important advantages of sample selection model is that the zero-censored data, selection bias, problem can be controlled by using the inverse Mills ratio which contains information about the unobserved factors that determine households success in innovation activity. The inverse Mills ratio would be obtained from the first stage of probit selection equation and it will be included in the second stage of OLS outcome equation. The significance of the inverse Mills ratio reflects the importance of selection bias for

which if it is statistically significant indicating the presence of sample selection biases. In fact, as Heckman demonstrated, if the processes are related, estimating a model of intensity of innovation without first estimating an equation of whether or not the household was successful innovative or not would lead to biased results.

One of the most important assumptions in sample selection model is that the error terms of both the selection and outcome equation are correlated which is indicated by the correlation coefficient,  $\rho_{\varepsilon u}$ . In additions the value of Lambda indicates whether there is selection bias or not. If the value of lambda is statistically significant there is selection bias in the sample and the use of sample selection model is encouraged and if it is not statistically significant it is possible to use either OLS or Tobit model. This condition was clearly observed during this analysis where Lambda has value of -0.13 with P value 0.041 which is statistically significant at 5% level. The negative sign of the lambda coefficient indicate that unobserved factors which affect the success of forage innovation are inversely related to unobserved factors which affect the intensity of forage innovation. Therefore base on the estimate of Heckman two stage model it is possible to say total training provided to the farming household, farm enterprise index, diversity of cooperative membership, total livestock unit, innovation index and forage technology with follow up will positively and significantly affects the intensity of forage innovations.

As indicated the result of Tobit and OLS models have vary comparable estimates. The reason could be due to the sizable number of zeros and therefore the likelihood maximization of Tobit is very similar to a linear model. But the Heckman two stage models gave results which are different from the two; in the sets of variables which are statistically significant parameter estimates and their signs. For instance in the case of Heckman the two stage model variable family sizes and diversification of fodder tree use is not statistically significant while they are statistically significant in both OLS/Tobit models. Generally the parameter estimates of OLS and Tobit models are closer than that of Heckman and OLS/Tobit models.

Although the Heckman two stage procedures sounds very good theoretically, its practical application is not easy including the interpretation of the result. Therefore, before going to the description of the practical estimation of the Heckman model researchers should understand the limitation of this model. Firstly, in order to get good result the selection equation should contain at least one variable which is not related to the dependent variable in the substantial equation. If such a variable is not present (and sometimes even if such a variable is present), there may arise severe problems of multicollinearity and addition of the correction factor to the substantial equation may lead to estimation difficulties and unreliable coefficients (Crouchley, 2002). The second important limitation is that results from the two stage estimator are sensitive to violations of normality assumptions other than other semi and parametric methods. But as parametric methods have also different limitations they often give different results and have less precision in a situation where there is selection bias and hence the Heckman approach continues to dominate the literature (Crouchley, 2002). Bascle (2008) also described that like other econometric methods the Heckman two-step procedure is sensitive to specification issues and whether some assumptions are met; however, it is still appealing as it provides the inverse Mills ratio which helps the analyst with a potential correction and an informative estimation of the self-selection bias.

Generally this indicates that for policy purposes one could suggest different recommendations based on the choice of either model. For instance, based on Heckman model someone can conclude that family size and diversification of fodder tree use have no significant role in improving the intensity of household innovation activity, while it was possible to reach on the opposite conclusion using OLS/Tobit model. In addition using Heckman two stage model it is possible to conclude that distance to nearest district centre has positive effect on intensity of innovation while it has negative effect in OLS/Tobit models. Therefore this analysis also indicated that the methodological approach that researchers use can totally affect the recommendation and conclusion which can be drawn

from a given study irrespective of the type information collected and researchers should carefully consider the type model in which they build based on their data.

The secondary data analysis accomplished to identify the factors which affect the success in Trypanomosis management indicated the possibility of extracting information from secondary data set for purposes other than the main objectives of the study. This type of analysis has different advantages such as assisting to define a research problem, provide information to formulate research questions, hypotheses and even to select a research design and sometimes it may be sufficient to answer a research question (Patel *et.al*, 2004). Mostly secondary data analysis also helps to obtain information easily and rapidly with minimum cost and time, if the archived data quality is good enough. The assembly and analysis of secondary data almost invariably makes an important contribution to the research process (Patel *et.al*, 2004). Therefore as a general rule, a thorough research of the secondary data is recommended prior to conducting any primary research.

This data analysis depicts the factors which affect the success in Trypanomosis management. Trypanomosis disease management can be significantly affected by different factors such as intensity of access to drugs, cut and carry animal management system, total time spent on treatment and market oriented crops production activity. However factors such as intensity of familiarity with different fly control methods will affect negatively the success in the disease management. This will help us to identify the factors which can play major role in the management of the disease and to plan the appropriate policy intervention to further strengthen the success in the disease management. The analysis result from the four different statistical packages also indicates the use of different statistical package cannot significantly change any data analysis results. If the researcher has the access and desired knowledge about a given statistical packages it is possible to use any statistical software, provided that the required parameter estimate are obtained.

The third data analysis task, summarizing baseline data for the impact assessment study of animal health intervention at Ghibe was also accomplished based on the data analysis plan prepared. However since the database had different variable definitions error, the task had

been really very challenging. This problem was even very worse for those data files in which the data collection instrument were not available.

One of the most common practices to assess the impact of an intervention is to use indicators before and after the intervention. Data collected to assess the impact of a given intervention before the start of the project is mostly described as baseline data. Baseline data are required to generate information on the status of the system before the intervention and to compared with the expected state of the system for ex ante assessment or with the impacts already seen for ex post assessments. Variables for the baseline indicators were selected carefully based on the objectives of the impact assessment. However, since these data were was not collected intentionally to use as a baseline reference there were some indicators which could not be obtained from these data sets and there might be a need to look for another sources for those variables or design additional methods to assess the impact. The absence of all the required information depicted one of the constraints of designing research projects based on secondary data.

Generally from the above two data analysis tasks it is possible to learn both the advantages and disadvantages of secondary data analysis and the role of proper data archiving systems for using data sets for secondary analysis.

## 6. Conclusion

- Challenges observed during the project indicated that a research methods professional should be involved in research activities starting from inception stage.
- The review of research reports and proposal depicted that researchers should work more on the organization of their manuscripts.
- There was limited capacity in the application of wide variety of statistical tools.
- There was limited data management practice among researchers and students. Therefore research institutions and universities should work more to avoid challenges related to data management.
- Using OLS for dependant variable observed at restricted non random sample will lead to biased estimates and Heckman two stage models is one of the best approaches to handle such type of data.
- Intensity of forage innovation will be positively affected by different factors such as total training, farm enterprise index, diversity of cooperative membership, total livestock unit, innovation index and forage technology with follow up or training.
- Based on the available data analysis result success in Trypanomosis disease management will be significantly and positively affected by cut and carry livestock management system, intensity of access to drugs, total time for treatment and market oriented crop production systems.
- Even though there are different options that could be used to extract information from secondary sources the validity and quality of analysis result will be constrained by the nature and quality of the data sets.

## **7. Recommendations**

- Researchers should have the basics skills and knowledge of data management in order to figure out common challenges of data management before running any data analysis. Especially if the data is entered by other supporting staff or assistants, without researchers close supervision and inspection; they should check the cleanness of the data at least by considering the most important variables.
- Research methods professional that are expected to support researchers from diverse disciplines should have the basic skills and knowledge of data management and data analysis options using different statistical tools.
- Universities and research institutions which are involved in the capacity building of graduate students should provide training on data management practices to students at the beginning of their research projects or devise a system which will let them to learn from senior professionals. This will help them to think and avoid common data management errors while analysing their data.
- Designing and analysis of Likert scale data is challenging to most research students and researchers. Therefore research methods professional should assist researchers by providing the appropriate knowledge.
- Training on statistical tools can be successful when it is given based on the identified need of the trainees and when the theoretical and practical sessions are balanced based on the knowledge and skill levels of the trainees. During the practical session, more time should be given to the trainees to do most of the activities by themselves with limited help from the trainers.
- Evaluation of training can help the trainer to improve future planning and implementation of training by giving the opportunity for both the trainer and trainees to review, adjust and revise goals, schedules, and procedures of the whole training process. Therefore training evaluation should be considered as a basic and non-negotiable component of the training and should be designed before the training.

- Peer review based on standardized guidelines will make the review process more objective therefore research institutions and reviewers should develop review guidelines before the review process based on the nature of the subject.
- Errors in data management can flaw the entire results of a given research projects and lead policy makers and practitioners to reach incorrect and biased decisions. Therefore scientists and researchers should strictly follow the data management aspect of their projects starting from the project planning stage by including appropriate data management plan in the study protocol.
- Data cleaning is one of the most important data management practices in the research process. Data cleaning should be undertaken by experienced researchers who have good understanding of the objectives of the study and subject of investigation. Scientist should not always leave the data cleaning activity to their assistants but should at least check the status of the data by assessing some of the most important variables randomly.
- Data collection instrument should be designed very carefully with appropriate procedures to collect the required information without wasting resources. Design of data collection instruments should consider the data management and data analysis options for all variables.
- As much as possible data files should be stored in a format that can be used by various statistical tools or that can easily be transferred to different statistical tools without losing much information
- Different variables affect the model in different ways. Hence during model specification the potential variables which affect the model should be identified based on their theoretical and practical importance.

## **8. Lessons learned**

This research project has given me enormous opportunity to develop my research methods knowledge and skills in different ways. Although there are many lessons which were learned during the different research methods and approaches tasks, only the most important ones are summarised below.

### **Research participation**

- If research methods professionals are involved in research projects from the inception stage, they will have significant contribution in ensuring robust research findings and results and improving the effectiveness of research resources invested. On the other hand, research methods professional should also be very dynamic and competent with adequate skills and knowledge of research methods and, have an aptitude to understand the other technical perspectives in the research team.
- Research consultancy based on collaborative approach will benefit both the consultant and the researcher. This approach empowers the researcher and gives opportunity to learn from the experience of the consultant and at the same time exposing the consultant to different research problems and challenges faced by researchers.

