APPLICATION OF FACTOR ANALYSIS TO THE 2009 GENERAL HOUSEHOLD SURVEY IN SOUTH AFRICA

by

SIMON MALESELA MONYAI

DISSERTATION

Submitted in (partial) fulfilment of the requirements for the degree of

MASTER OF SCIENCE DEGREE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Prof M Lesaoana

CO-SUPERVISORS: Mr P Nyamugure

Mr TB Darikwa

2015

DECLARATION

I declare that the dissertation hereby submitted to the University of Limpopo, for the degree of Master of Science in Statistics has not previously been submitted by me for a degree at this or any other university; that it is my work in design and in execution, and that all material contained herein has been duly acknowledged.

Monyai, SM (Mr) Date

DEDICATION

This project is dedicated to my Mom Agnes Monyai and my sister Lena Monyai, for all the support, motivation and contributions they have made in my life.

ACKNOWLEDGEMENTS

The author wishes to express his deepest gratitude to his supervisor Prof M Lesaoana who offered invaluable support and guidance. Her expertise, understanding, and ability to motivate the author. My special thanks to Mr P Nyamugure and Mr TB Darikwa for their patience and understanding in guiding me throughout the project. Without their assistance in various statistical techniques this dissertation would not have been possible. I would also like to thank the Faculty of Science and Agriculture Research Committees under Prof HJ Siweya for their assistance and insightful comments they provided at all levels of this research project. The author would like to convey a special word of thanks to the University of Limpopo and the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (MaSS) for the financial support. I also wish to extend special words of thanks to Statistics South Africa for providing the data, my family for their moral support and everyone who supported me directly or indirectly towards the completion of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Introduction: The high number of variables from the 2009 General Household Survey is prohibitive to do holistic analysis of data due to high correlations that exist among many variables, making it virtually impractical to apply traditional methods such as multinomial logistic regression. The purpose of this study to identify observed variables that can be explained by a few unobservable quantities called factors, using factor analysis.

Methods: Factor analysis is used to describe covariance relationships among 162 variables of interest in the 2009 General Household Survey (GHS) and 2009 Quarterly Labour Force Survey of South Africa (QLFS). Data for the respondents aged 15 years and above was analysed by first applying factor analysis to the 162 variables to produce factor scores and develop models for five core areas: education, health, housing, labour force and social development. Multinomial logistic regression was then used to model educational levels and service satisfaction using identified factor sores.

Results: The variability among the 162 variables of interest was described by only 29 factors identified using factor analysis, even though these factors are not measured directly. Multinomial logistic regression (MLR) analysis showed negative and significant impact of education factors (fees too high, violence and absence of parental care) on levels of educational attainment. “Historically advantaged” factor is the only factor significant and positively affects educational levels. Housing and social development factors were regressed against service satisfaction. Housing factors such as the home owners, age of a house and male household heads were found to be significant. Social development factors such as “no problem with health”, sufficient water, high income, household size and telephone access were found to be significant. Labour force factors such as employment, industrial business and occupation, employment history and long-term unemployment have positive and significant impact on levels of education.

Conclusion: It can be concluded that factor analysis as a data reduction technique has managed to describe the variability among the 162 variables in terms of just 29 unobservable variables. Using MLR in subsequent analysis, this study has managed to identify factors positively or negatively associated with educational levels and service satisfaction. The study suggests that educational, housing, social development and labour force facilities should be improved and education should be used to improve life circumstances.

Keywords: factor analysis, factors, multinomial logistic regression, logits, educational levels of attainment, service satisfaction, quality of service delivery.

ACRONYMNS

ANOVA	Analysis of Variance
CD	Census division
CFA	Confirmatory factor analysis
CPS	Current Population Survey
CSS	Central Statistical Service
*DisSa*t	Disatisfied
EFA	Exploratory factor analysis
EIU	Economist Intelligence Unit
ESG	Employment skill gaps
E1FesH	fees too high
E2Viol	violence
E3AbsP	absence of parental care
E4HisA	historically advantage
FA	Factor Analysis
FTE	Full-time education
H1BrickH	brick house
H2GovA	government assistance
H3HomO	home owners
H4AgeH	age of house
H5WaiL	waiting list
H6MalH	male household heads
GHS	General Household Survey
KMO	Kaiser-Meyer-Olkin
LL	Log likelihood
LRT	Likelihood Ratio Test
L1Emply	employment
L2IndbO	industrial business and occupational
L3EmplH	employment history
L4longU	long-term unemployment
MLR	Multinomial Logistic Regression
MSA	Measures of Sampling Adequacy
OR	Odds Ratio

PCA	Principal Component Analysis
PSUs	Primary Sampling Units
QLFS	Quarterly Labour Force Survey
QoL	Quality of Life
Sat	Satisfied
SPSS	Statistical Package for the Social Sciences
Stats SA	Statistics South Africa
S1No_pH	no problem with health
S2SuffiW	sufficient water
S3HighI	high income
S4PayS	payment of sewerage
S5TelepA	telephone access
S6AbseT	absence of toilet
S7HouseZ	household size
S8WaterI	water interruption
S9Pens	pensioners
VerySat	Very Satisfied
UK	United Kingdom
UNSC	University New Student Census

RESEARCH OUTPUTS / CONFERENCES

Peer Reviewed Journal Publication

Nyamugure, P., Lesaoana, M. and **Monyai, S.** 2011. *Application of factor analysis to the 2009 General Household Survey of South Africa.* ICASTOR Journal of Mathematical Sciences 5(1): 133-150.

Recently submitted for Publication

Monyai, S., Nyamugure, P., Lesaoana, M. and Darikwa, T.B. 2014. *Application of multinomial logistic regression to the 2009 General Household Survey of South Africa.*

Presentations at Conferences

Nyamugure, P., Lesaoana, M. and **Monyai, S.** *Application of factor analysis to the 2009 General Household Survey of South Africa*. 40th Annual Conference of the Operations Research Society of South Africa, 18-21 September 2011, Elephant Hills Hotel, Victoria Falls, Zimbabwe. http://www.orssa.org.za/wiki/upload_s/Conf/2011ORSSAConferenceProgramme.pdf

Monyai, S., Lesaoana, M. and Nyamugure, P. *Application of factor analysis to educational factors of the 2009 General Household Survey of South Africa*. 53th Annual Conference of the South African Statistical Association, 31 October to 4 November, 2011, CSIR, Pretoria, South Africa. http://www.sastat.org. za/sasa2011/images/ SASA 2011 PROGRAM.pdf

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

Some of the major challenges faced by the democratic government of South Africa include delivery of basic services such as clean water, energy, sanitation, education, health, housing, alleviation and eradication of poverty, and reducing high unemployment levels. Various policies and programmes on education, employment, health and social development have been introduced and developed to address some of the key challenges. However, these policies and programmes need to be evaluated and monitored on a regular basis. The data requirements to meet these challenges have also evolved over time, resulting in the establishment of the General Household Survey (GHS) in 2002. The GHS is an annual survey conducted by Statistics South Africa (Stats SA), and designed to determine the level of development as well as to measure the quality of service delivery in a number of key service factors such as employment, health and education (Statistics South Africa, 2010a). The GHS covers six broad areas, namely: (1) education, (2) health, (3) housing, (4) social development, (5) household access to services and facilities, and (6) food security and agriculture (Statistics South Africa, 2010a). Our study addresses the following core areas of interest: education, health, housing, social development and labour force. The Quarterly Labour Force Survey (QLFS) of the 4th quarter of 2009 was used for the labour force core area (Statistics South Africa, 2010b).

Despite this 8-year series (2002-2009), not much analysis of the GHS has been performed. The GHS of 2009 has 124 and 2009 QLFS has 38 variables of interest that describe the quality of life (QoL). These 162 variables are too many for analytical purposes. Thus, there is a need to investigate the major aspects that contribute to the QoL among South Africans. The challenge for this study is to identify a few but uncorrelated factors that best describe the QoL in South Africa and use them for subsequent analysis such as multinomial logistic regression.

QoL consists of different components. The components of interest in this study are the objective and subjective indicators as outlined by Brown *et al.* (2004). According to Brown *et al.* (2004) the objective indicators include the standard of living, health

and longevity, housing and neighbourhood characteristics, and they are typically measured with a cost of living, health service provision and education levels, among others. Subjective indicators include, among others, life satisfaction, individual fulfilment and happiness usually measured using indicators of balance of effect and self-worth (Brown *et al.*, 2004). The findings by Møller (2007) are that higher standard of services has strongest bearing on life satisfaction, hence on quality of life. Our research uses satisfaction with service delivery as a proxy to general life satisfaction. People satisfied with service delivery are likely to be satisfied with life.

In our study, data for the respondents aged 15 years and above was analysed by first applying factor analysis to the 162 variables to produce factor scores and develop factor analysis models for the five core areas of interest to our study. Factor analysis eliminated variables which were highly correlated. Multinomial logistic regression (MLR) technique was, in turn, used to determine the relationship between the independent factors (factor scores) and dependent variables (education levels and level of satisfaction with delivery of services).

Using MLR, the study has found that there is a relationship between education, labour force factors and education level of attainment. When considering education, the following factors are significant: fees too high, violence, absence of parental care and historically advantaged. When considering labour force, the most significant factors are employment, industrial business and occupation, employment history and long-term unemployment.

One of the objectives of this study is to determine housing and social development factors which have an impact on service satisfaction. The results revealed that housing factors such as the home owners, age of a house and male household heads, are important determinants of service satisfaction. This study has found that the most important determinants of social developments are: no health problem**,** sufficient water, high income, household size and access to a telephone factors. This study can be used for planning, more especially for education, health, housing, social development and labour force purposes.

1.2 RESEARCH PROBLEM

Our study investigates if the factors (unobserved variables) obtained from the five core areas are related to the observed variables such as educational level and service satisfaction which are not included in factor analysis. This entails testing the hypotheses that postulate that the unobserved factors of H1: education will have a positive association with educational level; H2: labour force will have a positive association with educational level; H3: housing will have a positive association with service satisfaction; and H4: social development will have a positive impact on service satisfaction.

1.3 PURPOSE OF THE STUDY

The main aim of this study is to identify factors of quality of life using the South African 2009 General Household Survey and 2009 Quarterly Labour Force Survey, and in turn use these factors to determine whether or not there is a relationship between:

i) educational level and educational factors;
ii) educational level and labour force factors;
iii) service satisfaction and housing factors; and
iv) service satisfaction and social development factors.

Note: Current health status does not have an effect on one's educational status and there is no data linking health status to service satisfaction.

1.4 CONTEXTUALIZATION

This dissertation is written using traditional thesis format and is divided into six chapters. Chapter one gives the introduction and background of the study. The chapter further focuses on the problem statement, aim and objectives of the study and the significance of the study, and also provides the layout of the study. Chapter two explores literature relevant to the study. Chapter three discusses in detail, the research methodology and procedures used in the study. Chapter four presents analyses and interprets the empirical data using factor analysis. Chapter five presents analyses and interprets the empirical data using multinomial logistic regression. Lastly, Chapter six summarises the study, draws conclusions and offers some recommendations.

CHAPTER 2: LITERATURE REVIEW

2.1 STATISTICAL TECHNIQUES

Factor analysis is a set of statistical techniques that are used to either explore or confirm the underlying structure among a set of variables so as to determine those variables that tap a factor (Nunnally and Bernstein, 1994). There are two types of factor analysis, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is used to create factor scores that are useful for regression analysis or other follow-up analyses (Gorsuch, 1983). CFA is a method used as an attempt to test hypotheses using factor analysis (Jöreskog, 1966; Gorsuch, 1983; Kline, 1994; Hair *et al.*, 2006). Like EFA, CFA is used to create factor scores to be used for subsequent analysis, but factor scores created using CFA might be used to identify ranking on latent variables, and to confirm the previous or existing indexes (Bollen, 1989). Kline (1994) views exploratory factor analysis as the ideal method where data is complex and it is uncertain what the most important variables in the field are, as is the case with the intended study. Factor analysis is used to analyse interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors) with minimum loss of information. Kline (1994) and Hair *et al.* (2006) acknowledge that when using factor analysis variables are somehow correlated. Therefore, those variables that share similar underlying dimensions should be highly correlated, and those that measure dissimilar dimensions should yield low correlation. These high/low correlation coefficients will become apparent in the correlation matrix because they form clusters indicating those variables which are associated and those which are not associated (Ho, 2006).

Tesfazghi *et al.* (2010) presented a case study where the urban QoL at small scale was measured, and its variability was evaluated for Kirkos sub-city of Addis Ababa, Ethiopia. The researchers addressed the issue of the variability at small scale and the relationship between subjective QoL and objective QoL, which according to them is not well known. They applied factor analysis to household survey secondary data to establish an index of objective QoL, which is an external condition of life or the observable facts derived from secondary data such as level of education, household

characteristics, crime and others. The subjective QoL were viewed as people's perception of their life measured by subjective indicators, usually derived from surveys of residents' perceptions, satisfaction and evaluation of their life. In their study, the researchers created an index based on the dimensions (factor scores) of the objective QoL, consisting of factors such as crowdedness, socio-economic status, safety and proximity, housing and demographic issues.

The subjective QoL score in Kirkos sub-city consisted of four dimensions, that is, physical, economic, social and proximity factors (Tesfazghi *et al.,* 2010). The analysis of variance, (ANOVA), was applied to establish and evaluate relationships between some variables of the QoL, while the coefficient of variation was applied to evaluate spatial variability. The results of their study revealed that the subjective QoL scores had large variations in the sub-city. The mean QoL score also indicated that on average the respondents in the sub-city were dissatisfied with the quality of their life. The study also revealed that respondents with higher education level and income were, on average, more satisfied with their QoL in the sub-city. The comparison between the subjective QoL and the objective QoL indicated a state of dissonance, adaptation, deprivation or well-being. Such results suggest that the two measures do not always indicate the same level of QoL. Their study will be useful when creating an index of QoL for the GHS using factor scores and testing the significance of the factors that would be extracted from the GHS 2009 data.

William *et al*. (1972) described the strategy for analysing the relationship between a dependent variable and a set of independent variables when the latter are not necessarily amenable to standard statistical treatment. This happens in a situation where full regression model contains the set of independent variables, most of which are highly correlated. The factor regression analysis was used by William *et al*. (1972), to solve the problem of multicollinearity (highly correlated variables) rather than using classical regression methods to estimate parameters. Their study concluded that regression upon factor analysis permits systematic analysis of data in situations where there is multicollinearity or singularity. The study by William *et al*. (1972) will inform our study on the creation of QoL indices. This is achieved in the current study by determining the unique contribution of each factor in explaining the dependent variable.

2.2 EDUCATION

Education is defined by Maliki *et al.* (2012) as the cornerstone of social development and a principal means of improving the character and pace of individual welfare. Maliki *et al*. (2012) revealed that higher levels of education and enlarged access will lead to productivity gains and income, hence reduced inequality and poverty. Their study also showed that there is one set of factor that is influencing learning which is schooling factor. In some countries such as Malaysia students' academic excellence is very much important due to the fact that parents assume that their child's academic success would guarantee life success (Hassan *et al.*, 2012). A study conducted by Hassan *et al.* (2012) considered the following unobserved factors affecting learning style: students' attitude before and after class (1); strategies used to comprehend the lecture (2); the importance of lecture (3); class size and condition (4); efforts outside class (5); classroom convenience (6); and importance on listening to lecture (7).

The study by Khorshidi and Rezaloo (2011) investigated the effective factors in creating prevalent high school in Tehran (capital of Iran) from school manager's point of view. Khorshidi and Rezaloo (2011) applied factor analysis and created an index of nine factors affecting creation of prevalence schools. These factors are: skill of manager (1); ability of teachers (2); equipping schools with technology and using it (3); objective celebration of religious and national ceremonies (4); acquiring international standards (5); success of students in entering higher education institutions (6); participation of students and their parents in school problems (7); dominance of human relationship in school (8); and using efficient tools for encouragement and punishment (9).

Researchers such as Majors and Sedlacek (2001) used factor analysis to organise students' services. The following eight factors emerged from 110 items of the University New Student Census (UNSC): religion/spirituality (1); help seeking (2); interracial relationships (3); academic self-concept (4); cultural tolerance (5); academic preparedness (6); shy/lonely (7); and cult approval (8). Few researches have determined factors related to highest educational attainment. Grade levels and students' achievement were found to be highly correlated with their level of adjustment in University (Mann, 2001; Sennett *et al*., 2003).

The purpose of the study by Strand and Winston (2008) was to assess the nature and level of pupils' educational aspirations and to elucidate the factors that influence these aspirations. Their study was motivated by poor pupil attendance, below average examination results and low rates of continuing in full-time education after the age of 16 years. Their concern was to extend the international data on young people's educational aspirations by adding data from the United Kingdom (UK). The target population was selected from inner city areas where educational aspirations such as rates of continuation in full-time education (FTE) are relatively low. They used factor analysis to create an index of eight factors from 34 items of the pupil questionnaire. The factors identified were: commitment to schooling (1); academic self-concept (2); teacher support (3); home-support for learning (4); positive peer support (5); disaffection-negative peers (6); laissez faire (7); and home-educational aspirations (8). These factors were used in subsequent analysis to explore any relations between the emerging factors and educational aspirations. The results of their study indicated that there is no significant difference in aspirations by gender or year group, but differences between ethnic groups were marked. Their study revealed that educational aspirations are strongly associated with some specific attitudes and influences that underlie the link between aspirations and attainment. Their study has assisted our study in creating an education index to be used in subsequent analysis.

2.3 HEALTH

Birhanu *et al*. (2010**)** assessed patient satisfaction with health care provider interaction and its influencing factors among out-patients at health centers in west Shoa, Central Ethiopia. The cross sectional facility-based study methods were conducted on 768 out-patients of six health centers based on patient flow during the 10 days prior to the start of data collection. The data was collected using pre-tested instrument, and factor analysis was used to identify factor scores for the items representing the satisfaction scale. Multivariate linear regression analysis was used to determine the influence of independent variables on the regression factor score. The independent variables used in their study were: perceived empathy, perceived technical competency, non-verbal communication, patient enablement, being told the name of one's illness, type and frequency of visit, knowing the providers and educational status. Regressing these factor scores, the results revealed that

interpersonal processes including perceived empathy, perceived technical competency, non-verbal communication and patient enablement significantly influence patient satisfaction.

Joshi *et al*. (2009) conducted a pilot study to determine the factor structure, reliability and validity of statements in a healthcare survey questionnaire as predictors of public perception of good healthcare system. Data on public perceptions of healthcare from national survey of 1434 adult Singaporeans was analysed using FA and MLR. Six factors extracted using FA were: national healthcare financing framework (1); service at public institutions (2); service at private institutions (3); individual responsibility for health (4); affordability at public institutions (5); and affordability at private institutions (6). The study by Joshi *et al*. (2009) revealed that factors 1; 2; 3; 4; and 5 were associated with good healthcare. The researchers noted further that snapshot surveys to assess perceptions of the healthcare system and the underlying reasons could be conducted with questionnaires abridged to include the five identified factors*.*

2.4 HOUSING

Hammill (2009) argued that an individual had an unsatisfied basic need in household population density if they lived within a household where the number of people per room in the housing space was greater than or equal to 3. The number of rooms in the house or equivalent did not include bathrooms, toilets, kitchens, passages, hallways or garages. The study by Hammill (2009) will be useful in creating housing index and identifying those factors related to service satisfaction. Simelane (2007) applied Principal Components Analysis (PCA) to compute an index of living. The household assets/characteristics data used in computing the index included: (1) ownership of telephone; (2) number of rooms (excluding toilet and bathroom) owned by the household (note that this variable was used to compute an index of household crowding because crowding is considered one of key indicators of housing quality); (3) source of energy for cooking (electricity/gas, paraffin, wood/coal/animal/other); (4) source of energy for heating (electricity/gas, paraffin, candles/others); (5) source of energy for lighting (electricity); (6) main source of domestic water supply; (7) type of toilet facility; and (8) type of main dwelling for the household (modern, tradition/informal, other/caravan/tent). Simelane (2007) applied standard multivariate regression technique for further analysis.

