Multiple Imputation Models for Missing Data:

A Case Study of Modeling Psychosocial Functioning among Adolescents in Ikere-Ekiti Local Government Area, Ekiti State.

Multiple Imputation Models for Missing Data: A Case Study of Modeling Psychosocial Functioning among Adolescents in Ikere-Ekiti Local Government Area, Ekiti State.

Ezekiel O. Olapade

B. Sc. (*Hons.*)(Statistics)

133767

ezekiel.olapade@gmail.com

+234 802 414 7342

A research project submitted in partial fulfilment of the requirements for the award of Master of Science degree in Biostatistics to

Department of Epidemiology and Medical Statistics
Faculty of Public Health
College of Medicine
University of Ibadan,
Ibadan, Oyo State, Nigeria

Supervisor

Dr. O. M. Akpa

May, 2014

CERTIFICATION

I hereby certify that this project work titled *Multiple Imputation Models for Missing Data: A Case Study of Modeling Psychosocial Functioning among Adolescents in Ikere-Ekiti Local Government Area, Ekiti State* was carried out under my direct supervision by **Mr. Ezekiel Oluwaseyi Olapade** in the Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Ibadan, and that it has been read and approved as meeting the requirements for the award of Master degree in Biostatistics.

Supervisor

Dr. O.M. Akpa

B.Sc. *(Hons.)*, M.Sc., Ph.D. (Ilorin)

Date: 27/08/2014

CERTIFICATION

I hereby certify that this project work titled *Multiple Imputation Models for Missing Data: A Case Study of Modeling Psychosocial Functioning among Adolescents in Ikere-Ekiti Local Government Area, Ekiti State* was carried out under my direct supervision by **Mr. Ezekiel Oluwaseyi Olapade** in the Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Ibadan, and that it has been read and approved as meeting the requirements for the award of Master degree in Biostatistics.

Supervisor

Dr. O.M. Akpa

B.Sc. *(Hons.)*, M.Sc., Ph.D. (Ilorin)

Date: 27/08/2014

DEDICATION

To the one I love — my wife.

ACKNOWLEDGEMENT

I thank my supervisor who enforced and thus impressed the act, ethics and conduct of good research throughout this exercise.

My parents, siblings and wife render helps when none seems coming. Motunrayo Shodimu is a rare lube lubricating the wheels upon which this program rode. I also thank Mrs Rachael A. Osasona for her incessant concern in the success of this program.

I boast in the wisdom that flows out of the abundance of grace and mercy of God. By these I am a coheir with His Son who has bestowed upon me undue assistance through the Spirit that teaches even the deepest things of this world and through these others and I proclaim Him Lord of all. All glory to Him forever.

Ezekiel Olapade

2014

TABLE OF CONTENT

LIST OF TABLES

ABSTRACT

Missing data present a challenge to health researchers in particular as incomplete data violate the complete-case assumption. A study about modeling Adolescents Psychosocial Functioning (APF) in Ekiti State presents such occurrence. Improper approaches to these missing data such as listwise deletion and mean imputation can lead to biased statistical inference using complete case analysis. This study presents the multiple imputation (MI) method, a technique based on Bayesian inference, and Fully Conditional Specification approach to imputing the missing values in the APF dataset.

A secondary dataset consisting of a random sample of 490 students from secondary schools in Ikere-Ekiti Local Government Area of Ekiti State participated in a study that seeks to know the effect of psychosocial well-being on depression using a combination of Rosenberg Self-Esteem Scale (RSES), Strength and Difficulty Questionnaire, and Center for Epidemiology Studies Depression Scale. Missing items on RSES ranges from 16 (3.3%) to 25 (5.1%). Hence, RSES was imputed using STATA mi command.

Pattern of missingness found in the dataset was arbitrary. Also, the data provided sufficient evidence against the MCAR assumption. Indeed, on the basis of their religion, students who were satisfied with themselves (item R1 of RSES) significantly differ from those without responses (χ^2=5.836, $p < 0.05$). Furthermore, a multiple logistic regression model estimation showed that the effects of religion (β = 1.549, $p < 0.05$) and father's education (β= 1.672, $p < 0.05$) on probability of nonresponse to R1 are significant. A linear regression model of self-esteem scores on the socio-demographic variables revealed more precise estimates when nonresponse is accounted for. For example, SSS 1 students had significantly higher self-esteem score before imputation (β = 6.930, s.e. = 1.217, $p < 0.01$) and after imputation (β= 6.671, s.e. = 1.138, $p < 0.001$) than the SSS 2 students with a relative reduction in standard error (s.e.) of about 6%. Also, effects that were not significant prior to imputation became significant after imputation.

