ASSESSMENT OF QUALITY OF DATA OF THE 2014 KENYA DEMOGRAPHIC AND HEALTH SURVEY (KDHS)

BY

ESTHER KAGENDO RUGENDO

Q50/70263/2011

A PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF ARTS IN POPULATION STUDIES AT THE POPULATION STUDIES AND RESEARCH INSTITUTE, UNIVERSITY OF NAIROBI

NOVEMBER 2016

DECLARATION

I declare that this research project is my own original work. It is being submitted for the degree of Master of Arts in Population Studies at the University of Nairobi. To the best of my knowledge, it has not been submitted before in part or in full for any degree or examination at this or any other university.

Candidate

ESTHER KAGENDO RUGENDO	**Signature**	**Date**
Q50/70263/2011	_____	_____

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

SUPERVISOR

	Signature	**Date**
MR. BEN OBONYO JARABI	_____	_____
	Signature	**Date**
PROF. MURUNGARU KIMANI	_____	_____

DEDICATION

I dedicate this project to my Husband, John Munene and our daughter Gabby Kawira for your presence, encouragement, moral support, patience and understanding throughout the period of this project. Your constant cheer and prayers inspired me to move on with courage and determination.

To my parents Gilbert Rugendo and Rose Keeru Rugendo, without who I could never have been where I am. You shaped my life through immense sacrifice, love, care and support. Through your guidance and role modeling, you taught me many things, instilled in me values and virtues that continue to shape my character.

To my brothers Eric Mwenda and Titus Murithi, for the encouragement and moral support.

ACKNOWLEDGEMENT

I am greatly thankful to the Almighty God, my provider and ever present help for the numerous blessings he has bestowed upon me.

I am thankful to the people who contributed to the successful completion of this project in different ways. I genuinely acknowledge the inspirational support and direction of my two university supervisors: Mr. Ben Jarabi and Prof. Murungaru Kimani for the insightful comments, suggestions and constructive criticisms during the whole process of the project. Their diligent endeavor, perseverance and understanding enabled me to complete this project successfully.

I will forever be appreciative to all lecturers at PSRI, Prof. Ikamari Lawrence, Prof. Agwanda Alfred, Prof. John Oucho, Mr. Ben Jarabi, Dr. Gichuhi Wanjiru, Dr. Anne Khasahala and Dr. Odipo George for their academic nourishment and influence that molded and provided me with a strong background knowledge in Population Studies and Research during my master's degree programme.

I am entirely grateful to my immediate family, classmates and friends for their absolute encouragement and support right through my academic life. May God bless you all.

To you all, I say "Thank you and God bless you abundantly".

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Good quality of data is important in reaching inferences and conclusions that are accurate, reliable and valid. Before analyzing demographic data, it is essential to be familiar with whether the data is accurate and provides acceptable answers. Age data evaluation allows the determination of the direction and magnitude of errors. The need for reliable, valid and complete sets of Kenyan data justified this study. The objective of the study was to assess the quality of the 2014 KDHS data. This was done with a special focus on assessing the completeness of reported date of birth for female interviewees and completeness of reported date of birth of children ever born. It also focused on assessing the degree of age heaping and digit preference by age and sex. The assessment was conducted using age ratios, sex ratios, Myer's blended index, and UN Age-sex Accuracy Index. The study also applied an appropriate adjustment method to smooth reported data. The adjustment was applied to remove short-term random variations, or outliers to reveal the important underlying quality data.

From the study, it emerged that the 2014 KDHS age data was incompletely reported by 22.5 percent for date of birth for women and children ever born. Some dates, months and years were missing, but since ages had been provided, the missing elements were imputed. Age data was inaccurately reported with high fluctuations in age ratios for males and females, which could be an indication of persons in various ages being carried across age group boundaries or persons misreporting their own ages for various reasons. This compromises the quality of data. Myer's index revealed digit preference of age ending with '0' and '5' where the digit ending with '0' is more preferred than '5' for both male and female respondents. Moreover, females showed higher age misreporting than male respondents. Ages ending with 1 have the highest digit avoidance, followed by those ending with 3, 7 and 9. At the same time, the UN age-sex accuracy index revealed general inaccuracy in the 2014 KDHS age data. Six regions recorded highly inaccurate data and two regions had inaccurate data. The findings are consistent with previous study findings on quality assessment on Kenyan data.

Although much has been done by agencies dealing with KDHS data collection to ensure that reported data is of the highest quality including rigorous training of interviewers on the most effective data collection methods, there is still room for improvement. This can be done by creating mass awareness and educating people on how false reporting during surveys and censuses distorts information derived from data and how their lives are affected. Since data anomalies could be arising from interviewers' motivation to reduce the length of the interviews leading to age misplacement and transfer, a more thorough supervision should be conducted during field work to ensure ages are recorded appropriately.

CHAPTER 1
INTRODUCTION

1.1 Background to the study

Data quality is of great significance when it comes to the Demographic and Health Surveys (DHS). DHS for a long time is regarded as the highest standard for collection of data of national representation in developing countries. Data quality is about having confidence in the data recorded and analyzed to make decisions (Moultrie et al., 2013). For data to be of high quality, it should meet the following five characteristics: completeness, accuracy, validity, reliability, precision and integrity.

Completeness means that all the requisite information is available and no data values are missing; accuracy is the fact that data accurately represent the real world or the actual situation (incorrect dates, ages, and even untimely or not current data could impact negatively on the operational and analytical applications); validity is a characteristic of data that measures what the demographer actually wants to measure; reliability is concerned with the consistency of data; timeliness is the relationship between the time of collection, collation, reporting and use of data for decision making; and precision is the margin of error or closeness of data when compared to the expected effect the project was set to get. Lastly, integrity is the measure of 'truthfulness' of the data (IRD, 1990). "In a perfect world, data would always be complete, accurate, current, pertinent, and unambiguous. In the real world, data is generally flawed on some or all of these dimensions" (Feeney, 2003). Just like as other statistics, population data, whether obtained through census, vital registration, surveys or others, are affected by errors that affect its quality (Moultrie et al., 2013).

Demographic data errors are categorized into two - content and coverage errors. Coverage errors relate to completeness as well as the quantitative aspects of data collection. Coverage errors occur due to omissions leading to under-enumeration or duplication leading to over-enumeration. There are several factors that lead to coverage errors. These factors include; inaccessibility of respondents, lack of co-operation, communication problems and improper boundary sketches or descriptions. Content errors on the other hand, refer to how qualitative traits of data such as sex, age, economic activity, and marital status among others are reported. Content errors occur as a

result of a respondent providing an incorrect answer or by an interviewer recording a wrong answer. For example, a question asked on current age can be sought by either asking for the "current age" or "completed number of years". These questions could raise different responses depending on the understanding of the interviewees (Moultrie et al., 2013).

Good quality of sex and age data is significant in reaching inferences and conclusions that are accurate, reliable and valid (Susuman et al., 2015). Therefore, at each stage of analysis, the person utilizing demographic data needs to look at the outcomes skeptically for possible signs of errors. Before analyzing any problem, a demographer should ascertain the level of accuracy of data (Moultrie et al., 2013). Data evaluation is a pre-requisite to determining the reliability of population estimates. Age data evaluation allows the determination of the direction and magnitude of errors. It forms a basis for applying adjustments to the data thus avoiding compounding of errors and to ensure data is of an acceptable standard (Pullum, 2006).

Most demographic and socio-economic analysis and variables are attributed to sex and age. Age is a crucial demographic data element used for analysis of various population dynamics. Age is also useful in analyzing the structure of populations and forecasting of growth rates (Bello, 2012). Age and sex errors greatly affect population statistics whether in basic summaries or in-depth analysis (Shryock et al., 1976). Age and sex are also utilized to provide population projections and to determine population forecasts (Siegel and Swanson, 2004). Other uses of age population statistics are in preparation of mortality, fertility, nuptiality studies, as well as construction of life tables (IRD, 1990). For instance, to analyze fertility rates, a researcher would require data on women of reproductive age, 15-49 years (Susuman et al., 2015).

Descriptive and in-depth analysis of demographic processes of mortality, fertility and migration depend on data by sex and age. Analysis of data by sex and age is applied by policy makers, actuaries, politicians, educationists, health practitioners and many others to make important decisions in their fields of operation (Susuman et al., 2015). Age and sex data provide information on the population size and growth, which are pre-requisites for sub-national and national planning, decision making and allocation of resources. Age data also provides annual updates on a country's population estimates, which are key to generation of many population

indicators. Population statistics, analyzed by age and sex, are interpreted to provide useful information for setting targets, planning, implementing and in monitoring and evaluation of social, developmental and economic programmes among others. In addition, such data is used in measurement of important demographic indicators of quality of life, such as life expectancy at birth and infant mortality rate (United Nations, 2014).

Despite its usefulness in demographic and epidemiological analysis, a few anomalies frequently arise in age data (Denic et al., 2004). According to KNBS (2012), age data is usually susceptible to such anomalies, unlike sex data (Pardeshi, 2010). Since age is one of the regular components in censuses and surveys, its misreporting constitutes one of the important demographic challenges (Spoorenberg, 2007). The most common age data problems that have been documented are age overstatement, digit preference and understatement which lead to such errors as age heaping or digit preference. Age heaping or digit preference is a situation where interviewees give numerical responses with specific preferred terminal digits. This mostly occurs due to ignorance of one's exact age, so the respondents or the interviewers tend to estimate and round off ages (A'Hearn et al., 2009), thus leading to heaping on terminal numbers of 0 or 5 at the expense of other ages (Pullum, 2006). It could also occur as a result of misinterpreting the nearest, the last, and the next birthday or rounding off to the nearest age ending in 0 or 5. For example, children below one year could be reported as one year old or zero years (Kidane, 2009).

Quality of sex and age data is influenced directly or indirectly by cultural, environmental, socio-economic, and demographic factors (Kidane, 2009). The patterns and causes of digit preference differ from one culture to the next. However, preference for ages ending with 0 and 5 is rather widespread, especially in Africa. Kidane (2009) further states that misreporting occurs as a result of lapse of memory or deliberate attempts by respondents or interviewers to under or overestimate. Other causes of age and sex errors include ignorance of the true age, low literacy levels, and data collection problems (Pullum, 2006). The quality of data varies over time and with surveys/censuses. The significance of the errors, considering their magnitude, depends on data uses. The errors may be large or small depending on the challenges experienced during the data collection and the efficiency of methods for data compilation (Moultrie et al., 2013).

Research has shown that age heaping tends to increase as age increases. Due to memory challenges, the older population may provide wrong age information. There is a tendency by individuals to overstate their ages thereby increasing the number of people who survive to old age (Moultrie et al., 2013).

Age heaping could occur as a result of cultural preference for or avoidance of certain digits (Stockwell et al., 1973). Age heaping has been observed in many cultures, especially for ages ending in 5 and 0 years. This happens mostly to people claiming to be aged 30 or 35 than 29, 34, etc (Pardeshi, 2010). Some cultures also report on multiples such as the Han Chinese who age heap on multiples of 12 years lining up with the zodiac cycle. In Korea, and other East Asian countries, sometimes prefer ages ending in 3 because it resonates with words or characters representing life. In other cultures, certain numbers are avoided such as thirteen because it is regarded as unlucky. In Korea and China, the number 4 is avoided because it has a sound of a death character (Jowett and Li, 1992).

Age heaping is distinct in populations with low educational levels. Previous studies demonstrate a relationship between the magnitude of age heaping and level of education. Age heaping is an informative measure of numeracy within the population and researchers continue to use it as a proxy to determine the quantitative reasoning ability (A'Hearn et al., 2009).

The characteristics of the interview could be important in misreporting. Studies have hypothesized that the actual time of the interview may induce some misreporting. For example, an interviewer may be motivated to shorten the last interview of the day. The earliest interviews may differ in quality from the last interviews of the day (Pullum, 2005). The other motivating factor is that, since the major aim of the household survey is to find women between age 15-49 eligible for the individual survey, the interviewers would want to decrease their workload through misstating some of the women's ages either as younger or older to take part in the individual survey. As a result, a woman aged 15-19 could be reported as age 10-14, or one who is age 45-49 could be reported as age 50-54 (Pullum, 2014).

Data completeness is essential in meeting requirements of current and future demographic data demands. It indicates whether or not all the necessary data is available in the data sets. It is the extent to which the important expected attributes of data are provided. Some data requirements are 'mandatory' while others are 'optional' (Pullum, 2006). Data variables such as, child alive or dead, which is reported by mothers about all their children is used to determine the child mortality and survival rates of a country or region.

There are several methods that have been developed and widely employed for assessing the accuracy of age-reporting to enhance understanding of data structure and anomalies (Moultrie et al., 2013). Demographers have developed several techniques to detect the degree of errors arising from age data such as age heaping (Kidane, 2009). Other techniques have been developed to correct or adjust the data. According to Moultrie et al. (2013), the most commonly used strategies for assessing accuracy of age reporting are: analysis of age and sex ratios, the UN age-sex accuracy index, the Myer's blended index, and the Whipple's index.