2.5 SOCIAL DEVELOPMENT

Analysis of the GHS data 2003-2007 by Statistics South Africa (2009) has identified an index for households eligible for child support grant (CSG), but not accessing it. The pattern of relationship among the dependent variables and the ratio indicators were identified for low earning households with children aged younger than 15 years who do not access the CSG. The respondents' questionnaires were subjected to factor analysis in order to assess the underlying structure in the responses. The factors identified were: (1) age/pension factor; (2) in-house dependency and support factor; (3) employment and educational institution attendance factor; and (4) level of education and wealth factor. A strong relationship was observed in factor (2) between total dependency ratio, child dependency ratio and in-house support ratio. In-house support ratio is defined as the extent to which unemployed household members aged 15-64 years are dependent on other household members for survival. Stats SA (2009) have not gone further to use multivariate technique such as regression to get the reliable estimates, which our study intends to do.

Sarstedt *et al*. *(*2009) examined the antecedent factors that explained the variations in overall service satisfaction judgments. They developed a measurement approach for satisfaction which was subsequently tested using a large-scale sample among soccer fans. A survey with 1054 participants was carried out and a total of 623 respondents yielded 600 usable responses after excluding 23 questionnaires that were incomplete. The researchers identified 108 items relevant for measurement of a fan's overall satisfaction. Only seventeen factors measuring 99 items were extracted by factor analysis. However, these factors explained less than half of fan satisfaction's variance. When estimating parameters, the analysis showed only seven factors exerting statistically significant influence on fan overall satisfaction judgement. Among these factors (i.e. stadium, team, fan-based support, club management, the atmosphere during a visit, club's internet site and accompanying entertainment) the stadium, team characteristics and fan-based support for the club and its management, were the most important factors influencing overall satisfaction.

Our research makes use of service satisfaction index as applied by Sarstedt *et al*. (2009) to create the regression models for the dependent variable "service satisfaction", and it uses EFA to explain the underlying factors. It further tests for the significance of the extracted factors using parametric methods. These factors are

also tested for the significance in creating the 2009 GHS QoL index based on the five broad core areas of interest.

2.6 LABOUR FORCE

The study by Mirza *et al.* (2014) was based on surveys conducted among 100 employers and 151 graduates from six universities and postgraduate colleges in the Gujrat-Sialkot-Gujranwala tri-cities in Pakistan. Factor analysis was used to group 24 specific skills into three broad categories and the disaggregated results were presented. Three factors identified were: communication and business specific skills, core employability skills, and professional skills. The different assessments of employers and students about job skills led to differences defined as skill, employability, and perception gaps. Mirza *et al.* (2014) revealed that professional skills include two individual skills (honesty and persistency) which could be considered a part of core employability skills. Factor scores were not further assessed.

Our research builds on a previous study by Blom and Saeki (2011) on the employment skill gaps (ESG) for Indian engineers through a survey of employers conducted in 2009. Blom and Saeki (2011) classified all ESG by factor analysis, into three skills groups: core employability skills, communication skills, and professional skills. Their results revealed that overall employers were dissatisfied with the quality of engineering graduates. According to rankings, soft skills (core and communication) were ranked more important than professional skills.

The study by Alasia (2004) assessed the degree of spatial diversity exhibited across Canada by using 1996 Census of Population data, aggregated at the census division (CD) level**.** The study was based on a range of commonly used and understood demographic, social and economic variables. A factor analysis was conducted in order to identify underlying dimensions that characterise each CD across Canada. The factor analysis resulted in six factors, each of which provided a profile of the CDs on a number of key attributes. Twenty-seven (27) variables used in the factor analysis were reduced to six factors which captured about 78% of the variance in the data set. Those six factors are: (1) labour force and economic attributes; (2) remote and agro-rural attributes; (3) demographic and labour force attributes; (4) employment attributes: complex manufacturing versus primary production attributes;

(5) employment attributes: traditional manufacturing versus government employment attributes; and (6) demographic dynamics attributes. The study revealed the multi-dimensional nature of the performance of regions and the variety of associated demographic, social and economic characteristics. The spatial pattern of the factor scores was not further assessed using subsequent analysis.

Houtman and Steijn (1990) undertook a study to test whether or not the relationship between unemployment and social participation should be conceived as direct effects, or as a result of cultural capital received from home. They examined causal effect of the independent variables of unemployment and its duration on the dependent variable, controlling cultural capital. In their research controlling variable 'cultural capital' was a factor which had been inferred from factor analysis on background characteristics and for which the factor scores were calculated. As a result, the researchers concluded that social isolation of the unemployed and minimum wage earners should be conceived as the continuation of life-style which stems from the period before dependency took place. They deduced that the connection between unemployment and social isolation might disappear when controlling for a number of background variables which are related to: (a) cultural capital (educational level, cultural and social participation of parents); and (b) the composition of the household (e.g. number of children, single parent status as well as age).

Economist Intelligence Unit (2005) developed a new quality of life index based on a unique methodology that links the results of subjective life satisfaction surveys to the objective determinants of quality of life across countries. The index was calculated for 111 countries in 2005. The nine QoL factors were: (1) material wellbeing; (2) health; (3) political stability and security; (4) family life; (5) community life; (6) climate and geography; (7) job security; (8) political freedom; and (9) gender equality. According to Economist Intelligence Unit (EIU), these indicators represent a country's quality of life index, or the "corrected" life-satisfaction scores, based on objective cross-country determinants. The total variance explained by these factors was 80%. The EIU developed a complete ranking of the countries worldwide and South Africa was ranked position 92. Most of the studies that regressed these factors revealed a small correlation between education and life satisfaction. Our study will

develop the QoL index based on the 2009 GHS of South Africa, though the issue of ranking will not be considered.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 FACTOR ANALYSIS

The current study uses factor analysis to extract important and fewer factors that contribute towards the QoL as envisaged in the 2009 GHS and QLFS core areas. In turn, the identified factors are used as input to multivariate techniques such as multinomial logistic regression model that explain service satisfaction and educational level.

Factor Analysis (FA) by definition, is a technique or method that is used to determine whether a finite number of observed variables $Y_1, Y_2, Y_3, ..., Y_n$, are linearly related to a smaller number of unobservable latent variables called factors, $F_1, F_2, F_3, ..., F_k$. The purpose of FA is to discover a simple relationship among the observed variables. In particular, FA seeks to discover if the observed variables can be explained largely or entirely in terms of the unobservable factors. FA also describes the covariance relationships among many variables in terms of a few underlying, but unobservable random quantities called factors (Johnson and Wichern, 2007). In FA the factors can be continuous, censored, binary, ordered categorical, counts or combinations of these variable types. Many statistical methods are used to study the relationship between independent and dependent variables. FA is used to study the patterns of relationship among many dependent variables, with the goal of discovering something about the nature of the independent variables (factors) that affect them, even though those independent variables were not measured directly. FA therefore obtains answers that are more hypothetical and tentative than is true when independent variables are observed directly. A typical FA suggests answers to four major questions:

1. How many different factors are needed to explain the pattern of relationships among these observed variables?
2. What is the nature of those factors?
3. How well do the hypothesised factors explain the observed data?
4. How much purely random or unique variance does each observed variable include?

The observed variables are modeled as linear combinations of the potential factors, plus error terms. FA is related to Principal Component Analysis (PCA), but the two differ in the sense that when PCA performs a variance-maximising rotation of the variable space, it takes into account all variability in the variables. In contrast, FA estimates how much of the variability is due to common factors (i.e. communality). The two methods become essentially equivalent if all the error terms in the FA model can be assumed to have the same variance.

Another advantage of FA over these other methods is that FA can recognise certain properties of correlations. For instance, if variables A and B each correlate 0.70 with variable C, and correlate 0.49 with each other, FA can recognise that A and B correlate zero when C is held constant because $0.70^2 = 0.49$. Multidimensional scaling and cluster analysis have no ability to recognise such relationships, since the correlations are treated merely as generic similarity measures rather than as correlations.

3.1.1 Factor Analysis Statistical Model

Suppose that we have a set of *n* observable variables $Y_1, Y_2, Y_3, ..., Y_n$ and *k* unobservable variables or factors, $F_1, F_2, F_3, ..., F_k$ where $k < n$. The set of observable variables $\mathbf{Y}$ has mean vector $\boldsymbol{\mu}$ and covariance matrix $\sum$. It is assumed that each of the *n* variables is linearly related to the *k* factors as follows:

$$\begin{aligned} Y_1 &= \mu_1 + \phi_{11}F_1 + \phi_{12}F_2 + \phi_{13}F_3 + \ldots + \phi_{1k}F_k + e_1 \\ Y_2 &= \mu_2 + \phi_{21}F_1 + \phi_{22}F_2 + \phi_{23}F_3 + \ldots + \phi_{2k}F_k + e_2 \\ &\vdots \\ Y_n &= \mu_n + \phi_{n1}F_1 + \phi_{n2}F_2 + \phi_{n3}F_3 + \ldots + \phi_{nk}F_k + e_n \end{aligned} \qquad \textbf{(3.1)}$$

The classical model of FA **(3.1)** can be expressed simply as:

$$Y_i = \mu_i + \sum_{j=1}^{k} \phi_{ij}F_j + e_i \quad i = 1, 2, \ldots, n \qquad \textbf{(3.2)}$$

where μ_i is the mean of the variable Y_i.

e_i is the error term which indicates that the hypothesised relationships **(3.2)** are not exact.

ϕ_{ij} is referred to as factor loading (or score) of variable i on factor j.

$\phi_{ij}F_j$ represents the contribution of the corresponding factor to the linear composite.

The fact that the factors are unobservable distinguishes the factor model from the multivariate regression model. The error terms $e_i's$ must be independent of each other and are such that $E(e_i)=0$ and $Var(e_i)=\sigma_i^2$. The factors $F_j's$ are independent of one another and also independent of the error terms and are such that $E(F_j)=0$ and $Var(F_j)=1$. The variance in terms of the factors expressed in **(3.2)** gives:

$$
\begin{aligned}
Var(Y_i) &= Var\left(\mu_i + \sum_{j=1}^{k}\phi_{ij}F_j + e_i\right) \\
&= Var\left(\sum_{j=1}^{k}\phi_{ij}F_j + e_i\right) && \text{since } \mu_i \text{ is a constant} \\
&= \sum_{j=1}^{k}\phi_i^2 Var(F_j) + Var(e_i) && \text{by independent assumption} \\
&= \sum_{j=1}^{k}\phi_i^2 + \sigma_i^2 && \text{since } Var(F_j)=1, \quad \forall_j \\
&= \underbrace{\phi_{i1}^2 + \phi_{i2}^2 + \ldots + \phi_{ik}^2}_{communality\ (e_i^2)} + \underbrace{\sigma_i^2}_{uniqueness(u_i^2)}
\end{aligned}
\qquad \textbf{(3.3)}
$$

$i = 1, 2, \ldots, n$

The communality measures the amount of variance a variable shares with all the other variables being considered and may be interpreted as the reliability of the indicator. The uniqueness may be further divided into two portions, (Harry, 1976).

1. The portion that is due to the particular selection of variables in the study *(specificity,* h_i^2).
2. The portion that is due to unreliability in measurement *(reliability,* r_i^2)

Total variance may be expressed as:

$$Var(Y_i) = c_i^2 + u_i^2 = c_i^2 + h_i^2 + r_i^2 \qquad \textbf{(3.4)}$$

If the *k* factors were perfect predictors of Y_i then $e_i = 0$ and $\sigma_i^2 = 1$. A large communality value indicates a strong influence by the underlying factors.

3.1.2 Correlation Analysis

As discussed above, each observed variable is a function of all the factors underlying the structure. The variances are unity or 1 because they are standardised without imposing additional constraints which enable the identification. This in a sense simply determines the units of measure of the unobserved construct. Let us now consider the consequences that these equations impose on the structure of the covariance matrix of the observed variables. Since $Var(Y_i) = \phi_{i1}^2 + \phi_{i2}^2 + \ldots + \sigma_i^2$ from equation **(3.2),** using the property that the factors are uncorrelated with a variance of 1, then for any two variables Y_i and Y_j:

$$\begin{aligned} Cov(Y_i, Y_j) &= E[(\phi_{i1}F_1 + \phi_{i2}F_2 + \ldots + \phi_{ik}F_k + e_i)(\phi_{j1}F_1 + \phi_{j2}F_2 + \ldots + \phi_{jk}F_k + e_j)] \\ &= \phi_{i1}\phi_{j1}Var(F_1) + \phi_{i2}\phi_{j2}Var(F_2) + \ldots + \phi_{ik}\phi_{jk}Var(F_k) + 0 \\ &= \phi_{i1}\phi_{j1} + \phi_{i2}\phi_{j2} + \ldots + \phi_{ik}\phi_{jk} \end{aligned} \quad \textbf{(3.5)}$$

Equation **(3.4)** follows from the fact that $Cov(F_i, F_j) = 0$, $Var(F_i) = Var(F_j) = 1$ and $Cov(e_i, e_j) = 0, \quad i, j = 1, 2, \ldots, n$

The commonalities are our center of interest because the error variance or unique variances do not contain information about the data structure. This demonstrates that the noise or measurement error needs to be removed, although it only affects the variances (the diagonal of the covariance matrix) and not the covariances. The correlation matrix should have fairly high correlations between the variables being investigated. Literature suggests that the correlation must be bigger than 0.33 for the factor model to be appropriate (Harry, 1976).

3.1.3 KMO and Bartlett's Test of Sphericity

3.1.3.1 Kaiser-Meyer-Olkin Measure of Sampling Adequacy

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy tests whether the partial correlations among variables are small, and it is used to examine the appropriateness of factor analysis. The KMO values greater than 0.5 but less than or equal to 1 indicate that factor analysis is appropriate. If two variables share a common factor with other variables, their partial correlation (a_{ij}) will be small, indicating the unique variance they share. KMO is defined as follows:

$$KMO = \frac{\underbrace{\sum\sum r_{ij}^2}_{i \neq j}}{\left(\underbrace{\sum\sum r_{ij}^2}_{i \neq j} + \underbrace{\sum\sum a_{ij}^2}_{i \neq j} \right)} \quad \textbf{(3.6)}$$

where r_{ij} is the correlation coefficient between variables i and j.

If $a_{ij} \cong 0$, the variables are measuring a common factor, and KMO ≅ 1.0. A value closer to 1 indicates that patterns of correlations are relatively compact and hence factor analysis should yield distinct and reliable factors (Field, 2005). Furthermore, values between 0.6 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are superb (Hutcheson and Sofroniou, 1999). If $a_{ij} \cong 1$ the variables are not measuring a common factor, and KMO ≅ 0.0, hence factor analysis is inappropriate.

A measure of sampling adequacy (MSA) can be calculated by taking only coefficients involving that variable. For the Y_i^{th} variable, the MSA:

$$MSA_i = \frac{\underbrace{\sum r_{ij}^2}_{j \neq i}}{\underbrace{\sum r_{ij}^2}_{j \neq i} + \underbrace{\sum a_{ij}^2}_{j \neq i}} \quad \textbf{(3.7)}$$

Again large values are needed for a good factor analysis.

3.1.3.2 Bartlett's Test of Sphericity

Bartlett's test of sphericity tests whether the correlation matrix $\mathbf{R}$ is an identity matrix, i.e. each variable correlates perfectly with itself (*r* = 1), but has no correlation with the other variables (*r* = 0) which would indicate that the factor model is inappropriate. The test calculates the determinant of the matrix of the sums of products and cross-products (*S*) from which the inter-correlation matrix is derived. This determinant of the matrix is converted to a chi-square statistic and tested for significance. The null hypothesis is that the inter-correlation matrix comes from a population in which the variables are non-collinear (i.e. an identity matrix) and that the non-zero correlations in the sample matrix are due to sampling error. If the approximate chi-square value is large and the significance level is small (less than 0.005), then the hypothesis that the variables are independent can be rejected. Bartlett's chi-square test can be used

to evaluate the correlation matrix (Bartlett, 1950). The Bartlett chi-square value is obtained by:

$$\chi^2 = -\left[(m-1) - \frac{1}{6}\left(2n+1+\frac{2}{n}\right)\right]\left[\ln(|S|) + n\ln\left(\frac{1}{n}\right)\sum e_j\right] \quad \textbf{(3.8)}$$

with degrees of freedom:

$$df = \frac{(n-1)(n-2)}{2}.$$

where n is the number of variables, k the number of factors, m is the sample size and $e_j = j^{th}$ eigenvalue of $\mathbf{R}$.

3.1.4 Matrix Method

For a large data set it is more convenient to represent the data structure in matrix form as follows:

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix},\ \mu_{n\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_n \end{bmatrix},\quad F_{k\times 1} = \begin{bmatrix} F_1 \\ F_2 \\ . \\ . \\ . \\ F_k \end{bmatrix},\quad \Lambda_{n\times k} = \begin{bmatrix} \phi_{11} & \phi_{12} & . & . & . & \phi_{1k} \\ \phi_{21} & \phi_{22} & . & . & . & \phi_{2k} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \phi_{n1} & \phi_{n2} & . & . & . & \phi_{nk} \end{bmatrix},\quad \in_{n\times 1} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Equation **(3.1)** can be expressed in matrix form as:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Lambda F} + \boldsymbol{\epsilon} \quad \textbf{(3.9)}$$

If the factors are independent then:

$$E[\in] = 0,\quad Cov(\in) = E[\in\in^T] = D,\quad E[FF^T] = \Phi = \mathrm{I} \text{ and } E[FF^T] = Cov(F) = \mathrm{I}$$

where $\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & . & . & . & 0 \\ 0 & \sigma_2^2 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & \sigma_n^2 \end{bmatrix}$ is a diagonal matrix.

The matrix **ΛF** is the matrix of common variances and covariances and $\in$ is the matrix of unique variances.

The covariance matrix of **Y** is given by:

$$Cov(Y) = E(YY^T)$$

$$\begin{aligned} E[YY^T] &= E\left[(\Lambda F + \in)(\Lambda F + \in)^T\right] = E\left[\Lambda F F^T \Lambda^T + \in \in^T\right] \\ &= \Lambda E\left[FF^T\right]\Lambda^T + E\left[\in \in^T\right] \\ &= \Lambda \Phi \Lambda^T + D = \Sigma. \end{aligned} \qquad \textbf{(3.10)}$$

Since the factors are independent, **(3.10)** simplifies to:

$$\begin{aligned} \Sigma &= \Lambda\Lambda^T + D \\ &= L + D \end{aligned} \qquad \textbf{(3.11)}$$

When $k = n$, the diagonal matrix becomes a zero matrix and $\Sigma = \Lambda\Lambda^T = L$. FA is most useful when *k* is small relative to *n*. The sample covariance matrix **S** is an estimator of the unknown population covariance matrix $\sum$. If the off-diagonal elements of **S** are small or those of the sample correlation matrix **R** essentially zero, then the variables are not related and a factor analysis will not prove useful. FA can be applied if matrix $\sum$ appears to deviate significantly from a diagonal matrix. The results of **(3.11)** lead to the following steps:

(i) Estimation of commonalities,
(ii) Extraction of initial factors,
(iii) Determination of the number of factors,
(iv) Rotation to a terminal solution,
(v) Interpretation of factors,
(vi) Calculation of factor scores,
(vii) Determination of model fit.