### **Training**

- It is not possible for the trainees to know all subtleties of several complex methods in few days, so it is advisable to focus on fundamental but practical concepts underlying the method instead of packing the session with many statistical methods.

## **Review of report and proposal**

- Reviews of research manuscripts by Research methods professionals which aimed at improving the quality of research outputs should focus mainly on the methodology aspect of the paper based on generally accepted guidelines and procedures. Well established internal and external peer review systems for research proposal and reports results enormous contribution to the quality of research projects.
- If peer review does not considered as an evaluation of the author or researchers work, but as one of the main activities in the research process to maintain the quality of research outputs, it will build the connections between researchers and reviewers and keeps them abreast of new research methods and approaches. In additions, it helps researchers to promote their work among fellow scientists and professionals and get new insights and understanding to further improve their work.

## **Data management**

- Because different scientists come up with different challenges including data stored in diverse statistical tools, a research methods professional who specializes only in very specific statistical tools may not be able to fully assist researchers working in different fields to solve their challenges.
- There is very little understanding about the analysis of Likert scale data collected from surveys. Data analysis which does not consider the assumptions of statistical principles, like analyzing likert scale data with t test will lead to wrong conclusions which may mislead policy makers and other users of the research outputs.
- Designing of likert type questions need knowledge of how to analyze the data. It is advisable to be consistent in valuing the response rate throughout the questionnaire such as using higher values or codes for positive responses such as “very important”, “strongly agree,” “good,” and lower values for negative responses such as “not very important”, “strongly disagree” “bad” etc. However coding in the reverse way, high

value to negative such as “strongly disagree” responses and low values for positive responses “strongly agree” is also possible as long as the researcher is clear and consistent with the coding throughout the data. But mostly mixed coding can create confusion during interpretation of the results.

- Setting up appropriate data management units with appropriate knowledge and skill in data management will offer an enormous benefit to research institutions.
- Handling of missing data and outliers are among the most important data management tasks which demand the attention of researchers. Both outliers and missing data can possibly lead to biased estimates especially in inferential analysis.
- Data management or analysis of qualitative data is more challenging than quantitative data and needs the skill and experience of the analyst in handling similar types of problem.
- A good data collection instrument developed based on the objectives of the study and the research questions will not only save resources by avoiding collecting unnecessary questions during survey but it also makes the data management and data analysis task more easy. A good data collection instrument can be developed by clearly identify the potential variables and their measurement scale and converting them to question based on the objectives of the study.
- Even though pilot test of data collection instrument is a very common practice before survey, there is limited effort or motivation to test the data collection instrument suitability for data entry. But if researchers pre-test the appropriateness of their data collection instrument during data entry before starting the data collection activity it will assist them to avoid potential challenges during data entry and analysis.

## **Data analysis**

- Multicollinearity was one of the common challenges in survey research especially when a model includes many variables. Multicollinearity can affect model parameter estimates, sign of coefficients and the overall goodness of model fit or the likelihood ratio of the model.
- Sometimes it is possible to change continuous variables into categorical variables if the original variables do not have significant meaning and the converted variables have meaningful contribution to the model than the original variable. However, in this situation the researcher should carefully decide which one to use. The choice should not be simply by looking the significance level of the variable but it should be based on the nature of the variable, its theoretical importance, and possible bias which can be resulted.
- Composite indicators can provide better information than individual variables, but researchers should carefully consider when and how to use the composite indicators. If the contribution of individual variables in the model is not important and only general information is needed it is better to construct composite indicators. However, the level of measurement and the information composed should reflect the true nature of the variables. In addition when a composite variable is included in the model both the method used to compute the variable and the variables used to construct the composite indicator should be explicitly described in the methods part the manuscript and the data description section of archived data.
- Using parametric tests such as T-test or ANOVA for Likert scale data can lead to biased parameter estimates which can result in wrong conclusions. The problem would be particularly worse if the scale does not follow the standard Likert scale format. Therefore, clearly understanding about how and when to use those Likert scale data and their design is very essential.

- The inclusion or exclusion of variables in the model will not only affect the overall goodness of fit of the model but it also affects the coefficient estimate and significance level of other variables in the model.
- The types of default statistical results obtained from different statistical software can differ based on the type of assumption used in the model specification. For instance Stata does not give the estimated R square for Heckman two stage models while R can give.

## 9. References

- Andèr,H, Mellenbergh, G & Hand, D 2008, *Advising on research methods: A consultant's companion*, Johannes Van Kessel Publishing, The Netherlands.
- Altman, M & McDonald, M 1997, 'Choosing Reliable Statistical Software', *Political Science and Politics* Vol. 34, No. 3 ,Sep., 2001, pp. 681-687
- Amherest College 2010, *Training and Development*, Viewed on 20/12/2010, [https://www.amherst.edu/offices/human\\_resources/training/](https://www.amherst.edu/offices/human_resources/training/)
- Australian Government 2007, 'Australian code for the responsible conduct of research', Viewed on July 1 2011, <http://www.ag.gov.au/ccca.>
- Barker, K , Pistrang, N &Elliot, R 2002, *Research Methods in Clinical Research: An introduction to students and practitioners*, John Wiley & Son, England.
- Bascle, G 2008, 'Controlling for endogeneity with instrumental variables in strategic management research' *Strategic organization* Vol 6(3): 285–327, Sage Publications (Los Angeles, London, New Delhi and Singapore)
- Batanero, C (Ed) 2000, 'Training researchers in the use of statistics', IASE Round Table Conference, Tokyo.
- Berthold, M, Berglot, C & Klawonn 2010, *and Guide to Intelligent Data analysis: How Intelligently Make Sense of Real Data*, Springer-Verlag, and London.
- Blake, R &Mouton, G 1976, *Consultation*, Reading MA, Addison –Wesley Publishing Co.
- Boen, J and Zahn, D 1982, 'The Human Side of Statistical Consulting', *Lifetime Learning*: Belmont, CA.
- Choi, S 2009, 'The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment', *Quarterly Journal of Political Science*, 2009 4: 153–165
- Clark, T 1989, *Managing outliers: Qualitative issues in handling of extreme observations in marketing research*, University of Notre Dame.
- Clayton, M 1996(a), 'The statistical consulting', *The statistical consultant*, Vol. 30 No.2
- Clayton, M 1996(b), 'The statistical consulting', *The statistical consultant*, Vol. 30 No.3
- CTI n.d, *Guidelines for Responsible Data Management in Scientific Research*, viewed on 23 July 2011, <http://ori.hhs.gov/education/products/clinicaltools/data.pdf>
- Coe, R,Stern ,R & Allian,E 2002, Data analysis of agro forestry experiment, lecture note, ICRAF, World agro forestry Center, Kenya.
- Cox, P 1968, 'Some Observations on the Teaching of Statistical Consulting', *Biometrics*, Vol. 24, No. 4. December, 1968.

- Crouchley , R , Bradley, S and Oskrochi, R 2002, *Evaluating the impact of missing data in social research: simulations and applications using the BHPS and the NCDS*, Lancaster University, Lancaster, LA1 4YF
- Crowther, D & Lancaster, G 2005, *Research methods: A Concise introduction to research methods in management and business consultancy*, Linacre House, UK.
- Davey, A & Savla, J 2010, *Statistical Power Analysis with Missing Data*, Taylor and Francis Group, LLC, New York.
- Dawson, C 2009, *Introduction to research methods, A practical guide to any one undertaking a research project*, How To Books, United Kingdom.
- Draugalis, J, Coons, S and Plaza, C 2008, 'Best Practices for Survey Research Reports: A Synopsis for Authors and Reviewers', *American Journal of Pharmaceutical Education* 2008; 72 (1) Article 11.
- Druckman, D 2000a, 'Frameworks, Techniques, and Theory: Contributions of Research Consulting in Social Science', *American Behavioral Scientist*, San Francisco, Jossey-Bass.
- Etling, A 1997, A case study on collaborative consultation: Outstanding research presentation, university of Nebraska, Lincoln.
- Eysenbach ,G 2004, 'Peer Review and Publication of Research Protocol and proposals, A Role for open access Journals', *Journal of Medical Internet Research*, Vol.6, No.3, viewed 20/10/1010 <http://www.jmir.org/2004/3/e37/>
- FAO 2010, Conducting Agricultural census and survey: Instructional manual: *FAO Statistical Development Series*, No.6 Viewed 22/12/2010, <http://www.fao.org/economic/ess/world-census-of-agriculture/conducting-of-agricultural-censuses-and-surveys/chapter-11-instruction-manuals/en/> ,
- Fathalla, M 2004, *A practical Guide to Health Researchers*, WHO Regional publication, Eastern Mediterranean Series 30
- Fowler, F.1984, *Survey research methods*, Sage Publication, London.
- Gallessich, J 1985, 'Toward a Meta-theory of Consultation', *The Counseling Psychologist*
- GAO 1992, *Quantitative data analysis: An introduction, Report to program evaluation and methodology Division*, USA.
- Godlee, F & Jefferson, T 2000, *Peer review in Health Science*, BJM Publishing Group.
- Gupta, K 2007, *A practical Guide to need Assessment*, 2nd.ed, Pfeiffer San Francisco,
- Hacett, P 2003, *Training practice*, Chartered Institute of Personal Development, London, UK.