This study assessed the level of completeness of reported date of birth for women interviewed, date of birth for children and reporting of children alive or dead. It helped identify any missing data within these variables, which could pose a significant challenge in data analysis.

1.2 Statement of the Problem

Although all required steps such as preparing questionnaires, pilot surveys, supervisors and enumerators trainings are adapted to decrease chances of population age errors, the quality of the Kenya Demographic Health Survey (KDHS) age data is often compromised (Mugo, 2012). This necessitates assessment of data quality and application of necessary adjustments to the data. Previous census data (Kodiko, 2014 & KNBS, 2012) and KDHS data of 2008-09 have had quality assessments (Mugo, 2012) revealing data quality issues in Kenyan data. The data errors were manifested by age heaping, digit preference and age displacement. The 2014 KDHS has not received such analysis, and because of previous data quality findings, it was vital for the assessment to be conducted.

Assessment of data quality of the 2014 KDHS, and its adjustment, was necessary to produce flawless data for programmatic use as well as to remove the effects of age-sex errors on population indicators generated. Flawed age-sex structures produce flawed future projections (Pullum, 2006). This study assessed the quality of the 2014 KDHS data and adjusted it accordingly. It applied adjustment methods to smooth the data in order to remove short-term random variations or outliers to reveal the important underlying unadulterated data.

1.3 Research Questions

The study sought to answer the following questions:

i. What are the data quality issues in the 2014 KDHS data?
ii. How are data quality issues addressed?

1.4 Objectives of the Study

The general objective of the study was to assess the quality of the 2014 KDHS data. The specific objectives of the study included:

i. To determine the quality of data of the 2014 KDHS
ii. To apply appropriate adjustment methods to the 2014 KDHS household age data.

1.5 Justification of the Study

For proper forecasting at national and sub-national levels, Kenya requires accurate information on demographic qualities of her population. Age and sex are important demographic variables utilized for providing population projections and to determine population forecasts. In Kenya, achievement of the national and county governments' development plans and proper resource allocation depends on availability of reliable, valid and accurate sex-age data as well as projections for sector-specific development, demographic patterns and dynamics, among others. Quality age data is also useful in monitoring and evaluating the impact of programmes such as health, education, agriculture and so on. This highlights the importance of having updated and accurate information on sex, age and population growth. Thus, the need to have valid, reliable, and complete Kenyan data sets substantiated this study. The study will help in appreciation of the quality of age and date of birth for mothers and children data of the 2014 KDHS. It will, therefore, contributed to improvement of data quality and provides data sets which, if analyzed,

can provide accurate estimates and projections of Kenya's population. Demographers can use the study findings to improve quality of their analysis and conclusions of population estimates and other variables of interest. Policy makers can make accurate decisions on the population using the study findings.

1.6 Scope and Limitations

The project focused on quality of data pertaining to three components from the 2014 KDHS: age-sex misreporting, completeness of date of birth for women, and completeness of date of birth for children ever born. The study was limited to assessing the quality of age-sex and date of birth data. The assessment and smoothing were performed on household age data, which is not the main focus for the DHS surveys. Another limitation is that methods used for smoothing of age data are commonly used for smoothing census data. Other characteristics such as contraceptives, social-economic, education, among others were not assessed because they were beyond the scope of this study.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter reviews previous work on role of age-sex data and assessment of quality of age and sex, assessment of DHS data and completeness of reported date of birth for mothers and children. It reviews what has been done to improve the errors encountered in working with age data. It gives a picture of the significance of data quality assessment and the methods used in previous studies in order to provide insights to the methods that are best applicable in such assessments. The chapter also points out areas that need careful consideration during data quality assessment.

2.2 Role of Age and Sex Data

Age and sex distribution is one of the most important pieces of information derived from censuses and surveys (ESCWA, 2013). Most demographic and socio-economic data are attributed to sex and age. Age is a vital variable in demography and is used for statistical and descriptive analyses of population dynamics. Age and sex not only provide the demographic make-up all over the world (Yazdanparast et al., 2012) but also act as a basis for future population estimates and projections on life expectancy, fertility, mortality and migration. Other estimates that rely on age and sex include: dependency rates which depend on data of children of 0-14 years, above 65 years and those between 15-64 years (Susuman et al., 2015) and mortality estimates that are affected by the quality of age at death data; the birth dates data, and birth histories data. Such data, if inaccurately reported cause a distortion of patterns and trends of mortality rates. For example, when there is selective omission of deaths in childhood, the deaths that occur during early infancy are most affected.

Most national developmental plans for the provision of public services and goods such as housing, employment, food, health, education, manpower etc depend on social and economic statistics disaggregated by sex and age (Bello, 2012). Therefore, age data, if not well collected and analyzed, lead demographers, statisticians, planners and researchers to make inaccurate inferences.

2.3 Data Quality

Age is one of the vital demographic variables that should be accurately collected and reported (West et al., 2005). However, in spite of its importance, it constitutes one of demography's most frustrating problems (Ewbank, 1981) as it often suffers from reporting errors and irregularities, which impacts negatively on its usage (Denic et al, 2004). Age tends to be more susceptible to anomalies (KNBS, 2012) and is more misreported than sex (Pardeshi, 2010). Since it is a constant element in censuses and surveys, its misreporting constitutes one of the most demographic challenges (Spoorenberg, 2009).

The most common age data problems that have been documented are age overstatement, digit preference and age heaping. Digit preference is more distinct in populations or sub-groups with low education status. Though the patterns and causes differ from one society to the next, age preference for those ending with "0" and "5" is rather wide spread and apparently this has existed for much of human history (Ewbank, 1981). In some developed countries, age misstatements of this type have drastically reduced to almost negligible in recent years. However, in most developing countries, data still suffers from digit preference (Byerlee and Terera, 1981). Studies on age preference and avoidance in developing countries have shown vast distortions (Caldwell, 1966; Ewbank, 1981; Byerlee & Terera, 1981; Caldwell & Igun, 1971). This occurs mostly when age is not known, and so respondents or interviewers tend to estimate which leads to heaping on 0 or 5 (Pullum, 2006).

Studies show that census in African countries usually suffer from digit preference also referred to as a content (or response) error or non-random measurement error (ESCWA, 2013; West et al, 2005; Yazdanparast et al, 2012). Since the late 1960s and 1970s, scholars have noted that there is a big impact of age heaping on demographic, economic and health statistics in Africa (Bocquier et al., 2011; Stockwell et al., 1973). Irregularities in age data from African and Asian samples have been noted by previous studies (Caldwell, 1966; Caldwell & Igun, 1971; Stockwell et al., 1973; Byerlee & Terera, 1981; Ewbank, 1981; Jowett & Li, 1992; Denic et.al, 2004; Palamuleni, 2013) and recent work has examined age heaping in Nigeria and Zambia (Bello, 2012). The quality of census in terms of age-reporting has, however, improved remarkably in Asia, but less so in African countries (Cleland, 1996). A study in Zambia shows that males were more inclined

towards reporting digit '5' compared to females who preferred '0'. In addition, besides digits 0 and 5, preference for ages with digits ending with 2 and 8 in 1990, and 8 in 2010 by both sexes was also common which is in agreement with other similar studies (Bello, 2012). In another study, data showed marked avoidance of ages ending with 1, 2, 3, 6 and 7 (Bwalya et al., 2015).

2.4 Reasons for Age Data Distortions

The reasons for age data distortions are varied and may arise from misreporting or mis-recording. More reasons include ignorance of one's true age as a result of uncertainty of date of birth, unclear instructions given by enumerators and methods used for data collection, social, cultural and political factors. Ignorance of exact age leads to approximation by the respondents (Ewbank, 1981; ESCWA, 2013) while cognitive biases towards reporting landmark ages leads to 'heaping' of data (Pardeshi, 2010; Stockwell et al., 1973). In addition, some people just choose not to reveal their actual age and end up rounding up ages. Moreover, other people try to meet age eligibility cut offs thus misreporting their ages. For example, young children below five may be reported as five years, females between the age of 10 to 14 who have undergone puberty could be reported to be age 15 to 19, particularly if they have children or are married. Women in their 40s who are breastfeeding their children could be reported as younger than they are. Age heaping increases with age and is more pronounced among elderly populations. This could be argued that some older people are more likely to forget their age. Elderly men tend to overstate their ages for purposes of getting prestige or so that they can be offered status of senior citizenship. In Africa, many elderly populations are illiterate and lack knowledge of their exact date of birth, thus they mostly estimate their current ages (West et al, 2005; Pardeshi, 2010).

Others argue that, it is due to social-cultural values that people in different societies attach either consciously or sub-consciously, to certain numbers which consequently results in age mis-statement. Stockwell et al. (1973) state that age heaping could occur as a result of cultural preference for or avoidance of certain digits. For example, among the elderly and younger ages, age may be over or under estimated just for preferences of certain numbers (West et al, 2005). Age heaping has been observed in many societies at terminal digits 0 and 5 years where ages have been reported as 40 or 45 rather than 39, 44, etc (Pardeshi, 2010). In other societies, age

heaping could occur at different multiples. For example, among the Han Chinese, age heaping occurs at multiples of 12 years going by the zodiac cycle (Jowett and Li, 1992).

Social factors also influence age reporting and contribute to age heaping. Societal incentives motivate people to distort their reported ages. For example, women above thirty years who are not married could feel pressured to report that they are younger for purposes of increasing their "marriageability". In addition, young people who want to go through esteemed stages of their cultural life such as initiation into adulthood, being granted more responsibilities and rights, may misreport their ages in order to fit the desirable group of being issued such privileges (A'Hearn et al., 2009).

Previous studies have established that there is a relationship between the levels of age heaping and population education. Age heaping is still used by researchers as a measure of population education, numeracy and quantitative reasoning capability (A'Hearn et al., 2009). More than fifty years ago, Bachi (1951) and Myer's (1954) studied the correlation between age heaping and education levels across and within various countries. Bachi (1951) analyzed the degree of age heaping among Muslims in Palestine and Jewish immigrants to Israel and found that the populations with some education reported better knowledge of their ages. On the other hand, Myer's (1976) found a correlation between exact age knowledge and income where majority of the poor were less educated. The study conducted by Herlihy and Klapisch (1985) found that distinct heaping was more pronounced among women who were poor and those living in small towns and rural areas. A'Hearn et al (2009) studied in more detail the correlation between age heaping and illiteracy levels in the population by analyzing data from 52 countries. The study reported that the magnitude of age heaping was significantly related to illiteracy and that the probability of reporting ages that are rounded increased so much with individual and regional levels of illiteracy.

Sex ratios and age ratios can be used to detect displacement of ages from one age group to the next. For example, from the household questionnaire, the sex ratio can be calculated at age 10-14 and at age 15-19. There can either be upward or downward transfer of ages. Net downward transfers of females across age 15-19 tend to make the sex ratio at 15-19 larger than the sex ratio

at 10-14. If there is no downward transfer of females in age group 15-19, then the ratio of females age 10-14 to females age 5-9 should be about the same as the ratio of females age 15-19 to females age 10-14. However, if females have been shifted downwards across age 15, the ratio of age 10-14 to age 5-9 should be larger than the ratio of age 15-19 to age 10-14 (Pullum, 2005).

The main purpose of the household survey is to pick out women who are eligible for individual female survey age. These are women of age 15-49 years. During the interview, interviewers could be motivated to bring down the amount of work they do by misreporting some ages through entering women as younger or older to take part in the individual survey. This could lead a woman to be reported as 10-14 yet her actual age is 15-19 or reported as 50-54 while she is actually 45-49 years. The probability of age displacement and age transfer is higher across age 50 than age 15. This could occur because age of younger groups is better known and better documented than that of the older cohorts. One is because, younger populations tend to use their ages more in their everyday life and their education levels are higher thus making age reporting be more accurate (Pullum, 2006). Two, it is easier to estimate the age of younger persons than for older ones. Three, older women answer a greater part of the questionnaire because they have longer birth histories. Thus interviewers could want to reduce their workload by displacing potential female respondents into age 50-54 than to 10-14 (Pullum, 2006).

2.5 Age Misplacement

The KDHS survey is subject to systematic age displacement. There is a tendency for ages in the household survey to be displaced across the boundaries of eligibility for individual surveys. For example, for women, there may be some downward displacement across age 15 or upward across age 50. This is believed to be the result of interviewers looking at ways of reducing their workload. It is also probably because the household respondent is not fully confident about their ages. Transfers happen and are noticeable within five years of the boundary (Pullum, 2006).

Misreporting of age or date of birth of children could be motivated by the interviewers' possible desire to reduce their workload in order to avoid the detailed questions about child health. These questions are to be asked about children born since January of a certain calendar year. The questions are asked about all children born since January of the fifth calendar year before the

year in which the fieldwork begins. Since the eligibility for the questions is based on date of birth, some births that occurred in for example 2014 might be transferred to 2013 to increase their ages. Apart from distorting measures of child health, such transfers can potentially distort fertility rates, leading to exaggerated fertility decline (Pullum, 2005).