3.1.5 Estimating Commonalities

The first step is to remove the unique component of the variance in order to keep the variance explained by the common factors only. In factor analysis, the diagonal elements of $\Lambda\Lambda^T$ are specified as the squared multiple correlations of each variable with the remainder of the variables in the set (the percentage of explained variance). The diagonal matrix **D** contains the residual variances. The principal factor (principal component) method and the maximum likelihood method are the most popular methods used to estimate the factor loadings $\phi_{ij}'s$ and the specific variances σ_i's.

3.1.6 The Principal Factor (Principal Component) Method

Principal Factor Analysis (PFA) differs from Principal Component Analysis (PCA) only in that the main diagonal entries of the correlation matrix **R** are replaced by communalities (Harris, 1975). The covariance matrix ∑ can be written in terms of eigenvalue-eigenvector $(\lambda_i, \mathbf{e}_i)$ pair form as:

$$\mathbf{\Sigma} = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_n e_n e_n^T = \begin{bmatrix} \sqrt{\lambda_1}\, e_1 & . & . & . & \sqrt{\lambda_n}\, e_n \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1^T \\ . \\ . \\ . \\ \sqrt{\lambda_n}\, e_n^T \end{bmatrix} \quad \textbf{(3.12)}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$

The preferred model is the one where the number of factors (*k*) are much less than the number of variables (*n*) since this is the one that explains the covariance structure in terms of the factors. If there are *n-k* small eigenvalues it is best to neglect their contribution and the covariance matrix written in terms of the remaining *k* eigenvalue-eigenvector as follows (Johnson and Wichern, 2007):

$$\mathbf{\Sigma} = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots + \lambda_k e_k e_k^T = \begin{bmatrix} \sqrt{\lambda_1}\, e_1 & . & . & . & \sqrt{\lambda_k}\, e_k \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1^T \\ . \\ . \\ . \\ \sqrt{\lambda_k}\, e_k^T \end{bmatrix} = \Lambda_{n \times k} \Lambda_{k \times n}^T \quad \textbf{(3.13)}$$

If the model has as many factors as the variables then the specific variance or the diagonal matrix will be equal to zero, and the covariance matrix will be as follows:

$$\mathbf{\sum = \Lambda \Lambda^T + 0 = \Lambda \Lambda^T} \quad \textbf{(3.14)}$$

This model assumes that the specific factors are of minor importance and can be ignored in the factoring of ∑. If the specific factors are included in the model then their variance may be taken to be the diagonal elements of the matrix $\mathbf{\sum = \Lambda \Lambda^T}$. If the specific factors are included then:

$$\sum = \Lambda\Lambda^T + D = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots + \lambda e_k e_k^T + D =$$

$$= \begin{bmatrix} \sqrt{\lambda_1}\, e_1 & . & . & . & \sqrt{\lambda_k}\, e_k \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1^T \\ . \\ . \\ . \\ \sqrt{\lambda_k}\, e_k^T \end{bmatrix} + \begin{bmatrix} \sigma_1 & 0 & . & . & . & 0 \\ 0 & \sigma_2 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & \sigma_p \end{bmatrix} \quad \textbf{(3.15)}$$

where $D_i = \sigma_{ij} - \sum_{j=1}^{k} \phi_{ij}^2$ for $i = 1, 2, \ldots, n$

When equation **(3.15)** is applied to the sample covariance **S** or the sample correlation matrix **R**, then the resulting solution is known as the principle component solution, since the factor loading are the scaled coefficients of the first few sample principle components.

3.1.7 Extracting Initial Factors

The initial factors are obtained by performing a principal component analysis on the matrix $\Lambda\Lambda^T$ .That is:

$$\Lambda\Lambda^T = C_{n\times n} = V_{n\times n} \Lambda_{n\times n} V_{n\times n}^T \quad \textbf{(3.16)}$$

where matrix V is such that:

$$V^T V = I \text{ and } \Lambda = V^T \Sigma V \quad \textbf{(3.17)}$$

3.1.8 Determining the Number of Factors

This step involves finding the number of factors $k < n$, that are necessary to represent the covariance structure. In this section we introduce two rules on how many factors to retain.

3.1.8.1 *λ*> 1 (Kaiser's rule)

This method eliminates values of the eigenvalue that are less than 1. The rationale for this rule is that each factor should account for at least the variance of a single variable. An eigenvalue is a ratio between the common (shared) variance and the specific (unique) variance explained by a specific factor extracted. The eigenvalue λ is such that $|\Sigma - I\lambda| = 0$ and is the variance of the linearly transformed variable Y_i. A factor with an eigenvalue greater or equal to one is considered to be

significant. An eigenvalue greater than one indicates that more common variance than unique variance is explained by the factor. From **(3.17)** since $\Lambda = V^T \Sigma V$ it follows that:

$$tr(\Lambda) = tr(V^T \Sigma V) = tr(V^T V \Sigma) = tr(\Sigma). \quad \textbf{(3.18)}$$

If the variables $Y_i's$ are normalised, the matrix Σ is the correlation matrix **R**. The trace of **R** (i.e., the sum of the diagonal terms) is equal to the number of variables *n*. It then follows from the equality in equation **(3.18)** that the sum of the eigenvalues of a correlation matrix is equal to the number of variables *n*. The problem is to find the number *k* so as to account for most of the covariance matrix $\boldsymbol{\Sigma}$.

3.1.8.2 The Elbow Rule

This test is used to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure (Hair *et al*, 2006). This method is based on a scree plot which is obtained by plotting the eigenvalues against the number of factors in their order of decreasing size. The elbow rule corresponds to finding the point on the scree plot, which makes an elbow, i.e. the point at which the curve first begins to straighten out. Those factors above this point of inflection are deemed useful and those below are not.

3.1.8.3 Rotation to Terminal Solution

Rotation helps to identify those variables that load on one factor and not on another factor. The ultimate goal of factor rotation is to come up with a simpler and more meaningful pattern of factors. The most commonly used method is the VARIMAX rotation suggested by Kaiser (1958). With this method, the rotation searches to give the maximum variance of the squared loadings for each factor (in order to avoid problems due to negative loadings). This results in obtaining extreme loadings (very high or very low).

If $\hat{\Lambda}$ is the $n \times k$ matrix of estimated factor loadings obtained by any method, then $\hat{\Lambda}^*$ which is $n \times k$ matrix of rotated loadings is given by:

$$\hat{\Lambda}^* = \hat{\Lambda} V \quad \textbf{(3.19)}$$

where $VV^T = V^T V = \mathrm{I}.$

The covariance or correlation matrix remains unchanged since:

$$\hat{\Lambda}\hat{\Lambda}^{T} + \hat{D} = \hat{\Lambda}VV^{T}\hat{\Lambda}^{T} + \hat{D} = \hat{\Lambda}^{*}\hat{\Lambda}^{T*} + \hat{D}. \quad \textbf{(3.20)}$$

This also shows that the communalities and the specific variance remain unchanged as well. For two factors or factors that are taken in pairs the rotation is given by:

$$\hat{\Lambda}^{*}_{2\times 2} = \hat{\Lambda}_{n\times 2} V_{2\times 2} \quad \textbf{(3.21)}$$

where $V = \begin{cases} \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} & \textit{for clockwise rotation} \\ \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} & \textit{for anticlockwise rotation} \end{cases}$

and β is the angle of rotation.

For more than two factors the VARIMAX method is used. VARIMAX rotation is the most common of the rotations that are available (Kaiser, 1958; Bonett and Price, 2005). It first involves scaling the loadings by dividing them by the corresponding communality as follows:

$$\tilde{\phi}_{ij}^{*} = \hat{\phi}_{ij}^{*} \Big/ \hat{h}_{i}$$

Here $\tilde{\phi}_{ij}^{*}$ is the loading of the i^{th} variable on the j^{th} factor after rotation, and $\hat{h}_{i}$ is the communality for variable i. The VARIMAX procedure selects the rotation to find this maximum quantity:

$$V = \frac{1}{p}\sum_{j=1}^{k}\left\{\sum_{i=1}^{p}\left(\tilde{\phi}_{ij}^{*}\right)^{4} - \frac{1}{p}\left[\sum_{i=1}^{p}\left(\tilde{\phi}_{ij}^{*}\right)^{2}\right]^{2}\right\} \quad \textbf{(3.22)}$$

Maximising V in **(3.22)** corresponds to spreading out the squares of the loading on each factor as much as possible (Lewis-Beck *et al.*, 2003). VARIMAX rotation is the sample variance of the standardised loadings for each factor, summed over the *k* factors. Our objective is to find a factor rotation that maximises this variance.

3.2 MULTINOMIAL LOGISTIC REGRESSION

MLR is the extension of the binary logistic regression when outcome variable is polytomous. In this section we generalise logistic regression for multinomial

(nominal) response variables. The model provides realistic and efficient estimates. The interpretation of an independent variable's role in differentiating dependent variable groups is the same as used in binary logistic regression. The good part about MLR is that there are multiple interpretations for an independent variable in relation to different pairs of groups. For the MLR with binary logistic regression, one category of the dependent variable is specified as the reference category and regression coefficient are estimated for each independent variable, for the contrast of each category of the dependent variable with the reference category.

3.2.1 Evaluating the Usefulness for Logistic Models

The benchmark that we will use to characterise a multinomial logistic regression model as useful is a 25% improvement over the rate of accuracy achievable by chance alone (Schwab, 2002; Tabachnick and Fidell, 2007; Petrucci, 2009). Without considering whether or not the independent variables had no relationship to the groups defined by the dependent variable, we would still expect to be correct in our predictions of group membership percentage of the time. This is referred to as by chance accuracy. The estimate of by chance accuracy is the proportional by chance accuracy rate, computed by summing the squared percentage of cases in each group. There are various reasons for lack of consensus indices of predictive efficiency. One of the reasons may be the fact that researchers are more often interested in goodness of fit than in accuracy of prediction or classification of the model indicated by classification table and proportional error measures. The proportional change in error measure of accuracy of prediction for selection of the model standard formula (2×2 prediction table with 1 degree of freedom) is shown by Menard (2002) as:

$$\phi_p = \frac{ad - bc}{0.5(a+b)(b+d) + (c+d)(a+c)} \qquad \textbf{(3.23)}$$

The best options for analysing the prediction (classification) tables provided by logistic regression packages involve proportional change in error measures of the form:

$$\text{predictive efficiency} = \frac{(error\ \text{without model}) - (error\ \text{with model})}{(error\ \text{without model})} \qquad \textbf{(3.24)}$$

For a classification model, an appropriate definition of the expected error without the model is:

Errors without model = $\sum_{i=1}^{N} f_i(N - f_i)/N$

Where N is the sample size and f_i is the number of cases observed in category i. If the model improves our prediction of the dependent variable, the formula for prediction of efficiency is the same as a proportional reduction in error formula. When it happened that the model actually does worse than it occurs, the predictive efficiency is negative and that is proportional increase in error (Menard, 2002).

3.2.2 Model fitting: Overall test of Relationship

We first have to describe the overall test of relationship, in this case a relationship between the dependent and independent variables. Model fitting mainly used to determine the presence of a relationship between the dependent and combination of independent variables. The overall test is based on the reduction in the likelihood values for a model which does not contain any independent variables and the model that contains the independent variables. This difference in likelihood followed a chi-square distribution and was referred to as the model chi-square. The significance test for the final model chi-square was the researcher's statistical evidence of the presence of a relationship between dependent and independent variables. After establishing the relationship, the next important thing to do is to establish the strength of multinomial logistic regression relationship.

3.2.3 Measuring Strength of Association (Pseudo-R^2)

The *pseudo* R^2 is defined by Borooah (2002) as $-LL_{R+F}/LL_R$ and is bounded from below by 0 and from above by 1. A zero value corresponds to all the slope coefficients being zero, and a value of 1 corresponds to perfect prediction (that is, $LL_R = 0$). The LL_R is the value of the log-likelihood function when the only explanatory variable was constant term. LL_{R+F} is the value of the log-likelihood function when all the explanatory variables were included.

There are three commonly used R^2 statistics (namely: Cox and Snell, Nagelkerke, and McFadden) to measure the strength of association between dependent variable and the explanatory (independent) variables (Cox and Snell, 1989; Nagelkerke,

1991; Reise, 2000; Agresti, 2002; Tabachnick and Fidell, 2007). Multinomial logistic regression does compute correlation measures to estimate the strength of the relationship (pseudo R^2 measures). A more useful measure to assess the utility of a multinomial logistic regression model is through the classification accuracy (Menard, 1995).

3.2.4 Relationship of Independent and Dependent Variables

The next step after evaluating the usefulness for logistic models is to determine the relationship of independent and dependent variables. There are two important types of tests for individual independent variables, that is the likelihood ratio test and the Wald test. The likelihood ratio test evaluates the overall relationship between an independent variable and dependent variables, while, the Wald test evaluates whether or not the independent variable is statistically significant in differentiating between two groups in each of embedded binary logistic comparisons. One should be careful in the sense that, if an independent variable has an overall relationship to the dependent variable, it does not necessarily suggest statistical significance. In fact, it might or might not be statistically significant in differentiating between pairs of groups defined by the dependent variable.

3.2.5 Methods of Fitting and Interpreting Parameters: *Nominal Response*

When the categorical dependent outcome has more than two levels, for instance in-state of predicting only 1=satisfied or 0= dissatisfied with service, we may have three groups such as 0=dissatisfied, 1=satisfied, and 2=very satisfied. In this case we have more than two levels that means the reference category has to be chosen in comparison. Multinomial logistic regression is good because; it does not assume normality, linearity, or homoscedasticity. These assumptions are part of multiple regression (Tabachnick and Fidell, 2007). Variable selection or model specification methods for multinomial logistic regression are similar to those used with standard multiple regression. MLR assumes that the choice of or membership in one category is not related to the choice or membership of another category. If the groups of the outcome variables are perfectly separated by the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated (Tabachnick and Fidell, 2007).

3.2.5.1 Baseline-Category Logits for Nominal Response

For nominal response variables, an extension of logistic regression forms logit model by pairing each category with a baseline category and each logit equation results in separate parameters. Y is a categorical (polytomous) response variable with J categories, taking on values $0, 1, \ldots, J\text{-}1$. The assumptions for multinomial logistic regression are as follows: (1) Observations Y are statistically independent of each other; (2) Y are random sample from a population where Y has a multinomial distribution with probability parameters $\{\pi_0(\mathbf{x}), \ldots, \pi_{J-1}(\mathbf{x})\}$; and (3) One category has to be set aside as a base category (hence $J-1$ parameters).

For the group data it will be convenient to introduce auxiliary random variables representing counts of responses in various categories. For instance, if there are J response categories, then for the ith observation there will be J binary response variables, $Y_{i0}, \ldots, Y_{ij}$ where:

$$Y_{ij} = \begin{cases} 1 & \text{if case } i \text{ response is in category } j \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(3.25)}$$

Since only one category can be selected for response i, then

$$\sum Y_{ij} = 1$$

Let n_i denote the number of cases in the ith group and y_{ij} denote the number of response from the ith group that fall in jth group, with observed value y_{ij}. The probability distribution of the counts y_{ij} given the total n_i is given by multinomial distribution:

$$P(Y_1 = y_1, \ldots, y_{J-1} = \begin{cases} \dfrac{n!}{y_j! \ldots y_{j-1}!} \pi_0(y), \ldots, \pi_{J-1}(y) & \text{when } \sum_{j=1}^{J\text{-}1} y_j = n \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(3.26)}$$

Given a certain choice of $J-1$ of these, the rest are redundant. For dependent variable with J categories, this requires calculation of $J-1$ equations, one for each category relative to reference category, to describe the relationship between the dependent variable and the independent variable. The logit model pairs each response category with a baseline category, often the last or the most common one. The group coded $Y = 0$ will serve as the reference outcome value to form logit

comparing $Y = j, \ldots, J\text{-}1$ to it. To develop the model, let us assume that we have p covariates and a constant term, denoted by the vector of covariate, $\mathbf{x}$, of length $p+1$. If the first category is the reference or $Y = 0$ the general formula for the logit function is given by:

$$g_j(\mathbf{x}) = \ln\left[\frac{P(Y = j \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})}\right] = \beta_{j0} + \beta_{j1}X_{J1} + \ldots + \beta_{jk}X_{JP} = \mathbf{x'}\boldsymbol{\beta}_j \quad \textbf{(3.27)}$$

Hence, for each case, there will be $J\text{-}1$ predicted log odds. The intercept parameter (β_j) is the logits for success when X_j is zero and the slope parameter β_j is the logit difference in indicating how much the log-odds change with unit on the predictor (Reise, 2000). If we consider Agresti (2002) and Hosmer and Lemeshow (2000) baseline-category logits are given by:

$$\pi_j(\mathbf{x}) = P(Y = j \mid \mathbf{x}), \text{ for } j = 0, 1, \ldots, J\text{-}1 \text{ with } \sum_{j=0}^{J-1}\pi_j(\mathbf{x}) = 1 \quad \textbf{(3.28)}$$

Each of which is a function of the vector of $p+1$ parameters $\boldsymbol{\beta}' = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})$. For the observation, the counts at J categories of Y are treated as multinomial with probabilities $\{\pi_0(\mathbf{x}),\ldots,\pi_{J-1}(\mathbf{x})\}$. A general expression for conditional probability in J category model is given as follows:

$$\pi_j(\mathbf{x}) = P(Y = j \mid \mathbf{x}) = \frac{e^{g_j(\mathbf{x})}}{1 + \sum_{k=1}^{J-1} e^{g_k(\mathbf{x})}} \quad \textbf{(3.29)}$$

where the vector $\beta_{00} = 0$ and hence $g_0(\mathbf{x}) = 0$.