Consequently, MI is a missing data technique that allows for valid statistical inference with complete case statistical analysis. Therefore, health researchers should consider conducting proper missing value analysis so as to achieve substantial inference.

Keywords: Multiple Imputation, Fully Conditional Specification, Multivariate Normal Imputation.

Number of words: 353.

ABSTRACT

Missing data present a challenge to health researchers in particular as incomplete data violate the complete-case assumption. A study about modeling Adolescents Psychosocial Functioning (APF) in Ekiti State presents such occurrence. Improper approaches to these missing data such as listwise deletion and mean imputation can lead to biased statistical inference using complete case analysis. This study presents the multiple imputation (MI) method, a technique based on Bayesian inference, and Fully Conditional Specification approach to imputing the missing values in the APF dataset.

A secondary dataset consisting of a random sample of 490 students from secondary schools in Ikere-Ekiti Local Government Area of Ekiti State participated in a study that seeks to know the effect of psychosocial well-being on depression using a combination of Rosenberg Self-Esteem Scale (RSES), Strength and Difficulty Questionnaire, and Center for Epidemiology Studies Depression Scale. Missing items on RSES ranges from 16 (3.3%) to 25 (5.1%). Hence, RSES was imputed using STATA mi command.

Pattern of missingness found in the dataset was arbitrary. Also, the data provided sufficient evidence against the MCAR assumption. Indeed, on the basis of their religion, students who were satisfied with themselves (item R1 of RSES) significantly differ from those without responses (χ^2 =5.836, $p < 0.05$). Furthermore, a multiple logistic regression model estimation showed that the effects of religion ($\beta = 1.549$, $p < 0.05$) and father's education ($\beta = 1.672$, $p < 0.05$) on probability of nonresponse to R1 are significant. A linear regression model of self-esteem scores on the socio-demographic variables revealed more precise estimates when nonresponse is accounted for. For example, SSS 1 students had significantly higher self-esteem score before imputation ($\beta = 6.930$, s.e. = 1.217, $p < 0.01$) and after imputation ($\beta = 6.671$, s.e. = 1.138, $p < 0.001$) than the SSS 2 students with a relative reduction in standard error (s.e.) of about 6%. Also, effects that were not significant prior to imputation became significant after imputation.

Consequently, MI is a missing data technique that allows for valid statistical inference with complete case statistical analysis. Therefore, health researchers should consider conducting proper missing value analysis so as to achieve substantial inference.

Keywords: Multiple Imputation, Fully Conditional Specification, Multivariate Normal Imputation.

Number of words: 353.

CHAPTER ONE

INTRODUCTION

1.1 Background

The theory of most classical statistical analyses of datasets employed in most researches, particularly in health-related inquiries, is built on an assumption that the datasets used provide valid values on all variables in consideration so that the intention of such analysis, making valid inferences regarding a population of interest, is attainable. However, a frequent occurrence in practice is the problem of missing values or nonresponse, a situation where valid values are not available on one or more variables. Indeed, rarely does a researcher avoid some form of missing data problem (Rubin, 1987; Allison, 2012).

The problem of missing data is often pronounced in studies that make use of self-report instruments such as Rosenberg Self Esteem Scale (RSES). In a study among 1931 surgical patients, Shrive et al (2006) measured level of depression using Zung Self-rated Depression Scale (SDS). Among these patients, 351 failed to respond to all of the questions. The challenge therefore is to address, the issues raised by missing data, especially those that affect the generalizability of inferences arising from the analysis.

Several approaches to missing values exist in practice. Some simply throw away data. For example in regression analysis complete-case analysis excludes all cases with missing outcome or response. Two problems arise in connection with this practice: The results of a statistical analysis may be bias due to the systematic difference that exists between cases with missing value and the completely observed cases. Also, if many variables are included in a model and for the sake of a simple analysis a large number of incomplete cases are discarded then there may be insufficient number of complete cases.

Furthermore, one or more variables with sufficiently large amount of missing data may be dropped from the analysis. A potential problem associated with this practice is dropping variables that are highly correlated with the response. Another simple approach is to use subset of cases with complete information on all variables included in a particular analysis. This approach, usually called available-case analysis, is prone to the problem that different analysis will be based on different subset of the data so that results are inconsistent over such different analysis. In addition, as with complete-case analyses, inferences may be bias if respondents differ systematically from non-respondents.

Some techniques do not discard any data: Mean imputation simply replaces each missing data with the mean of the fully observed values for that variable, random imputation draws random values with replacement from the observed component of the variable, iterative regression imputation sequentially replaces missing values in a variable by conditioning on the fully observed variables in the dataset, matching funds for all units with a missing value on a variable, a unit with similar values on other variables and replaces the missing component of the variable by the corresponding value assumed by the match (Gelman and Hill, 2006).