2.6 Data Incompleteness

The assessment of missing data is a vital gauge of data quality. If data has a high level of unstated or missing responses, it may imply that the interviewers have not been trained adequately, the questions are poorly worded, the interviews are too demanding, and so on. Unstated answers show misused time, fieldwork, resources and it leads to statistical inefficiency. It reduces the number of valid responses thus limiting the analysis to a smaller number of responses (Pullum, 2006). Completeness of variables is a requisite to getting high quality results of analysis. Thus, in order to make accurate estimates and predictions, all data should be completely reported. Date of birth for women interviewees could be incompletely reported and this would pose a risk of leading to poor quality of analysis that depends on this variable. To ascertain the quality of KDHS data, the study examines two aspects of the survey. One is the completeness of reported age of women interviewed, and two, is the completeness of reported age of children ever born.

The variables this study used to assess the quality of age-sex and date of birth reporting were found in the household and individual questionnaires of the 2014 KDHS. The first one was the age of the household head which was found in the household questionnaire. The date of birth for women and children ever born was extracted from the individual questionnaire. The date of birth of women respondents provided three items related to age: her complete years of age, birth year, and birth month. Minimally, a woman was expected to report her age or her birth year. Various respondents provide less information on these data items, or inconsistencies exist in cases where all the information is reported. Such data may require imputation of one or more items to make the data complete. Completeness assessment determines the percentage of respondents who provided all three data items and those who did not provide all. It also determines if there was inconsistency of data that would require imputation.

2.7 Assessment of Quality of the DHS data

The role that age data plays in demographic analysis cannot be overemphasized. Awareness of age data distortions and inaccuracies should trouble demographers a great deal that data quality assessment and adjustment should be a fundamental process of demographic data analysis. Therefore, irregularities should be assessed, identified and adjusted before users of demographic data could apply meaningful analysis. As a matter of general policy, DHS data is not adjusted to compensate for inconsistencies such as omission and displacement. This is done in order to maximize the analytical possibilities for researchers who would like to use the data sets outside the DHS scope. Therefore, the data files are distributed without conducting adjustments. Generally, there is no adjustment for data quality issues even when DHS produces indicators for the main reports on surveys and for comparative reports (Pullum, 2014). This calls for users of the data to conduct adjustment of data before they delve into the analysis.

Many studies have been conducted to assess the quality of DHS data for various years and countries. Schoumaker (2011) assessed the levels of birth omissions in DHS in sub-saharan Africa. The study showed that birth omission was a common error in all countries and more so for Guinea, Cameroon, Niger, Mali, Mozambique and Burkina Faso. These omissions introduced biases in fertility rates and indicated significant variations of fertility estimates between countries. Birth omissions and underreporting were more prevalent among recent births than longer period births. It showed that DHS questionnaires that were more complex and had longer reference periods produced significantly higher levels of omissions. Yet some countries such as Kenya in 1998, Mozambique in 1997, Nigeria in 1999, and Cameroon in 1998 that had shorter reference periods produced poor quality of birth histories. In countries where education levels are high, there were much lower levels of omissions. Omissions were higher among uneducated women (Schoumaker, 2011).

Pullum (2008) assessed the DHS age and date reporting using household data of various countries to determine the level of missing dates, data completeness, age heaping, digit preference and age transfers across age groups. Kenya's 1993 DHS was among the ones with more than ten percent of age displacement, especially for females across ages 15 and 50, at 16.1 percent displacement rate. In 1993 KDHS, the estimated level of women of age 40-45 who were

misreported at age 50-54 was 28.5 percent. The country was fourth after Burkina Faso in 1993, Uganda in 1995 and Nigeria in 1990 which misreported at 32.0, 31.3 and 28.8 percent respectively.

Pullum (2006) conducted an age data assessment to determine how data is reported across various DHS surveys. He assessed three indicators from the household data with the Myer's index for age heaping, downward displacement of female respondents across age 15, and their upward displacement across age 50. Further, he assessed the completeness of data reported on births, marriage and age. Results show that sub-Saharan Africa had an above average level of misreporting, especially on age transfers. This included downward transfers for females at age 15-19, their upward transfer at age 45-49, and upward transfers for children below 1 year. The data was characterized by incompleteness of various dates.

In assessment of omission of births by Pullum (2014), the 1993 KDHS data revealed a report of less boys than were expected at birth with a deviation of -5.7 within 10 years prior to the survey. The highest level of incompleteness of date of birth was recorded by Guinea with 52.7 percent and Yemen with 45.9 percent. Omission of births was also found to be high in Dominican Republic and Armenia at 8.7 and 7.7 percent respectively. In Kenya, omission of births was at 3.6 percent.

Shireen et al. (2015) studied the quality of anthropometric data in the DHS data of various countries. The study found that Kenya's age heaping for children age 0-59 months was 6.7%, which means that if 6.7% of the responses were shifted to proper reporting, there would be no age heaping.

Wafula and Ikamari (2007) determined that there was age heaping in the 1998 and 2003 KDHS data sets on ages 20 and 25. It revealed that there was high underreporting of age amongst married women below 25 years, which suggested that there was age shifting from the other nearby ages. The same applied for ages 40 and 45. The 1998 KDHS data showed digit preference for "0" and "5", and digit avoidance for odd numbers like 3, 7 and 9. The 2003 KDHS data had less fluctuation despite the fact that there was under reporting at ages below 25 years.

Randall and Coast (2016) showed that the Kenyan DHS data from 1993 to 2008-2009 was characterized by inaccuracies where the Whipple's index score was consistently high. Over the years, the study reported that the Kenyan population census data improved in quality (from 160 to 150 Whipple's score) yet the KDHS data quality remained stable, oscillating around high levels of 130-150, especially for younger Kenyan adults.

Studies on assessment of the Kenya Population and Housing Censuses of 1979, 1989, 1999 and 2009 reported UN accuracy indices of 28.1, 24.9, 26.4 and 23.7 respectively (KNBS, 2012). Kodiko (2014) assessed the quality of reported data in the 2009 Kenya Population and Housing Census and indicated a national UN accuracy index of 23.72. Mugo (2012), on the assessment of the quality of data of the 2008-09 KDHS, showed that national UN accuracy index was 40.93 and that data was characterized by systematic quality issues and errors. The UN accuracy index ranged from being highly inaccurate to inaccurate as follows; North Eastern-139, Nairobi-119, Coast-77, Eastern-57, Nyanza-57, Rift Valley-57, Central-54, and Western-50. This data necessitated application of strong smoothing techniques. Thus, the need to conduct an assessment of the 2014 KDHS data to determine the extent of errors then adjust accordingly.

There are several methods that have been developed and widely employed for evaluating the accuracy of age-reporting to enhance understanding of data structure and anomalies (Moultrie et al., 2013). Demographers have developed several techniques to detect the level of digit preference errors. Other techniques have been developed to correct or adjust the data. According to Moultrie et al. (2013), the most commonly used summary indices used to evaluate the quality of age reports are analysis of age and sex ratios, the United Nations age-sex accuracy index, Myer's blended index, and Whipple's index.

CHAPTER 3
METHODOLOGY

3.1 Introduction

This chapter examines the data used in the study as well as methods applied to assess quality. It describes the data sources, selection of appropriate methods of analysis, and the actual assessment of data quality with focus on age, sex, date of birth of women, and date of birth for children ever born. This study employed several methods that have been developed to evaluate the accuracy of age-reporting to enhance complete understanding of data structure and anomalies. In addition, it describes the strategies used to adjust the data in order to address the anomalies identified.

3.2 Data Sources

The study analyzed data from the 2014 KDHS. Data was sourced from the household file, which contains information on usual household members as well as visitors with their characteristics such as age and sex. For assessment of data completeness, data was sourced from the female questionnaire.

3.3 Sample Size

The 2014 KDHS covered a total of 36,430 households where the household questionnaire was administered to 43,898 respondents. A total of 31,079 females and 12,819 males were interviewed. The data collected from these respondents formed the sample size for age and sex quality assessment in the study.

Information on completeness of date of birth for women interviewed and for children came from the individual woman questionnaire. A total of 16,419 women and 21,896 children records were assessed for completeness.

3.4 Techniques of Assessing Data Quality

This section describes the methods that have been developed and used to assess the accuracy of data on sex and age using frequencies. It also assessed the completeness of age and sex data as well as children reported alive or dead. The techniques for assessing accuracy include the use of

graphical methods, sex and age ratios, Whipple's index, Myer's blended index and UN Joint score. Frequencies were used to detect completeness of date of birth of household heads, date of birth for children and child living status.

3.4.1 Graphical Methods

Graphical methods are quick ways of evaluating age data. Graphing of single year population age and sex data helps to make visible the occurrences of age heaping from an early stage of assessment. This diagrammatic assessment of heaping is as good as age heaping detected through measures such as Whipple's index, Myer's blended index or the United Nations age-sex accuracy index. Graphical analysis of sex and age composition includes the population pyramid, pie charts, line graphs and bar charts (Siegel and Swanson, 2004). A line graph was drawn plotting ages in single years of the population on the x-axis and the number of people was plotted on the y-axis. The graph derived is expected to be smooth if it has not experienced sharp declines or increases in fertility, mortality, and minimal migration. Troughs on the graph show avoidance of specific ages while high points indicate preference of specific ages. However, drawing graphs of the entire population by sex and age presents some limitations because it masks the errors for older age populations in populations where there are more numbers of younger people. The initial data assessment needs to be done using age in single years, after which one can examine the five-year age distributions (Moultrie et al., 2013).

3.4.2 Sex Ratios

Sex ratios are useful in exploring probable data errors for all ages, especially when there are larger populations of younger ages (Moultrie et al., 2013). Sex ratio is widely used to determine the sex composition of a population (Shryock et al., 1976). Sex ratio is usually defined as the number of males per every 100 females in a population. One hundred is the point of balance of the sexes. A sex ratio above 100 shows more males in the population while a sex ratio below 100 shows an excess of females (Siegel and Swanson, 2004). This ratio can be disaggregated by age, which is ratio of male population between age's n and x to females in the same age group. This is represented as $\frac{P_m}{P_f} \times 100$ where P_m is the total number of males and P_f is total number of

females. The overall sex ratio depends on the age structure, patterns of fertility, mortality and migration of the population.

Given the differences between female and male mortality patterns, especially at older ages, the sex ratio will be strongly determined by the age structure of the population. In real life, however, most vital events can be predictably proportioned between males and females. Generally, males outnumber females at birth, but higher rates of male mortality with advancing age offset this pattern. A sex ratio at birth, therefore, usually ranges between 95 and 102 (KNBS, 2010). Thus, failure to observe these typical sex distributions may signify either errors in the data or unusual population characteristics. To obtain a more accurate assessment, researchers normally compare the sex ratio estimated from the data with that obtained in previous years. It will tend to be higher for younger populations and lower for older populations. Due to higher mortality among males, the sex ratio in the total population switches to 95-97. For populations with high levels of sex-selective outmigration (such as male soldiers leaving a country for war), particularly in certain age groups (e.g. aged 15-29), the sex ratio may be even smaller (Moultrie et al., 2013).

In developing countries, the sex ratio at birth (SRAB) is typically around 105. For example, SRAB in Kenya is 102 (KNBS, 2010). As population gets to middle ages of 45 years, the sex ratio declines gradually, and thereafter, declines rapidly because rates of male mortality supersede female mortality rates. Sex ratio graphs should follow the corresponding mortality patterns unless there are outlier occurrences such as over out-migration or immigration of either sex. Sex ratio patterns differ in populations where there is a high level of sex-selective migration as a result of job placements, especially among the youths. If more young males migrate, there is a distinct decline in sex ratios, then a steady increase among the older males as migrants go back home (Moultrie et al., 2013).

Sex ratios were used in this study as an approach of identification of transfers of ages across 15 and 50. This was based on the comparison between sex ratios of ages 10-14 and 15-19 as well as 45-49 and 50-54. If there is no net displacement, it would be expected that these two sex ratios are approximately equal (Pullum, 2005).

3.4.3 Age Ratios

Age ratios are useful in assessing the quality of age-sex data. It is done by comparing age ratios of data under assessment with the standard age-sex ratio values. Calculation of age ratios is used to determine the level of displacements of reported age across age groups (Pullum, 2006). Age ratio is the ratio of total population in a certain age group divided by the sum of populations in that age group and in age group before and after multiplied by a third. This is then multiplied by a hundred. Age ratios are stated for 5-year age groups as shown below:

$$\text{Age ratio} = \frac{{}_5P_a}{\frac{1}{3}({}_5P_{a-5} + {}_5P_a + {}_5P_{a+5})} \times 100$$

Where ${}_5P_a$ is population in a certain age group; ${}_5P_{a-5}$ is population in the age group before; and ${}_5P_{a+5}$ is population in the age group after. When data does not have abnormal variations, there are no irregularities or if changes in the population are quite minimal, age ratios are equal to 100. This is the standard age ratio. The average absolute deviation from 100 indicates the deviation of an age ratio from the standard. The lower the age-ratio deviation, the higher the accuracy of the reported age data. Deviations from 100, without plausible external factors such as migration or calamities that could affect particular age groups, indicate undercounting or displacement errors in the data. Spikes in age ratios could arise if one age group is smaller than the adjacent age groups, resulting to it being below adjacent age groups (Moultrie et al., 2013). Previous studies have shown that mortality increases with age in Africa and so age ratios are expected to decline to below 100 as age increases (Pullum, 2006).