3.2.5.2 Maximum Likelihood Estimation

For n independent observations the joint probability for the likelihood is given by:

$$\prod_{i=1}^{n}\left(\pi_j(\mathbf{x})\right)^{Y_{ij}} \quad \textbf{(3.30)}$$

The conditional likelihood function for sample of n independent observations and J categories is:

$$l(\beta) = \prod_{i=1}^{n}\prod_{j=0}^{J-1}\left[P(Y = j \mid \mathbf{x})\right]^{Y_{ij}} = \prod_{i=1}^{n}\prod_{j=0}^{J-1}\pi_j(\mathbf{x})^{Y_{ij}} \quad \textbf{(3.31)}$$

Taking the log and using the fact one category is selected for response i the log-likelihood function is:

$$L(\beta)=\sum_{j=1}^{n}\left(\sum_{j=1}^{J-1}\left(Y_{ij}\mathbf{x}'\boldsymbol{\beta}_j\right)-\log_e\left[1+\sum_{j=1}^{J-1}e^{(\mathbf{x}'\boldsymbol{\beta}_j)}\right]\right) \quad \textbf{(3.32)}$$

The likelihood equations are found by taking the first partial derivatives of $\mathbf{L}(\boldsymbol{\beta})$ with respect to each of the $p+1$ unknown parameters (Kutner *et al.,* 2005). For simplicity of the notation let

$$\pi_{ij}=\pi_j(\mathbf{x}) \quad \textbf{(3.33)}$$

The general form of this equation is as follows:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}}=\sum_{i=1}^{n}X_{ki}=\left(Y_{ij}-\pi_{ij}\right) \quad \textbf{(3.34)}$$

for $j=0,1,\ldots,J\text{-}1$ and $k=0,2,\ldots,p$. We recall that $x_{0i}=1$ for each subject. The maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$, of these parameters are chosen to be those values which maximise **(3.32)** and is obtained by setting the derivatives of the log-likelihood equation to zero and solving for $\boldsymbol{\beta}$ (Hosmer and Lemeshow, 2000). The $J-1$ response function may be obtained by substituting the maximum likelihood estimates of the $J-1$ parameters vectors into the expression in **(3.29)**. As shown by Kutner *et al.* (2005), the estimator is expressed as:

$$\hat{\pi}_j(\mathbf{x})=\frac{e^{g_j(\mathbf{x})}}{1+\sum_{k=1}^{J-1}e^{g_k(\mathbf{x})}}=\frac{e^{\mathbf{x}'\hat{\boldsymbol{\beta}}_j}}{1+\sum_{k=1}^{J-1}e^{\mathbf{x}'\hat{\boldsymbol{\beta}}_k}} \quad \textbf{(3.35)}$$

The estimators of the variance and covariance are obtained by evaluating $\mathbf{var}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$ (Hosmer and Lemeshow, 2000). The standard errors of the estimated coefficients are:

$$SE\left(\hat{\beta}_j\right)=\mathrm{var}\left(\hat{\beta}_j\right)^{\frac{1}{2}} \text{ where } j=0,1,\ldots,J\text{-}1 \quad \textbf{(3.36)}$$

Multicollinearity in the multinomial logistic regression solution is detected by examining the standard errors for the $\hat{\beta}$ coefficients. A standard error larger than 2.0 indicates numerical problems. The *exp*($\hat{\beta}$) are the z-ratio of the estimated coefficients to their estimated standard errors. The z-ratios are asymptotically distributed as N(0,1) under the null hypothesis that associated coefficients are zero.

3.2.5.3 Interpreting and Assessing the Significance of the Estimated Coefficient

3.2.5.3.1 Odds ratios

In order to include the outcomes being compared as well as values of the covariate, the odds ratios in multinomial outcomes setting are generalised in this form: Assume that the outcome labelled with $Y = 0$ is the reference outcome. The subscript on the odds ratios is being compared to the reference outcome (Hosmer and Lemeshow, 1989). That is, the odds ratio of outcome j versus 0 for covariate values of $x = a$ versus $x = b$ is:

$$OR_j = (a,b) = \left[\frac{\frac{p(Y = j \mid x = a)}{P(Y = 0 \mid x = a}}{\frac{P(Y = j \mid x = b)}{P(Y = j \mid x = b)}} \right] \quad \textbf{(3.37)}$$

3.2.5.3.2 Confidence interval for $\hat{\beta}$

The general large-sample formula for $100(1-\alpha)\%$ confidence interval for comparison of outcome level j versus the reference category, for any i levels of the independent variable is:

$$\exp\left[\hat{\beta}_j \pm z_{1-\alpha/2} SE\left(\hat{\beta}_j\right)\right] \text{ where } j = 0, 1, \ldots, J\text{-}1 \quad \textbf{(3.38)}$$

One of the important suggestions imposed by Menard (1995) and Greene (2003) about measuring goodness-of-fit is that one should report the maximised value of the log-likelihood function. Since the hypothesis that all the slopes in the model are zero is often interesting, the results of comparing the full model with an intercept only model should be reported (Borooah, 2002).

3.2.5.3.3 Likelihood ratio comparison tests

The likelihood ratio test (LRT) is used to test hypotheses about the significance of the predictor variables (interaction terms). For MLR $J-1$ parameter estimates are tested simultaneously for each independent variable. The effect of individual or groups of explanatory variables on response can be assessed by comparing the deviance statistics (-2LL) for two nested models. The resulting statistic is tested for significance using chi-square distribution with $J-1$ degrees of freedom comparing

the reduced model (model without variables) and full model (model with variables). The hypothesis test is:

$H_0 : \beta_j = 0 \quad j = 0, 1, \ldots, J\text{-}1$

H_0: There is no difference between the fitted and full (intercept only) model

The test statistic:

$$-2LL_{diff} = -2LL_R - \left(-2LL_{R+F}\right) \sim \chi^2 \qquad \textbf{(3.39)}$$

where R is the reduced nested model and $R+F$ is the full model. The degrees of freedom are equal to the number of slope coefficients estimated. If the H_0 is rejected, the conclusion is that at least one of the $J-1$ coefficients are significantly different.

3.2.5.3.4 Wald tests for $\hat{\beta}$

The alternative to LRT, is Wald test that tests for the statistical significance of individual coefficients to determine which logits are significantly affected by X. For MLR, there are $J-1$ coefficients to be tested for each and every independent variable. The set of coefficients must either be retained or dropped. The hypothesis test:

$H_0 : \beta_j = 0 \quad j = 0, 1, \ldots, J\text{-}1$

i.e. null hypothesis that a given X has no effect on the odds of $Y = j$ versus $Y = 0$.

The Wald statistic (Magee, 1990) may be calculated as:

$$z^2 = \left[\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}\right]^2 \sim \chi^2 \text{ where } j = 0, 1, \ldots, J\text{-}1 \qquad \textbf{(3.40)}$$

Alternatively it follows the standard normal distribution, that is:

$$z = \frac{\hat{\beta}}{SE\left(\hat{\beta}_j\right)} \text{ where } j = 0, 1, \ldots, J\text{-}1 \qquad \textbf{(3.41)}$$

Equation (**3.41**) is parallel to the t-ratio for coefficients in linear regression. Therefore, the test has one (number of restrictions) degree of freedom.

3.3 DATA COLLECTION

3.3.1 Introduction

The 2009 General Household Survey (GHS) data employed in this study was collected by Statistics South Africa. Although the survey covers six core areas, our study focuses on the following five areas: education, health, social development, housing and labour force (Statistics South Africa, 2010a; 2010b). These indicators would determine the level of development in the country and evaluate the performance of existing programmes and projects on a regular basis. The survey is also aimed at benchmarking the quality of service delivery in all the major sectors of the South African economy. “A multi-stage, stratified random sample was drawn using probability proportional to size principles. First level stratification was based on province and second tier stratification on district council” (Statistics South Africa, 2010a). For the GHS, a total of 25 361 households were sampled across the entire country yielding 94 263 responses, and for the QLFS a sample of 3 080 primary sampling units (PSUs) was generated using stratified two-stage design. A total of 162 variables across the five core areas are considered in this study. The two datasets were collected across all the nine provinces of South Africa and both cover cross-cutting variables such as gender, age group, population group, marital status and highest level of education. We next discuss the four core areas in the 2009 GHS and one core area in the 2009 QLFS covered in our study.

3.3.2 Core areas in the 2009 General Household Survey

3.3.2.1 Education

Under education information on literacy level (ability to write own name, letters, filling in a form, reading and calculating ability of respondents), was collected. Information on educational institutions, distance to educational institutions, means of transport used and amount of fees paid, was also collected. Problems at educational institutions: violence and nature of violence at educational institutions, absenteeism and their reasons, were also among some of the variables that were being investigated. Data on various reasons for not attending any educational institution was also collected.

3.3.2.2 Health

Information such as the nature of illness that respondents suffered or are suffering from, was gathered. Illness due to abuse of alcohol or drugs, flue or respiratory tract infection, depression or mental, sexual transmitted including HIV/AIDS, vehicle accidents, gunshots wounds, trauma due to violence and assault beatings, were also collected among several other illnesses. Chronic illness that includes asthma, diabetes, cancer, hypertension and arthritis, were also collected. Reasons on medication for these illnesses were investigated. Difficulties in seeing, hearing, walking, remembering, concentrating, self-care and communication, were also collected. Degree of permanent disability, use of corrective aids for disability such as eye glasses, hearing aid, walking stick and wheelchairs, were collected.

3.3.2.3 Housing

The major objective of the 2009 GHS (Statistics South Africa, 2010a) was to collect information from households about different aspects of the people's living arrangements and finding the type of dwellings South Africans live in. The type of dwelling is divided into: fully owned or partially owned, renting and other unspecified types of dwellings. Other variables such as wall materials, roofing materials, number of rooms, years spent on waiting list and year the house was built, were also investigated in the 2009 GHS (Statistics South Africa, 2010a).

3.3.2.4 Social Development

Information on social welfare namely: visits by community care givers, services of victims of domestic violence, social work services for drug abuse, child protection services and correctional services, was collected. Variables also investigated in 2009 GHS (Statistics South Africa, 2010a) include: access to clean water and its source, sanitation and refuse collection, access to telecommunications, transport, sources of energy and food access. Sources of energy include electricity, paraffin, wood and gas. Several water sources in the 2009 GHS (Statistics South Africa, 2010a) were also investigated, but the main sources are, piped water, bore water, water from rivers, well, spring and dams.

3.3.2.5 Labour Force

Data on whether people have their own business, do farm work, look for work, accept job if offered, who they work for, government job creation programmes, types of occupation and if they receive child support grants among other variable are being investigated in the 2009 QLFS (Statistics South Africa, 2010b).

3.3.2.6 Summary

This section of data collection has provided the specific variables contained in the 2009 GHS and 2009 QLFS data set under each of the core areas that is, education, health, housing, social development and labour force. We note that this research employed secondary data for which Statistics South Africa is the primary collector. In Chapter 4 we apply factor analysis to analyse the data, while in Chapter 5, multinomial logistic regression is applied. Like the others (Blom and Saeki, 2011; Khorshidi and Rezaloo, 2011; Penny, 2011; Simelane, 2007; Statistics South Africa, 2009; Strand and Winston, 2008) the variables used in this data sets are either categorical or continuous (numeric).

CHAPTER 4: PRESENTATION AND INTERPRETATION OF FINDINGS

4.1 INTRODUCTION

This chapter analyses and interprets secondary data collected by Statistics South Africa (Stats SA) in order to explore or understand the quality of life among South Africans. Data collected through the 2009 GHS questionnaire and 2009 QLFS were analysed using the Statistical Package for the Social Sciences (SPSS) (Statistics South Africa, 2010a; 2010b). The inferential statistics (factor analysis) was used to obtain factor scores. Multinomial logistic regression (MLR) models are done in Chapter 5 to determine if there is any relationship between factor scores and educational levels, and to determine if there is any relationship between factor scores and service satisfaction. Descriptive statistics were done to determine the marginal frequencies. Furthermore, to facilitate analysis and interpretation, the 2009 GHS and 2009 QLFS questionnaires cover personal and demographic information.

4.2 DATA ANALYSIS

Secondary data obtained from Stats SA was analysed. The dependent variables (educational level and service satisfaction) comprise of more than two categories. The dependent variables in this case are levels of education, namely: “no school”, “less grade 12”, “grade 12”, “above grade 12”. For this modelling the reference category is chosen as “no schooling” and service satisfaction during household visits was used as a dependent variable or variable of interest scored on a five-point likert scale. The new three-point likert scales created were “very satisfied=1”, “satisfied=2” and “dissatisfied=3”. The survey respondents who are in the “very satisfied” group are compared to those who are in the “dissatisfied” group; and those who are in the “satisfied” group are also compared to those who are in the “dissatisfied” group.

4.2.1 Factor Analysis

The purpose of factor analysis (FA) is to describe covariance relationships among many variables in terms of a few but unobservable random quantities called factors. If variables within a particular group are highly correlated among themselves and have small correlations with variables in a different group, then each group of variables represents a single factor that is responsible for the observed correlations.

Variables in the same group do not necessarily imply that they are similar but that they have the same effect to the response in question. FA is a data reduction technique that is used to simplify a large number of inter-correlated variables or measures to a few representative factors which will be used for subsequent analysis. Looking at the size and number of variables of interest (162) from the 2009 GHS and 2009 QLFS data collected by Stats SA it is justifiable to use FA to identify the unobservable factors.

4.2.1.1 Education

The first step in FA is to identify unobservable factors that are being faced at educational institutions. A total of 29 variables relating to education issues were considered in this section with the hope of reducing this number of variables by grouping those that are correlated and explain the same problem. The next step is to determine whether the 29 observed variables are linearly related to a smaller number of unobservable factors, and to discover if these 29 observed variables can be explained in terms of few unobservable factors. As discussed in Section 3.3.1 a Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity are conducted first to examine the appropriateness of factor analysis. The two tests were conducted and the results are as shown in Table 4.1.

Table 4.1: KMO and Bartlett's test for the problems at educational institutions

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.974
Bartlett's Test of Sphericity	Approx. Chi-Square	2.620E9
	df	406
	Sig.	.000

According to Field (2005) the KMO value of 0.974 obtained in Table 4.1, strongly indicates that FA is appropriate, and since the value is closer to 1, this further indicates that patterns of correlations among the 29 variables are relatively compact and FA will yield reliable factors. This is also supported by the approximate Bartlett's test of sphericity which tests the hypothesis that the correlation matrix formed by these 29 variables is an identity matrix. The corresponding Bartlett's chi-square test is very large in this case $\left(2.620 \times 10^{9}\right)$ and its significant level is smaller than 0.005,

justifying the use of FA. Therefore the hypothesis that the 29 variables are independent is rejected.

Table 4.2 shows the total variance explained or accounted for by each extracted factor. The principal component analysis (PCA) method explained in Section 3.3.4 is used for the extraction. Although four factors have been generated (under the component column in Table 4.2), not all of them are useful in representing the 29 variables. Those with eigenvalues greater than one (which are four in this case) are considered (see Section 3.3.6.1 under initial eigenvalue). The four factors will contain as much information as the 29 variables. These four factors account for 84.283% of the total variance attributed to these variables. This shows that the model with only four factors is adequate to represent the data.

Table 4.2: Total variance for the extracted factors of problems at educational institutions

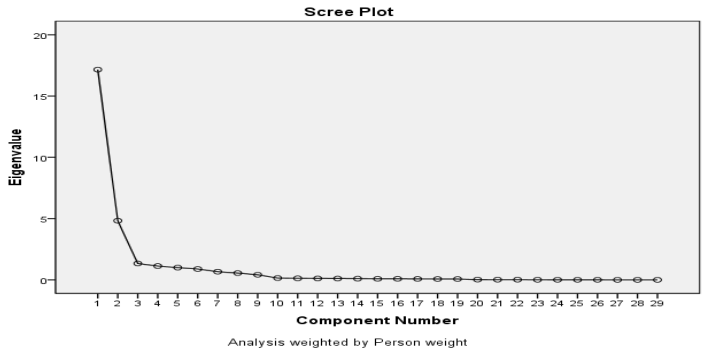
Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	17.157	59.162	59.162	15.516	53.504	53.504
2	4.826	16.641	75.803	6.237	21.506	75.010
3	1.334	4.601	80.404	1.553	5.357	80.366
4	1.126	3.881	84.285	1.136	3.919	84.285
5	.999	3.444	87.729			
6	.885	3.051	90.780			
7	.666	2.295	93.075			
8	.553	1.908	94.983			
9	.413	1.425	96.408			
27	.001	.004	99.996			
28	.001	.003	99.999			
29	.000	.001	100.000			

Extraction Method: Principal Component Analysis.

NB components 10 to 26 have been deleted so that the table can be reduced

The scree plot shown in Figure 4.1 also suggests that a model with four factors is sufficient to represent the data.

Figure 4.1: Scree plot for the extracted factors of the problems at educational institution



The component transformation matrix (Table 4.3) presents the correlations that relate the 29 variables to the four extracted factors after varimax orthogonal rotation. The coefficients called the factor loadings, indicate how closely the 29 variables are related to each of the four factors. Table 4.3 shows the component transformation matrix of the extracted four factors. The factor correlations indicate that the four factors are not strongly correlated as shown by their low coefficients. This implies that the extracted four factors are independent of each other and that they are explaining different aspects of the 29 initial variables. Table 4.3 also shows the rotated component or factor matrix after a varimax orthogonal rotation.

Table 4.3: Component (factor) transformation matrix

Component	1	2	3	4
1	**.934**	.338	.116	.026
2	-.335	**.941**	-.046	.000
3	-.125	.004	**.992**	.026
4	.021	.009	.028	**.999**

Extraction Method: Principal Component Analysis.

The rotated factor matrix (Table 4.4) contains the rotated factor loadings which are the correlations between the variables and the factors. Because these are correlations, possible values range from -1 to +1. All the correlations that are less than 0.33 have been excluded as they are not significant. In addition, by removing the clusters of low correlations that are probably not meaningful anyway, makes the output easier to read.

Table 4.4: Rotated component (factor) matrix for the problems at educational institution

Rotated Component Matrixa				
	Component			
	1	2	3	4
Total amount of tuition fees paid	.972			
Educational institution	.971			
Public or private institution	.967			
Distance learning classes	.966			
Means of transport	.951			
Problems at educational institution: Classes too large/too many learners	.944			
Problems at educational institution: Lack of teachers	.944			
Problems at educational institution: Fees too high	.944			
Problems at educational institution: Facilities in bad condition	.944			
Problems at educational institution: Poor quality of teaching	.941			
Time taken to school	.940			
Problems at educational institution: Other	.933			
Problems at educational institution: Teachers are often absent from school	.927			
Problems at educational institution: Book	.923			
Problems at educational institution: Teachers were involved in a strike	.921			
Violence, corporal punishment or verbal abuse	.850			
Absent from school	.846			
Reason no fees payment	.595			
Nature of violence: Other		.977		
Nature of violence: Physical violence by teacher		.976		
Nature of violence: Verbal abuse by teachers		.976		
Nature of violence: Physical abuse		.976		
Nature of violence: Verbal abuse by learners		.976		
Nature of violence: Corporal punishment by teacher		.976		
Mother part of the household			.800	
Father part of the household			.786	
Highest education level			.446	
South African Province				-.748
Population group				.710
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.				
a. Rotation converged in 4 iterations.				

The four factors that were extracted are the factors that we are most interested in, as they are factors that are explaining our data. For example, the first factor is relating to 18 variables which are all pointing to the *fees and teacher's behaviour*. The

behaviour of teachers, such as going on strike, teachers absent from school and poor quality of teaching, load highly on this factor. Variables such as high tuition fees, fees too high, etc. all point to monetary issue. We might call this first factor *fees too high*, since the students' fees are too high.

The second factor might be called *violence*, as most of the 6 variables which load highly on this second factor are all referring to violence between teachers and students. The violence is in the form of verbal abuse, physical abuse or corporal punishment of students by teachers. The third factor has to do with parental care, and might be called *absence of parental care*, as this factor is referring to problems accounted for by students who either do not have the father, or mother, or both parents as part of the family.

The fourth factor which has two variables loading highly on it might be called *historical advantage*. This is so because it is referring to the population group and also the province where the students come from. It is a fact that historically the financial assistance that was given to or received by students was different within population groups and also different within provinces.