- Hames, I 2003, *Peer review and manuscript management in Scientific Journal*, Malden, USA, Blackwell publishing.
- Hand, D & Everitte, B 1987, *The Statistical Consultant in Action*, Cambridge University, Cambridge, UK.
- Harvey, L 2004–9, ‘Analytic Quality Glossary, Quality Research International’, viewed on 24/09/2011 <http://www.qualityresearchinternational.com/glossary/>
- Haitovsky, Y 1968, ‘Missing data in regression analyses, *Journal of the Royal Statistical Society, B30*, 67-82.
- Health link worldwide.2003, *Exploring the Training Process*, viewed on December 10 2011, <http://www.healthlink.org.uk/PDFs/training.pdf>
- Heckman, James 1979, ‘Sample Selection Bias as a Specification Error’ *Econometrica* 47:153–161.
- ILRI 2011, Independent impact assessment of the World Bank funded Nigeria Avian influenza control and Human pandemic preparedness and response project (NAICP), Report submitted to project management unit.
- JKUAT 2010, Guide to ARM3109: Research project, Kenya.
- Kabiru, C, Izugbara ,C , Wambugu, S and Ezeh ,A 2010, ‘Capacity development for health research in Africa: experiences managing the African Doctoral Dissertation Research Fellowship Program’, *Health Research Policy and Systems*, Viewed on July 20 -2011, <http://www.health-policy-systems.com/content/8/1/21>
- Kiggundu , R 2006, Innovations systems and development, PhD thesis United Nations University , Maastricht
- Kuzon,W ,Urbanek ,M McCabe, S 1997, ‘The seven deadly sins of statistical analyses, *Ann Plast Surg. Pub Med PMID*.
- Liwins, A and Silver C, *Choosing a CAQDAS Package: A working paper*, Viewed on 27 December 2010, <http://caqdas.soc.surrey.ac.uk/Resources/QUICworkingpapers.html>.
- Mark Ware Consulting 2008, *Peer review in scholarly journals, Perspective of the scholarly community, an international study*, Viewed on 10 December 2010, <http://mrkwr.wordpress.com/articles/>
- Miller, S 2009, *Statistical practices: The seven deadly sin*, viewed on June 24 2011, <http://www.tandfonline.com/doi/abs/10.1076/chin.9.3.221.16455?journalCode=ncny20#prview>

- Monash University 2011, *Data management planning*: Australia, viewed on July 1 2011, <http://www.researchdata.monash.edu.au/guidelines/planning.html>
- Muller, H 2010, Statistical Consulting: Lecture note distributed in the department, SIA: 401, UC Davis, winter 2010.
- Muraya, P, Garlick, C & Coe, R 2002, *Research data management: study guide*, World Agro Forestry Center, ICRAF, Kenya.
- Nelson, L & Beedy, D 2008, 'Research Team Training: Moving Beyond the Job Description', *Applied Nursing Research*, Vol.121, Issue 3, PP.159-164.
- ORI 2011, *Responsible Conduct of Research, visit the Office of Research Integrity's*, viewed on 20 /08/2011 <http://ori.dhhs.gov>
- Pallant, J 2007, *SPSS survival Manual*, New York, USA.
- Patel, K, Muir-Leresche, K, Coe, R and Hainsworth, SD. (Eds.) 2004, *The Green Book: A Guide to Effective Graduate Research in African Agriculture, Environment, and Rural Development*. The African Crop Science Society, Kampala, Uganda.
- Pautasso,C and Pautasso ,M 2010, 'Peer reviewing Interdisciplinary papers', *European Reviewer*, Vol 18.No.2 ,227-137,Academic Eurpoea.
- POST 2002, 'Peer Review', *Post Notes*, September, Number 182, [www.parliament.uk/post/home.htm](http://www.parliament.uk/post/home.htm)
- Puskur R 2010, *Innovation in livestock systems* , ILRI, viewed on 11/11/2010  
<http://www.ilri.org/InnovationInLivestockSystems/>
- Ramos, M, Morneo, M, Valdes, B &Catena, A 2008, 'Criteria for the peer review process for publication for experimental and Quasi experimental research in psychology: A guide for creating research paper', *International Journal of clinical and health Psychology*, vol.8 no.3 ,751-764.
- Schafer, L 1997, *Analysis of Incomplete Multivariate Data* , *Monographs on Statistics and Applied Probability* 72, Boca Raton, FL: Chapman & Hall/CRC.
- Schoenbatch, V and Rosamond, W 2000, *Understanding the fundamentals of Epidemiology: an evolving text*, University of North Carolina, USA.
- Schleicher, A and Satio, M 2005, *Data Preparation and Management: Module 10, Quantitative research Methods in Educational planning*, UNESCO, International Institute for educational planning.
- Seals, D & Tanka, H 2009, 'Manuscript peer review: A helpful check list for students and Novice referees', *Advance in Psychology Education*, Vol.23, pp.52-58
- Seglen, P 1997, 'Why the impact factor of journals should not be used for evaluating research: Education and Debate, *BJM*, VOL314.

- Smith, R 2006, 'Peer review: a flawed process at the heart of science and journal', *JRSM*, vol.99 no.4, pp 178-182, viewed on 2010, November 23, [www.ncbi.nih.gov/pmc/article/PMC1420789/](http://www.ncbi.nih.gov/pmc/article/PMC1420789/)
- Strickland, H 1996, 'The nature of statistical consulting', *The Statistical Consultant*, Vol.13, pp .2-5.
- Tavakoli, A , Jackson, K, Moneyham, L, Phillips, K, Murdagh, C,and Medin, G, 'Data Management Strategies', *SESUG proceedings*, viewed on 10/08/2011 ,<http://www.sesug.org>
- Tefeera., D & Altbach, P 2004, *African higher education: Challenges for the 21st century*, Center for International Higher Education', Campion Hall, Boston College, Chestnut Hill,MA 02467
- Trochim, M 2006, *The Research Methods Knowledge Base*, 2nd Edition. Viewed on 23/07/2011, <http://www.socialresearchmethods.net/kb/>.
- UNDP 2006, 'United Nations Development Programme: Capacity Development Group Knowledge Services and Learning', *Working paper #3*, United Nation.
- Wayman, J 2003, 'Multiple Imputation For Missing Data: What Is It And How Can I Use It?' *Paper presented at the 2003 Annual Meeting of the American Educational Research Association*, Chicago, IL.
- Wiles, R , Durrant, G , Broe, S & Powell, J 2005, *Assessment of Needs for Training in Research Methods in the UK Social Science Community*, ESRC National Centre for Research Methods, University of Southampton.
- World Bank 2008, *Using training to build capacity for development: Independent evaluation group*, USA Washington Dc.

## 10. List of Appendices

### **Appendix 1: Review of proposal comments to principal investigator**

Mail To: Innovation in livestock systems

Subject: Review of your proposal

**Title: Household Livelihood and socio organizational impact of animal health intervention study.**

I have tried to review your protocol and I really appreciate the idea and the methodology proposed. Generally the protocol is good. However, I am suggesting you to consider the following points in each section of your protocol for further improvements.

#### ***A. Title***

The title is concise, descriptive and specific. It reflects the general purpose of the study and where the study will be undertaken. It is also possible to rephrase the purpose of the study from the title.

However the word ‘household livelihood’ indicates broad and complex issue which can be measured by different indicators such as Economics Security, Food Security, Nutritional Security and etc... Therefore it is preferred to indicate the specific indicator than the broad terms that the study is going to address (which can be the most important variable). In addition, most of the time it is recommended that the title should show the main research objectives of the study. Hence by considering the above two points I suggest you to substitute the word ‘Household Livelihood’ with Household food security.

The other most important issue in the title is ‘socio-organizational’ which indicates the impact of the intervention on the social organizations of the households. Is this issue clearly indicated in the objective and methodology part of the study? How are you going to address this? If not it is good to substitute this with other issue that will be addressed more specifically in the assessment. And finally to make the title catchier, I suggest changing the order of the sentence and start with “Impact of animal Health intervention....”

#### ***B. Background and research problem***

Generally the background and research problem section is organized very well. But I suggest the following points

- In this section the rationale of your proposed study should be clearly indicated including why it is worth doing. Even though it is briefly mentioned at the end of

the 4th paragraphs, I suggest adding more clarification on these issues for reader who is not familiar with the study.

- Since the study is an assessment on mainly the impact of a given intervention on the food security status of households and food security status of households can be affected by different factors, it is better to mention the limitation and scope of the study by considering other confounding factors which can affect food security.
- The review of some other peoples work carried out on the subject is one of the major parts of a good research proposal but this is not reflected in this proposal why? Is that due to the absence of similar work carried out on the subject previously or.

### C. *Objectives and research questions*

- Objectives 3, which says “Develop a better understanding on the factors that explain the effectiveness of Community Based Trypanosomosis & Tsetse Control (CBTTC) intervention designed to enhance household livelihood” is a very good and important objective for such types of study. But when we say developing understanding it looks that those factors are already identified and the research is trying to add some value on their roles. However, if the factors are not properly identified with similar studies carried out by different people previously and if they are not still known very well, the objective will be more informative if it can be modified as, to identify the factors that explain the effectiveness of community based trypanosomes & tsetse control (CBTTC) intervention designed to enhance household livelihood at Gibhe.
- Land use change can be attributed to different factors such as climatic condition, human activities through agriculture, deforestation and reforestation, natural disasters like landslides from earthquakes and soil erosion, technologies, government policies and etc. Therefore it may be very difficult to attribute the land use change in this area only to the project intervention, although the project intervention can be one of the major component of the causes for land use in the area, so it is better to consider all these things and design the appropriate strategy before the study and explain this briefly in the methodology section.
- Research question 1, “what is the effectiveness of CBTTC interventions in enhancing household livelihood as measured by household food security compared to no or other interventions?” indicates the study will compare the effectiveness of the intervention with without intervention or other interventions. But the word **other intervention** is too broad and general while a research question should be specific that should indicate the specific intervention the assessment will use for comparison. Therefore for this study it is better to compare with only the without intervention situation because comparison with other intervention will be very complicated and challenging with limited resources and time.

#### ***D. Methods***

- The organization of the methods section is good.
- In the method section it is mentioned that “ In addition, land use change ( cropping systems) and time series aerial imagery analysis will be undertaken to determine the trends in land use changes overtime as a landscape level and estimate the distribution in space and time of ecosystems in the intervention area, respectively”, Which needs imageries data taken starting from the project intervention or in some time intervals during the project time , but the availability and source of these datasets is not explained very well in the methods section. So I suggest including this information in this section briefly.
- The data analysis part only talks about the inferential and qualitative data analysis and it does not say anything about descriptive analysis which helps to describe the sample households and different variables. So it is advisable to mention the descriptive part also in the data analysis section.
- The statistical package that will be used for data analysis is not mentioned and it is better to mention it here.

#### ***E. Reference***

Mostly ‘et.al’ is applied in text references of the work of four or more authors in **Harvard** style and after using all authors' names and year in the first time of the reference occurs and in subsequent citations or in six or more authors in **APA** styles, but here it is applied for three references such as , Clark N., Smith J., and Hirvonen M., 2007 or Van Mierlo B., Arkesteijn M. and Leeuwis C. (2010). Which reference style is used here?