3.4.4 Myer's blended index

Myer's blended index is one of the most commonly used methods of measuring the magnitude of age heaping in single years distribution (Kidane, 2009). Myer's index (1940) has been used by demographers to measure the magnitude of digit preference errors in single year age data. This index measures the concentrations at digits especially, '0' and/or '5'. It considers preference (or avoidance) for ages ending in any digits from zero to nine. The index is obtained by first calculating a "blended" population where all digits are expected to be equal (United Nations, 1956). The index is based on the assumption that the numbers of people vary linearly by age, i.e. the age distribution is an arithmetic progression from 0 onwards, the number of persons of subsequent ages decreases by the same number. This principle indicates that, if there is no age

heaping the population of each terminal digit marks 10 percent of the population. Myer's blended index is obtained by adding up the absolute deviations between the aggregate and theoretical distribution (Pullum, 2006). The value of the index is one-half the sum of the absolute deviations. It includes the division into half the total deviations from the summed up observed proportions. It can be inferred as the lowest proportion of the responses that need to be transferred from one terminal digit to the other in order to get a harmonized division on ages (Siegel and Swanson, 2004). The ideal range of Myer's index is between 0 and 100, where 0 shows no age heaping while 100 shows that every age reported ends in the same digit. Myer's blended index is calculated using an age range of 23-62 years.

Among the measures developed to assess errors of age heaping, Myer's blended index (MBI) is most widely used. The most important upgrading on the MBI is the aspect of blending the population to prevent biases arising from influences of mortality on terminal ages of "0". The MBI has been utilized to quantify the scale of age heaping of censuses, surveys and other population data (Siegel and Swanson, 2004). A large value of the index can result from large deviations at any digit; the preferred digits will be those with the largest positive deviations (Pullum, 2006). Myer's blended index takes this into consideration that in order to identify the amount of age heaping, the first step would be to look at the relative frequency of each final digit 0, 1, 2…9. Most age distributions have fewer and fewer numbers as age increases because of the combined effects of mortality and a history of population growth. As a result, there tends to be more cases with final digit 0 than final digit 1; more cases with final digit 1 than final digit 2, and so on (Pullum, 2005).

3.4.5 Whipple's index

The Whipple's index is a method used to quantify the extent of age misreporting by detecting avoidance or prevalence of a particular terminal digit. It is calculated as a ratio of persons aged 25, 30, 35, 40, 45, 55, and 60 as one fifth or one tenth of the people reported in ages between 23 and 62. It excludes early childhood ages and extreme old ages which are influenced by other reporting errors that do not include age preference (U.S. Bureau of the Census, 1985). The expected values of Whipple's index range from 100 to 500 due to the assumption that the number of persons either increase or decrease linearly as age advances. When age reporting is

accurate (no heaping or avoidance at any ages), then it is expected that the sum of persons with ages ending in 0 and 5 would be exactly 1/5th of the total persons in the ages between 23 to 62. If all people reported their ages with terminal digits of 0 and 5, this index would be 500. However, if there is dislike or avoidance for terminal digits of 0 and 5, then this index would vary between 0 and 100. The data is interpreted as highly inaccurate when this index is more than 175, it is inaccurate when the index falls between 125 and 174.9, approximate if it is between 110 and 124.5, fairly accurate if it is 105 to 109.5 and highly accurate if it is less than 105 (United Nations, 1973).

Myer's blended index is theoretically comparable to the Whipple's index and they provide results that are alike in identifying the levels of age misreporting. Both indices are used to test the accuracy of single year data. The Whipple's index evaluates the influence of preference on ages ending with digits 0 and 5 only. It is also an effective method of assessing accuracy of digit preference, and it is advantageous because it is computed easily. Myer's index assesses the avoidance and preference of all terminal digits. Myer's model evades the biases presented by Whipple's model that does not consider the fact that mortality affects terminal ages where 0 is more than other numbers 1 to 9 (Nasir and Hinde, 2014). The Myer's model avoids this form of inconvenience by first computing the "blended" population where close to equivalent sums are expected for all digits. In that case, the "blended" sums of each digit need to be close to ten percent of the final total. The deviations are summed up irrespective of the negative or positive signs and the final sum makes the Myer's index (Siegel and Swanson, 2004).

The major shortcoming of Whipple's index is that it assesses digit preference of only two terminal digits, which are 0 and 5 (United Nations, 1955). The strength of Myer's blended index is that it covers the shortcoming of Whipple's index by assessing the avoidance and preference of all terminal digits from 0 to 9. For that reason, this study used Myer's blended index to examine age heaping in single years.

3.4.6 UN Age-Sex Accuracy Index

The United Nations age-sex accuracy index is used to evaluate the quality of five year age groups (United Nations, 1956). Five year age grouping reduces errors due to misreporting or age

shifting within the age groups. The index combines measures of accuracy of age-group data for both sexes separately measuring the accuracy of sex ratios of various age groups. Calculations entail dividing the population in a specific 5-year age group by the average population of the two adjacent 5-year age groups, multiplied by 100. Sex-ratio differences are calculated as the successive differences in sex-ratios between one age-group and the next one.

If there were fewer variations in births, deaths and migration, the population in the three consecutive age groups need to make a close to linear pattern, and hence age ratios must be approximately 100. The deviations from 100 indicate the extent of misreporting in the age group and the sum of deviations (irrespective of sign) gives a measure of accuracy or misreporting of age data. Similarly, the index considers the sum of deviations (irrespective of the sign) in reported sex ratios of consecutive age-groups. The index consists of the totals of female and male age ratio scores added onto three times the sex ratio scores for ages 0-14 through 65-69. It is based on the assumption that differences in sex ratios according to age should approximate to zero. Large variations from 100 could occur as a result of populations experiencing a lot of gender-specific migration. Without abnormal gender-specific events, variations from 100 indicate possible data errors.

The UN age-sex accuracy index is then the total of (a) average variations of male age ratios from 100, (b) average variation of female age-ratios from 100, and (c) three times the average of age variations in reported sex ratios. This index allows one to compare datasets by sex. This is because it evaluates age ratio scores and sex ratio scores together. The UN score classifications are put in three classes: One is accurate, where the index is < 20; two is inaccurate where the score is 20 and 40; and three is highly inaccurate where the score is > 40. It should be noted that when interpreting the UN index, the age-sex evolving from demographic changes should be considered carefully since the index is not able to detect inaccuracies occurring from unnatural changes.

Shryock et al. (1976) observed that the main limitation of the index is that it fails to consider expected reduction in sex ratio as a result of age increase, age abnormalities due to epidemics, migration and wars and usual variations in deaths and births. Another limitation they observed is

that the index uses age ratio that does not include the central age group thus exposing it to upward bias. It should also be noted that a lot of weight is put to sex ratio within the formula (Shryock et al. 1976). The joint scores may be affected by differentials in sex ratio that is in favor of females because males experience high mortality rates.

The United Nations age-sex accuracy index was used in this study because of its use of 5-year age group data for males and females to assess the quality of age and sex data in a given population. The index utilizes age ratio and sex ratio scores to produce a merged score which indicates the level of age-sex data quality of a population (Shryock et al. 1976; Arriaga et al. 1994).

This study used the Population Analysis Spreadsheet (PASEX), which is an easy to use MS Excel spreadsheet developed by the US Census Bureau to assess the quality of data. The PASEX AGESEX spreadsheet was used to generate age ratios, sex ratios and UN accuracy indices while the SINGAGE spreadsheet generated the Myer's index. Data was subjected to smoothing using the AGESMTH spreadsheet.

3.5 Adjustment of Age Data

Data adjustment is done to minimize errors found in data to approximate a more accurate set for analysis. There is no generalized adjustment solution for all populations. The smoothing technique used depends on the age-sex data errors, thus age structure should be assessed prior to making a decision on the appropriate method for adjustment. A graph of the age and sex distributions should be made before making any decision about whether or not smoothing is required and which technique would be appropriate for the particular country's situation. In general, a regular saw-tooth pattern across successive age groups provides a good rationale for smoothing. Comparisons among successive surveys and knowledge of past trends of mortality, fertility, and migration also helps in appraising the accuracy of age and sex data of the population (Arriaga et al., 1994).

The method applied for adjusting data should depend on the degree of inaccuracy. While slightly accurate data is modified minimally, inaccurate data is modified drastically. There are different

techniques of smoothing data, those that smooth without modifying the totals such as Karup-King-Newton, Carrier-Farrag, Arriaga formula, as well as those that modify the totals such as United Nations and Arriaga's strong smoothing formula. The United Nations methods, Karup-King Newton, Carrier-Farrag, and the Arriaga method are applied to achieve light smoothing and the strong method achieves stronger smoothing such as. Light or "slight" smoothing gently modifies irregularities in the age structure while strong smoothing modifies most irregularities, and, therefore, more likely to modify features which may represent actual facts instead of errors. Light smoothing formulas give the greatest weight to what was reported for the age group in question and smallest weight to adjacent age groups. Strong smoothing formulas give greater weight to adjacent age counts and/or over wider age intervals. The resulting pattern does not follow the contours of reported data as well as lighter smoothing (Siegel and Swanson, 2004).

The techniques which use enumerated population data in all 10-year age groups produce almost the same results. The major variation is that some do not smooth the initial and final 10-year age groups within the population. The Karup-King-Newton and Carrier-Farrag (Carrier and Farrag, 1959) techniques cannot detach the initial or final 10-year populations, but the Arriaga technique does that. The Strong and Arriaga techniques utilize the whole population from age 0 - 79 to assess quality. The Arriaga method conducts light smoothing through the modification of populations in all 10-year age groups. The Strong technique conducts modifications of 10-year groups with the assumption that age could be wrongly reported as far as 10-years. Light smoothing techniques such as United Nations, Karup-King Newton and Carrier-Farrag techniques adjust data falling between 10-69 years, while the Arriaga method adjusts data between 0 to 79 years (Arriaga et al., 1994).

3.5.1 Carrier-Farrag Technique

The Carrier-Farrag technique is found on a theory that the connection of one 5-year age group and to its constituent 10-year age group is an average of similar relationships in three consecutive 10-year age groups (Carrier and Farrag, 1959). The formula is as below;

$${}_5P_{x+5} = {}_{10}P_x / [1 + ({}_{10}P_{x-10} / {}_{10}P_{x+10})^{1/4}] \text{ and}$$

$${}_5P_x = {}_{10}P_x - {}_5P_{x+5}$$

Where:

${}_5P_{x+5}$ represents the population at ages x+5 to x+9;

${}_{10}P_x$ represents the population at ages x to x+9; and

${}_5P_x$ represents the population at ages x to x+4.

3.5.2 Karup-King-Newton Formula

The Karup-King-Newton formula takes the assumption of a quadratic interaction of three successive 10-year age groups in the data. Like the Carrier-Farrag and UN techniques, Karup-King-Newton formula accepts the reported population of each 10-year age group hence produces similar results (Siegel and Swanson, 2004).

${}_5P_x = ½ ({}_{10}P_{x)} + 1/16 ({}_{10}P_{x-10} - {}_{10}P_{x+10})$ and

${}_5P_{x+5} = {}_{10}P_x - {}_5P_x$

Where:

${}_5P_x$ is the first of two 5-year age groups comprising a 10-year age group ${}_{10}P_{x.}$

3.5.3 The Arriaga Formula

The Arriaga's formula is based on a model that applies 2 degree polynomial through the middle part of three successive 10-year age groups and then presents results as 5-year age groups (Arriaga et al., 1994). If 5-year age groups are used, the following formulas are utilized;

${}_5P_{x+5} = (- {}_{10}P_{x-10} + 11 {}_{10}P_x + 2 {}_{10}P_{x+10}) / 24$ and

${}_5P_x = {}_{10}P_x - {}_5P_{x+5}$

Where:

${}_5P_{x+5}$ is the population ages x+5 to x+9;

${}_{10}P_x$ is the population ages x to x+9; and

${}_5P_x$ represents the population at ages x to x+4.

To smooth the oldest and youngest age groups in the population, the 10-year age groups are divided into outermost age groups (the oldest and the youngest). To assess the young age groups, the formula below is applied:

${}_5P_{x+5} = (8 {}_{10}P_x + 5 {}_{10}P_{x+10} - {}_{10}P_{x+20}) / 24$ and

${}_5P_x = {}_{10}P_x - {}_5P_{x+5}$

To assess older age groups, the coefficients in the formula are reversed as shown below:

$_5P_x = (-\ _{10}P_{x-20} + 5\ _{10}P_{x-10} + 8\ _{10}P_x) / 24$ and

$_5P_{x+5} = {}_{10}P_x - {}_5P_x$

3.5.4 Arriaga's Strong Formula

If more aggressive smoothing is desired (Arriaga et al., 1994), this can be achieved with this formula. An easily applied strong smoothing procedure is as follows: (a) first step involves smoothing the 10-year age groups; (b) adjusting the outcome of the population into smoothed ages; and (c) separating the adjusted 10-year age groups to 5-year age groups by utilizing any of the techniques described above such as Arriaga. The following is Arriaga's Strong Smoothing Formula:

$_{10}P'_x = (_{10}P_{x-10} + 2\ _{10}P_x + {}_{10}P_{x+10}) / 4$

Where: $_{10}P'_x$ is smoothed population of ages x to x+9.