4.2.1.2 Health

A total of 45 variables are considered under this section. The KMO value of 0.947 (Appendix A) is obtained, which justifies the use of FA since it is very close to 1. The use of FA is also supported by the approximate Bartlett's test for sphericity which tests the hypothesis that the correlation matrix is an identity matrix and therefore the 45 variables are independent. The corresponding Bartlett's chi-square test of sphericity (5.196×10^{9}) and the significant level which is very small (smaller than 0.005) all prove that the hypothesis that the 45 variables are independent, can be rejected (Appendix A). It is important to consider the fact that if the probability value is greater than 0.005, then FA should not be performed on the data. A total of six factors were extracted, contributing 88.366% (Appendix B) of the total variance. This again shows that the model with only six factors is adequate to represent the health data. A positive loading from Table 4.5 shows a positive relationship on the variable,

and a negative relationship shows an inverse relationship between the variable and the factor.

In Table 4.5, the first factor relates to the 18 variables, most of which are referring to illness that can be controlled or managed if behavioural change is encouraged. Illness such as abuse of alcohol or drugs, severe trauma due to violence, assault beatings, gunshot wounds, minor trauma, sexually transmitted disease, motor or vehicle accident injuries, flu or acute respiratory tract infection, diarrhoea, depression or mental illness, TB or severe cough with blood and diabetes, can be reduced if behavioural change is encouraged. We might call this factor "*manageable illness*". Behavioural change campaigns need to be encouraged if these illnesses are to be eradicated.

The second factor is referring to availability of medication of mainly chronic illness. This factor can be referred to as "*medication*". The third factor loads highly on chronic illnesses and can therefore be called "*chronic illness*". The fourth factor consists of illnesses which are related to permanent injury and this factor can be called "*physical and social disability*". The fifth factor is mainly referring to pregnancy illness and can be referred to as "*pregnancy*". The last and sixth factor is relating to one's status in the community and the corresponding population group. This factor can be referred to as "*social*". All the 45 health variables can be expressed in, or compressed into six factors that contain all the information that is explained by all the 45 variables.

Table 4.5: Rotated component matrix for health

Rotated Component Matrixa						
	Component					
	1	2	3	4	5	6
Nature of illness/injury: Abuse of alcohol or drugs	.992					
Nature of illness/injury: Severe trauma due to violence, assault beatings	.992					
Nature of illness/injury:Gunshot	.992					
Nature of illness/injury: Cancer	.992					
Nature of illness/injury: Minor trauma	.992					
Nature of illness/injury: Sexually transmitted disease	.992					
Nature of illness/injury: Motor or vehicle accident injuries	.992					
Suffer illness/injuries	.992					
Nature of illness/injury: Flu or acute respiratory tract infection	.992					
Nature of illness/injury: Diarrhoea	.992					
Nature of illness/injury: Do not know	.991					
Nature of illness/injury: Depression or mental illness	.991					
Nature of illness/injury: TB or severe cough with blood	.991					
Nature of illness/injury: Diabetes	.990					
Nature of illness/injury: Other illness or injury	.989					
Nature of illness/injury: High blood pressure	.988					
consult a health worker	.909					
Why not consult	.399					
Medication for chronic illnesses: Cancer		.970				
Medication for chronic illnesses: HIV and AIDS		.959				
Medication for chronic illnesses: Asthma		.947				
Medication for chronic illnesses: Arthritis		.945				
Medication for chronic illnesses: Diabetes		.944				
Medication for chronic illnesses: Other		.939				
Medication for chronic illnesses: Hypertension/high blood pressure		.930				
Chronic illness: Cancer			.992			
Chronic illness: HIV and AIDS			.984			
Chronic illness: Asthma			.971			
Chronic illness: Other			.969			
Chronic illness:Arthritis			.969			
Chronic illness: Diabetes			.968			
Chronic illness: Hypertension/high blood pressure			.924			
Difficulties: Concentrating				.947		
Difficulties: Remembering				.941		
Difficulties: Self-care				.925		
Difficulties: Hearing				.916		
Difficulties: Walking				.901		
Difficulties: Communication				.894		
Difficulties: Seeing				.796		
Pregnancy					.978	
Gender					-.961	
Current status of pregnancy					.409	
Marital status						.827
Age group						-.712
Population group						-.598
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.						
a. Rotation converged in 6 iterations.						

4.2.1.3 Housing

One of the goals of factor analysis is to assess how much of variance is due to the causal influence of latent factors that have been extracted. When the extracted factor accounts for variance in a variable, that variable should be retained by the extracted factors. On the basis of KMO value of 0.630 (Appendix C), it appears that the data set is suitable for FA. This value strongly indicates that FA is appropriate. KMO looks not only at the correlations, but their patterns between variables. Similar to Table 4.1 the corresponding Bartlett's chi-square test (7.057×10^7) as it appears in Appendix C is very large and its significant level is smaller than 0.005, also justifying the use of FA. The hypothesis that the 16 variables are independent can be rejected and we can therefore conclude that there are some interrelationships among the variables.

These arguments all collectively indicate that the set of variables is appropriate for FA. Six factors with an eigenvalue greater than 1 are extracted. All the extracted factors as appearing in Appendix D account for 67.480% of the total variance, a figure which is significantly high. We can conclude that these six factors can represent the housing data in the 2009 GHS. In order to identify which variable belong to which component or factor, a varimax orthogonal rotation analysis was carried out and the results are shown in Table 4.6. Factor 1 loads highly on wall material, roof material, number of rooms, roof material, and market value of the property. This factor can be called *brick house*. Factor 2 loads highly on original beneficiary of the dwelling, RDP or state subsidised dwelling and house subsidy received. This factor might be called *government assistance*. The third factor (factor 3) loads strongly on variables such as ownership of dwelling and monthly rent or mortgage. This factor can be interpreted as *home owners*. Factor 4 loads on dwelling originally built, age of household head and population group of household. This factor might be referred to as *age of house*. Factor 5 loads strongly on members on the waiting list and RDP or state subsidized house waiting list and therefore it might be called *waiting list*. Factor 6 loads highly on sex of household head. This sixth and last factor can be interpreted as *male household heads*.

Table 4.6: Rotated factor (component) matrix for housing

	Component					
	1	2	3	4	5	6
Walls material	.756					
Number of rooms: Open plan dining rooms/sitting rooms/TV rooms	.730					
Roof material	.707					
Market value of the property	.611					
Original beneficiary of the dwelling		.853				
RDP or state subsidized dwelling		.842				
House subsidy received		.635				
Ownership of dwelling			.958			
Monthly rent or mortgage			.948			
Dwelling originally built				.678		
Age of household head				.665		
Population group of household head				.564		
Main dwelling				-.557		
Members on the waiting list					.858	
RDP or state subsidized house waiting list					.817	
Sex of household head						.911

4.2.1.4 Social Development

The 2009 GHS has generated 35 social development variables that are analysed under this section. The KMO value of 0.880 (Appendix E) indicates that FA is appropriate, and since the value is closer to 1 this indicates that patterns of correlations among the 35 variables are relatively compact and FA will yield reliable results. This is also supported by the approximate Bartlett's test of sphericity which tests the hypothesis that the correlation matrix is an identity matrix. KMO and Bartlett's tests show that the hypothesis that these 35 variables are independent can be rejected (0.00<0.005), hence concluding that some of these variables are correlated. Thus, FA would be useful in reducing and grouping the correlated variables into unobservable factors. Grouping the initial 35 variables into nine factors, altogether accounts for 75.488% (Appendix F) of total variability. This shows that the model with only nine factors is adequate to represent the data since this value is significantly large. The rotated component matrix (Table 4.7) presents the

correlations that relate the 35 variables to the nine extracted factors after varimax orthogonal rotation. Rotation is a process that involves redistribution of the variation for the different variables between components such that each variable is more or less clustered in one factor than being spread throughout the factors.

All the correlations or values of factor loading below 0.33 on either direction were suppressed because they are insignificant. The suggested cut-off point is appropriate for the interpretative purpose (i.e. the loadings greater than 0.33 represent a substantive value). The nine factors that were extracted are the factors that we are interested in and the strength (value of component loading) of the relationship between the factors or components and initial variables allow us to interpret the results as follows: The rotation of the factor structure (varimax) indicates that ten variables loaded very highly onto factor 1. Factor 1 seems to represent problems experienced while visiting the health worker/facility. We will call this factor the *no problem with health*. The four variables that load highly on factor 2 all seem to relate to water supply and we call this factor *sufficient water*. All the four variables that load highly on factor 3 all seem to be related to source of income such as rental income, interest, income from a business and sales of farm products. Factor 3 is then called *high income*.

Factor 4 is positively attributed to payment of social services and problems that arise from payment of sewerage system and type of toilet facilities. We might call this factor *payment of sewerage*. Factor 5 is called *telephone access* since it consists of telephone, use of telephone in past five years, and cellular. Factor 6 might be regarded as *absence of toilet* because it is characterised by toilet facility shared and location of the toilet facility. The variables that load highly on factor 7 all seem to be related to the number of household members, especially children. This factor might be given a label *household size*. Factor 8 may be called *water interruption* since it is related specifically to problems with water supply. Finally factor 9 might be attributed to old aged people who are getting grants and pensions and as such this factor can be called *pensioners*.

Table 4.7: Rotated component matrix for social development

	Component								
	1	2	3	4	5	6	7	8	9
Problems experienced while visiting the health worker/facility: Other	.970								
Problems experienced while visiting the health worker/facility: Incorrect diagnosis	.963								
Problems experienced while visiting the health worker/facility: Facilities not clean	.952								
Problems experienced while visiting the health worker/facility: Have never been	.941								
Problems experienced while visiting the health worker/facility: Opening times not convenient	.937								
Problems experienced while visiting the health worker/facility: Too expensive	.933								
Problems experienced while visiting the health worker/facility: Staff rude or uncaring or turned patient away	.906								
Problems experienced while visiting the health worker/facility: Drugs that were needed not available	.874								
Problems experienced while visiting the health worker/facility: Long waiting time	.822								
Medical help	.494								
Free basic water		.905							
Water supply interruption		.887							
Water supply		.799							
Reason for non payment		.571							
Source of income: Sales of farm products and services			.851						
Source of income: No income			.849						
Source of income: Other income sources e.g. rental income, interest			.732						
Source of income: Income from a business			.546						
Payment for the sewerage system				.703					
Type of toilet facility				.663					
Population group of household head				-.641					
Source of income: Remittances				-.432					
Telephone					.805				
Use of telephone in the past five years					.779				
Cellular telephone					.697				
Toilet facility shared						.918			
Location of the toilet facility						.885			
Causes of piped water interruption							.945		
Duration of water interruption							.942		
Children 17 and younger								.901	
Children 5 and younger								.895	
Age of household head									.812
Source of income: Pensions									-.711
Source of income: Grants									-.554
Source of income: Salaries/wages/commission									.499

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

4.2.1.5 Labour Force

The KMO value of 0.910 (Appendix G), indicates that FA is an appropriate method, and since the value is closer to 1 this indicates that patterns of correlations among the 38 variables are relatively compact and FA should yield different and reliable factors. This is also supported by the approximate Bartlett's test which tests the

hypothesis that the correlation matrix is an identity matrix. The Bartlett's chi-square test (5.717×10^9) in this case has significant level less than 0.005. This implies that the hypothesis that the 38 variables are independent can be rejected. Four factors with eigenvalues greater than 1 are extracted. The extracted four factors explain 84.283% of the total variation (Appendix H). This four-factor model should be sufficient to summarise total sample variance. The cells in Table 4.8 are factor loadings, and indicate the strength of the relationship between each factor and each variable. These factor loadings can be interpreted as correlation between each factor and each variable. The correlation coefficients range between -1 and 1, with 0 indicating no correlation. The factor analysis transforms a set of correlated observed variables into a set of uncorrelated factors (i.e. factors in Table 4.8 are uncorrelated).

The first factor as it appear in Appendix H has maximum variance of 34.3%, subject to the constraints that the sum of the squared factor loadings is equal to 1. The second factor has the next highest variance (28.824%), given the constraints that the sum of the squared factor loadings is equal to 1 and it is uncorrelated with factor 1. The subsequent factors or dimensions have decreasing order of variance, subject to the same constraints as applied to the second factor. As shown in Table 4.8, Factor 1 can be roughly interpreted as *employment* because it is positively correlated with variables such as employment status, look for work, produce goods, do construction, etc. We can interpret factor 2 as *industrial business and occupation*, since we observe positively high correlations with type of business or enterprise, sectors excluding and including agriculture to formal and informal sectors, underemployment, employment contract and work status, number of employees and industry. This factor is negatively correlated with variables such as main industry and main occupation.

We can interpret factor 3 as *employment history*, since we observe negatively high correlations with previous occupation (group) and previous industry (group). This factor is strongly positively correlated with variables such as main reason you stopped working, previous occupation, employer and previous industry. The fourth factor is *long-term unemployment*, clearly correlated to variables such as duration of trying to find work, long-term unemployment and placing adverts.

Table 4.8: Rotated component (factor) matrix for the labour force

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Status	.977			
Look for work	.933			
Have paid work	.924			
Catch food	.916			
Charity	.915			
Do construction	.915			
Welfare grants	.914			
Unemployment status	.905			
Do farm work	.903			
Child support grant	.901			
Involvement in at least one non-market activity	.884			
Have a job or start a business	.871			
Pension	.860			
Have own business	.801			
Have unpaid work	.785			
Inactivity reason	.723			
Age	.573			
Main industry		-.971		
Main occupation		-.969		
Underemployment		.963		
Type of business or enterprise		.933		
Occupation		.889		
Sector (includes agriculture in the formal and informal sectors)		.842		
Sector (excludes agriculture from formal and informal sectors)		.842		
Number of employees		.841		
Main work		.841		
Work status		.833		
Industry		.802		
Informal employment		.694		
Previous occupation (grouped)			-.969	
Previous industry (grouped)			-.969	
Main reason you stopped working			.933	
Previous occupation			.925	
Whom did you work for			.919	
Previous industry			.853	
How long been trying to find work				.856
Long-term unemployment				.795
Placed adverts				.698

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

The component transformation matrix of the labour force shows that four factors extracted, are independent of each other and that they are explaining different aspects of the 38 initial variables. The components are also perfectly correlated with themselves and perfectly uncorrelated with the others.

4.3 Summary

In this chapter we have applied FA to the 29 variables in education, 45 variables in health, 16 variables in housing, 35 variables in social development and 38 variables in labour force, from which 4, 6, 6, 9 and 4 factors respectively extracted. Thus, from a total of 162 variables 29 factors were extracted. The extracted factors are summarised in Table 4.9.

Table 4.9: Summary of the factors extracted by core area

Core area	Factor No.	Factor name	No. of vars/factor
Education	1	fees too high	18
	2	violence	6
	3	absence of parental care	3
	4	historical advantage	2
Health	1	manageable illness	18
	2	medication	7
	3	chronic illness	7
	4	physical and social disability	7
	5	pregnancy	3
	6	social	3
Housing	1	brick house	4
	2	government assistance	3
	3	home owners	2
	4	age of house	4
	5	waiting list	2
	6	male household heads	1
Social Development	1	no problem with health	10
	2	sufficient water	4
	3	high income	4
	4	payment of sewerage	4
	5	telephone access	3
	6	absence of toilet	2
	7	household size	2
	8	water interruption	2
	9	pensioners	4
Labour force	1	employment	17
	2	industrial businesses & occupation	12
	3	employment history	6
	4	long-term unemployment	3

CHAPTER 5: MULTINOMIAL LOGISTIC REGRESSION

5.1 INTRODUCTION

Chapter 5 focuses on applying multinomial logistic regression (MLR) on education, housing, social development and labour force. The following issues are covered: significance test of the model, log likelihood (change in -2LL), measures analogous to R^2: Cox and Snell R^2 and Nagelkerke R^2, classification matrices as a measure of model accuracy and parameter estimations. The marginal percentage has been used to calculate proportion by chance of accuracy and compares it to prediction of accuracy rate. The MLR was used ignoring the fact that there is an ordinal nature in the categories of the dependent variables (Chan, 2005). This implies that ordinal regression may not be suitable because Parallel lines test assumptions were not met when using our datasets.

5.2 APPLYING MLR TO EDUCATION FACTORS OF GHS

MLR is used to regress the education outcomes against the education factors of GHS identified by factor analysis. The response variable (dependent variable) is the educational levels. A multinomial variable, originally encoded on a scale from grade 0=1 to tertiary levels=29 with no schooling=98, was recoded into four categories: “no schooling”, “less grade 12”, “grade 12”, “above grade 12”. For this modelling the reference category is chosen to be “no Schooling”. The independent variables are: E1Feshh (*fees too high*)**,** E2Viol (*violence*), E3AbsP (*absence of parental care*) and EHisA (*historically advantaged*). The results are shown in the sections that follow.

5.2.1 Overall Test of Relationship

The first step is to study the overall relationship between the education level and the independent factors identified by factor analysis. Table 5.1 is used to test the presence of such a relationship based on the chi-square distribution. The null hypothesis for this test is that there is no difference between the null model (that is, model without the independent variables) and the final full model (that is the model that includes the independent variables), versus the alternative hypothesis that there is a difference between the null model without the independent variables and the final model with the independent variables. The difference between these two

measures is the model chi-square value 23 819 596.724 (65 374 295.141- 41554 698.417) that is tested for statistical significance. The test statistic value, 23 819 596.724 has a significant level less than 0.05 (p=0.00<0.05). This implies that the null hypothesis that there is no difference between the null and final models is rejected. It is concluded that there is evidence to support the fact that there is a relationship between the levels of education outcomes and the associated identified independent factors.

Table 5.1: Model fitting information for education

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Null	65 374 295.141			
Final	41 554 698.417	23 819 596.724	12	.000

5.2.2 Strength Overall Test of Relationship

The strength of the dependent and independent variable in the MLR model will be assessed in this subsection by considering the pseudo R^2 correlations, classification of accuracy measures, likelihood ratio tests and the Wald test.

5.2.2.1 Pseudo R-squared

The correlation measures provided by MLR analysis are the pseudo R^2. The pseudo R^2 as it appears in Table 5.2 accounts for the amount of variance explained in the outcome variable by the independent variables. The Cox and Snell and Nagelkerke pseudo R^2 in Table 5.2 suggest that the variation in the level of education outcomes explained by the education factors ranges between 51% and 59%. Thus, a relatively high level of variation is explained by the model.

Table 5.2: Pseudo R^2 for education

Cox and Snell	Nagelkerke
0.509	0.593

5.2.2.2 Evaluating the Usefulness of the MLR Model

The pseudo R^2 has provided us with a measure of the extent of association between the independent and dependent variables of education but what it fails to do is give us the extent of the accuracy or the errors inherent in the model. This is exactly what

the accuracy of classification measures do. This measure, which is also a measure of the extent of the strength of the relationship between the independent and dependent variables, assesses the accuracy of the model by comparing the predicted values of the model to the observed values.