Almost all the reference cited in the paper is listed in the reference section except Maxwell and Frank Berger, 1992 and Maxwell et.al 2000. Conversely Maxwell, S., and Smith, M. 1992, and Maxwell, S., 2001 are cited in the reference section but they are not mentioned somewhere in the text of the paper.

#### **F. General comments (additions)**

- If possible it is better to include ***Executive summary*** of the protocol for busy readers. It is also the most important part of the entire document, because the executive summary is the sales pitch to convince the funder that your project should be considered for funding.
- Even though it was not common practice to include ***data management plan*** in the research protocol currently it has become one of the evaluation criteria for funding research projects. However, in this protocol there is no any information about the data management plan which describes the procedures and responsibilities to assure data quality, data documentation and archiving, and efficient data processing. Therefore it is more advisable to include brief data management plan in the method section.

## **Appendix 2: Review of report comments to author**

### **Title: The value of innovation system analysis for livestock research and policy: Case study of an international research institute program on Trypanosomosis control in Ethiopia's Ghibe Valley**

The baseline innovation assessment report was reviewed according to the guidelines developed for the review of the research report. Each of the sections in the report was reviewed and, comments and suggestion are provided separately for the different sections. In generally the report was well written and highlights some key lessons that are relevant for livestock research and development organizations.

#### **A. Title**

The title of the study- “The value of innovation system analysis for livestock research and policy: case study of an international research institute program on Trypanosomosis control in Ethiopia's Ghibe Valley” is catchy. However, it does not fully reflect the content of the study if looked at in isolation from the complete report. The title of a study should reflect the main objectives of the study and the reader should be able to guess what is in the body of the text by reading the title. In this case, the main objectives of the study which were “to develop an understanding of factors underlying process of change, draw lessons from the innovation process in Trypanosomosis control in the area and identify factors contributing to the reported success in control of the disease” were not fully reflected in the title. After looking at the title, a reader would expect that the paper will focus on the role/ importance of innovation system analysis for livestock research and policy, but the study was largely an assessment of the innovation capacity of farming households in a specific context (Trypanosomosis control) and area (Ghibe) . So an alternative title which reflected this better would have been preferable.

In addition, the title should have indicated the main variables of the study. However, this is not very obvious in the present title. In my opinion the main variable of the study is the intensity of innovation related activities, and key factor that affects this, the capability of farming households to innovate among others. Finally, the phrase ‘an international research institute program’ does not add much value to the title.

#### **B. Abstract**

The abstract briefly summarizes the content of the study. It describes the problem under investigation, the purpose of the study which is “to assess innovation performance and identify lessons and principles that could be strengthened...” and the methods adopted- “diagnostic qualitative analysis, household level survey, and joint assessment exercise and research workshop.” It also succinctly describes the results and conclusion part, “households enjoyed improved access to research technologies and high quality Trypanosomosis control service.” The conclusion emphasizes that the finding could benefit

many other researchers, policy and development agencies seeking to strengthen agricultural innovation in poor areas of developing countries. The language used is very simple and concise.

### **C. Introduction/literature review**

The author has provided adequate background about the study sites, previous studies and result obtained from the intervention. All previous researches related to this study are clearly described in this part. The importance of the problem was clearly explored as it is described, to answer “how and why the mentioned positive change processes had occurred, what capabilities needed to further strengthen and, what principles and lessons could be scaled out elsewhere.” However, no specific research hypothesis was formulated or stated in this paper because the study was conducted mainly to explore the experience and lessons in relation to innovation in the area and to understand the circumstances surrounding the positive change processes. The approaches used to understand the process of change and the lessons learned are briefly stated.

### **D. Methods**

The methodology section clearly describes how the study was conducted and the specific procedures or approaches used. The author used a combination of methods which included desk review, key informant interviews, PRA, household survey and research workshops. Multiple and complementary data collection methods were used which helped to obtain in depth information. The methods allowed triangulation of the findings of the study leading to a more thorough and meaningful interpretation of quantitative household survey data and to substantiate the qualitative findings collected from PRA and key informant interviews.

The methods used and their combination are very appropriate for this type of study and allow obtaining comprehensive and complete information about the subject under investigation. The organization of the description of methods was good and it was divided appropriately into subsections which included a description of the sample design. Most of the sections were described in detail such that it is easy for a reader to understand how the study was conducted. However, this section of the report has the following limitations:

- Sampling method for key informant interviews was not clearly described and the sample size for Key informant interviews and PRA was not clearly indicated. Some pertinent questions which arise in this regard include: who were the potential key informants for this specific study, what was the expected result from the interview, were different groups of people interviewed using the same set of questions?
- The definition of variable (INNOVA) was provided in part three of the paper; however, there is no clear description about the type of proxy indicators used to estimate INNOVA.

- In addition, the methods part should have described why the researcher used Tobit model: this would help the reader understand the appropriateness of the model for that specific data set.

### **E. Results**

The results section of the report includes both the descriptive and inferential statistics and appears to be comprehensive and complete. However the organization of this section can be improved to make it more readable and keep the reader captive. This section contains a subsection called “the data” which mainly talks about the descriptive part and definition of variables; another section called the ‘analytical results’ talks about the regression model used. The result of the analysis can support the justification which would be used to make the conclusion.

### **F. Discussion**

The discussion part of the study was not clearly and separately indicated in the manuscript. Part of the discussion was included under the result section. Even though the interpretation for the regression output was presented, the link to the results and the objectives of the study was not clear in the discussion part. The interpretation of the result did not consider the sources of potential bias and other limitations due to precision of measurements. For instance the data was collected from two areas in which ILRI was actively involved in research and other areas in which there was no research intervention, which is called intervention or exposure bias. Therefore the interpretation should have considered the specificity of the data and the effect of ILRI intervention in the 5 Kebeles. The external validity of the result is properly explained, while the internal validity was not explained very well and to the extent that it was warranted.

### **G. Conclusion**

The conclusion part of the report was presented under the fourth section named “What can this kind of diagnosis are used for and how?” and it provides a summary of the main findings and their implication for use. However the limitation of the study and potential further research in the same or similar areas was not clearly indicated in this section.

### **H. References**

The reference system follows APA style which includes both the author and date of publication. The in text citations cause minimal disturbance to the flow of the writing as most of them are cited at the beginning or end of the sentences. The references seem appropriate to the subject of the study and the citations are in the correct format. However, all the references cited in the text, (such as ILRI, 2007; Mytelka, 2000 and Kiggundu, 2010) were not listed in the reference section and; there are also references listed (Mulatu. W 2005; ILRI, 2006 and World Bank 2006b) which were not mentioned anywhere in the text.

## **I. General comments;**

### **1. General comment on the Tobit model used**

Before estimating the level of innovation related activity and identifying the key factors which affect the capability of farming household to innovate there should be clear information whether that farming households are innovative or not. One cannot expect all the farming households to involve in innovative activities. Some farming households may not be involved in innovation activity at all, while some may involve in limited activities. And therefore an index of innovation is meaningful only for those households which are totally or partially involved in the expected innovative activities. There could be clear theoretical reasoning why some households do not innovate at all or others innovate only to a limited extent. It is also possible to have a theorize why some households do not innovate at all - it could be due to lack of motivation to innovate or lack of the required knowledge or resources or incentives.

The use of the Tobit model in such a situation, where there can be sample selection bias, has some notable limitations that can be remedied with the use of another model.

- Firstly, in the Tobit model the same set of variables and coefficients determine both the probability that the farmer is innovative or not (to be involved in innovative activity) and the value of the intensity of innovative related activity.
- Secondly, the model does not allow a full theoretical explanation of why the observations that are censored or why those farming households who were not innovative did not like to participate in the innovative activities. But it is possible to see why this may happen and can be demonstrated with sample selection models which can address these shortcomings by modifying the likelihood function.

With sample selection model a different set of variables and coefficients can be used to determine the probability of farming household to involve in innovative activity and the value of index of innovation, given that the farming household was involved in innovative activity. This allows us to take account of the censoring process because selection and outcome are not totally independent in this case; however, these variables may overlap to a point or may be completely different.

In the framework of the sample selection model, you could specify one equation for whether or not a household is innovative, and a different equation for the probability of intensity of innovation. In fact, as Heckman demonstrated, if the processes are related, estimating a model of intensity of innovation without first estimating an equation of whether or not the household was innovative would lead to biased results. Therefore the use of sample selection model should have been tried to avoid the selection biases

### **2. General comment on the organization of the report**

The organization of the report is generally good. But it is possible to organize this report better to help the readers understand the report easily and fully explain the results. Each

section of the report should have clear and coherent content. For instance the methods section and the results section were not separate and resulted in a pot pourri of ideas. In addition, there was no clear discussion and conclusion part to link the results to the objectives of the study.

### **Appendix 3: Review of report guideline questions**

To standardize the review process the following guiding question will be used for each of the main content of the report.

#### **A. Title**

- Is it fully explanatory when it stands alone?
- Does it summarize the main idea of the study?
- Does it indicate the main variables under investigation and their relationships?
- Are all the words in the title useful?

#### **B. Abstract**

The abstract should briefly summarize the content of the study to help readers to survey quickly the main content of the article.

- Does it correctly reflect the purpose and content of the study? (Does it describe the problem under investigation, the purpose of the study, the methods, result and conclusion concisely)?
- Is the language simple and concise?

#### **C. Introduction/literature review**

The introduction should present the problem under investigation clearly.

- Does the author have researchable questions or hypothesis and are they clearly stated?
- Does it explore the importance of the problem?
- Does it explain previous relevant literatures?
- Is the approach to solve the problem clearly stated?

#### **D. Methods**

The method should briefly describe how the study was conducted. It should be stated in detail so that the reader can evaluate its appropriateness and any interested investigator can replicate it without any difficulty.

- Have the author selected appropriate methodology to answer the research question?
- Is it properly organized? Does it contain the appropriate sub sections?
- Are the research participants identified appropriately?
- Are the sampling procedure, sample size, power and precision of the study described?
- Does the design, measurement and covariate specified clearly?
- Is the research design appropriate? What about the sampling technique?
- Are the proposed methods of data analysis appropriate for the research question and objectives?