3.5.5 The UN 5-point Formula

The United Nations formula is a normal procedure applied to age data of developing countries for smoothing. It is an application of curves that are mathematical in nature to enumerated age data and making use of curves to get approximate populations for all age groups. The UN came up with a formula for smoothing 5-year age groups beginning from ages 10 to 74. The formula assumes that 5-year age distributions are obtained till age 85. The United Nations procedure has assumptions that total losses and gains of adjacent age groups are stable. Five different ages are used to obtain an adjusted/smoothed age for the middle age group (United States Bureau of Census, 1985). This formula produces light smoothing, and if age misreporting is severe, light smoothing probably is not sufficient. The difference between totals of unadjusted populations and that of adjusted populations is quite minor (Carrier and Farrag, 1959). The UN 5-point smoothing formula of data adjustment is preferably used to adjust data that is less accurate. The formula involves taking data for two preceding age groups, and data for the two following age groups and the age group being adjusted. United Nations formula is a follows:

$_5P'_x = (1/16)\ (-\ _5P_{x-10} + 4\ _5P_{x-5} + 10\ _5P_x + 4\ _5P_{x+5} - {}_5P_{x+10})$

Where: $_5P'_x$ represents the smoothed population ages

P_X represents the population in age range x to x+4

Arriaga and Associates (1994) found that there were minor differences for results produced by various procedures. Even if smoothing produces more plausible age distributions, it may not improve distortions in sex ratios by age (and vice versa). The population age distribution may not need to be fully smoothed across all age groups if only part of it is considered problematic. If underreporting exists at a particular age, instead of smoothing, one may need "filling". The formulas work to reduce errors and irregularities to put data into credible forms for analysis. Errors could be also rollover from past errors or fluctuations of demographic data or from irregularities of current data. Therefore, data adjustment must not be applied for data that is fairly accurate. If applied in such a situation, smoothing formulas could introduce more errors when they eliminate minor irregularities that truly reflect the population composition. In the same way, if data is inaccurate yet previous demographic trends show irregularities, it is appropriate to adjust figures for a few but not all age groups (Arriaga et al., 1994).

The Arriaga method was used in this study because its model is applicable to adjust African data, more so Kenyan data. This is because, unlike other methods of data adjustment, it adjusts data from 0 to 79 years by running a second degree polynomial through the centre of three consecutive 10-year age groups. It adjusts the first and last 10-year age groups and incorporates the results into 5-year age groups which makes it easy for interpretation and use in African demographic studies. In addition, the Arriaga method detects errors resulting from digit preference and age misreporting, which characterize most African data (Pullum, 2006) and adjusts accordingly (Arriaga et al., 1994).

The calculations in these techniques are quite complex. However, there are provisions for computer software that produce smoothing for all age distributions. The most familiar software used is PASEX built by the US Census Bureau. The software is an Excel spreadsheet of Windows. The spreadsheet used for smoothing is referred as AGESMTH. It contains a worked-out example of the applications of five procedures of age smoothing, that is, by Carrier-Farrag, Karup-King Newton, Arriaga as well as the UN method, and the Strong method. Among these, the most commonly used are the UN and the Strong methods (Arriaga et al., 1994).

CHAPTER 4

ASSESSMENT OF QUALITY OF DATA

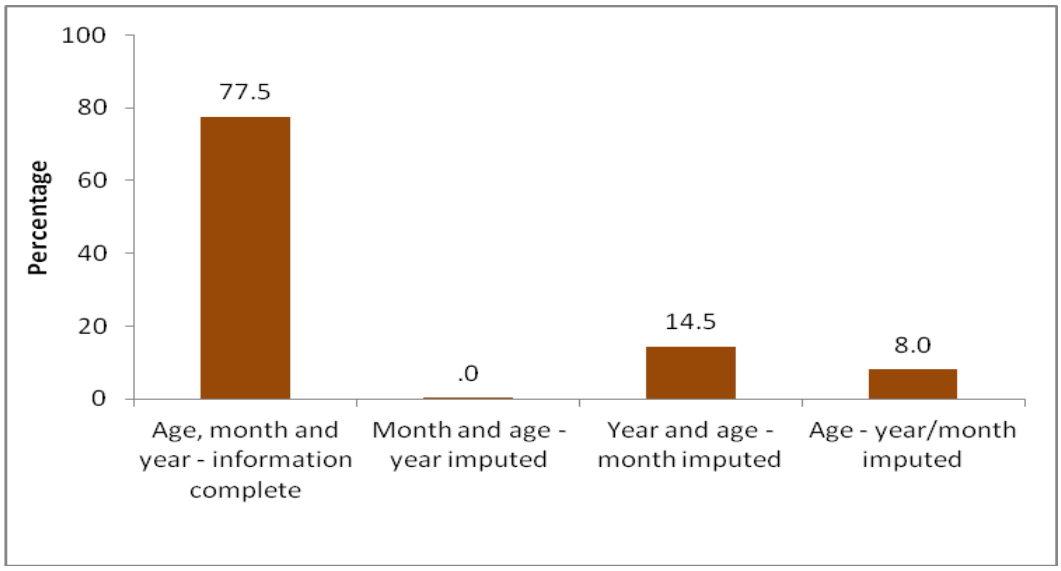
4.1 Introduction

This chapter presents the results of assessment conducted on completeness of birth date of females and date of birth of children ever born. It presents results of age and sex data quality assessment of all household members of the 2014 KDHS. The total population in the sampled households was 153,822, of which females were 78,096 and males were 75,726. The Myer's index, age ratios, sex ratios and United Nations age-sex accuracy index were applied to assess the data quality of sex and age. Data was adjusted using the Arriaga light smoothing method.

4.2 Completeness of Reporting of Data on Date of Birth

Three variables that detail the age and date data elements are found in the individual female questionnaires of the KDHS. The respondent is asked to provide three items related to age: her age in whole years, birth year, and birth month. Minimally, women need to report their age, month or a birth year. Some respondents give less than the three items. Sometimes when all items are given, inconsistencies are found which need data to be imputed for either one or more items. Completeness assessment determines the percentage of respondents in a survey who failed to give all three data items or those with data had inconsistencies. Figure 4.1 shows the percentage of completeness of date of birth reported by females interviewed.

Figure 4.1: Percentage of completeness of date of birth, 2014

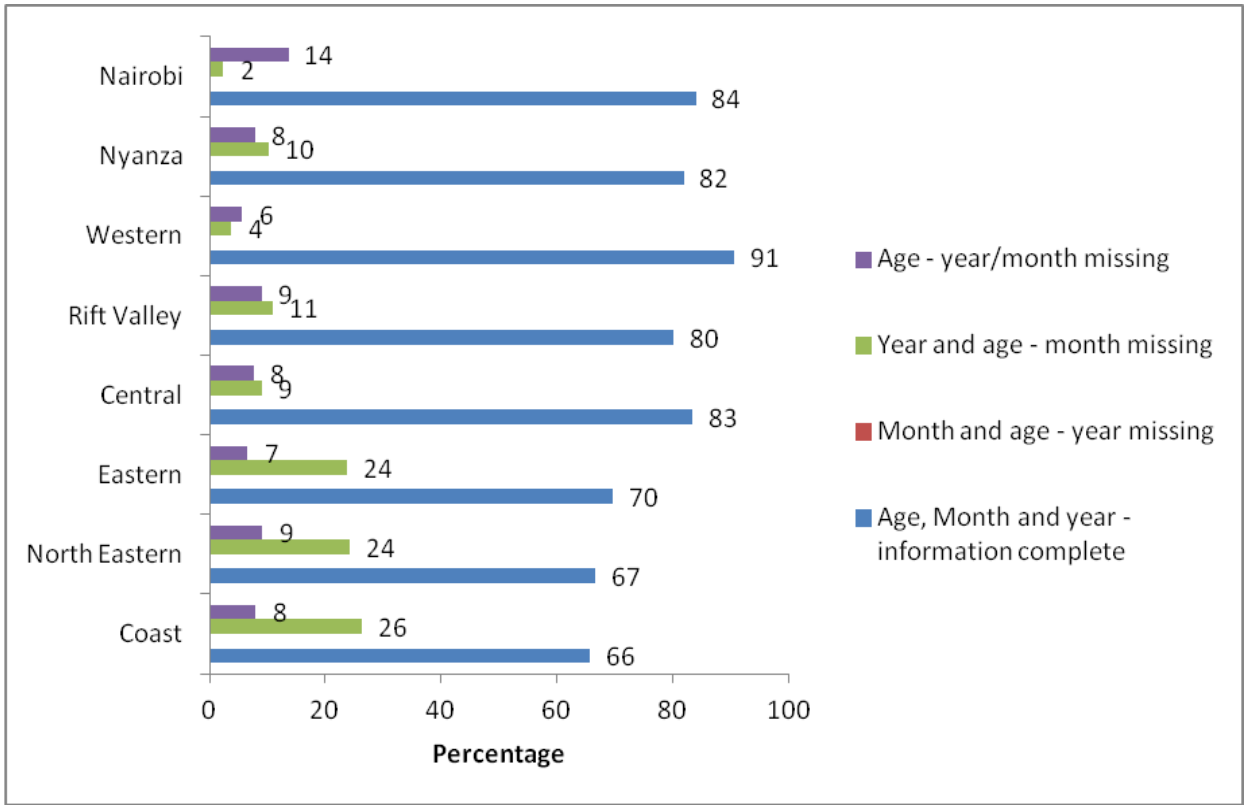


About 78 percent of the females interviewed reported complete date of birth with month and year. While the other 23 percent missed either the year or month but reported their ages, with majority of them (15%) missing the month of birth. For most respondents who reported incomplete date of birth, they did not know the month they were born. For respondents who missed an element of date either month, year yet had reported age, the data was imputed to provide a complete and valid date of birth.

4.3 Completeness of Reporting of Data on Date of Birth by region

The same assessment was conducted for completeness of reporting of date of birth by region as shown in Figure 4.2 below.

Figure 4.2: Completeness of date of birth for women interviewed by region, 2014



Coast, North Eastern and Eastern regions showed high levels of month missing or both year and month. Although some dates, months and years were missing, ages had been provided and thus the missing elements were imputed.

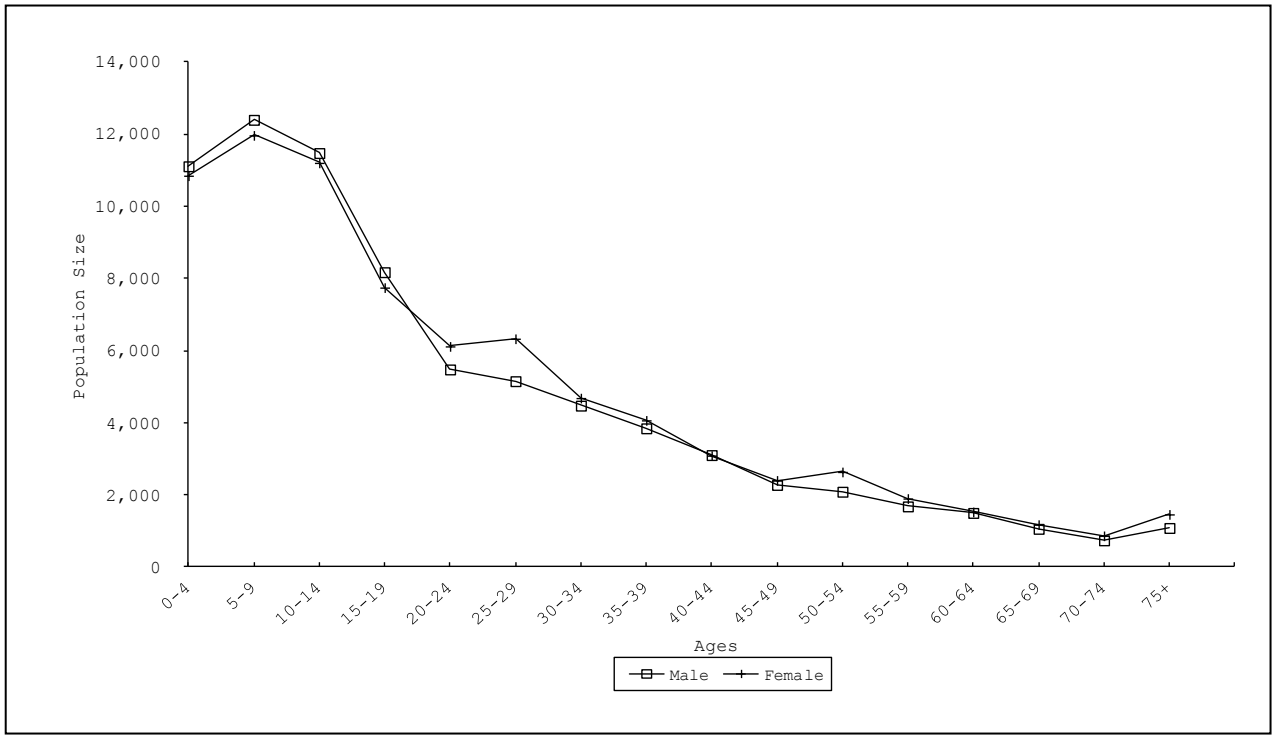
4.4 Completeness of Reporting of Date of Birth for Children Ever Born

The completeness levels of date of birth reported for children ever born was assessed. Ideally, a woman provides information about their children, whether alive or dead as well as their date of birth. The mother provides three items related to age: child's age, birth month and birth year. At a minimum, there should be an age or a birth year reported. About 99 percent of respondents reported complete date of birth with month and year for their children. While the other one percent missed either the year or month but reported their ages. Since 99 percent respondents reported month, year or age, the data could be considered complete. Generally, there was good reporting of date of birth for children ever born by the female interviewees in all regions. Although some dates, months and years were missing, ages had been provided and thus the missing elements were imputed.