To calculate the accuracy of classification we first consider the marginal frequencies for “no schooling”, “less grade 12”, “grade 12” and “above grade 12” which are 6.4%, 60.2%, 29.1% and 4.3%, respectively (Appendix I). These are then used to calculate the proportion by chance accuracy rating which was found to be 45.3% (i.e. $0.064^2+0.602^2+0.291^2+0.043^2=0.453$). The benchmark that is used to characterise a multinomial logistic regression model as useful is a 25% improvement over the rate of accuracy achievable by chance alone. Thus, the classification accuracy rate should be at least 25% more than the proportion by chance accuracy rate of 45.3%, i.e. it must be at least 57% for the MLR model to be adequate. Table 5.3 shows the comparison of the observed and the predicted levels of education outcomes and the extent to which they can be correctly predicted.

Table 5.3: Classification of accuracy for education

Observed	Predicted				
	no Schooling	less grade 12	grade 12	above grade 12	Percent Correct
no Schooling	**1993290.80055103**	55870.78500309	101415.23682665	1444.34908080	92.6%
less grade 12	6207.18783736	**19205913.62670121**	923700.03412630	2412.39669195	95.4%
grade 12	48386.75876271	6420652.92802807	**3214222.01375571**	46601.01303766	33.0%
above grade 12	37598.02957889	531015.76882438	861051.56959624	**13369.22594909**	0.9%
Overall Percentage	6.2%	78.3%	15.2%	0.2%	73.0%

There are two groups that have high levels of accurate prediction at 92.6% for “no schooling” and 95.4% for “less grade 12”. Correct classification is only 33.0% for “grade 12” and 0.9% for “above grade 12”. The correctly classified cases are on the diagonal in Table 5.3 and are shown here in bold font. The overall correct classification for all cases is 73.0% and the groups of the dependent variables with the strongest predictions are “no schooling” and “less grade 12”. This means that this model is more useful for those individuals whose highest level education is “no schooling” and “less grade 12”. This makes sense since beyond matriculation some

factors such as violence do not apply. The overall classification accuracy rate displayed in Table 5.3 is 73.0% which is greater than the proportional by chance accuracy criteria of 57%. It means the model improves on the proportion by chance accuracy rate by 25% or more so that the criterion for classification accuracy is satisfied and the model is adequate.

5.2.2.3 Relationship between Independent and Dependent Variables

The likelihood ratio test (LRT) in Table 5.4 presents the significance of each of the factors individually. It tests the improvement in the model fit with each of the factors. Each of these factor scores shown in Table 5.4 has a p-value of 0.00 which is less than 0.05. This means that there is a relationship between the dependent variables and the independent education factors; hence all the four education factors should be included in the model.

Table 5.4: Likelihood ratio test for education

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	85560035.236	44005336.819	3	.000
E1FesH	51671794.690	10117096.273	3	.000
E2Viol	50486798.649	8932100.232	3	.000
E3AbsP	54138915.640	12584217.223	3	.000
E4HisA	46083545.172	4528846.754	3	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

The LRT may have confirmed that the levels of education outcomes have an association with each of the education factors but this does not necessarily mean that each of the factors is statistically significant as far as distinguishing any of the two classified education level outcome variables. The Wald test, discussed in Section 5.2.3, is the one used to make this distinction.

5.2.2.4 Test for Statistically Significant Factors and Parameter Estimation

The results for fitting the MLR models for the four education factors are shown in Table 5.5. Summarised in Table 5.5 are coefficient estimates (β), the standard

errors, the Wald statistic, the Odds Ratio (OR) represented by $Exp(\beta)$ and their corresponding 95% confidence intervals for each level of education outcome. The Wald statistic along with the significance value (p-value) is used to test the significance of each factor. In this case it is to test if the education factor can significantly distinguish each education level against the reference level which is “no schooling”. The Wald test is used to measure the improvement brought about by adding each education factor on the intercept-only (null) model. All the Wald test p-values in Table 5.5 are equal to 0.00, hence all less than 0.05, which means that all the education factors are statistically able to differentiate between each of the education levels and the reference variable of “no schooling”. Thus all the factors must be retained in the model.

Table 5.5: Parameter estimates for education

Highest level of education[a]		B	Std. Error	Wald	df	Sig.	Exp(β)	95% Confidence Interval for Exp(β)	
								Lower Bound	Upper Bound
less grade 12	Intercept	49.350	.040	1506481.141	1	.000			
	E1FesH	-58.140	.051	1279200.104	1	.000	1.000E-013	1.000E-013	1.000E-013
	E2Viol	-221.511	.191	1347016.518	1	.000	1.000E-013	1.000E-013	1.000E-013
	E3AbsP	-10.009	.007	1824724.287	1	.000	4.501E-005	4.436E-005	4.567E-005
	E4HisA	-.928	.003	104447.745	1	.000	.395	.393	.397
grade 12	Intercept	38.359	.040	918495.793	1	.000			
	E1FesH	-44.041	.051	742303.797	1	.000	1.000E-013	1.000E-013	1.000E-013
	E2Viol	-163.485	.189	745649.133	1	.000	1.000E-013	1.000E-013	1.000E-013
	E3AbsP	-8.729	.007	1394707.255	1	.000	.000	.000	.000
	E4HisA	.064	.003	504.227	1	.000	1.066	1.060	1.072
above grade 12	Intercept	24.509	.041	349461.903	1	.000			
	E1FesH	-27.879	.053	276232.015	1	.000	8.802E-013	8.031E-013	9.656E-013
	E2Viol	-102.819	.196	274146.677	1	.000	1.000E-013	1.000E-013	1.000E-013
	E3AbsP	-6.580	.008	748729.929	1	.000	.001	.001	.001
	E4HisA	.920	.003	97136.100	1	.000	2.509	2.495	2.524

a. The reference category is: no schooling.

According to the significance of each category of the independent variable the resulting model equations are as follows:

$$\log\left(\frac{<grade12}{no\_school}\right) = 49.350 - 58.140(\text{E1FesH}) - 221.511(\text{E2Viol}) - 10.009(\text{EAbseP}) - 0.928(\text{E4HisA}) \quad \textbf{(5.1)}$$

$$\log\left(\frac{grade12}{no\_school}\right) = 38.359 - 44.041(\text{E1FesH}) - 163.485(\text{E2Viol}) - 8.729(\text{E3AbseP}) + 0.064(\text{E4HisA}) \quad \textbf{(5.2)}$$

$$\log\left(\frac{>grade12}{no\_school}\right) = 24.509 - 27.879(\text{E1FesH}) - 102.819(\text{E2Viol}) - 6.580(\text{E3AbseP}) + 0.920(\text{E4HisA}) \quad \textbf{(5.3)}$$

where

Dependent variable (Education levels) categories are:
<grade 12 = less grade 12; grade 12 ; > grade 12= above grade 12

Independent variables are:
E1FesH=fees too high; E2Viol=violence; E3AbsP=absence of parental care; E4HisA=historically advantage.

The reference category of the Dependent Variable is “no school =0”. Maximum likelihood estimates determine the effect for all pairs of categories. In the four outcome category models there are three logit functions (Table 5.5).

5.2.3 Check for multicollinearity and numerical errors

The standard errors of the coefficient estimates (β) are used to check for numerical errors or multicollinearity in the solution of the multinomial logistic regression. A standard error that is greater than 2.0 indicates numerical problems, such as multicollinearity among the independent education factors. For this model no standard error is greater than 2, (Schwab, 2002). The model loss precision when the confidence interval is wider. For the significant variables, small confidence intervals suggest greater precision of the variable (Kleinbaum *et al*., 2008). Confidence intervals displayed in Table 5.5 are very small, suggesting great precision of factors. Confidence intervals that include 1, mean that there is no significant relationship between educational level and factor scores (Hosmer and Lemeshow, 1989). We have to account for the increased Type I error because of the large number of

statistical tests being run (Petrucci, 2009). This means that we have to avoid using standard $p<0.05$ critical value, the correct value to be utilised is obtained by dividing 0.05 by the total number of predictors (Tabachnick and Fidell, 2007; Petrucci, 2009). For our model the statistical significance should be determined at a $p<0.0125$ obtained by dividing 0.05 by 4.

5.2.4 Interpretation of the MLR Results

Looking at E4HisA (which is the *historically advantaged*) it can be deduced that, when holding other factors constant, the odds for someone who is historically advantaged of proceeding to complete tertiary (above grade 12) rather than remain with no formal schooling are 2.51 times than the historically disadvantaged person. Holding other factors constant, it can be seen that the historically disadvantaged are 2.53 (1/0.395) times more likely than the historically advantaged to have no formal schooling at all than achieving at most grade 11 education. The odds are almost the same though slightly higher, at 1.07, for the historically advantaged to complete grade 12 than those with no formal schooling. Respondents are significantly more unlikely to have any form of education compared with no schooling at all on E1FesH (*Fees too high*), E2Viol (*Violence*) and E3AbsP (*Absence of parental care*). This is because the odds ratios between three groups (less grade 12, Grade 12 and above grade 12) when compared to no schooling give odds ratios that are close to zero on these factors. Thus, there is a huge difference between these three groups and no schooling on these three factors.

5.3 APPLYING MLR TO HEALTH FACTORS OF GHS

The current health status does not have an effect on one's educational status and there is no data linking health status to service satisfaction. So testing significant of health factors should be ignored.

5.4 APPLYING MLR TO HOUSING AND SOCIAL DEVELOPMENT FACTORS OF GHS

The response variable is the level of service at the time of the household visit, a multinomial variable, originally encoded on a scale from "very satisfied"=1 to "very dissatisfied"=5, was recoded into three categories: "very satisfied"=1, "satisfied"=2

and “dissatisfied”=3. The predictors (independent factors) are housing and social development factors.

5.4.1 Overall Test of Relationship

The model fit analysis obtained in chi-square statistic with value 5.598×10^{6}, with a p-value of 0.00 shows that the model is significant as indicated in Table 5.6. We reject the null hypothesis which states that there was no difference between the model without explanatory factors and the model with explanatory factors. Hence the existence of a relationship between housing-social development factors and service satisfaction was supported.

Table 5.6: Model fitting information for housing and social development

	Model Fitting Criteria	Likelihood Ratio Tests		
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	23175893.290			
Final	17577671.394	5598221.896	30	.000

5.4.2 Pseudo R-squared

The pseudo R^2 statistics as used before were as follows: Cox and Snell, 0.343 and Nagelkerke, 0.416 (Table 5.7). The Cox and Snell and Nagelkerke pseudo R^2 suggest that the variation in the service satisfaction outcomes explained by the housing and social development factors range between 33.4% and 40.5% (Table 5.7).

Table 5.7: Pseudo R^2 for housing and social development

Cox and Snell	Nagelkerke
.343	.416

5.4.3 Evaluating the Usefulness of the MLR Model

The classification of accuracy rate or the overall percentage as shown in Table 5.8 was 72.4% greater than the proportional by chance accuracy rate of 62.1% ($1.25\times49.66\%=62.1\%$), hence satisfying the criterion for classification accuracy. The service satisfaction model gives better accuracies for “very satisfied” group only with 95%. Hence our model would not be a “good” model if we want to predict the “satisfied” with 25.4% and “dissatisfaction” with 31.9%.

Table 5.8: Classification of accuracy for housing and social development

Observed	Predicted			
	Very satisfied=1	Satisfied=2	Dissatisfied=3	Percent Correct
very satisfied=1	**8367394.7966998**	310171.3915766	113231.2656579	95.2%
Satisfied=2	1773255.1938214	**704940.4267347**	296621.0477505	25.4%
dissatisfied=3	547943.8274113	635052.2481115	**584051.5228805**	33.1%
Overall Percentage	80.2%	12.4%	7.5%	72.4%

5.4.4 Relationship between Independent and Dependent Variables

Each of these factor scores has a p-value of 0.00 which is less than 0.05, implying that all housing and social development factors have a significant effect on service delivery satisfaction. Hence all the six housing and nine social development factors should be included in the model.

5.4.4 Test Statistically Significant Factors and Parameter Estimation

The results for fitting the MLR models for the housing and social development factors are shown in Table 5.9. The Wald test shows that all the factors are statistically significance. Our statistical significance should be determined at a $p<0.003$ (0.05 divide by 15) for housing and social development factors.

5.4.6 Check for Multicollinearity and Numerical Errors

Just like in the previous sections we have to check the standard errors for the parameters' explanatory variables that are larger than 2. Looking at Table 5.9 there are no standard errors which are greater than 2, hence there is no factor causing numerical problem (multicollinearity).

5.4.7 Interpretation of the MLR Results

Two groups of parameter estimates are produced in Table 5.9 to compare categories "very satisfied" and "satisfied" with the reference category "dissatisfied". The first six factors are for housing and the remaining nine factors are for social development. We start by interpreting parameter estimates for housing, followed by those of social development as shown in Table 5.9.

We first look at the parameter table comparing "satisfied" and "dissatisfied". All but

two variables are statistically significant. Some, although statistically significant, have odds ratios close to one and do not make interesting interpretations. Thus, for those parameters which are significant we will interpret those which are giving us some interesting inferences and are not close to one (1±0.05). The rest of the factors are included for modelling and reference purposes.

We finally look at the parameter table comparing “very satisfied” and “dissatisfied”. All the variables are statistically significant. Some variables, although statistically significant, have odds ratios close to one and do not make interesting interpretations. Thus, for those parameters which are significant we will interpret those which gave us some interesting inferences and are not close to one (1±0.05).

In the four outcome category models (see Table 5.9), there are two logit functions of housing factors as follows:

$$\log\left(\frac{VerySat=1}{DisSat=3}\right)=2.711$$
$$-0.025(\text{H1BrickH})+0.073\text{H2(GoverA)}-0.0990(\text{H3HomeO})+0.152(\text{H4AgeH}) \quad \textbf{(5.4)}$$
$$-0.008\text{H5(WaitL)}-0.014\text{H6(MaleH)}$$

$$\log\left(\frac{Sat=2}{DiSat=3}\right)=1.058+0.007(\text{H1brickH})+0.049(\text{H2GoverA})-0.002(\text{H3HomeO}) \quad \textbf{(5.5)}$$
$$+0.024(\text{H4AgeH})-0.027(\text{H5WaitL})+0.031(\text{H6MaleH})$$

where

Dependent variable (Service Satisfaction) categories are:

$VerySat$ = very satisfied; Sat =satisfied; $DiSat$ =dissatisfied

Independent variables are:

H1BrickH=brick house; H2GovA=government assistance; H3HomO=home owners; H4AgeH=age of house; H5WaiL=waiting list; H6MalH=male household’s heads

In the four outcome category models there are two logit functions of social development factors (see Table 5.9**)**.

$$\log\left(\frac{VerySat=1}{DisSat=3}\right)=2.711+9.490(\text{S1NopH})+0.420(\text{S2SuffW})+0.877(\text{S3HighI})$$
$$-0.626(\text{S4PayS})+0.751(\text{S5TelepA})+0.192(\text{S6AbseT})+0.155S7(\text{HouseZ}) \quad \textbf{(5.6)}$$
$$-0.117(\text{S8WaterI})-0.170(\text{S9Pens})$$

$$\log\left(\frac{Sat=2}{DiSat=3}\right)=1.058+2.414(\text{S1NopH})+0.051(\text{S2SuffW})+0.223(\text{S3HighI})$$
$$-0.097(\text{S4PayS})+0.181(\text{S5TelepA})+0.009(\text{S6AbseT})-0.022(\text{HouseZ}) \quad \textbf{(5.7)}$$
$$-0.082(\text{S8WaterI})-0.041(\text{S9Pens})$$

where

Dependent variable (Service satisfaction) categories:

VerySat =very Satisfied; *Sat* =satisfied; *DiSat* =dissatisfied

Independent variables:

S1No_pH=no problem with health; S2SuffiW=sufficient water; S3HighI=high income; S4PayS=payment of sewerage; S5TelepA=telephone access; S6AbseT=absence of toilet;S7HouseZ=household size;S8WaterI=water interruption; S9Pens=pensioners

5.4.7.1 Housing

Looking at H3HomeO, the *home owners* are 9.7% less likely to be very satisfied than dissatisfied when compared to non-owners. H4AgeH **(***age of a house***)**, the odds of being “very satisfied” rather than “dissatisfied” increased by a factor 1.165 by living in a new house than living in old houses (mud or wood house).H4AgeH **(***age of a house***)** was 17% more likely to influence the “very satisfied” group than “dissatisfied” group.

5.4.7.2 Social Development

Holding other factors constant, the odds for someone with S1NopH (*no problem with health*) were 11.18 times more likely to be in the “satisfied” group instead of “dissatisfied” than those with health problem. This makes sense since there was free treatment for public hospitals. Persons with S3HighI (*high income***)** are 1.25 times more likely to be satisfied than dissatisfied when compared to those with low income. Those who are not paying for sewerage 1.10(1/0.907) times more likely to be “satisfied” with service delivery than those who are paying. Looking at S5TelepA (*telephone access*), holding other factors constant respondents with access to

telephone were 1.199 times more likely to be “satisfied” instead of “dissatisfied” than those with access to telephone. The odds of being “dissatisfied” is 1.09(1/0.921) times more than “satisfied” with each unit increase in S7HouseZ (household size).

Table 5.9: Parameter estimates for housing and social development

Parameter Estimates

Service satisfaction during the visit[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Bound	Bound
very satisfied=1	Intercept	2.711	.001	3587249.157	1	0.000			
	H1BrickH	-.025	.001	312.508	1	.000	.975	.973	.978
	H2GoverA	.073	.001	4589.255	1	0.000	1.076	1.074	1.078
	H3HomeO	-.099	.001	6518.480	1	0.000	.905	.903	.908
	H4AgeH	.152	.002	8650.818	1	0.000	1.165	1.161	1.168
	H5WaitL	-.008	.001	67.074	1	.000	.992	.990	.994
	H6MaleH	-.014	.001	140.016	1	.000	.986	.984	.988
	S1NopH	9.490	.006	2433732.768	1	0.000	13222.129	13065.426	13380.711
	S2SuffiW	.420	.001	146953.463	1	0.000	1.521	1.518	1.525
	S3HighI	.877	.003	108637.547	1	0.000	2.404	2.391	2.416
	S4PayS	-.626	.001	186980.522	1	0.000	.535	.533	.536
	S5TelepA	.751	.001	286106.795	1	0.000	2.119	2.113	2.124
	S6AbseT	.192	.001	39388.695	1	0.000	1.211	1.209	1.214
	S7HouseZ	.155	.001	22749.491	1	0.000	1.167	1.165	1.170
	S8WaterI	-.117	.001	14104.355	1	0.000	.889	.888	.891
	S9Pens	-.170	.001	14015.203	1	0.000	.843	.841	.846
satisfied=2	Intercept	1.058	.002	462977.124	1	0.000			
	H1BrickH	.007	.001	27.417	1	.000	1.007	1.005	1.010
	H2GoverA	.049	.001	2174.268	1	0.000	1.050	1.048	1.052
	H3HomeO	-.002	.001	3.121	1	.077	.998	.995	1.000
	H4AgeH	.024	.002	219.547	1	.000	1.024	1.021	1.028
	H5WaitL	-.027	.001	719.705	1	.000	.974	.972	.976
	H6MaleH	.031	.001	678.025	1	.000	1.031	1.029	1.034
	S1NopH	2.414	.004	301613.754	1	0.000	11.183	11.087	11.280
	S2SuffiW	.051	.001	2165.795	1	0.000	1.052	1.050	1.054
	S3HighI	.223	.003	7431.009	1	0.000	1.250	1.244	1.256
	S4PayS	-.097	.001	4624.780	1	0.000	.907	.905	.910
	S5TelepA	.181	.001	18237.237	1	0.000	1.199	1.196	1.202
	S6AbseT	.009	.001	85.256	1	.000	1.009	1.007	1.011
	S7HouseZ	-.022	.001	477.315	1	.000	.978	.976	.980
	S8WaterI	-.082	.001	7145.133	1	0.000	.921	.920	.923
	S9Pens	-.041	.001	833.967	1	.000	.959	.957	.962

a. The reference category is: dissatisfied=3.