Missing data that occur in at least two variables present a special challenge. Some of these are alleviated by multiple imputation, a technique first introduced by Rubin (1987), and its two paradigms namely fully conditional specification (FCS) and multivariate normal imputation (MVNI). This involves "filling in" missing data with $m > 1$ values randomly drawn from an imputation model.

Data arising from psychometric applications often involve variables with missing components. For example, Crawford et al (2004) investigated personality disorder symptoms in a community sample of 714 young people to assess their relationship over time with well-being during adolescence and the emergence of intimacy in early adulthood. Youth and parent interviews were conducted at Time 3 (T3) (1985-1986) and Time 4 (T4) (1991-1993).

Approximately 3.9% of the data assessed at T3 and T4 were missing, although missing data occurred mostly in cases where parents had not been interviewed. Accordingly, complete scores were imputed using multiple regression equations based on the available youth reports and youth gender.

1.2 Problem Statement

At a point in the analysis stage a researcher is perhaps very often faced with what to do about missing data. Improper missing value analysis, such as deleting cases with missing observation, may bias result of statistical inference and cause loss of statistical power because of relatively large reduction in sample size (and hence, loss of information) particularly if the units with missing values differ systematically from the completely observed cases. The problem may become more acute when the reason for the missing data is directly related to the missing value itself, (Gelman and Hill, 2006). This may occur, for example, when the magnitude of the data to be provided influences respondent's attitude to giving genuine response to the question asked.

Indeed, some approach to missing value simply remove variable with most missing values. Should this be done in the context of a regression analysis, or more generally a causal inference analysis variables relevant to the model may be excluded from the analysis (Rubin, 1987).

In relation to health studies, when missing data are not properly dealt with, data analysis samples may not reflect the full population of interest. A study done by Stuart et al. (2009) with 9,186 youths participating in the United States national evaluation of the Community Mental Health Services survey, most variables have missing values for 30% - 70% of the children. A method of missing data analysis that removes these variables or cases with missing data reduces the sample size by a factor of at least three, resulting in a sample that may not be representative unless data is missing completely at random.

1.3 Justification

Despite the revolution experienced in the last two decades in the methods for handling missing data many researchers have either barely heard of the modern and superior methods for handling missing data or they are not well vast and grounded in the implementations of their methodologies. Perhaps because of the several technical difficulties in their implementations in terms of time and computational effort, some researchers resort to the use of rather simpler but more problematic method without checking whether the assumptions underlying such practice are valid.

Most epidemiologists and medical researchers usually interested in drawing causal inferences pertaining to risk-factor and disease evaluation are better enhanced with multiple imputation as a missing data analytic tool. Multiple imputation provides a good balance between quality of inference and ease of use. Indeed, it has been shown that it produces unbiased and almost asymptotically efficient parameter estimates that are robust to departures from normality assumptions, presence of high missing data rates or low sample size (Graham et al, 1997; Graham and Schafer, 1999; Schafer and Graham, 2002). Hence, this study explores the possibility of multiple imputation technique as a solution to the problem of missing data.

1.4 Objectives of the Study

Main Objective

Our main objective in this study is to impute and present the multiple imputation models for the missing values in the Adolescent Psychosocial Functioning (APF) survey using the fully conditional specification approach.

Specific Objectives

1. To determine the type and extent of missing value in the Adolescent Psychosocial Functioning (APF) dataset
2. To specify and apply the appropriate imputation models for the missing data in the APF dataset
3. To impute the missing values in the APF dataset, in particular, the Rosenberg Self-Esteem Scale.
4. To compare results of linear regression modelling of self-esteem score before and after imputation.

1.5 Notation

R1	On the whole, I am satisfied with myself
R2	At times I think I am no good at all
R3	I feel that I have a number of good qualities.
R4	I am able to do things as well as most other people
R5	I feel I do not have much to be proud of
R6	I certainly feel useless at times
R7	I feel that I'm a person of worth, at least on an equal plane with others
R8	I wish I could have more respect for myself
R9	All in all, I am inclined to feel that I am a failure
R10	I take a positive attitude toward myself

Specific Objectives

1. To determine the type and extent of missing value in the Adolescent Psychosocial Functioning (APF) dataset
2. To specify and apply the appropriate imputation models for the missing data in the APF dataset
3. To impute the missing values in the APF dataset, in particular, the Rosenberg Self-Esteem Scale.
4. To compare results of linear regression modelling of self-esteem score before and after imputation.