4.5 Population Distribution

Data was analyzed to examine the population distribution by sex and age. Figure 4.3 shows the sex and age characteristics of the population.

Figure 4.3: Population by age and sex, 2014



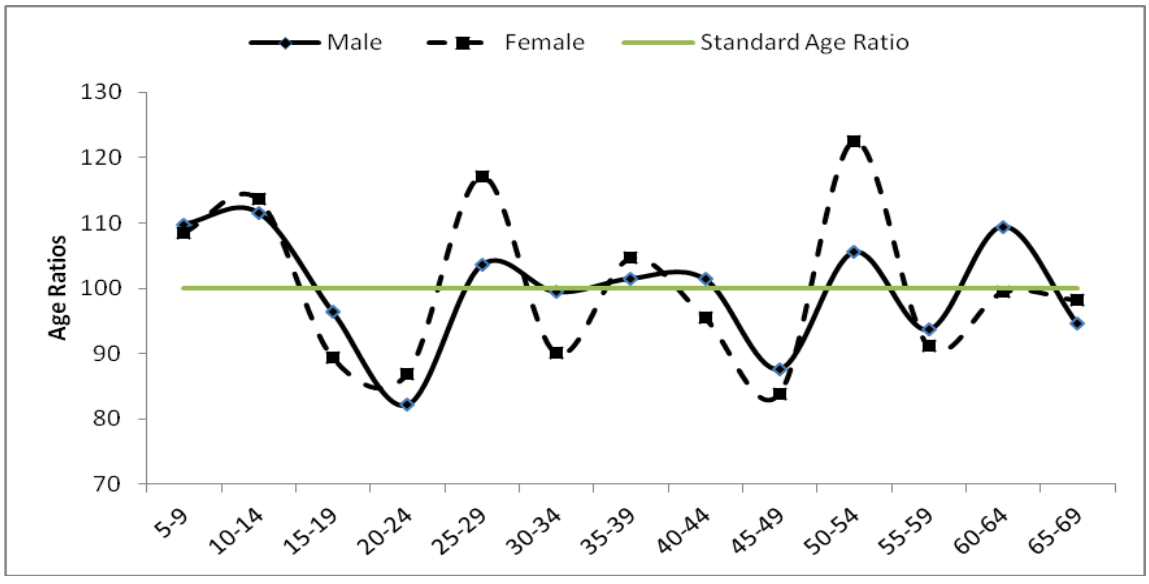
The population below 20 years constituted about 55 percent of the total, which shows a young population. About 6 percent of the population was above 60 years. This is consistent with the 2009 population census data which showed the population below 20 years as 54 percent and that above 60 years as 6.3 percent (KNBS, 2012).

4.6 Age and Sex Ratios

4.6.1 Age Ratios

As detailed in Chapter 3, age ratio deviations from 100 indicate the extent of misreporting in an age group and the sum of deviations (irrespective of sign) gives a measure of accuracy or age misreporting. It is assumes that differences in sex ratio according to age should approximate to zero. An age ratio under or over 100 could imply that persons were misclassified to an age group that is adjacent to their actual age. Figure 4.4 shows the age ratios by age and sex.

Figure 4.4: Age Ratios by age and sex, 2014



The average age ratio score deviation from 100 was 6.8 for males and 10.1 for females (Table 4.1 on page 37). The data can, therefore, be considered fairly accurate since it deviates slightly from 100. Figure 4.4 shows fluctuations in the age ratios for both male and female age ratios. For male age ratios, high age ratio points were found in the age groups 10-14 and 60-64 and low points in the age groups 20-24 and 45-49. For female age ratios, fluctuations are found in almost all age groups although with varying degrees of deviations. The high age ratio points in females

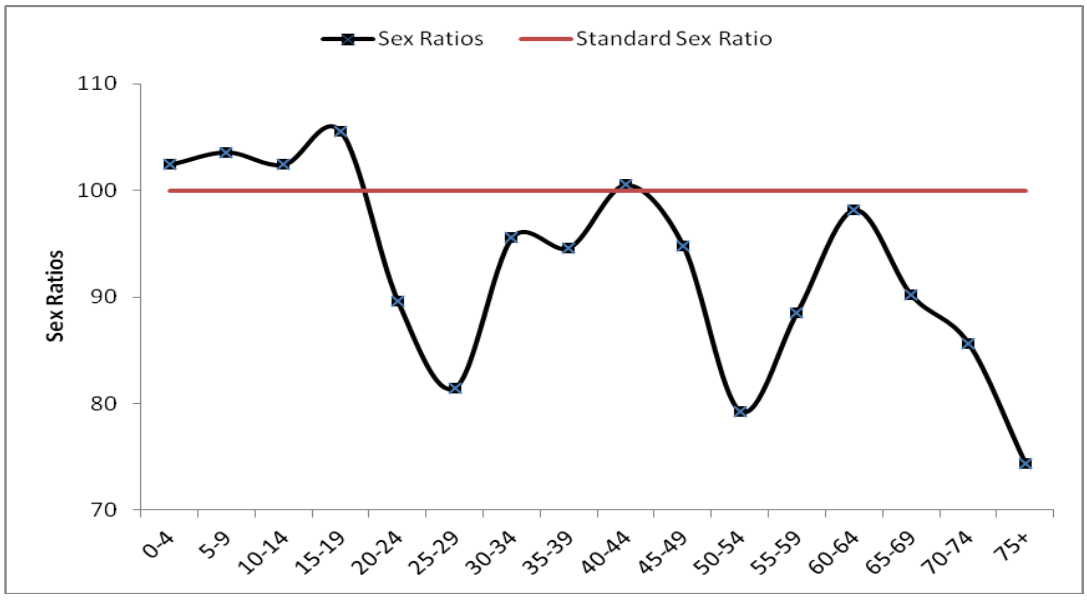
were in the age groups 10-14, 25-29 and 50-54 and low points were in 20-24 and 45-49. The fluctuations indicate major variations of population frequencies of nearby age groups. The maximum age ratio for males is 111.6 in the age group 10-14, and for females is 122.6 in the age group 50-54. Then the minimum age ratio for males is 82.2 at age 20-24 and for females is 83.9 in the 45-49 age group.

Irregularities of these age ratios could be as a result of displacement from 15-19 to 10-14 and from 45-49 to 50-54. This could be attributed to misreporting of ages by interviewers misreporting age data in order to minimize their work of administering questionnaires to respondents age 15-19 and 45-49. For females, there could have been transfers at age 50 and less at age 15. This could probably occur if interviewees were ignorant of their true ages or that interviewers were avoiding them since they have more births and hence more data to report.

4.6.2 Sex Ratios

As detailed in Chapter 3, sex ratio is the number of males for every 100 females in the population. One hundred is the balancing point of both sexes. A sex ratio more than 100 shows more males while a lower than 100 shows more females. Figure 4.5 shows the sex ratios by age group.

Figure 4.5: Sex ratios by age, 2014



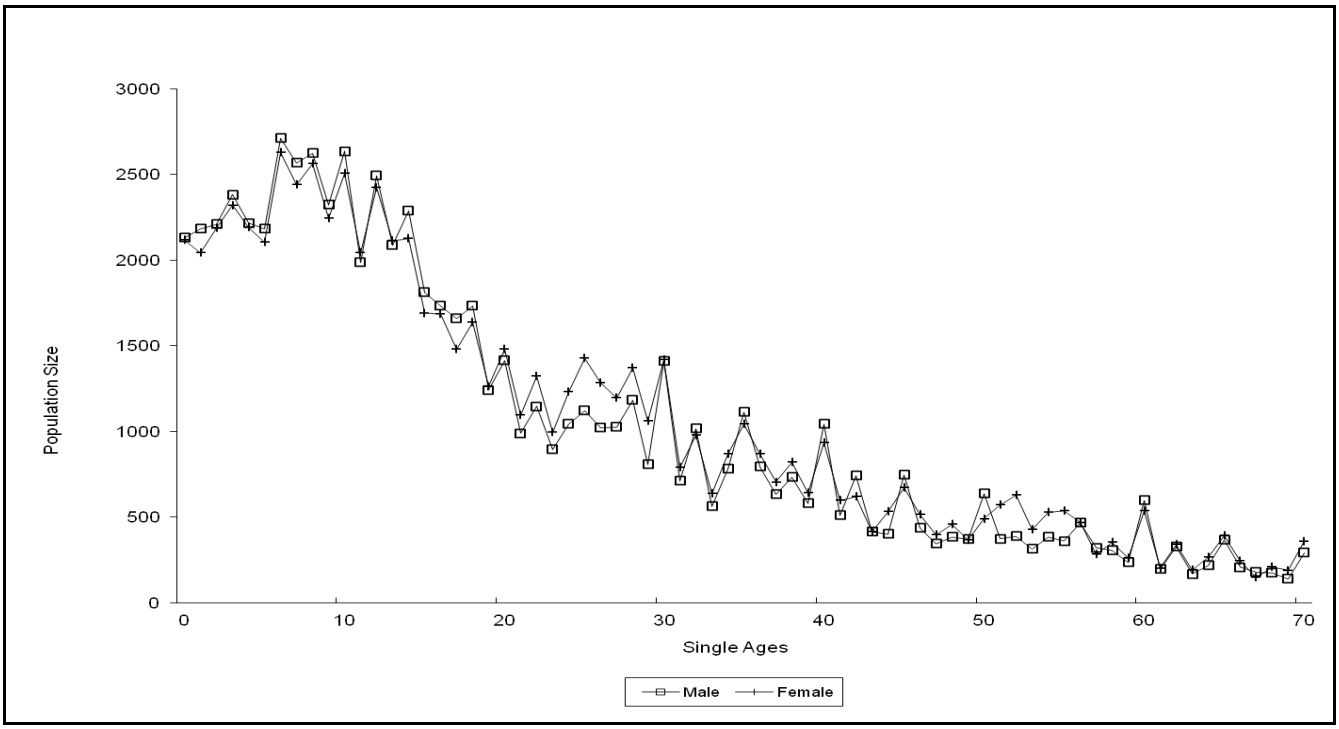
The overall sex ratio was 92.6 (Table 4.1 on page 37). Figure 4.5 shows that the reported sex ratios at younger ages are higher than those in older ages. The sex ratio for the ages 0-4 was 102 and reaches its highest in age 15-19 at 106. This indicates that more males than females reported their ages to be in these age groups. A dip was experienced in ages 25-29 with sex ratio of 81.4, 50-54 with sex ratio of 79.3 and 75+ with sex ratio of 74.3. This indicates that the population had more females than males at older ages. The rapid highs and lows show sex distribution that signifies inaccuracy of data or abnormal population traits. Overall, these scores are an indication of some errors in the reported data. These numbers suggest a net deficit of women in age group 15-19, and an excess of women in age group 25-29, 50-54 and above 75. The more negative or positive the sex ratio is, the greater the indication of misreporting.

4.7 Assessment of Age Heaping

4.7.1 Population Distribution

On assessment of age heaping, the 2014 KDHS data does not follow a smooth linear graph, as shown in Figure 4.6.