### **E. Results**

The result should summaries the output of the analysis in table and graphical formats. The result part of the manuscript will be evaluated based on the following guiding questions

- Are the data reported sufficient enough to justify the conclusion?
- Is it complete (covers all the descriptive and inferential), accurate and insightful?

### **F. Discussion**

The discussion part should evaluate and interprets the result of the study and provides their implication according to the hypothesis and objectives of the study. The discussion part of the manuscript will be evaluated based on the following questions,

- Does it emphasize both on the theoretical and practical consequence of the result?
- Does the interpretation of the result takes in to account source of potential bias, imprecision measures and other limitations of the study?
- Are the external and internal validity explained properly?
- Are there any possible errors, misinterpretation and over interpretation?

### **G. Conclusion**

The conclusion part of any research report should not be a mere repetition of the information presented but it should indicate the summary of the main important findings and their implication for use. It could also indicate the limitation of the study and where further research would be useful if necessary.

### **H. References**

The reference section will be reviewed according the following criteria

- Are all the reference cited in the paper listed in the reference section?
- Are all the reference cited in the reference section mentioned somewhere in the text of the paper?
- Does the reference seem appropriate to the subject of the study? How?
- Are the citations of the reference in the correct format?

### **I. General comments**

The report will also be assessed base on the following questions for general comments on the quality of the article.

- Is there any missing information?
- How is originality, the quality and the importance of the report?
- Indicate areas in which you are not qualified or comfortable to critique?

## **Appendix 4: Review of proposal guideline questions**

### **Review of proposal guidelines**

To standardize the review process, the following guiding question will be used to assess each of the main components of the proposal.

#### **A. Title**

The title of the proposal should be SMART. It will be evaluated with this criteria and whether it is possible or not to rephrase the purpose and subject matter of the study by reading the title before reading the other sections of the proposal. Since the title should reflect the main ideas of the introduction, it will also be evaluated whether the main idea of the introduction is reflected in the title or not. The important information missed in the title and the possible ways that the author can revise to make it more informative and clear will be suggested. If the title of the proposal does not need any revision, appropriate explanation will be given why it does not need and how much is it informative.

A good title should have the following properties:

- The title needs to be very specific in nature. So the title need to be very clear and it should provide an idea to the readers about what to expect from the research paper.
- In spite of being specific it should also have the expressive power to show the full gamut of the research study in those few words.
- It should tell the total nature of the subject.
- It needs to be very definite and clear.
- The title needs to be attractive and interesting enough to catch the attention of the readers. (<http://www.articlesbase.com/writing-articles/research-paper-title-957927.html#ixzz1HEtAOkSO>)

#### **B. Introduction**

The introduction part of the proposal will be reviewed according the following guiding questions.

- Does the introduction give adequate background information, definition of terms and citation necessary to understand the statement of the problem, the hypothesis and the research questions?
- Is it possible to clearly and easily identify the statement of the problem from the introduction?
- How is the introduction organized?

### **C. Materials and method**

- How is the method organized? Is the organization appropriate? If not what type of organization make the methods more understandable?
- Does the method provide enough information so that other researcher can repeat it without any additional information?
- Will the statistical analysis be used provide the expected information from the data if not why? Which is the best possible way?

### **D. Result**

Does the author indicate the type of observation plan to make? Does the type and nature of data clearly indicated?

### **E. Reference**

The reference section will be reviewed according the following guiding questions

- Are all the reference cited in the paper listed in the reference section?
- Are all the reference cited in the reference section mentioned somewhere in the text of the paper?
- Does the reference seem appropriate to the subject of the study? How?
- Are the citations of the reference in the correct format?

### **F. The overall organization of the proposal**

- Is the overall organization of the proposal in scientific format?
- Does the author present his or her idea in logical format?

After reviewing, the proposal based on the above guiding questions, general and summarized comment will be given to the principal investigator and the response of the investigator and lesson learned from the review will be documented.

## **Appendix 5: Consultancy report summary format**

**Content of Consultancy report:** The consultancy report will contain the following main points (Muller, 2010).

### **A. Problem to be addressed**

The background, the main question to be addressed from client's perspective, any related previous studies by same investigator or other investigators in the same field and the researcher will be asked the main challenge that he/ she has faced and need assistance.

### **B. Design**

Number of independent cases or units (power calculation), type of study (experimental or observational study), sampling procedure, type of measurement (repeated measurements), the representativeness of the sample for the population, the randomization, presence of special designs (ANOVA, Split plot, Case-Control, Sequential, Equivalence study etc.) and researcher suggestion about the presence of any hidden biases or dependencies in the data or Measurements.

### **C. Measurements and Variables**

Number and type of measurements made per subject/unit, the types of variables, physical units and statistical distributions, Variation of measurements, measurement process, possible transformations and predictor and response variables

### **D. Preliminary Analysis**

Kind of analysis the researcher has done before: exploratory data analysis, basic plots, and data cleaning. Any specific procedures in researcher mind? Are they adequate? Have similar data been analyzed by the researcher or others?

### **E. Recommendations**

Data checking and assessing data quality Data exploration pre-processing: Outlier removal, transformations, data cleaning, Statistical model, hypothesis test, confidence regions, Diagnostics and residual analysis, Computational implementation and software and possible interpretation of results.

### **F. Conclusions**

The conclusion section will include additional consulting schedule if applicable, who will do what? And computing issues

## **Appendix 6: Data management plan**

### **Data management plan**

The data management task will be accomplished for Gibe baseline study data, which was undertaken in four different phases from 1997 to 1998, for Positive Deviance in the use of improved livestock Feed Resource in Developing countries and Household livelihood and socio-organizational impact of animal health intervention study, which will be undertaken in the first quarter of 2011 and analysis of projects/networks working on land and water management in the Nile basin of Ethiopia.

Generally, the data management task includes the following studies.

#### **Task 1**

**Title of the study:** Household livelihood and socio-organizational impact of animal health intervention study

#### **Objective of the study:**

- Generate empirical evidence on the effectiveness of animal health intervention to enhance household livelihood as measured by household food security
- Map the changes in landscapes as a result of the intervention
- Develop a better understanding on the factors that explain the effectiveness of community based trypanosomosis & tsetse control (CBTTC) intervention designed to enhance household livelihood

The data management task for this study will include the following main activities

#### **A.Participation in designing of data collection instrument, data collection, and data entry**

- Designing of the data collection instrument will be carried out together with the principal investigator, the project team leader, and other scientists.
- During data collection, in additions to any assignment given by the team, supervision of the data collection process and data checking activities will be accomplished.
- Inconsistencies and errors in each questionnaire will be checked every evening with enumerators.
- In additions to the above tasks, data entry format and checking procedure will be prepared to the study data using CSPro statistical package.

## **B.Data organization**

The data will be organized in appropriate format, which helps for further analysis and data archiving. This will include defining and entering the appropriate variables name for column headings and formatting the data sheet with appropriate styles.

## **C. Data storage and access**

Preparation of Meta data

Meta data will be prepared will be prepared by using the questionnaire, project proposal and by consulting the project team. The Meta data will include the following information;

1. Title. The name of the data set and the project
2. Authors. Names of principal investigators and others, with their contact addresses.
3. Data set overview. General introduction to the data set, location of where the data was collected, when it was collected, and any references.
4. Instrument description. Brief description of data capture instrument with references
5. Data collection and processing. Description of how data were collected, computed values, and quality control procedures
6. Data format. Structure of data files and naming conventions, codes (if used), data format and layout, version number and date.

## **D.Data files description**

Data file description will be prepared for each type of data sets. The description of the data file will include the following information,

- Format of the file
- Number of records in each file
- Number of variable in each file
- If there is any derived variable, the calculation and means of derivation

## **E.Organizing the data for archival**

The data will be organized with appropriate format for archival according to the regulation of the organization. Generally, the archival data will include the following information

- Data Files and their description
- Data documentation – meta-data
- Sampling scheme
- Project protocols/concept notes
- Justification documents
- Data collection instruments/questionnaires

- Analysis syntax files

Finally, at least one copy of the completed, error-checked of the original datasheets and a digital copy of the data with an appropriate organization or entity information will be submitted to livestock innovation system research team.

From the data management task the role of different participant, lessons learned, description of the data management problem, the approach used to solve it and an assessment of how the approach was successful will be documented for future reference.

## **Task 2**

**Title of the study:** Positive deviance in the use of improved livestock feed resource in developing countries.

**Status of the study:** Completed but not reported

### **Objectives of the study:**

The general objectives of the study is to identify successful cases of forage innovation in the country or area with focus on characterizing the process , factors and conditions underlining their success and drawing lessons and principles that could be scaled out through ongoing and new forage promotion initiatives

### **Objective of the data management**

The data is in SPSS format which has 603 cases and 1652 variables. The main objectives of the data management will include the following points;

- To check the data for any possible errors in the data coding and entry activities
- To organizing the data for analysis and archival

### **Major activities**

- 1.Checking variable definition and coding by using the questionnaire.** Before proceeding to any data management and data analysis activity variable definition and coding system of the data will be checked using any other important documents such as code book, audit trial and questionnaire.
- 2.Checking for outliers and other data entry errors:** Since the initial analyses are a continuation of the checking process and should include a first look at summaries of the data. Useful things that will be produce at this stage are:  $\square\square\square$  *extreme values*, in particular the minimum and maximum observations; *box plots*, to

compare groups of data and highlight outliers; *tables* of the data in treatment order.

**3.Organizing the data for analysis:** Hence the other important step in the data organization or management stage often involves calculations to restructure the data into the appropriate form for analysis. This will be performed in the software used for the data entry: The data organization activity will include

- Computing new variables
- Recoding variables in to the same or different variables

During data organization

- A record will be kept of all changes to the data which is going to be the part of the database, and is kept in the audit trail.
- There will be a single “master copy” of the data. This will help to preserve data integrity. The master copy will increase in size as data accrues.