1.5 Notation

R1	On the whole, I am satisfied with myself
R2	At times I think I am no good at all
R3	I feel that I have a number of good qualities.
R4	I am able to do things as well as most other people
R5	I feel I do not have much to be proud of
R6	I certainly feel useless at times
R7	I feel that I'm a person of worth, at least on an equal plane with others
R8	I wish I could have more respect for myself
R9	All in all, I am inclined to feel that I am a failure
R10	I take a positive attitude toward myself

CHAPTER TWO

LITERATURE REVIEW

Consequent upon more recent researches that critically examined the problem of missing data, there is considerable amount of literature devoted to this problem whose approaches range from the parametric to nonparametric and semiparametric, most of which advocate for exploring reasons for missing data.

Regardless of the reasons for missing data: attrition, refusal, ignorance, or measurement errors, missing observations still present a problem in all areas of research (Allison, 2001). To attend to this problem researchers often make implicit or explicit assumptions about the missing data process besides confirming that missing data are really missing (Schafer and Graham, 2002). The ignorable missing data process assumption simplifies the analysis of missing data since the mechanism causing the missing observations need not be modeled explicitly. Two conditions have to be met for missing data mechanism to be ignorable: Data is missing at random (MAR) and parameters in the missing data model are distinct from those in the complete data model.

Furthermore, examination of the missing data pattern, a description of which observations in the data are missing, may be of interest when dealing with incomplete data. A monotone missing data pattern (MMP) offers more flexibility in the choice of missing data method than an arbitrary missing data pattern (AMP) (Little and Rubin, 2002).

This chapter gives a brief review of literatures on missing data mechanisms (Section 2.1), assumption of ignorable missing data mechanism (Section 2.2), and missing data pattern (Section 2.3). Also, an account of several approaches to missing values in general is given in (Section 2.4), in particular, multiple imputation (Section 2.5), its Bayesian approach (Section 2.6), and its two paradigms - FCS (Section 2.7) and MVNI (Section 2.8) are also discussed. Finally, we present the mi command in STATA (Section 2.9).

2.1 Missing Data Mechanism

Given an $n \times p$ data matrix $Y = (y_{ij})$ consisting of p variables (y1, … , yp) measured on a sample of size n that would occur in the absence of missing values, where yij is the value of variable yj ; j= 1; … ; p for unit i; i = 1, … , n: With missing data, define the missing data indicator matrix R = (rij), such that rij = 1 if yij is missing and rij = 0 if yij is observed. The matrix R then defines the missing data pattern. We write Y = (Yobs;Ymis); where Yobs denote the observed components or entries of Y , and Ymis denote the missing components.

We denote the jth variable of the observed component Yobs by yobs j and similarly ymis j denote the jth variable of the missing component Yobs. The missing data process models the probability that the data at hand is observed as a function of the observed variables in Yobs and unobserved variables in Ymis. It is written as a conditional probability density P(Rij = 1|Yobs;Ymis) for some i and j.

We also introduce notations for Bayesian discussion: The joint probability distribution of Yobs; Ymis and R is denoted by f(Yobs;Ymis; R | φ, ϕ) which is indexed by the unknown parameters. The likelihood and the prior distribution of these parameters are denoted by l(φ, ϕ|Yobs, R) and $\pi(\varphi, \phi)$, respectively. Missing data processes are classified into several types in accordance with the diffeerent assumptions concerning the relation between R on the one hand and Yobs; Ymis on the other. In this work we follow Rubin's classification into missing at random (MAR), missing completely at random (MCAR) and not missing at random (NMAR) also called nonignorable (NI).

2.1.1 Missing Completely at Random

A variable is missing completely at random (MCAR) if the probability of nonresponse is the same for all units, for example, if each respondent tosses a coin and refuses to answer if a head shows up. In this instance the cases with missing data are indistinguishable from cases with complete data.

More formally, the observed values of Y are truly a random sample of all Y values with no underlying process that lends bias to the observed data. MCAR is a special stricter case of MAR. It occurs when the distribution of missingness does not depend on Ymis and Yobs:

$$P(Rij = 1 \mid Yobs; Ymis) = P(Rij = 1) = r$$

where r is the proportion of responses estimated by r = nobs=n. The assumption of MCAR is rather strong, yet reasonable under certain condition as when data are missing by the study design, that is when the missing data are not intended to be collected in the firrst place. In these instances, specific remedies for missing data are not needed because the allowance for missing data are inherent in the design used (Little and Rubin, 2002; Schafer, 1997). The missing data are sometimes referred to as ignorable missing data.