Figure 4.6: Population by single ages and sex, 2014

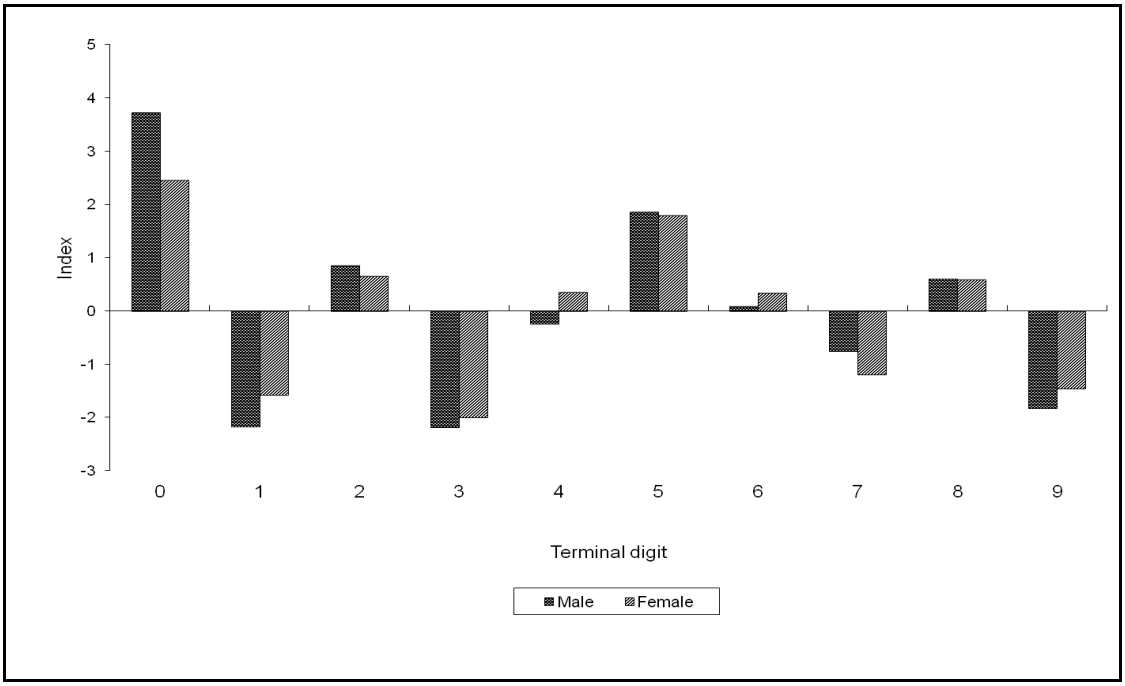


Data fluctuations were experienced for both sexes. The peaks in Figure 4.6 indicate age heaping, for example, at ages 10, 30, 35, 40, 45, 50, 60, 70 and 80.

4.7.2 Digit Preference

The Myer's blended index is applied to assess single year age data by determining the magnitude of digit avoidance or preference for digits ending with 0, 1, 2 … 9. According to this index, age preference scores fall between 0 and 90. When there is no age heaping, the total population in each terminal age digit 0 to 9 denotes 10 percent of the population. Directions of the indices are an indication of digit preference or avoidance while the numbers show the strength of preference or avoidance. The results of assessment of digit preference using Myer's index for the 2014 KDHS data are presented in Figure 4.7.

Figure 4.7: Myer's Preference by Digit, 2014



From the assessment, the overall Myer's index score or digit preference was 13.1. Figure 4.7 shows the variations of proportions of blended population of every terminal digit or the Myer's index scores. Males recorded a higher rate of digit preference with an overall score of 14.3 than females with 12.4. For both males and females, the digit 0 was more preferred than digit 5.

Digits that were avoided by both sexes include 1, 3, 7 and 9. These results are consistent with previous studies (Bwalya et al., 2015; Pardeshi, 2010).

4.7.3 National UN Age-sex Accuracy Index

As detailed in Chapter 3, the United Nations age-sex accuracy index is used to evaluate the quality of enumerated sex and age data in 5-year age groups. It combines measures of accuracy of age group data for both sexes with the accuracy of sex ratio scores of various age groups. The United Nations age-sex accuracy index classifies population age-sex structures into three categories: 1) accurate – if score is < 20; 2) inaccurate – if score is 20 - 40; and 3) highly inaccurate – if score is > 40. Table 4.1 shows the results of age-sex accuracy index for males and females for the 2014 KDHS data.

Table 4.1: UN Age-sex Accuracy Scores, 2014

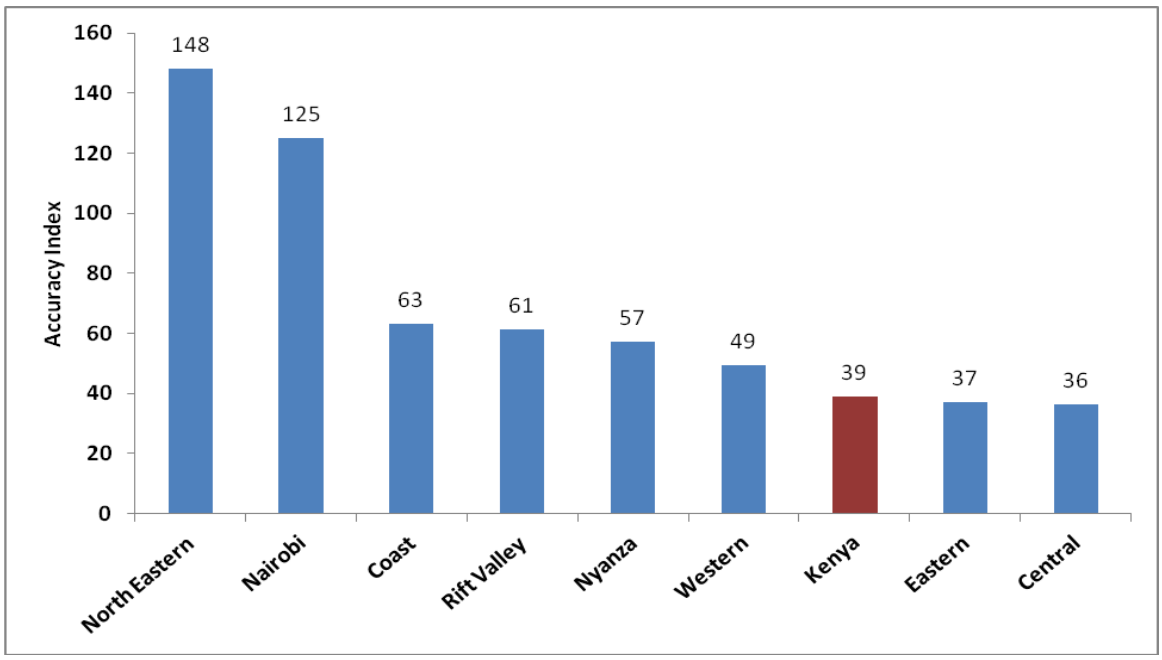
Indicator	Score
Males age ratio score	6.8
Females age ratio score	10.1
Sex ratio score	92.6
UN age-sex accuracy index	39.1

The UN age-sex accuracy score is 39.1, which, according to the classification, shows inaccurate data. The inaccuracy is slightly higher in female age data than for males. It is recommended that data is subjected to smoothing if the accuracy index is above 20. Therefore, this data was subjected to light smoothing in order to make it more accurate and produce a data set that can be analyzed to produce reliable results. The Arriaga method was employed for smoothing.

4.7.4 UN Age-sex Accuracy Index by Region

Figure 4.8 shows the UN age-sex accuracy index by region in descending order and in comparison to the Kenyan index score.

Figure 4.8: United Nations age-sex accuracy index by region, 2014

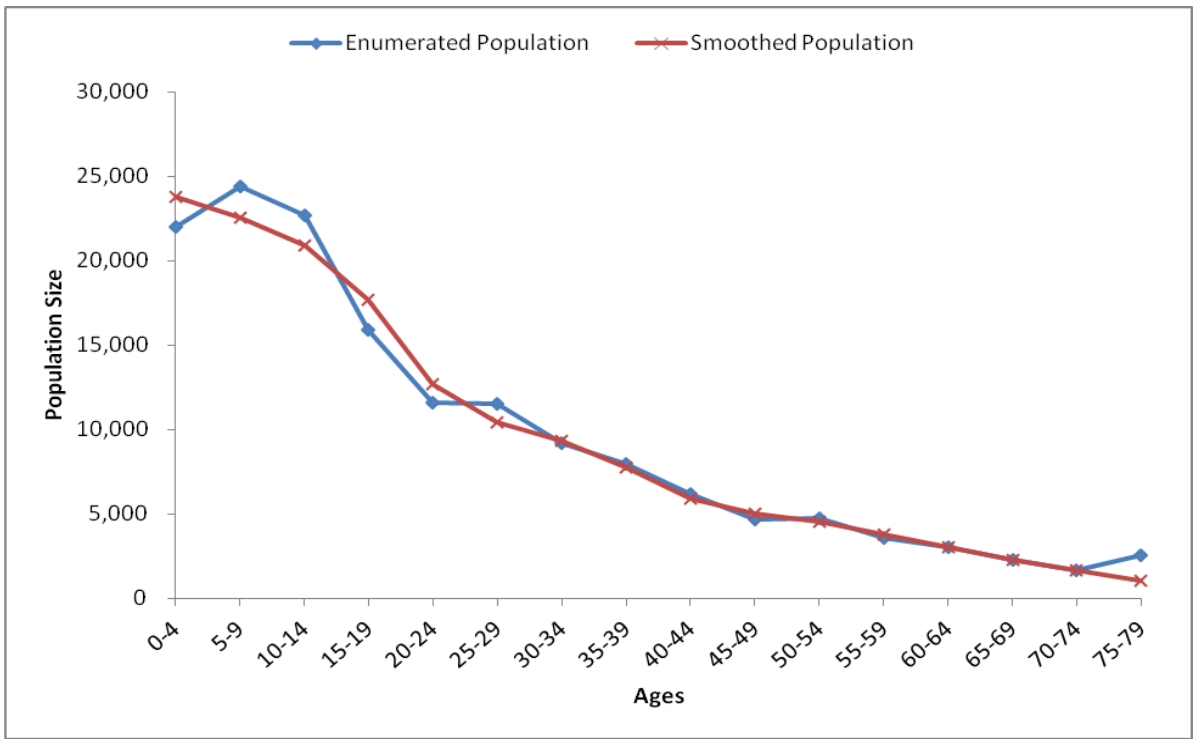


The national accuracy score was 39. The region with the highest accuracy index is North Eastern followed by Nairobi. Eastern and Central regions recorded indices that are close to the overall national index. Figure 4.8 indicates that reported age data of six regions was highly inaccurate and that age data from Eastern and Central regions was inaccurate.

4.8 Data Smoothing

After the assessment of age and sex data quality by applying the UN index, it was found that data was inaccurate. Myer's index also showed digit preference at terminal digits 0 and 5. As per the UN classification, smoothing is applied to data that has a score of above 20, which is termed as inaccurate. Based on these findings, a light smoothing formula was applied to improve the quality of data. The formula used was the Arriaga method. Figure 4.9, shows the difference in population distribution of enumerated and smoothed data.

Figure 4.9: Enumerated versus smoothed population data, 2014



Data showed some changes after smoothing. The population reduced after smoothing for age 5-9 by 1844, 10-14 by 1765, 25-29 by 1091 and 75-79 by 1536. This is an indicator of reporting errors during enumeration. After smoothing, the resulting line graph showed a smoother curve and an even distribution of ages than the enumerated data.

There was age misreporting in the age data of the 2014 KDHS making it inaccurate. The findings are consistent with previous study findings on quality assessments on Kenyan data. This gives credibility to the methodologies used in this study.

CHAPTER 5

SUMMARY, CONCLUSION AND RECOMMENDATIONS

This chapter gives the summary of the results, conclusion and recommendations.

5.1 Summary

From this study, it emerged that the 2014 KDHS age data was, to a small extent, incompletely reported for date of birth for women and children ever born. Some dates, months and years were missing, but since ages had been provided, the missing elements were imputed. Age data was inaccurately reported with large fluctuations in age ratios for males and females, which could be an indication of persons in various ages being carried across age group boundaries or persons misreporting their own ages for various reasons. This compromised the quality of data. The Myer's index revealed digit preference of age ending with '0' and ' 5' where the former digit is more preferred than the latter for both male and female respondents. Moreover, females showed higher age misreporting than male respondents. Ages ending with 1 have the highest digit avoidance followed by those ending with 3, 7 and 9.

The UN age-sex accuracy index revealed general inaccuracy in the 2014 KDHS age data with an index score of 39.4, which indicated inaccurate data and hence required smoothing. Six regions recorded highly inaccurate data while two regions had inaccurate data. North Eastern and Nairobi regions had a high inaccuracy level while Central and Eastern regions were inaccurate. The Arriaga smoothing method was applied to adjust the data. Application of smoothing methods resulted in slight changes of sex and age data in various age groups. A new data set was generated, one that is more reliable for use by demographers and other stakeholders for further analysis.

5.2 Conclusion

Quality data is fundamental in producing accurate inferences in surveys. However, it is difficult to get high quality data, especially in Africa. This study provided evidence that quality issues exist in the 2014 KDHS data, characterized by incompleteness and inaccuracy of date of birth and current ages reported. The study findings suggest that additional strategies are required to improve the reporting of the future KDHS data. Only highly accurate data would help Kenya

provide accurate population estimates and projections for policy making, decision making, resource allocation and other uses.

5.3 Recommendations

Based on the above findings, the following policy and research recommendations are proposed.

5.3.1 Policy Recommendations

The study shows evidence of age misreporting that could have arisen due to reporter or interviewer bias. For respondents' bias to be addressed, agencies involved in the KDHS work need to also engage the public in understanding the importance of accurate reporting. This can be done by creating policies aligned to supporting mass awareness and educating people on how false reporting during surveys and censuses distorts information derived from data and how their lives are likely to be affected.