**4.Data storage and access:** Similar to task 1

**5.Data file description:** Similar to task 1

**6.Organization of the data for archival :** similar to task 1

### **Task 3**

**Title of the study:** Ghibe baseline study

**Status of the study:** Completed and reported

**Objectives of the study:**

Objective of the data management task

- To check the data for any possible errors in the data coding and entry activities
- To organizing the data for baseline summary analysis and archival

### **Major activities**

**1.Checking variable definition and coding by using the questionnaire.** Before proceeding to any data management and data analysis activity variable definition and coding system of the data will be checked using any other important documents such as code book, audit trial and questionnaire.

**2.Checking for outliers and other data entry errors:** Since the initial analyses are a continuation of the checking process and should include a first look at summaries

of the data. Useful things that will be produced at this stage are:  $\square\square\square$  *extreme values*, in particular the minimum and maximum observations; *box plots*, to compare groups of data and highlight outliers; *tables* of the data in treatment order.

**3.Organizing the data for analysis:** Hence the other important step in the data organization or management stage often involves calculations to restructure the data into the appropriate form for analysis. This will be performed in the software used for the data entry: The data organization activity will include

- Computing new variables for baseline summary
- Recoding variables into the same or different variables for baseline summary

During data organization

- A record will be kept of all changes to the data which is going to be the part of the database, and is kept in the audit trail.
- There will be a single “master copy” of the data. This will help to preserve data integrity. The master copy will increase in size as data accrues.

**4.Data storage and access:** Similar to task 1

**5.Data file description:** Similar to task 1

**6.Organization of the data for archival :** similar to task 1

#### **Task 4:**

**Title: ANALYSIS OF PROJECTS/NETWORKS WORKING ON LAND AND WATER MANAGEMENT IN THE NILE BASIN OF ETHIOPIA**

**Status of the study:** On going

#### **Objectives of the study**

One of the main objectives of the program is to facilitate an institutional structure for joint learning and exchange between different key players working on land and water management in Ethiopia at the national level.

To understand roles, relationships and interactions between actors, to identify barriers to interactions among actors and forums, and develop ways of addressing those challenges

#### **Objective of the data management task**

- To collect and summarize the data for analysis.

#### **Major activities**

**1.Data collection together with international water management research assistants.**

**2.Organizing the collected data for analysis:** Hence analysis of qualitative data is not an easy task like that of quantitative data organizing the data for analysis in excel is an important task. Therefore the collected data will be organized using excel for further analysis.

**During data organization**

1.A record will be kept of all changes to the data which is going to be the part of the database, and is kept in the audit trail.

2.There will be a single “master copy” of the data. This will help to preserve data integrity. The master copy will increase in size as data accrues.

3.Data storage and access: Similar to task 1

4.Data file description: Similar to task 1

5.Organization of the data for archival: similar to task 1

## **Appendix 7: Data analysis plan**

### **Task 1**

**Title of the study:** Positive deviance in the use of improved livestock Feed resource in Developing countries

**Status of the study: Completed but not reported**

### **Objectives of the study**

The general objectives of the study is to identify successful cases of forage innovation in the country or area with focus on characterizing the process , factors and conditions underlining their success and drawing lessons and principles that could be scaled out through ongoing and new forage promotion initiatives.

### **Analysis objective**

- To determine the factors which affect the intensity of forage innovation

### **Research questions and expected lesson from the question**

- What are the different factors which affect the intensity of Innovation among the positive deviant?
- From this research question different factors and their impact on the intensity of households' innovation will be identified.

### **Main Parts of this analysis**

The appropriate methods of data analysis are determined by the data types and variables of interest, the actual distribution of the variables, and the number of cases. Different analyses of the same dataset may reflect or represent different aspects of the underlying data structure. The data analysis will contain any combination of the following types of data analysis strategies;

#### **A. Exploratory:**

Exploratory data analysis uses numerical and graphical methods to display important features of your data set. Exploratory data analysis helps us to highlight general features of your data to direct future analyses. This type of data analysis will help to clarify what will be expected from the data. It answers the following main questions: What do the data look like? Should outliers be included or excluded in the analyses? Do the data need cleaning for consistency? How much missing data is there and how should it be handled? What do the distributions look like for key variables? It also pinpoints problem areas in the data.

**It also helps to identify the distribution of the data, outliers and missing value:**

What's the "shape" of the data? Where do most of the values lie? Are they clumped around a central value, and if so, are there roughly as many above this value as below it? We look at the distribution for each variable to determine which analyses would be most appropriate. Sometimes it is necessary to examine distributions of data partitioned by other key variables.

In a survey, missing values correspond to skipped questions or unendorsed options. A discussion between the student and the research team will take place in determining how missing values should be handled. In some cases, missing values might be perfectly normal. However, in some cases missing values for important variables might exclude a record from certain analyses. Sometimes it is appropriate to place normalized values in place of missing values.

**Outliers:** "Unusually" large or small values that are dramatically separated from the rest of the data might be: 'out-of-range' or physically impossible values that resulted from entry or processing error. Merely "weird" values might represent entry error

**B. Descriptive**

The most common type of data analysis which will summarize the findings and describe the sample. The descriptive statistics will tell us how the data look, and what the relationships are between the different variables in the data set. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. It also helps to present quantitative descriptions in a manageable form. Each descriptive statistic reduces lots of data into a simpler summary. With different limitations, descriptive statistics provide a powerful summary that may enable comparisons across households or other units.

**C. Inferential:**

The Inferential statistics allow us to draw conclusions about the larger population from which the sample is drawn. It mainly helps to determine the factors which affect the intensity of innovation.

**Analysis technique and its appropriateness**

The main purpose of the analysis will be to explain the intensity of innovation as a function of different explanatory variable. But the intensity of innovation only observed for those households that are successful in innovation and those household who are unsuccessful have zero innovation index. There could be a very clear theoretical reason why these households do not become successful after starting to adopt the technology. In

additions zero may not be the lowest value but my understanding on this is that perhaps censoring occurs at some point where the households are unsuccessful during participation or not. There could be a theory that specifies these unsuccessful households who has below zero and zero innovation index primarily due to with lack of skill or lack of the required knowledge and other factors. In the framework of the sample selection model, you could specify one equation for whether or not a household is successful or not and a different equation for the intensity of innovation. Because only those households who are successful have try to intensify their innovation activities. In fact, as Heckman demonstrated, if the processes are related, estimating a model of intensity of innovation without first estimating an equation of whether or not the household was successful or not would lead to biased results.

Hence by considering the relation between the success of the households and intensity of innovation a Heckman two step sample selection model will be used to avoid selection biases by considering both the successful and unsuccessful households. The model will have two parts; the first selection model, which is the probit model which indicate the probability of success in forage innovation and the second outcome model, which indicates the factors which determine the intensity of innovation among positive deviant.

Assumption of the model and its equation

The Heckman model also uses the following assumptions:

1. That is both error terms are normally distributed with mean 0, variances as indicated and the error terms are correlated where  $\rho_{\varepsilon u}$  indicates the correlation coefficient.  
 $(\varepsilon, u) \sim N(0, 0, \sigma^2_{\varepsilon}, \sigma^2_u, \rho_{\varepsilon u})$
2. The error terms are independent of both sets of explanatory variables.  
 $(\varepsilon, u)$  is independent of  $\mathbf{X}$  and  $\mathbf{Z}$   $\text{Var}(u) = \sigma^2_u = 1$

Heckman's correction involve normality assumption, provides a test for sample selection bias and formula for bias corrected model. The Heckman correction takes place in two stages (Heckman, 1997). First, the model will be formulated based on economic theory using Probit regression model and the estimated model gives the probability that intense innovation are undertaken (the dependent variable). The second stage corrects the purposive sampling by incorporating a transformation of these predicted individual probabilities as an additional explanatory variable. The data will be analyzed using R statistical package.

The model can be given as follows,

**Prob(D=1/Z)= $\phi(Z\gamma)$ .....(1)**

Where indicates the success ( $D = 1$  if the respondent is successful in forage technology and  $D = 0$  otherwise),  $Z$  is a vector of explanatory variables,  $\gamma$  is a vector of unknown parameters, and  $\Phi$  is the cumulative distribution function.

$$f^* = X\beta + u \dots \dots \dots (2)$$

Where  $f^*$  is intensity of innovative activity which is not observed if the individual is not participated in forage innovation activities.

Then the conditional expectation intensity of innovation activity can be given as,

$$E[f|X, D=1] = X\beta + \rho\sigma_u\lambda(Z\gamma) \dots \dots \dots (3)$$

$X$  = Independent variable

Where  $\rho$  = is the correlation between unobserved determinants of propensity to become successful and unobserved determinants of intensity of innovation  $u$ ,  $\sigma_u$  = is the standard deviation of  $u$ , and

$\lambda$  = is the inverse Mills ratio evaluated at  $Z\gamma$ .

The equation demonstrates Heckman's insight that sample selection can be viewed as a form of omitted variable bias as conditional on both  $X$  and on  $\lambda$ , as if the sample is randomly selected. The intensity of innovation equation can be estimated by replacing  $\gamma$  with Probit estimates from the first stage, constructing the  $\lambda$  term, and including it as an additional explanatory variable in linear regression estimation of the intensity of innovation equation. Since  $\sigma_u > 0$ , the coefficient on  $\lambda$  can only be zero if  $\rho = 0$ , so testing the null hypothesis that the coefficient on  $\lambda$  is zero is equivalent to testing for sample selectivity.

**Task 2**

**Title of the study: Ghibe baseline innovation study.**

**Status of the study: Completed and reported**

**Objective of the study**

- Develop understanding the factor underling process of change and draw lesson from the innovation process in Trypanosomosis control in the area and identify factors contributing the recorded success in control of the disease
- Generate baseline data and material that could be used to monitor and learn from process of change.

- Identify and validate the technological, organizational, policy and institutional factors and capabilities that would have to be addressed and strengthened in future scaling up and scaling out activities.

### **Analysis objective**

- To identify the factors which affect the probability of observing improvement in Trypanomosis management?

### **Main parts of the data analysis**

**Inferential:** The main inferential analysis in this study will concentrate on the model logit model which will be used to identify the factors which affect the probability of observing improvement in Trypanomosis management.

### **Task 3**

**Title of the study: Household survey for Ghibe vally( Ghibe baseline study)**

**Status of the study: Completed and reported**

### **Objective the study**

**Analysis objective:** To summarize baseline indicators to Household livelihood and impact of animal health intervention study.