2.1.2 Missing at Random

Most nonresponses are not MCAR and can be noticed from the dataset. For example, the different nonresponse rates for students whose parents are educated and those whose parents are not educated indicate that the questions on self-esteem among adolescents is not missing completely at random. A variable is missing at random (MAR) if the probability of missingness depends only on available information.

Formally, Rubin (1976) defined missing data to be missing at random if the distribution of missingness does not depend on Ymis.

$$P(Rij = 1 \mid Yobs, Ymis) = P(Rij = 1 \mid Yobs)$$

for some i and j. In other words, the observed values represent a random sample of the actual Ymis values for each value of Yobs, but the observed data for Ymis do not necessarily represent a truly random sample from all Ymis values. It has a drawback that values are not generalizable to population even though missing data process is random in the sample.

It is seldom possible to test whether the assumption of MAR is met except by obtaining the follow-up data from non-respondents. However, an erroneous assumption of MAR may often have only a minor impact on estimates and standard errors as demonstrated by Collins et al (2001) using many realistic cases.

2.1.3 Not Missing at Random

When the probability of missingness depends on the (potentially missing) variable itself, this is called nonignorable missing data mechanism (MDM). Formally, this occurs when the distribution of missing data depends on Ymis. This mechanism for some i and j is typified by

$P(Rij = 1|Yobs;Ymis)$

If missing data is nonignorable, properly accounting for this mechanism required external information about the distribution of Ymis that is typically beyond the data so that the missing data generating mechanism is modelled to get good enough estimates of the parameters of the parameter of interest.

Apart from the assumptions about missing data mechanism, assumptions also have to be made regarding the parameters of the missing data mechanism, in relation to those of the data. The distinctness of parameters assumptions differ in meaning from both the frequentist and the Bayesian perspective. The frequentists interpret it to means that the joint parameter space of and must be the product of the two individual parameter spaces, while for the Bayesian it means that a joint prior distribution applied to the parameters must factor into the independent marginal distributions (Schafer, 1997).

2.2 The Ignorable Missing Data Assumption

To properly analyze, at least approximately, a dataset with missing values, not only does the researcher need to select an appropriate course of action and remedy the nonresponse if possible, but also the researcher inevitably must understand the reasons for nonresponse. However, since the missing observations are indeed unknown, examination of assumptions about the missing observations is inherently difficult. Tests for the MCAR assumption have been suggested in the literature (see Little, 1988; Park and Lee, 1997 and Chen and Little, 1999), but no feasible way exist to test the MAR assumption (Schafer and Graham, 2002).

In some situations missing data is known to be at least MAR so long as the process leading to the missing data is under the control of the researcher, for example, with help of double sampling or randomized experiments with unbalanced design. This situation arises when the data is missing due to the study design (Schafer, 1997). However, one can increase the plausibility of the MAR assumption, and hence explain the missingness, by including auxiliary variables and variables that are known to be highly correlated with the variables containing missing data in the imputation model.

Auxiliary variables will also remove nonresponse bias that can be accounted for by the observed data, thereby reducing possible bias due to deviations from the MAR assumption (Collins, Schafer and Kam, 2001). Still, even though MAR is impossible to test for, it is the most commonly assumed missing data mechanism (Stuart et al., 2009).

2.3 Missing Data Pattern

To aid the choice of missing data techniques examination of the missing data pattern, a description of the values in the data matrix that are actually missing, can be of importance. Usually, missing data patterns are divided into monotone missing pattern (MMP) and arbitrary missing patterns (AMP).

A MMP arises when the data for a variable in a data set can be ordered in such a way that having a missing value on that variable also means having missing values on all following variables. MMP often occurs in longitudinal studies due to attrition, where dropping out by definition means that all the following observations will be missing. When only a variable in the data set contains missing observations, a special case of MMP, the univariate missing data pattern (UMP) arises. An AMP on the other hand arises when the data matrix cannot be ordered as in MMP.

Table 2.1: Missing Data Patterns

i	y_1	y_2	y_3	y_4
1	20	73	71	81
2	96	87	54	80
3	26	71	68	
4	71	69	54	
5	83	36		
6	80	35		
7	98	55		
8	65			
9	54			
10	95			

(a) MMP

i	y_1	y_2	y_3	y_4
1	40	75	79	27
2	30	81	56	34
3	44	82	45	60
4	90	33	91	83
5	54	97	37	
6	21	31	55	
7	96	58	42	
8	25	59	45	
9	81	72	71	
10	58	96	37	

(b) UMP

i	y_1	y_2	y_3	y_4
1	85	75		
2	83	21	34	
3		79	57	58
4			29	55
5	66	71	58	54
6	44	99		62
7	33			47
8	38		82	35
9	74	58	32	26
10	33	26	36	

(c) AMP

Item nonresponse in surveys is an example of AMP where for some reasons respondents fail to answer one or more questions. However, missing values in one variable does not necessarily implies that all following variables are missing. (Little and Rubin, 2002). The analysis of incomplete data may be greatly simplified if the missing data pattern is MMP in the sense that it may allow for the likelihood function to be factorized into factors for each block of cases with missing observations in the same variables, which can then be maximized separately. Often, methods constructed solely for MMP demand less computations than those designed to handle AMP. It may sometimes even be worth considering removing a small number of cases or impute values for some variables using an arbitrary missing data method in order to create a data set with a "monotone" missing data pattern (Little and Rubin, 2002).