Since the results showed data anomalies from age misplacement and transfers, a more thorough supervision should be conducted during field work to ensure ages are recorded appropriately. This would be useful in reducing interviewer biases.

5.3.2 Research Recommendations

This study has focused on assessing completeness of date of birth, age and sex data quality of the KDHS data. Further in-depth studies should be conducted to assess the role of interviewers in influencing misreporting, especially of age and date of birth accordingly. For example, the extent to which interviewers bear responsibility for age transfers, the probability that an interviewer would shift an age or birth date is dependent on the respondent having low, no education or is less certain about the true age.

REFERENCES

A'Hearn, B., Baten, J., & Crayen, D. (2009). Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital. *The Journal of Economic History, 69*(03), 783-808.

Arriaga, E. A., Johnson, P. D., & Jamison, E. (1994). Population Analysis with Micro-Computers, Volumes I and II. Washington D.C.

Bachi, R. (1951). The tendency to round off age returns: measurement and correction. *Bulletin of the International Statistical Institute, 33*, 195-221.

Bello, Y. (2012). Error Detection in Outpatients Age Data using Demographic Techniques. *International Journal of Pure and Applied Science and Technology*, *10*(1), 27 – 36.

Bocquier, P., Madise. N., & Zulu, E. (2011). Is there an Urban Advantage in Child Survival in Sub-Saharan Africa? Evidence from 18 Countries in the 1990s. *Demography, 48*(2), 531-558.

Bwalya, B. B., Phiri, M., & Mwansa, C. (2015). Digit preference and its implications on population projections in Zambia: Evidence from the census data. *International Journal of Current Advanced Research, 4*(5), 92-97.

Byerlee, D., & Terera, G. (1981). Factors Affecting Reliability in Age Estimation in Rural West Africa: A Statistical Approach. *Population Studies, 35*(3), 477-90.

Caldwell, J., & Igun, A. A. (1971). An Experiment with Census-type Age Enumeration in Nigeria. *Population Studies, 25*(2), 287-302.

Caldwell, J. (1966). A Study of Age Misstatement among Young Children in Ghana. *Demography, 3*(2), 477-490.

Carrier, N. H., & A. M. Farrag. (1959). The Reduction of Errors in Census Populations for Statistically Underdeveloped Countries. *Population Studies, 12*, 240–285.

Central Bureau of Statistics. (2002). Kenya 1999 Population and Housing Census Volume VII: Analytical Report on Population Projections. *Ministry of Finance and Planning*, Kenya.

Cleland, J. (1996). Demographic Data Collection in the Less Developed Countries. *Population Studies, 50*, 433-450.

Denic, S., Khatib, F., & Saadi, H. (2004). Quality of age data in patients from developing countries. *Journal of Public Health, 26*(2), 168-171.

Edward, G.S., & Wicks, J.W. (1974). Age heaping in recent national censuses. *Bio-demography and Social Biology, 21*(2), 163-167.

Ewbank, D.C. (1981). Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis, Committee on Population and Demography. *National Academy Press*, Washington, D.C.

Economic and Social Commission for Western Asia (ESCWA). (2013). Report of the Committee on Social Development on its ninth session. *E/ESCWA/SDD/2013/ IG.1/6 Report*.

Feeney, G. (2003). Data assessment: Encyclopaedia of Population. Demeny, P. & McNicoll, G. *(Ed). (Vol. 1).* New York: Macmillan Reference USA.

Herlihy, D., & Klapisch-Zuber, C. (1985). Tuscans and their Families: A Study of the Florentine Catasto of 1427. New Haven: Yale University Press.

Institute for Resource Development (IRD), (1990). An Assessment of DHIS-1 Data Quality. Demographic and Health Surveys Methodological Reports 1. *Institute for Resource Development (IRD) & Macro Systems*. Columbia, Maryland. USA.

Jowett, J., & Li, Y.Q. (1992). Age heaping: China Contrasting Patterns from China. *GeoJournal, 28*(4), 427-442.

Kenya National Bureau of Statistics (KNBS) and ICF Macro. (2015). *Kenya Demographic and Health Survey 2014.* Calverton, Maryland: KNBS and ICF Macro.

Kenya National Bureau of Statistics (KNBS). (2012). *Kenya 2009 Population and Housing Census Analytical Report on Population Projections.* Kenya National Bureau of Statistics, Ministry of Planning, National Development and Vision 2030, Nairobi.

Kenya National Bureau of Statistics (KNBS) and ICF Macro. (2010). *Kenya Demographic and Health Survey 2008-09*. Calverton, Maryland: KNBS and ICF Macro.

Kidane, A. (2009). Digit Preference in African Survey Data and Their Impact on Parametric Estimates African Econometric Society Conference. *For presentation at the African Econometric Society Conference, (July 11-13 2009)*. Abuja, Nigeria.

Kodiko, H.O. (2014). Sub-national Projections Methods: Application to the Counties in the Former Nyanza Province, Kenya. Masters project for Master of Science in Population Studies. *University of Nairobi*.

Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., and Zaba, B. (eds). (2013). Tools for Demographic Estimation. *International Union for the Scientific Study of Population*. Paris. http://demographicestimation.iussp.org.

Mugo, E.W. (2012). Assessment of quality of data: The case of the 2008-2009 KDHS. M. A Population Studies. *University of Nairobi*.

Myer's, R. (1976). An instance of reverse heaping of ages. *Demography 13* (4), 577-580.

Myer's, R. (1954). Accuracy of age reporting in the 1950 United States census. *Journal of the American Statistical Association XLIX*, 826-831.

Myer's, R. (1940). Errors and bias in the reporting of ages in census data. *Transaction of the Actuarial Society of America; XLI*(2):104).

Nasir, J. A., & Hinde, A., (2014). An Extension Of Modified Whipple Index– Further Modified Whipple Index. Pak. *J. Statist. 30*(2), 265-272.

Palamuleni, M.E. (2013). Age reporting in the North West Province, South Africa, 1996- 2007. Paper presented at the 2013, *Annual Meeting of the Population Association of America*, New Orleans, April 11- 13.

Pardeshi, G.S. (2010). Age heaping and accuracy of age data collected during a community survey in the Yavatmal district, Maharashtra. *Indian Journal of Community Medicine, 35*(3), 391-395.

Pullum, T. W., and Stan B. (2014). Evidence of Omission and Displacement in DHS Birth Histories. *DHS Methodological Reports No. 11*. Rockville, Maryland, USA: ICF International.

Pullum, Thomas W. 2008. An Assessment of the Quality of Data on Health and Nutrition in the DHS Surveys, 1993-2003. *Methodological Reports No. 6*. Calverton, Maryland, USA: Macro International Inc.

Pullum, T. W. (2006). An Assessment of Age and Date Reporting in the DHS Surveys, 1985-2003. *Maryland: Macro International Inc.,* 5. Calverton, Maryland: Macro International Inc.

Pullum, T.W. (2005). A statistical reformulation of demographic methods to assess the quality of age and date reporting, with application to the Demographic and Health Surveys. Paper presented at the 2005, *Annual Meeting of the Population Association of America*, Philadelphia, March 31-April 2.

Randall, S., and Coast, E., (2016). The quality of demographic data on older Africans. *Demographic Research, 34*. pp. 143-174. ISSN 1435-9871

Schoumaker B., (2011). Omissions of Births in DHS Birth Histories in Sub-Saharan Africa: Measurement and Determinants. *Washington DC, Centre de recherche en démographie et sociétés, Université catholique de Louvain*

Shireen, A., Kothari, M., & Pullum T. (2015). An Assessment of the Quality of DHS Anthropometric Data, 2005-2014. *DHS Methodological Reports No. 16*. Rockville, Maryland, USA: ICF International.

Shryock H. S., Siegel J. S., & Associates (1976). The Methods and Materials of Demography. *New York: Academic Press*.

Siegel, J., and Swanson, D. (2004). The methods and materials of demography (2nd edition). *San Diego, California: Elsevier Academic Press.*

Spoorenberg, T. (2009). Is the Whipple's index really a fair and reliable measure of the quality of age reporting? An analysis of 234 censuses from 145 countries. *New York: United Nations Department of Economic and Social Affairs*.

Spoorenberg, T. (2007). Quality of Age Reporting: Extension and Application of the Modified Whipple's index. *Population (2nd ed.). 62* (4), 729-741.

Stockwell, E.G., Nagi, M.H., & Snavley, L.M. (1973). Digit Preference and Avoidance in the Age Statistics of Some Recent African Censuses: Some Patterns and Correlates. *International Statistical Review, 41*(2), 165-174.

Susuman, A.S., H. F. H., Lougue, S., Ogujiuba, K., & Mwambene, E. (2015). An assessment of the age reporting in Tanzania population census 2012. *Journal of Social Sciences Research, 8* (2).

United Nations (2014). *Principles and Recommendations for Population and Housing Censuses,* Revision 3. New York, Statistical Papers, Series M, No. 19 Rev. 3. Sales No. E.13.XVII.10

United Nations (1983). Indirect Techniques for Demographic Estimation, Manual X. *Population Studies, No. 81.* Sales No. E83.XIII.2.

United Nations. (1973). *Population Census Statistics III: Demographic Yearbook. (25),* Sales No. E/F.74.XIII.1

United Nations. (1956). *Manuals on Methods of Estimating Population. Manual II: Methods of Appraisal of Quality of Basic Data for Population Estimates*. Sales No. E.56. XIII.2.

United Nations, (1955). Manuals on methods of estimating population: Manual II Methods of Appraisal of Quality of Basic Data for Population Estimates, *Department of Economic and Social Affairs, Population Branch, New York.*

United States Bureau of the Census, (1985). Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, *Washington, D.C.*

Wafula, S., & Ikamari, L. (2007). Patterns, levels and trends in unmet need for contraception: a case study of Kenya. *https://www.researchgate.net/publication/281937670*

West, K.K., Robinson, J.G., & Bentley, M. (2005). Did Proxy Respondents Cause Age Heaping in the Census 2000? *National Academies of Sciences, Washington, DC 3*(6), 58–65.

Yazdanparast, A., Pourhoseingholi, M.A., & Abadi, A. (2012). Digit preference in Iranian age data. *Iranian Journal of Public Health. 9*(1).

APPENDICES

Appendix 1: 2014 KDHS Population in Single Years

Age	Male	Female	Both sexes
0	2129	2115	4,244
1	2181	2041	4,222
2	2210	2188	4,398
3	2381	2320	4,701
4	2215	2190	4,405
5	2181	2102	4,283
6	2712	2626	5,338
7	2567	2441	5,008
8	2623	2561	5,184
9	2321	2243	4,564
10	2634	2507	5,141
11	1985	2043	4,028
12	2491	2422	4,913
13	2085	2114	4,199
14	2288	2124	4,412
15	1813	1691	3,504
16	1733	1685	3,418
17	1659	1478	3,137
18	1734	1635	3,369
19	1241	1263	2,504
20	1413	1479	2,892
21	988	1094	2,082
22	1145	1323	2,468
23	893	995	1,888
24	1044	1231	2,275
25	1120	1425	2,545
26	1021	1284	2,305
27	1026	1195	2,221
28	1183	1370	2,553
29	808	1060	1,868
30	1411	1418	2,829
31	710	791	1,501
32	1016	978	1,994
33	564	635	1,199
34	779	867	1,646
35	1114	1043	2,157
36	792	867	1,659
37	633	704	1,337
38	734	818	1,552
39	578	639	1,217
40	1043	932	1,975
41	510	597	1,107
42	740	620	1,360
43	414	413	827
44	402	531	933
45	747	671	1,418
46	437	515	952
47	344	397	741
48	382	456	838
49	371	367	738
50	636	490	1,126
51	372	569	941
52	388	628	1,016
53	312	428	740
54	385	526	911
55	359	537	896
56	467	468	935
57	320	284	604
58	306	354	660
59	234	261	495
60	598	535	1,133
61	197	202	399
62	325	339	664
63	167	190	357
64	218	267	485
65	365	392	757
66	206	242	448
67	179	148	327
68	174	209	383
69	140	189	329
70	291	359	650
71	102	126	228
72	161	154	315
73	84	82	166
74	107	149	256
75	168	197	365
76	84	101	185
77	57	70	127
78	109	116	225
79	59	79	138
80+	611	901	1512
Total for all ages	**75,726**	**78,096**	**153,822**

Appendix 2: 2014 KDHS Population in Five-Year Age groups

Age	Male	Female
0-4	11,116	10,854
5-9	12,404	11,973
10-14	11,483	11,210
15-19	8,180	7,752
20-24	5,483	6,122
25-29	5,158	6,334
30-34	4,480	4,689
35-39	3,851	4,071
40-44	3,109	3,093
45-49	2,281	2,406
50-54	2,093	2,641
55-59	1,686	1,904
60-64	1,505	1,533
65-69	1,064	1,180
70-74	745	870
75-79	477	563
80+	611	901
Total for all ages	**75,726**	**78,096**