### **Main part of analysis**

The main analysis in this study is mainly descriptive in nature and the analytical frame work will includes the following main components.

#### **i). Analysis of household demographics, composition and, living conditions**

The information collected on these indicators during the baseline study will be analyzed and presented under the following heads in percentage terms.

- Extent of male-and-female headed households
- Education levels of household heads and members

#### **ii). Analysis of household ownership of productive assets**

It includes information on the size and type of agricultural land holding, status of land ownership, crop variety planted, crop yield sale, crop input, crop rotation livestock production. The results of the analysis will be provided under the following heads in terms of percentages

- Agricultural land-holding details and crop diversity of annual staples
- Household production of livestock

### **iii). Analysis of household food security**

Since the concept of household food, security is multi-dimensional which has complex interactions with various indicators it is difficult to capture using any single/specific indicators. The dynamic interactions between different components of food security at the household level, and the ways in which people of Gibe gain access to food and income will be investigated and analyzed using the following different perspectives:

- Household food in/security as analyzed from households' own production
- Household food in/security according to household food insecurity access scale
- Households' income diversity - access to different sources of income
- Household food utilization – food consumption pattern, type of food consumed

**Appendix 8: Household Livelihood and socio organizational impact of animal health interventions study first draft questionnaire**

Status of questionnaire

Impact Assessment Survey Questionnaire

---

*Household livelihood and socio-organizational impact of animal health intervention study*

**CONFIDENTIAL**

All the information collected in this questionnaire will be completely confidential and will not be used to assess the respondents' personal benefits or challenges.

Identification

		CODES	
District	PA		
Household Number			

---

## Section 1: Household Information

---

### 1.1 Household Head

1.1.1 Sex ----- Male  Female

1.1.2 Age -----

1.1.3 Ethnicity: 1. Oromo 2. Amhara 3. Guraghe 4. Others

1.1.4 Religion: 1. Islam 2. Orthodox 3. Catholic 4. Protestant 5. Others

1.1.5 What is the level of education of the household head? 1. Illiterate 2. Read and Write 3. 1- 4 grade 4. 5-8 grade 5. 9-12 grade 6. Vocational 7. Others

1.1.6 Have you always lived in this part of the country?  
1. Yes 2. No

### 1.2 Household Characteristics

1.2.1 Household size .....

1.2.2 Number of Economically active people in the household  
.....


### 1.2.3 Information on HH- members

NO	Name (Permanent HH Members)	Sex	Age	Relation to HH-H	Level of education

Code: Sex 1=Male 2=Female

•Relationship 1= Self 2=Husband 3=Wife 4=Daughter 5= Son 6 = Grandchild  
7 = Parent 8=Labourer 9=Sister 10=Brother 11=Step child 12 = others

•Level of education 1. Illiterate 2. Read and Write 3. 1- 4 grade 4. 5-8 grade 5. 9-12 grade 6. Vocational 7. Other

## Section 2: Resource Endowment and use

---

### 2.1. Land resources

2.1.1 **Do you have your own land?** 1. Yes      2. No

2.1.2 **If no to Q.2.1.1, source of land for cultivation is** \_\_\_\_\_

1. Rented in   2. Share cropped   3. Received as a gift   4. Others (specify)\_\_\_\_\_

2.1.3 **What is the total size of your land?** \_\_\_\_\_in hectare or local units

2.1.4 **What is the total area of land you cultivated in 2009/2010?** \_\_\_\_\_

1. Owned \_\_\_\_\_ 2. Rented in \_\_\_\_\_ 3. Share cropped \_\_\_\_\_ 3. Received as a gift \_\_\_\_\_  
4. Others (specify) \_\_\_\_\_

2.1.5 **What is the total area of land you cultivated before 24 years of the project?**

2.1.6. Indicate the amount of the following resources you have before 24 years and now

Type of resource	Before 24 years	Now	Remark
<b>Total area converted from cultivated land to permanent grassland</b>			
<b>Total permanent grassland cultivated</b>			
<b>Total Forest planting</b>			
<b>Total Forest removal</b>			
<b>Number ponds Digging</b>			
<b>Total area for Hedgerow planting</b>			
<b>Total area for Hedgerow removal</b>			

**2.2. Livestock**

2.2.1 Do you own livestock? 1. Yes 2. No

2.2.2 If yes, indicate type and number of livestock owned currently and before 24 years.

Type of Livestock		Before 24 years	Now(Current)	Remark
Cattle	Oxen			
	Cows			
	Heifer			
	Bull			
	Calves			
Sheep and goat	Sheep			
	Goat			
Equines	Horses			
	Mules			
	Donkeys			
Poultry	Chicken			
Beekeeping	Beehives			

2.2.3. If you don't have enough oxen, how do you get additional oxen you need?

1. Hire from someone 2. Coupling with other farmer 3. Borrow from friends 4. By contributing labour to a person who has oxen 5. Others (specify) \_\_\_\_\_

2.2.4 Do you have enough feed for your animals? 1. Yes 2. No

2.2.5 **If yes what are the sources?** \_\_\_\_\_ (multiple answers possible)

1. Own grazing land 2. Communal grazing land 3. Crop by-products 4. Others (specify)

2.2.6 **If no how do you cover the deficit?** 1. Limit number of livestock 2. Purchase of additional fodder from other source 3. Leave the problem as it is 4. Others

2.2.7. Is **animal disease a problem to you now?** \_\_\_\_\_ 1. Yes 2. No 3. Does not concern

2.2.8 If yes to question number 2.2.7, **what is the type of disease?** -----  
-----

2.2.9. If yes to Q. no 2.2.7, **do you get enough drugs to treat your animals?** \_\_\_\_ 1. Yes 2. No

2.2.10. **What are the most important drug your are using?** .....

2.2.11. **Did any of your animals receive pour on this year?** 1. Yes 2. No

2.2.12. If your answer is Yes, How many animals from each category receive the pour on?

	Calves	Heifers	Cows	Oxen	Bulls	Others
January						
February						
March						
April						
May						
June						
July						
August						
September						
October						
November						
December						

**SECTION3: FOOD SECURITY**

---

**3.1 What crops does your household grow for home consumption?** For each crop, indicate its average acreage and its total production for the past 3 (2008-2010) years

Crop grown	Area	Total production

**3.2 What are the types of input used for different crops?**

Type of input	Total amount used	Total cost

**3.3. How has production of food crops in your household been in the years 2009 and 2010 compared to before 24 years: would you say it has increased, decreased, or remained the same?**

1= Increased By how much..... (Bags/tins/kg) [Go to 3.4]

2= Decreased By how much..... (Bags/tins/kg)) [Go to 3.5]

3= Remained the same [Go to 3. 6]

4= Do not know [GO TO 3.6]

**3.4 What explains the increase in the production of food crops in your household in the 2009 and 2010 years?** [*PROBE TO GET AN EXHAUSTIVE LIST OF FACTORS CONTRIBUTING TO THE INCREASE*]

**3.5 What explains the decrease in the production of food crops in your household in the 2009 and 2010 years?** [*PROBE TO GET AN EXHAUSTIVE LIST OF FACTORS CONTRIBUTING TO THE DECREASE*]

**3.6 During the 2009 and 2010 period, would you say the number of food shortage months in your household across the year, has increased, decreased, or has remained the same compared to the period before the intervention?**

1= Increased By how many months..... (Include fractions of months) (GO TO 3.7)

2= Decreased By how many months?..... (Include fractions of months) (GO TO 3.8)

3= Remained the same (GO TO 3.8)

9= Do not know

**3.7 Can you please explain how the increase has in the number of months of food shortage in your household during 2009 and 2010 in comparison to the period before 24 years?** [*PROBE TO GET AN EXHAUSTIVE LIST OF FACTORS CONTRIBUTING TO THE INCREASE*]

**3.8 Can you please explain how has the decrease in the number of months of food shortage in your household during 2009 and 2010 in comparison to the period before 24 years?** [*PROBE TO GET AN EXHAUSTIVE LIST OF FACTORS CONTRIBUTING TO THE DECREASE*]

**3.9 During the 2009 and 2010 period, would you say the number of meals per day in your household has increased, decreased, or remained the same when compared the period before 24 years?**

1= Increased From .....times/day to.....times/day

2= Decreased From .....times/day to.....times/day

3= Remained the same

**3.10 During this 2009 to 2010 period, has your household's coping strategies for food deficiency changed or remained the same?** 1= Changed 2= Remained the same

**3.11 What are the new strategies for coping with food deficiency that your household now employs? [PROBE TO GET COMPLETE LIST]**

**3.12 Indicate the type and amount of food consumed in the last 14 days in your home (Questions to be forwarded for house wife)**

S.N	Food items	Unit	purchase d	Own production	From relatives	From any other source	Total consumed
<b>1</b>	Maize						
	Bread						
	Injera						
	Roasted						
	boiled						
	porridge						
<b>2</b>	Wheat						
	bread						
	boiled						
	porridge						
<b>3</b>	Barley						
	bread						
	injera						
<b>4</b>	Teff						
	Injera						
	bread						
<b>5</b>	Sorghum						
	bread						

	injera						
	boiled						
	porridge						
<b>6</b>	Sweet potato						
<b>7</b>	taro						
<b>8</b>	Haricot bean						
<b>9</b>	kotcho						
	bread						
	bullu						
<b>10</b>	Livestock production						
	Milk						
	Meat						
	Egg						

**SECTION 4: INCOME OF THE HOUSEHOLDS**

**4.1 During the 2009-2010 periods, what are the major sources of income for your household?** Rank the sources in terms of the amount of income accruing from the source. *(PROBE TO GET A COMPLETE LIST AND FILL THE INFORMATION IN TABLE BELOW)*

Source of income	Amount in cash	Amount in kind
Household income		
Remittance		
Gift		
Inheritance		
Donations		
Aid		
Dowry		
Other specify		

**4.2 In comparison to 24 years before, would you say your income for the 2009-2010 periods has increased, decreased, or remained the same?**