In the next section we present an overview of some methods available to handle incomplete data, relying on different assumptions about the data missing.

Table 2.1: Missing Data Patterns

i	y_1	y_2	y_3	y_4
1	20	73	71	81
2	96	87	54	80
3	26	71	68	
4	71	69	54	
5	83	36		
6	80	35		
7	98	55		
8	65			
9	54			
10	95			

(a) MMP

i	y_1	y_2	y_3	y_4
1	40	75	79	27
2	30	81	56	34
3	44	82	45	60
4	90	33	91	83
5	54	97	37	
6	21	31	55	
7	96	58	42	
8	25	59	45	
9	81	72	71	
10	58	96	37	

(b) UMP

i	y_1	y_2	y_3	y_4
1	85	75		
2	83	21	34	
3		79	57	58
4			29	55
5	66	71	58	54
6	44	99		62
7	33			47
8	38		82	35
9	74	58	32	26
10	33	26	36	

(c) AMP

Item nonresponse in surveys is an example of AMP where for some reasons respondents fail to answer one or more questions. However, missing values in one variable does not necessarily implies that all following variables are missing. (Little and Rubin, 2002). The analysis of incomplete data may be greatly simplified if the missing data pattern is MMP in the sense that it may allow for the likelihood function to be factorized into factors for each block of cases with missing observations in the same variables, which can then be maximized separately. Often, methods constructed solely for MMP demand less computations than those designed to handle AMP. It may sometimes even be worth considering removing a small number of cases or impute values for some variables using an arbitrary missing data method in order to create a data set with a “monotone" missing data pattern (Little and Rubin, 2002).

In the next section we present an overview of some methods available to handle incomplete data, relying on different assumptions about the data missing.

2.4 Approaches to Missing Data

There are several different approaches to missing data analysis. The good ones are identified by three conditions: The method should produce unbiased parameter estimates, the method should provide a means to access the uncertainty about the parameter estimates, and the method should possess good statistical power (Graham, 2009). Moreover, the aim of such technique is not to recreate missing values but to retain the characteristics of the data and the association between variables, in such a way that valid and efficient inferences can be made (Schafer and Graham, 2002).

Probably the most common approach is simply to "ignore" missing values and run models without doing anything about missingness. In effect, what is done depends on the defaults of the statistical analysis software used. Usually, this corresponds to complete-case analysis (CCA) - an approach that simply throws away data by excluding all cases with missing response variable (in regression context for example). This method suffers from a loss of information in the incomplete cases and at risk of bias if the missing data is not MCAR. Furthermore, one or more variables with sufficiently large amount of missing data may be dropped from the analysis. A potential problem associated with this practice is dropping variables that are highly correlated with the response.

Another simple approach is to use subset of cases with complete information on all variables included in a particular analysis. This approach, usually called available-case analysis (ACA), is prone to the problem that different analysis will be based on different subset of the data so that results are inconsistent over such different analysis. In addition, as with complete-case analyses, inferences may be bias if respondents differ systematically from non-respondents. In ACA there is also a of risk producing correlations outside the natural bound of [-1, 1] (Little and Rubin, 2002).

Single imputation (SI) involves filling in the missing value once, creating one "complete" dataset. SI methods range from ad-hoc methods like mean imputation, hot-deck or mean matching, to more complex methods like regression imputations, predictive mean matching and stochastic regression imputation (Little and Rubin, 2002). Other inappropriate methods include missing data indicator, and last observation carried forward. Imputing the conditional mean would probably be the best guess for every missing value if the goal of imputation is to recreate the missing data as good as possible. However, to preserve associations between variables and provide valid parameter estimates, Little and Rubin (2002) conclude that the imputations should be conditional on the observed data, rather than the means of the conditional distribution. Failure to incorporate imputation uncertainty in the standard errors as well as inefficiency of parameter estimates are the two major disadvantages of SI. Failing to take into account the uncertainty caused by the fact that the imputed values are estimated from the data may produce too small standard errors, narrow confidence intervals (CI) and low p-values (Little and Rubin, 2002).