All the factors for social development are significant, p<0.003. Holding other factors constant, the odds for someone who had S1NopH (*no health problem***)** were many (13065) times more likely to be in the “very satisfied” group instead of “dissatisfied” than those with health problem. Holding other factors constant, respondents having S2SuffiW (*sufficient water***)** as one of the basic need were 1.52 times as likely to be

“very satisfied” group instead of “dissatisfied” group than those with no water at all. Those who have got S3HighI (*high income*) are 2.40 times more likely to be “very satisfied” than “dissatisfied” when compared to those with low income.

Those who are not *paying for sewerage* are 1.87(1/0.535) times more likely to be “very satisfied” than “dissatisfied” with service delivery than those who are paying (S4PayS). Looking at S5TelepA (*telephone access*), holding other factors constant, respondents with access to telephone were 2.12 times more likely to be “very satisfied” as opposed to “dissatisfied” than those without access to telephone.The odds of being “dissatisfied” is 1.13 (1/0.889) times more than “very satisfied” with each unit increase at S7HouseZ (*household size*)**.**

5.5 APPLYING MLR TO LABOUR FORCE FACTORS OF QLFS

MLR is used to regress the educational levels against the labour force factors identified by factor analysis. The response variable (dependent variable) is the educational levels. A multinomial variable, originally encoded on a scale from grade 0=1 to tertiary levels=29 with no schooling=98, was recoded into four categories: “no schooling”, “less grade 12”, “grade 12”, “above grade 12”. For this modelling the reference category is chosen as “no schooling”. The independent variables are employment, industrial business and occupational, employment history and long-term unemployment.

5.5.1 Overall test of relationship

The model fit analysis in Table 5.10 obtained a chi-square statistic value, 12271454.548 with a p-value of 0.00 showing that the model is significant. It follows that the null hypothesis that there is no difference between the model without independent variables and the model with independent variables is rejected. In other words there exists a relationship between the labour force factors and educational levels.

Table 5.10: Model fitting information for labour force

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	59250805.717			
Final	46979351.169	12271454.548	12	0.000

5.5.2 Pseudo R-squared

The Nagelkerke and Cox and Snell pseudo R^2 in Table 5.11 suggest that the variation in the level of education outcomes explained by the labour force factors ranges between 22% and 29%.

Table 5.11: Pseudo R^2 for labour force

Cox and Snell	Nagelkerke
.222	.286

5.5.3 Evaluating the Usefulness of the MLR Model

The overall classification accuracy rate is given by 75.3 % and a 25% increase over the proportional by chance accuracy rate is 59.7% (Table 5.12). This means that the model accuracy rate of 75.3% meets this criterion. Hence our model is good or suitable for “less grade 12” with 99.9% level of accurate prediction as compared to other levels of education (Table 5.12).

Table 5.12: Classification of accuracy for labour force

Observed	Predicted				
	no schooling	less grade12	grade 12	above grade12	Percent Correct
no schooling	**0**	8689944.8115	0	0	0.0%
less Grade12	23706.8022	**36812886.1615**	0	0	99.9%
grade 12	1471.7825	2031088.0334	**0**	0	0.0%
above grade12	959.7895	1308827.3864	0	**0**	0.0%
Overall Percentage	.1%	99.9%	0.0%	0.0%	75.3%

5.5.4 Relationship between Independent and Dependent Variables

The LRT as it appears in Table 5.13 was applied to test the significance of each of the factors individually. All the factors had a p-value of 0.00 which is less than 0.05. This means that there is a relationship between the dependent variables and the

independent education factors. Hence all the four education factors should be included in the model.

Table 5.13: Likelihood ratio test for labour force

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	112820288.648	65840937.479	3	0.000
L1Emply	49503063.661	2523712.492	3	0.000
L2IndO	51694929.487	4715578.318	3	0.000
L3EmplH	47499651.538	520300.369	3	0.000
L4LongU	48622493.353	1643142.184	3	0.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

5.5.5 Test for Statistically Significant Factors and Parameter Estimation

The results for fitting the MLR models for the labour force factors are shown in Table 5.14. The Wald test shows that all the factors are statistically significance. In order to account for the increased Type I error due to multiple significance tests, our statistical significance should be determined at a $p<0.0125$ (0.05 divided by 4).

5.5.6 Check for Multicollinearity and Numerical Errors

There are no standard errors that are greater than 2, which imply that there is no multicollinearity (Table 5.14).

5.5.7 Interpretation of the MLR Results

Looking at Table 5.14 L1Emply (*employment)*: Employed respondents were much more likely to have a “less grade 12” qualification over “no schooling” than non-employed respondents given that other factors are held constant. Similar trend is noticed for the same variable among “grade 12” and “above grade 12” over “no schooling” group (odds ratio (OR)>1). This is not unexpected. L2IndO (*Industrial business and occupational*): Respondents working in industries were much more likely to have a “less grade 12” than “no schooling” qualification than those not working in industries all other factors being equal/constant. Reverse trend is noticed for the same factor among “grade 12” and “above grade 12” over “no schooling” group (OR<1).This make sense because those who were matriculated and attain tertiary education are not likely to work as industrial workers. Looking at L3EmplH

(employment history), it can be deduced that, when holding other factors constant, the odds for someone who had history of employment of attaining or have attained tertiary education remain with no formal schooling are 1.653 times than the person with no history of employment.

Respondents with (L4LongU) *long-term unemployment* were 4.042 times more likely to have a "less grade 12" qualification than "no schooling" than those not with long-term unemployment when holding other factors constant. Similar trend is noticed for the same variable among "grade 12" and "above grade 12" over "no schooling" group (OR>1).

Table 5.14: Parameter estimates for labour force

Highest education level[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
less grade12	Intercept	2.199	.001	7284138.560	1	0.000			
	L1Emply	.724	.001	1770113.223	1	0.000	2.063	2.061	2.065
	L2IndO	.765	.001	1290134.013	1	0.000	2.150	2.147	2.153
	L3EmplH	.502	.001	342112.506	1	0.000	1.653	1.650	1.656
	L4LongU	1.397	.002	677508.858	1	0.000	4.042	4.029	4.056
grade 12	Intercept	-1.328	.001	808992.754	1	0.000			
	L1Emply	.897	.001	358235.697	1	0.000	2.452	2.444	2.459
	L2IndO	1.804	.001	2261071.150	1	0.000	6.074	6.060	6.088
	L3EmplH	.619	.001	286926.985	1	0.000	1.857	1.853	1.861
	L4LongU	1.573	.002	714750.313	1	0.000	4.821	4.803	4.838
above grade12	Intercept	-1.660	.002	1050873.808	1	0.000			
	L1Emply	.901	.002	307379.308	1	0.000	2.461	2.453	2.469
	L2IndO	1.706	.001	1688324.407	1	0.000	5.504	5.490	5.518
	L3EmplH	.505	.001	129855.100	1	0.000	1.656	1.652	1.661
	L4LongU	1.273	.002	351387.330	1	0.000	3.573	3.558	3.588

a. The reference category is: no schooling.

From three outcome category models (Table 5.14**)**, two logit functions of labour force factors are:

$$\log\left(\frac{< grade12}{no\_school}\right) = 2.199 + .724(\text{L1Emply}) + 0.765(\text{L2IndO}) + 0.502(\text{L3EmpH}) - 1.397(\text{L4LongU}) \quad \textbf{(5.8)}$$

$$\log\left(\frac{grade12}{no\_school}\right) = \text{-}1.328 + .897(\text{L1Emply}) + 1.804(\text{L2IndO}) + .619(\text{L3EmpH}) + 1.573(\text{L4LongU}) \quad \textbf{(5.9)}$$

$$\log\left(\frac{> grade12}{no\_school}\right) = \text{-}1.660 + 0.901(\text{L1Emply}) + 1.706(\text{L2IndO}) - 0.505(\text{L3EmpH}) + 1.273(\text{L4LongU}) \quad \textbf{(5.10)}$$

where

Dependent variable (Education levels) categories are:
<grade 12 = less grade 12; grade 12 ; > grade 12= above grade 12

Independent variables are:

L1Emply=employment; L2IndbO=industrial business and occupational; L3EmplH =employment history and L4longU=long-term unemployment.

CHAPTER 6: SUMMARY, CONCLUSION AND RECOMMENDATIONS

This study focused on understanding the relationship between factors of QoL and two variables of educational level and level of satisfaction with household service delivery in the context of South Africa. There were 162 variables of interest to us pertaining to QoL in the 2009 GHS and the 2009 QLFS. The main purpose of the study was to devise a means of reducing the 162 variables of interest into manageable few factors with minimal loss of information, for which factor analysis was found to be appropriate. These factors were identified using exploratory factor analysis by combining those variables that are correlated.

It was also the objective of this study to identify QoL factors that may have an impact on the level of education and on the satisfaction level in household service delivery. MLR was used to analyse the identified QoL factors related to education level and satisfaction level in household service delivery and determine the extent to which the relationships between these variables vary across the entire South Africa.

A summary of how the aim and objectives were achieved through factor analysis and multinomial regression analysis is given in the subsequent two sections. In addition, the limitations of the study are discussed in Section 6.3 and recommendations for further research are given in Section 6.4, before arriving at the final conclusions in Section 6.5.

6.1 FACTOR ANALYSIS

The 162 QoL variables of interest to our study were reduced to 29 unobservable factors for five core areas namely: education, housing, health, social development and labour force. Since factor analysis is a statistical method used to uncover the underlying structure of a relatively large set of variables and to describe variability among observed variables in terms of a potentially lower number of unobservable variables called factors, this research has managed to identify these factors in each of the five core areas that are being investigated. The information gained from factor analysis about the interdependencies between observed variables can be used to reduce the set of variables in a data set in later surveys. The identified factors fall into five categories of core areas, namely education, health, housing, social development and labour force. Twenty-nine (29) of the variables studied pertained to

education and can be represented by four factors, namely: (1) *fees too high***,** (2) *violence***,** (3) *absence of parental care* and (4) *historically advantaged*. These four factors account for 84.285% of the total variance, meaning that they are a true representative of the education component of the GHS. In order to improve on education, the responsible authorities should concentrate on these four core factors of education. In the subsequent GHS, Statistics South Africa in its data collection or questionnaire design may want to have their education questions directed at addressing these factors.

Forty-five (45) variables that were considered in the 2009 GHS pertained to **health**, from which a total of six factors were extracted. These six factors contributed 88.366% of the total variance which is an indication that the six factors are sufficient to represent the health component of the GHS. The six factors are: (1) *manageable illness*, (2) *medication*, (3) *chronic illness*, (4) *physical and social disability*, (5) *pregnancy* and (6) *social*.

Sixteen (16) of the variables in the 2009 GHS that have been analysed in this study, pertain to **housing**, and these were reduced to six factors. These factors are: (1) *house bricks*, (2) *government assistance*, (3) *home owners*, (4) *age of house*, (5) *waiting list*, and (6) *male household heads*. These factors contribute 67.480% of the total variance. These are the crucial issues that need to be addressed when looking at the housing issue.

Thirty-five (35) of the variables studied in the 2009 GHS pertaining to **social development** yielded nine factors. These factors contribute 75.488% of the total variance. The high number of extracted factors can be attributed to the diversity in the variables studied on social development. The nine extracted factors are: (1) no problem with health, (2) *sufficient water*, (3*) high income*, (4) *payment of sewerage*, (5) *telephone access*, (6) *absence of toilet*, (7) *household size*, (8) *water interruption* and (9) *pensioners*.

Thirty-eight (38) variables of interest pertaining to **labour force** in the 2009 QLFS generated four factors. These factors are: (1) *employment*, (2) *industrial business and occupational*, (3) *employment history*, and (4) *long-term unemployment*. These factors contribute 84.283% of the total variance, which is an indication that the extracted four factors are sufficient to represent 38 variables in labour force survey.

6.2 MULTINOMIAL REGRESSION

It was also the objective of this study to identify QoL factors that may have an impact on the level of education and on the satisfaction level in household service delivery. MLR was used to analyse the identified QoL factors related to education level and satisfaction level in household service delivery and determine the extent to which the relationship between these variables varies across the entire South Africa.

Starting with **education,** multinomial logistic regression analysis was used to determine the influence of independent variables (factor scores) on levels of education. The factors for education considered are: fees too high**,** violence factor, absence of parental care and historically advantaged factor. It may be observed that these four factors are classified up to 73.0% into four categories: “no schooling”, “less grade 12”, “grade 12”, “above grade 12”. The proportion by chance of accuracy rate (probability of classified cells) was 57%, but the model equation has predicted 73.0% correctly classified. Since prediction accuracy is above 57%, the current study used this prediction for model accuracy. Analyses of parameter estimates have shown that high school fees, violence and absence of parental care have negative influence on levels of education. As expected, respondents are significantly more unlikely to have any form of education compared with no schooling at all on these factors. This is because the odds ratios between three groups (less grade 12, grade 12 and above grade 12) when compared to no schooling give odd ratios that are close to zero on these factors. Historically disadvantaged people are the most featured with no formal schooling at all than the historically advantaged ones, which may be am impact of apartheid (lack of facilities or limited resources, unemployment etc.). The results of earlier researches support this finding (Ngcobo and Tikly, 2010).

The MLR was applied to the following factors of **housing**: *house bricks*, *government assistance*, *home owners*, *age of house*, *waiting list*, and *male household heads*. Housing and social development were analysed simultaneously, so the classification of accuracy rate or the overall percentage was 72.4% greater than the proportional by chance accuracy rate of 62.1%. The author used this prediction (72.4%) for model accuracy since it met the standard. Analysis revealed that, age of a house and home owners are the key factors likely to influence “very satisfied” and “satisfied” groups than the “dissatisfied” group.

The analysis revealed that home owners are less likely to be “very satisfied” than “dissatisfied” when compared to non-home owners. *Age of a house* is likely to influence the “very satisfied” group than the “dissatisfied” group. This seems to make sense as people in relatively new houses are more likely to enjoy their surroundings.

Considering **social development**, factors such as *no problem with health*, *sufficient water*, *high income*, *payment of sewerage*, *telephone access*, *absence of toilet* ,water interruption, *households size* and *pension* showed an association with service satisfaction by two categories of respondents, viz. respondents who rated these factors as “very satisfied” and those who rated them as “satisfied” using a base category of “dissatisfaction”. More than 80% of factors were statistically significant but interpretation was mainly based on those factors which gave us some interesting inferences and are not close to one (1±0.05), as a rule of thumb.

Our analysis (as expected) has indicated that the odds of respondents who had *no health problem* were likely to be in the “very satisfied” group instead of “dissatisfied” than those with health problem. Generally, most people without good health are likely to be “dissatisfied” with health services. The study by Joshi *et al*. (2009) revealed that factors such as service at private institutions and affordability at public institutions were not associated with a good healthcare. Our analyses have shown that communities with *sufficient water* were likely to be satisfied than those not having access to water at all.

Our study revealed that those with *high income* are likely to be “very satisfied” than “dissatisfied” when compared to those with low income. This observation may be because persons with high income often reside in more affluent places, and that high income benefited those who are historically advantaged than those who are historically disadvantaged. Annan (2000) indicated that poor health, disease and disability can prevent people from working full time, limiting their income and their ability to work to move out of poverty. He went further to emphasise that living in areas that have no sewage or clean water, poor people are much more susceptible to illness and disease. Although living without sewage or paying it, is common among black people because of their location in rural areas. It has been exposed by this study that people not paying for sewerage are likely to be “very satisfied” instead

of “dissatisfied” as compared to those who are paying for sewerage. Generally most people do not like paying services.

Our study has shown that the odds of being “dissatisfied” is 1.09 times more than “satisfied” with each unit increase in *household size.* Earlier researches support this finding by arguing that many of the poor are concentrated in large households with many dependants and this crowding in certain households leads to lack of income sources on the part of some members (Van der Berg, 2003). Holding other factors constant, respondents with access to a telephone were 2.12 times more likely to be “very satisfied” as opposed to “dissatisfied” than those with no access to telephone. The dissatisfaction occurs because the majority of people especially blacks could not afford paying for their bills. Møller (2007) has indicated that the latest 2005 survey results by South African Research Foundation have shown that the use of telephones in homes decreased from 30% in 1996 to 22% in 2004 and 2005. The cellular phones usage has grown from 2.4% in 1996 to 41.6% in 2005 and Black usage grew from 0.4% in 1996 to 36% in 2005.

Of the cases used to create the model for **labour force**, 75.3% (overall) of them are classified correctly. This compares favourably to the proportional by chance accuracy rate that of 59.7%. This prediction of 75.3% was useful for high accuracy of labour force model. The analysis showed that employed respondents were much more likely to have a “less grade 12”, “grade 12” and “above grade 12” qualification over “no schooling” qualification than non-employed respondents. This makes sense because education increases the chances of one to find a job (Faridi and Basit, 2011). As expected respondents with *long-term unemployment* were more likely to have a “less grade 12” instead of “no schooling” qualification than those without long-term unemployment. Respondents *working in industries* were much more likely to have a “less grade 12” instead of “no schooling” qualification than those not working in industries all. Thus, we can conclude that lack of formal schooling disadvantages an individual on the job market.

6.3 LIMITATIONS OF THE STUDY

A stratified random sample of households was taken when Stats SA conducted the 2009 GHS. Thus, any interpretations made using data from this survey should be understood to have some random errors inherent in the sampling plan. Any survey is

prone to response and non-response bias, but Stats SA has ensured that this is minimal (Statistics South Africa, 2010a).

Yalcin and Amemiya (2001) argued that factor analysis has attracted rather limited attention from statisticians because of reasons such as identification ambiguity, heavy reliance on normality and limitation to linearity. But in this current study, Bartlett's chi-square test significant levels are small (0.00<0.005), justifying the use of factor analysis. Nevertheless, interpretation of the results of factor analysis is based on a "heuristic" nature since more than one interpretation can be made of the same data factored the same way.

MLR includes larger sample size for accurate estimation of parameters. MLR Models tend to be difficult or impossible to interpret with more groups (i.e. four or more) to compare in the dependent variable. In MLR the coefficient estimates do not maximise any goodness of fit measure. Our study have shown that goodness of fit test significance level is small (0.00<0.05) suggesting that our model does not adequately fit the data. Our study ignored the goodness of fit test because of many cells with zero frequencies. Goodness of fit is strongly influence by the sample size. The sample size and effect size both determine significance (Homer *et al*., 1997; Hosmer and Lemeshow, 2000; Acher and Lameshow, 2006; Garson, 2012). Since our sample size consists of 25 361 households, it is hard to have the model that fit the data. Large sample size does not guarantee adequate cell size. Another reason for goodness of fit test to be significant is because of overdispersion (D'Souza *et al*., 2013; Field, 2013). Goodness of fit, like any other significant test, tells us whether the model fits or not, and does not tell us about the extent of the fit. The evidence from other research is that this measure can be misleading and analysis can proceed without compromising the results (Borooah, 2002; Chan, 2005).