1=Increased By how much? ..... birr [Year/Month/Week/Day] *GO TO 4.3*

2=Decreased By how much.....birr [Year/Month/Week/Day] *GO TO 4.4*

3=Remained the same (*GO TO 4.5*)

**4.3 How has the increase in your household’s income in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE INCREASE*]

**3.4 How has the decrease in your household’s income in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE DECREASE*]

**4.5 Has your strategies for coping with cash deficiency in your household changed or remained the same during the 2009-2010 periods when compared to the period before?**

1= Changed 2= Remained the same (*GO TO 4.7*)

**4.6 What are the new strategies for coping with cash that you have been using during the 2009-2010 period?** [*PROBE TO GET COMPLETE LIST OF THE NEW COPING STRATEGIES*]

**4.7 During the 2009-2010 periods, what are the major sources of credit to your household? Rank the sources in terms of the amount of the credit from the source.** (*PROBE TO GET A COMPLETE LIST AND FILL THE INFORMATION IN TABLE BELOW*)

**4.8 Of the mentioned sources of credit, which ones are new in that they were not available as sources of credit for your household during the period before 2009-2010?** [*PROBE TO GET COMPLETE LIST*]

**4.9 How would you compare the number of credit providers during the 2009-2010 periods to the period before: would you say that the number of credit providers for the 2009-2010 period has increased, decreased, or remained the same when compared to the period before?**

1=Increased By.....? 2=Decreased By.....? [*GO TO 4.11*] 3=Remained the same [*GO TO 4.12*]

**4.10 How has the increase in the number of credit providers in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE INCREASE*] [*GO TO 4.12*]

**4.11 How has the decrease in the number of credit providers in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE DECREASE*]

**4.12 How would you compare the amount of credit available for the 2009-2010 periods with the period 24 years before, would you say it has increased, decreased, or remained the same?**

1=Increased By how much? ..... Birr

2=Decreased By how much...Birr [*GO TO 4.14*]

3=Remained the same

**4.13 How has the increase in the amount of credit available to your household in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE INCREASE*]

**4.14 How has the decrease in the amount of credit available to your household in the 2009-2010 periods come about?** [*PROBE FOR EXHAUSTIVE EXPLANATION FOR THE DECREASE*]

### Appendix 9: Description of independent variable for Heckman two stage model

Variable	Descriptions
Intensity of Collaboration	It is a composite indicator which indicates the extent of farmers' collaboration in different activities. It is computed from 10 different variables which have binary values 1 if the household is collaborating with other actors and 0 otherwise. It is expected to have positive effect in innovation.
Total Training	It is the sum of normalize values of households' participation in government and nongovernment organizations training and extension services. For each of training from government and non government organization the categorical response variables are normalized as follows, More than 20 trainings=score=3 ,Normalized score=1, 10-20 times=score=2,Normalized score=0.67 Less than 10 =score=1, Normalized value=0.33Never participated =score= 0, Normalized value=0.00It is expected to have positive influence on innovation.
Age of household head	A continuous variable indicating the age of the households head in years. It is also used as a proxy to measure the average farming experience. It is expected to have a range of influence on the dependant variable, thus the expected sign is ambiguous.
Distance from nearest district center	It is a continuous variable which indicates the distance of the household from the nearest district center. As households are near and nearer to urban area it is expected to participate in more intensive farming activities which need more technologies than others. On the other hand farmers near to urban area may participate more on non farming / off farm activities which decrease participation in farming activities. Therefore the expected sign is ambiguous.
Farm enterprise index	A composite indicator calculated from three variables which indicate the engagement of the household in dairy farming, fattening of cattle and fattening of sheep and goat. The index is calculated from the mean value of binary responses of the three variables which takes 1 if the household is engaged in the enterprise and 0 otherwise.
Family size	It is the number of people in the households which are available to agricultural labor and innovation activities. The expected sign is ambiguous.
Total land owned privately	The total amount of private land owned by the household. Farm size is often correlated with farm income which increases the probability of innovation. It is expected to affect innovation positively.
House of the	It is the house of the respondent which indicates the asset base of the rural households. It is binary dummy variable which takes 1 if the household owns

respondent	Corrugated iron sheet house and 0 for Thatched house. Its expected sign is ambiguous.
Sex of household head	This is the sex of the household head male/female which helps to differentiate male from female headed households. It is a binary dummy variable which takes 1 for Male headed households and 0 for Female headed households. Male headed households are expected to be more innovative than female headed households.
Intensity of forage use Diversity	It is a composite indicator calculated from a range of constitute element or proxy indicators which indicate the current use of fodder tree for different purposes. It is calculated from binary responses variables which takes 1 if the household is using fodder tree for a given purpose and 0 otherwise. It is expected to affect innovation positively.
Livestock management system  X <sub>11_1</sub> : Open grazing X <sub>11_2</sub> : Mixed system	It is the type of livestock management system for which the household is practicing. It is a categorical variable which includes open grazing system, Mixed Type and Cut and carry or total confined system. It is coded as a dummy variable where Cut and Carry =1 and 0 otherwise, Open grazing =1 and 0 otherwise and Mixed type =1 and 0 otherwise. Cut and carry system is taken as a reference variable and it was not included in the model. Both variables are expected to have negative effect on forage innovation compared to the reference.
Dairy cooperative membership	It shows whether the household is a member of any dairy cooperative or not. It is a binary variable which takes 1 if the household is a member of dairy cooperatives and 0 otherwise. It is expected to have positive influence in innovation.
Total Livestock Unit	This is the total number of livestock owned by the farmer measured in livestock unit. It is a continuous variable calculated according to FAO guidelines. Households who have more livestock unit are expected to be more innovative than those who have less.
Innovation index	It is a composite indicator calculated from a range of constitute element or proxy indicators of innovation related activities and normalized their values for aggregation. It is expected to have positive effect for forage innovation.
Forage technology with follow up or training	It is a variable which shows whether the household was introduced forage technologies with training or follow up. It is dummy binary variable which takes 1 if the household was introduced the technology with training or follow up and 0 otherwise. It is expected to have positive sign.
Education level of Household	This is the level of non formal and formal education completed by the household head. It is continuous variable and expected to have positive

Head	influence on the innovation activities of households.
Altitude category	It is the category of Altitude where the household is living. It is recoded as a dummy binary variable which takes 1 for Midland, and 0 for Highland. The expected sign is ambiguous.
Age of household head	A continuous variable indicating the age of the households head in years. It is also used as a proxy to measure the average farming experience. It is expected to have a range of influence on the dependant variable, thus the expected sign is ambiguous.
Migration	It is a variable which shows whether the household lives in that part of the country or not. It is binary variable which takes 1 if the household lives in that part of the country and 0 otherwise. Its sign is ambiguous.
Intensity of Access to forage technology	It is a composite indicator computed from 5 different variables which indicates the extent and availability of technologies in the right time, place, quantity, quality and reasonable price to the household. The variables are constructed from the mean value of the binary response which takes 1 if the household has access to technology on right time, quantity, quality, place, price and 0 otherwise. It is expected to have positive influence in innovation.
Consultancy about forage	It shows whether the household is consulting anybody about forage technology or not. It is a dummy binary variable which takes 1 if the household is consulting and 0 otherwise. It is expected to have positive effect on forage innovation.
Forage technology with traders connection	It is a variable which indicates whether the household was introduced forage technologies with traders' connection. It is dummy binary variable which takes 1 if the household access the technology with traders' connection and 0 otherwise. It is expected to have positive effect on innovation.
Intensity of Adverse Effect	It is a composite indicator calculated from a range of constitute element or proxy indicators of adverse effects the household faced previously. It is calculated from different adverse effects which takes 1 if the household faced the risk and 0 otherwise. The expected sign is ambiguous.

### Appendix 10: Description of explanatory variables for binary logit model

Variables	Description
Sex of household	This is the sex of the household head male/female which helps to differentiate male from female headed households. It is a binary dummy variable which takes 1 for Male headed households and 0 for Female headed households. Male headed households are expected to be more successful than female headed households.
Age	A continuous variable indicating the age of the households head in years
Distance from nearest Town	It is a continuous variable which indicates the distance of the household from the nearest district center. As households are near and nearer to urban area it is expected to participate in more forage adoption activities and on the other hand farmers near to urban area may participate more on non farming / off farm activities which decrease participation in farming activities. Therefore the expected sign is ambiguous.
Cut and carry= $Z_{4\_1}$ Confinement = $Z_{4\_2}$	It is the type of livestock management system for which the household is practicing. It is a categorical variable which includes Confinement, Communal and Cut and carry. It is coded as a dummy variable where Cut and Carry =1 and 0 otherwise, Communal =1 and 0 otherwise and Confinement =1 and 0 otherwise. Communal grazing system (open grazing) is taken as reference.
Gundi cooperative memberships	It is a binary variable which indicates the membership status of households. It takes 1 for members and 0 otherwise
Intensity of Access to drug	It is a composite indicator computed from 5 different variables which indicates the extent and availability of drug in the right time, place, quantity, quality and reasonable price to the household. The variables are constructed from the mean value of the binary response indicators which takes 1 if the household has access to treatments on right time, quantity, quality, place, price and 0 otherwise. It is expected to have positive influence in innovation.
Intensity of familiarity with technologies	It is composite indicators which constructed from the intensity of households familiarity with pour on, insect trap and target fly control. The variables takes take 1 if the household has familiarity and 0 otherwise.
Cost for treatment	It is a continuous variable which indicates the household expenditure to treat animals for Trypanomosis.

Combined use treatments	It is a binary variable indicating whether the household is using the combination pour on and insect trapping technology. I takes 1 for Yes and 0 for No
Injection practice	It is a binary variable which indicates household practice on treating animal on injection. It takes 1 for Yes and 0 otherwise
Crops for market	It is a binary variable which indicates household involvement in surplus crop production. It takes 1 if Yes and 0 otherwise.
House of the respondent	It is the house of the respondent which indicates the asset base of the rural households. It is binary dummy variable which takes 1 if the household owns Corrugated iron sheet house and 0 for Thatched house. Its expected sign is ambiguous.
Total land owned privately	The total amount of private land owned by the household.
Treatment year	It is the year which indicates when the households starts treating their animals
Time for treatment	It is the total amount of time the household treats the animal