According to the criteria given by Graham, (2009) criteria, case deletion and SI can only be used in special limited cases. Case deletion has low power due to unnecessary wide CIs and biases most parameter estimates unless the data are MCAR. SI may bias covariances and correlations, equivalently underestimating the variances and standard errors of the estimates.

Two generally recommended methods do meet Graham's criteria which are maximum likelihood (ML) and MI (Schafer and Graham, 2002). Under the MAR assumptions both methods yield consistent, asymptotically efficient and normally distributed estimates.

As with ordinary ML with complete data, the likelihood function is maximized with respect to the parameter. With complete data the likelihood is the product of the likelihood for all observations. The difference, for the incomplete-data case, is that the likelihood function is factorized into different parts according to the missing observations. For example, suppose the ith elements of continuous variables Y1 and Y2 contain missing observations that satisfies MAR assumption but the rest are complete.

An extension of the likelihood can include missing data on several variables by factorizing the likelihood into more than two parts. Among the different methods to maximize the likelihood function, the EM-algorithm (Dempster et al. 1977) is perhaps the most common.

MI is a general approach to deal with incomplete data. In contrast to SI, several plausible values are imputed for each missing observation. By imputing m > 1 random draws from a posterior distribution for every missing observation, m "complete" datasets are created. Each dataset is analysed using standard complete-data method producing m point estimates that are then combined into one single estimate with their standard error consisting of both a within- and between-imputation variation component, properly reflecting the imputation uncertainty. Hence, by imputing several plausible values, the inefficiency problem in SI is resolved (Little and Rubin, 2002).

When comparing ML and MI, both their advantages and disadvantages should be considered. The greatest advantage of ML over MI is that ML is efficient while MI is only almost efficient (Allison, 2012). MI however has the great advantage that the imputations and the analysis can be done separately without putting the burden of dealing with the incomplete data on the researcher. In ML, handling the missing observations and performing the analysis have to be done simultaneously, putting a strain on the researcher who may not be familiar with the ways of dealing with incomplete data. Further, once the imputed data sets are constructed by MI, various statistical analyses can be conducted using the multiply imputed data sets. In the next section MI will be considered in more detail. How to combine the estimates from the imputed data sets into one by the rules of Rubin (1987) and how to construct the imputation model by using fully conditional specification (FCS) will be described.

2.5 Multiple Imputation

In multi-variable analysis, general purpose techniques exist for handling the problem of missing value of which MI seems to be one of the most attractive. Proposed by Rubin (1977) and further elaborated by Rubin (1987), the basic idea of MI is simple and quite attractive:

1. Impute missing values using appropriate imputation model that incorporates random variation into the model
2. Do this m times to generate m "complete" datasets, m > 1

3. Perform the desired analysis on each of the m "complete" data set using standard complete-data methods
4. Average the values of the parameter estimates across the m samples to produce a single point estimate
5. Calculate the standard errors by
   1. averaging the squared standard errors of the m estimates
   2. calculating the variance of the m parameter estimates across samples, and
   3. combining the two quantities using a simple formula.

Multiple imputation has several desirable features. It introduces appropriate random error into the imputation model which makes it possible to obtain unbiased estimates of parameters. Also, it provides a good estimates of the standard errors, which is achieved through repeated imputation. It can also be used with any kind of data and any kind of analysis without specialized software (Allison, 2000).

To obtain these desirable properties from MI, Rubin (1987, 1996) describes certain assumptions which must be met. First, data must be missing according as a MAR process. Second, the imputation model must be \correct in" some sense. Third, the analysis model must be similar, in some sense, with the model used in the imputation.

However, it is easy to violate these assumptions in practice. In particular, there are often strong reasons to suspect that data are not MAR. Even if MAR condition is satisfied, often times it is not easy to generate random imputations that provides unbiased estimates of the desired parameters. Also, we expect simulated imputations to give adequate and reasonable prediction of the missing data and the variability among the set of simulated imputations reflect an appropriate degree of uncertainty in the imputation mechanism. A proper imputation5 method satisfies some technical conditions provided by Rubin (1987) under which MI method leads to frequency-valid answers. These conditions, although useful for evaluating some properties of a given method, provides little guidance as to creating a method in practice (Schafer, 1999). To subvert this problem Rubin argues that imputation be done by employing Bayesian methods.

2.6 Bayesian Approach to Multiple Imputation

A linear regression imputation predicts the value of a missing variable using a regression on fully observed predictors of missingness. These imputed values have too small a variance because the model does not account for the fact that parameters in the imputation model are only estimate subject to sampling variability. Schafer, (1997) uses a Bayesian approach to multiple imputation that requires a non-informative prior reflecting little or no belief about the parameters. Separate random draws of imputation parameters are then made from the resulting posterior distribution. However, when values are missing on one or more predictors iterative procedures are necessarily applied. For general missing value pattern two major iterative techniques are used.