6.4 FURTHER RESEARCH

Groups of interrelated variables in the four core areas of the 2009 GHS plus the one core area of the 2009 QLFS have been discovered and further research was used to establish how these variables are related to each other. It will be interesting to see if similar results can be obtained with subsequent surveys of the GHS and/or QLFS.

6.5 CONCLUSION AND RECOMMENDATIONS

It can be concluded that factor analysis as a data reduction technique has managed to describe the variability among the 162 variables considered in this study in terms of just 29 unobservable factors. It is much easier to concentrate on the smaller and manageable 29 factors compared to the larger 162 variables. The purpose of factor analysis of discovering simple relationships among the observed variables has been achieved in this study. The covariance relationships among the 162 variables has been described by the fewer underlying, but unobserved 29 factors even though these factors were not measured directly. If Stats SA were to adopt the reduced number of variables, then subsequent GHSs will be much easier to handle as researchers will be concentrating on these fewer but important areas. Stats SA, the primary collector of the data used in this research is recommended to use the results of factor analysis in their GHS in both questionnaire designing and data collection. The fewer factors that have been discovered in this study may be used as an indicator on the areas that need to be addressed. It is much easier and cheaper to concentrate on critical and fewer areas that have been identified by factor analysis. Government, Non-Governmental Organisation (NGO) and the private sector who are the implementers of various developmental projects studied in this research may also make use of the findings of factor analysis in this research.

Several factors were identified which significantly affect educational levels and service satisfaction among South Africans. Using MLR, the study has found that there is a relationship between labour force factors, educational factors and education level of attainment. This finding I supported by Faridi *et al.* (2010). Also there is a relationship between housing factors, social development factors and service satisfaction. The author suggests that education and housing and social development facilities should be improved. South Africans should have access to services to boost their satisfaction egos. Labour market participation should be given the first priority.

REFERENCES

Acher, K.J. and Lameshow, S. 2006. *Goodness of fit test for a logistic regression model fitted using survey sample data*. Stata Journal 6(1): 97-105.

Agresti, A. 2002. *Categorical Data Analysis*. John Wiley & Sons. Inc., New York.

Alasia, A. 2004. *Mapping the socio-economic diversity of rural Canada: A multivariate analysis*. Statistics Canada, Agriculture Division.

Annan, U.S.G.K. 2000. *Millennium Report*. Chapter 3, 34-44.

Bartlett, M.S. 1950. *Tests of significance in factor analysis*. British Journal of Statistical Psychology 3(2): 77-85.

Birhanu, Z. Assefa, T. Woldie, M. and Morankar, S. 2010. *Determinants of satisfaction with health care provider interactions at health centres in central Ethiopia: A cross sectional study*. BMC Health Services Research 10(1): 78.

Blom, A. and Saeki, H. 2011. *Employability and skill set of newly graduated engineers in India*. Policy Working Paper 5640. World Bank.

Bollen, K.A. 1989. *Structural Models with Latent Variables*. New York: Wiley.

Bonett, D.G. and Price, R.M. 2005. *Inferential methods for the tetrachoric correlation coefficient.* Journal of Educational and Behavioral Statistics *30*(2): 213-225.

Borooah, V.K. 2002. *Logit and Probit: Ordered and Multinomial Models*. No.138 Thousand Oaks, CA: Sage.

Brown, J. Bowling, A. and Flynn, T. 2004. *Models of quality of life: A taxonomy, overview and systematic review of quality of life.* In: (Proceedings) European Forum on Population Ageing Research, 2004. Dept. of Sociological Studies, University of Sheffield, Sheffield, UK.

Chan, Y.H. 2005. *Biostatistics 305. Multinomial logistic regression*. Singapore Medical Journal 46(6): 261.

Cox, D.R. and Snell, E.J. 1989. *The Analysis of Binary Data.* 2nd edn. London: Chapman and Hall.

D'Souza, K. A. Maheshwari, S. K. and Banaszak, Z. A. 2013. *Research framework for studying driver distraction on Polish city highways*. Management and Production Engineering Review 4(2): 12-24.

Economist Intelligence Unit. 2005. The Economist Intelligence Unit's quality of life index. http://www.economist.com/media/pdf/QUALITYofLIFE.pdf (Accessed on 07 July 2014).

Faridi, M.Z. and Basit, A.B. 2011. *Factors determining rural labour supply: A micro analysis*. Pakistan Economic and Social Review 49(1): 91-108.

Faridi, M.Z. Malik, S. and Ahmad, I. 2010. *Impact of education and health on employment in Pakistan: A case study*. European Journal of Economics, Finance and Administrative Sciences 18: 58-68.

Field, A. 2005. *Discovering Statistics using SPSS (Introducing Statistical Methods)*. 2nd edn. London: Sage.

Field, A. 2013. *Discovering Statistics using IBM SPSS Statistics.* 4th edn. London: Sage.

Garson, G.D. 2012. *Testing statistical assumptions*. North Carolina: Statistical Associates Publishing.

Gorsuch, R.L. 1983. *Factor Analysis*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Greene, W.H. 2003. *Econometric Analysis*. 5th edn. *Upper Saddle River,* New Jersey.

Hair, J.F. Black, W.C. Babin, B.J. Anderson, R.E. and Tatham, R.L. 2006. *Multivariate Data Analysis.* 6th edn. Upper Saddle River, NJ: Pearson Education Inc.

Hammill, M. 2009. *Income poverty and unsatisfied basic needs*. Social Development Unit, Subregional Headquarters of ECLAC/Mexico.

Harris, R. J. 1975. *A Primer of Multivariate Statistics*. Academic, New York.

Harry, H.H. 1976. *Modern Factor Analysis*. 3rd edn. Chicago: University of ChicagoPress, 1960HarmonModern Factor Analysis1960.

Hassan, S. Ismail, N. Jaafar, W.Y.W Ghazali, K. Budin, K. Gabda, D. and Samad, A.S.A. 2012. *Using factor analysis on survey study of factors affecting Students' Learning Style.* International Journal of Applied Mathematics and Informatics Styles 1(6): 33-113.

Ho, R. 2006. *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS*. Boca Raton: Chapman and Hall/CRC Press.

Hosmer, D. W. Hosmer, T. Le Cessie, S. and Lemeshow, S. 1997. *A comparison of goodness-of-fit tests for the logistic regression model*. Statistics in medicine 16(9): 965-980.

Homser, D.W. and Lemeshow, S. 1989. *Applied Logistic Regression*. *New York:* John Wiley and Sons.

Hosmer, D.W. and Lemeshow S. 2000. *Applied Logistic Regression*. 2nd ed. New York: John Wiley and Sons.

Houtman, D. and Steijn, A. 1990. *Are Non-Working People Socially Isolated?* In: J. J., Tacq, and J. Tacq, (1997). Multivariate analysis techniques in social science research: From problem to analysis. London: Sage.

Hutcheson, G. and Sofroniou, N. 1999. *The Multivariate Social Scientists*. London: Sage.

Johnson, R.A. and Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*, 4th edn. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Jöreskog, K.G. 1966. *Testing a simple structure hypothesis in factor analysis.* Psychometrika 31(2): 165-178.

Joshi, V.D. Chen, Y.M. and Lim, J.F.Y. 2009. *Public perceptions of the factors that constitute a good healthcare system*. Singapore Medical Journal 50(10): 982-989.

Kaiser, H.F. 1958. The *varimax criterion for analytic rotation in factor analysis.* Psychometrika *23*(3): 187-200.

Khorshidi, A. and Rezaloo, M. 2011. *Recognizing Effective Factors in Creating Prevalent High-Schools in Tehran*. J. Appl. Environ. Biol. Sci. 1(12): 688-694.

Kleinbaum, D.G. Kupper, L.L. Nizam, A. and Muller, K. E. 2008. *Applied Regression Analysis and Other Multivariable Methods Thompson Higher Education.* 4th edn, Belmont*, CA*: Duxbury Press.

Kline, P. 1994. *An Easy Guide to Factor Analysis*. London: Routledge.

Kutner, M.H. Nachtsheim, C.J. Neter, J. and Li, W. 2005. *Applied Linear Statistical Models*. 5th ed. Toronto: McGraw-Hill Irwin.

Lewis-Beck, M.S. Bryman, A.E. and Liao T.F.F. 2003. *The Sage Encyclopedia of Social Science Research Methods*. London: Sage Publications.

Magee, L. 1990. *R2 measures based on Wald and likelihood ratio joint significance tests*. American Statistician 44(3): 250-253.

Majors, M.S. and Sedlacek, W.E. 2001. *Using factor analysis to organize student services*. Journal of College Student Development 1(2): 272-278.

Maliki, S.B. Benhabib, A. and Bouteldja, A. 2012. *Quantification of the poverty-education relationship in Algeria: A multinomial econometric approach*. Topics in Middle Eastern and African Economies, 14.

Mann, S.J. 2001. *Alternative prospectives on the student experience: alienation and engagement.* Studies in Higher Education 26(1): 7-19.

Menard, S. 1995*. Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage

Menard, S. 2002. *Applied Logistic Regression Analysis*. 2d edn. Quantitative Applications in the Social Sciences Series, Vol. 106.

Mirza, F.M. Jaffri, A.A. and Hashmi, M.S. 2014. *An assessment of industrial employment skill gaps among university graduates: In the Gujrat-Sialkot-Gujranwala industrial cluster, Pakistan*. Vol. 17. Intl. Food Policy Res Inst.

Møller, V. 2007. *Satisfied and dissatisfied South Africans: Results from the General Household Survey in international comparison.* Social Indicators Research 81(2): 389-415.

Nagelkerke, N.J. 1991. *A note on a general definition of the coefficient of determination*. Biometrika 78(3): 691-692.

Ngcobo, T. and Tikly, L.P. 2010. *Key dimensions of effective leadership for change: A focus on township and rural schools in South Africa*. Educational Management Administration & Leadership 38(2): 202-228.

Nunnally, J.C. and Bernstein, I.H. 1994. *Psychometric Theory*. New York: McGraw Hill.

Penny, K. 2011. *Factors that influence student e-learning participation in a UK higher education institution*. Interdisciplinary Journal of E-Learning and Learning Objects *7*(1): 81-95.

Petrucci, C.J. 2009. *A primer for social worker researchers on how to conduct a multinomial logistic regression*. Journal of Social Service Research 35(2): 193-205.

Reise, S.P. 2000. *Using Multilevel Logistic Regression to Evaluate Person-Fit in IRT Model.* Multilevel Behavioral Research.

Sarstedt, M. Schwaiger, M. Ringle, C. M. and Gudergan, S. 2009. Satisfaction with *services: An impact-performance analysis for soccer-fan satisfaction judgements*. In Australian and New Zealand and Marketing Academy

Conference.

Schwab, J.A. 2002. *Multinomial logistic regression: Basic relationships and complete problems*. From: www.utexas.edu/courses/schwab/sw388r7/SolvingProblems /Analyzi. *(*Accessed 07 February 2013*).*

Sennett, J. Finchilescu, G. Gibson, K. and Strauss, R. 2003. *Adjustment of black students at a historically white South African university*. Educational Psychology 23(1): 107-116.

Simelane, S.E. 2007. *Poverty in post-apartheid South Africa: Measurement, trends and the demography of the poor*. Dissertation available from ProQuest*.* Paper AAI3271816.

Statistics South Africa. 2009. *General Household Survey Series Volume I Social grants In-depth analysis of the General Household Survey data 2003*-2007. Statistical release. P0318.1. Pretoria: Statistics South Africa.

Statistics South Africa. 2010a. *General Household Survey 2009.* Statistical Release. P0318. Pretoria: Statistics South Africa.

Statistics South Africa. 2010b*. Quarterly Labour Force Survey Quarter 4, 2009*. Statistical release. P0211. Pretoria: Statistics South Africa.

Strand, S. and Winston, J. 2008. *Educational aspirations in inner city schools.* Journal of Educational Studies 34(4): 249-267.

Tabachnick, B.G. and Fidell, L.S. 2007. *Using Multivariate Statistics*. 5th edn. New York: Allyn and Bacon.

Tesfazghi, E.S. Martinez, J.A. and Verplanke, J.J. 2010. *Variability of quality of life at small scales: Addis Ababa, Kirkos sub-city*. Social Indicators Research 98(1): 73-88.

Van der Berg, S. 2003. *Poverty in South Africa – An analysis of the evidence.* University of Stellenbosch. Department of Economics.

William, E.C. Samuel, R.H. and Dale, G.S. 1972. *The use of factor regression in data analysis.* From: http://mlrv.ua.edu/1972/VOL_2_4/V2_N4_A4.PDF (Accessed 04 April 2013).

Yalcin, L. and Amemiya, Y. 2001. *Nonlinear factor analysis as statistical method*. JSTOR: Statistical Science 16(3): 275-294.

APPENDICES

APPENDIX A: KMO and Bartlett's test for health

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.947
Bartlett's Test of Sphericity	Approx. Chi-Square	5.196E+09
	df	990
	Sig.	.000

APPENDIX B: Total variance for the extracted factors contributing towards health

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	17.901	39.781	39.781	17.901	39.781	39.781	16.905	37.567	37.567
2	7.336	16.302	56.082	7.336	16.302	56.082	6.793	15.096	52.664
3	6.116	13.590	69.673	6.116	13.590	69.673	6.633	14.741	67.405
4	4.912	10.915	80.588	4.912	10.915	80.588	5.756	12.792	80.196
5	2.044	4.543	85.131	2.044	4.543	85.131	2.053	4.561	84.758
6	1.456	3.235	88.366	1.456	3.235	88.366	1.624	3.608	88.366
7	.917	2.038	90.403						
8	.868	1.928	92.332						
9	.825	1.834	94.166						
10	.454	1.009	95.175						
41	.001	.001	99.997						
42	.000	.001	99.998						
43	.000	.001	99.999						
44	.000	.001	99.999						
45	.000	.001	100.000						

Extraction Method: Principal Component Analysis.

NB components 11 to 40 have been deleted so that the table can be reduced

APPENDIX C: KMO and Bartlett's test for housing

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.630
Bartlett's Test of Sphericity	Approx. Chi-Square	7.057E7
	df	120
	Sig.	.000

APPENDIX D: Total variance explained for the extracted factors of housing

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.136	19.598	19.598	3.136	19.598	19.598	2.281	14.259	14.259
2	2.248	14.050	33.649	2.248	14.050	33.649	2.119	13.242	27.501
3	1.726	10.786	44.435	1.726	10.786	44.435	2.018	12.612	40.113
4	1.463	9.144	53.579	1.463	9.144	53.579	1.648	10.301	50.414
5	1.208	7.549	61.128	1.208	7.549	61.128	1.614	10.088	60.502
6	1.016	6.352	67.480	1.016	6.352	67.480	1.116	6.978	67.480
7	.823	5.143	72.623						
8	.735	4.592	77.215						
9	.708	4.427	81.642						
10	.682	4.261	85.903						
11	.609	3.806	89.709						
12	.563	3.519	93.228						
13	.425	2.657	95.885						
14	.382	2.390	98.275						
15	.183	1.146	99.421						
16	.093	.579	100.000						

APPENDIX E: KMO and Bartlett's test for social development

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.880
Bartlett's Test of Sphericity	Approx. Chi-Square	4.414E+08
	df	595
	Sig.	.000

APPENDIX F: Total variance for the extracted factors contributing towards social development

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	Variance	Cumulative %
1	8.779	25.084	25.084	8.779	25.084	25.084	8.129	23.227	23.227
2	4.159	11.882	36.966	4.159	11.882	36.966	2.954	8.440	31.667
3	3.150	9.000	45.966	3.150	9.000	45.966	2.840	8.114	39.781
4	2.893	8.266	54.232	2.893	8.266	54.232	2.444	6.982	46.764
5	1.936	5.530	59.762	1.936	5.530	59.762	2.152	6.149	52.912
6	1.668	4.766	64.528	1.668	4.766	64.528	2.084	5.955	58.868
7	1.425	4.073	68.601	1.425	4.073	68.601	2.012	5.749	64.617
8	1.272	3.635	72.236	1.272	3.635	72.236	1.969	5.625	70.242
9	1.138	3.252	75.488	1.138	3.252	75.488	1.836	5.246	75.488
10	.981	2.803	78.292						
11	.860	2.458	80.749						
12	.736	2.103	82.852						
13	.624	1.782	84.634						
14	.578	1.652	86.287						
15	.518	1.481	87.767						
16	.461	1.318	89.085						
17	.429	1.225	90.310						
18	.388	1.110	91.420						
19	.380	1.085	92.505						
20	.357	1.021	93.526						
21	.335	.957	94.483						
22	.303	.867	95.349						
23	.275	.785	96.134						
24	.234	.667	96.802						
25	.186	.530	97.332						
26	.179	.512	97.844						
27	.175	.499	98.343						
28	.146	.418	98.761						
29	.123	.352	99.113						
30	.091	.260	99.373						
31	.079	.227	99.599						
32	.054	.155	99.754						
33	.046	.131	99.886						
34	.021	.059	99.945						
35	.019	.055	100.000						

Extraction Method: Principal Component Analysis.

APPENDIX G: KMO and Bartlett's test for Labour force Survey

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.910
Bartlett's Test of Sphericity	Approx. Chi-Square	5.717E+09
	df	741
	Sig.	.000

APPENDIX H: Total variance for the extracted factors of labour force

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	17.318	44.404	44.404	17.318	44.404	44.404	13.378	34.302	34.302
2	8.415	21.577	65.981	8.415	21.577	65.981	11.241	28.824	63.126
3	5.091	13.054	79.035	5.091	13.054	79.035	5.745	14.731	77.857
4	2.047	5.248	84.283	2.047	5.248	84.283	2.506	6.426	84.283
5	.997	2.556	86.839						
6	.979	2.510	89.349						
7	.834	2.139	91.488						
8	.646	1.656	93.144						
9	.575	1.475	94.619						
10	.305	.782	95.401						
11	.290	.744	96.145						
12	.262	.672	96.817						
13	.204	.524	97.341						
14	.183	.468	97.809						
15	.150	.385	98.194						
16	.106	.273	98.467						
17	.100	.256	98.724						
18	.081	.208	98.932						
19	.072	.185	99.117						
20	.064	.165	99.282						
21	.055	.142	99.424						
22	.044	.112	99.536						
23	.037	.094	99.631						
24	.031	.080	99.710						
25	.029	.074	99.785						
26	.023	.059	99.844						
27	.017	.044	99.888						
28	.011	.029	99.917						
29	.010	.025	99.942						
30	.006	.016	99.958						
31	.005	.013	99.971						
32	.004	.011	99.982						
33	.003	.008	99.990						
34	.002	.004	99.994						
35	.001	.002	99.997						
36	.001	.001	99.998						
37	.000	.001	99.999						
38	.000	.001	100.000						

Extraction Method: Principal Component Analysis.

APPENDIX I: Case processing summary for education

		N	Marginal Percentage
Highest level of education	no schooling	2152021.17146157	6.4%
	less grade 12	20138233.24535681	60.2%
	grade 12	9729862.71358409	29.1%
	above grade 12	1443034.59394860	4.3%
Valid		33463151.72435083	100.0%
Missing		427634.90298476	
Total		33890786.62733559	
Subpopulation		14880[a]	

a. The dependent variable has only one value observed in 14880 (100.0%) subpopulations.