2.7 Fully Conditional Specification

Limitations occur in practice concerning the specification of joint distribution for an entire dataset due to complex relations between variables that are hard to capture in the distribution - since datasets often consist of variables measured on different scales in practice. By implementing MI under a FCS, a multivariate distribution is assumed. However, it is unnecessary to specify explicitly the form of the joint model. Instead of drawing the imputations from a pre-specified joint distribution, imputations are generated on a variable-by-variable basis using a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, FCS draws imputations by iterating over the conditional densities. This even makes it possible to specify models for which no known joint distribution exist (van Buuren, 2007).

Let Y be the partially observed complete sample from the multivariate distribution P(Y j_), where the vector of unknown parameters completely specifies the distribution. Also, consists of parameters specific to the respective conditional distribution and are not necessarily the product of the factorization of a \true" joint distribution. Further, let Y □j be all variables in the the data except yj ; j = 1; _ _ _ ; p: The posterior distribution is obtained by iteratively drawing from the conditional marginal distributions, that are assumed to completely specify the joint distribution. Starting with an initial imputation, FCS draws imputations by iterating over the conditional densities, thereby constantly filling in the current draws of every variable. The tth iteration is thus the t-th draw from the Gibbs sampler (van Buuren, 2012).

As the cycle reaches convergence, the current draws are taken as the _rst set of imputed values. The cycle is then repeated until the desired number of imputations have been achieved (van Buuren et al., 2006).

FCS has many practical advantages over JM. Dividing the multidimensional problem into several one dimensional problems allows for more flexible models than if a joint model would be used. The joint distributions available for MI are rather limited while there exist many univariate distributions that can be used for imputation purposes. Hence, bounds, constraints and interactions between variables that may be difficult to include as a part of a multivariate model, can be more easily incorporated. Further, generalizations to data with nonignorable missing data mechanisms might be easier. Finally, different imputation models specified for every variable is easier to communicate to the practitioner (van Buuren et al., 2006).

FCS, however, suffers from the lack of theoretical justification. Incompatibility of the conditional distributions may be a problem.8 Convergence, and the distribution to which the conditionals converge, may or may not depend on the order of sequence of variables. This lack of theoretical justification may further cause problems because of difficulties when examining the quality of the imputations as the joint distribution may or may not exist, and convergence criteria.

2.8 Multivariate Normal Imputation

MVNI is a kind of joint modelling that involves specifying a multivariate normal distribution for missing data and drawing imputation from their conditional distributions by Markov Chain Monte Carlo (MCMC).

Suppose that we know how to combine the estimates from multiple imputations and how many imputations to estimate, we then determine the proper way to simulate imputations. Schafer, (1997) gives an excellent presentation of these methods.

Assuming MAR assumption for the nonresponse, the approach is to simulate the missing data under some assumptions (Rubin 1987). For MI to allow valid inference, the imputations must be proper. It is pertinent to note that MI does not require an ignorability assumption. The assumption is required when it seems reasonable so as to simplify the problem of specifying a nonresponse mechanism. The posterior distribution of R is an average over the repeated draws from f(YmisjYobs), the posterior predictive distribution of the missing data given the observed data. Since the imputations is independent of the response matrix R, we are treating nonresponse as MAR. Schafer, (1997) treats these results as Bayesianly proper, de_ned as multiple imputations which are independent draws from f(YmisjYobs).

The multiple imputations treated here are repeated imputations, repeated draws from the posterior predictive distribution f(YmisjYobs). Proper imputations must include all sources of modelling uncertainty including B, between imputation variability. Schafer, (1997) provides data augmentation (Tanner & Wong, 1987) as a method for generating Bayesianly proper imputations which include B. We present here the adaptation for the multivariate normal model.

Markov chain to draw MIs (which introduces the risk of dependency between the data sets) or running m independent chains. If one used one Markov chain, one would choose some k sufficiently large, say 500, such that one would draw from the distribution only after it has stabilized and at that point, draw after every k cycles of the IP procedure. One can examine diagnostic autocorrelation functions (ACF) to see if the autocorrelation across iterations is sufficiently low to treat the draws from one Markov chain are independent. m independent Markov chains are preferable since there is no autocorrelation by construction, but the cost is running m1 additional MCMC simulations using the IP algorithm. As computation costs decline, this becomes less of an issue. This tradeoff is probably best addressed by running independent chains. Independent chains also should give the analyst a more reliable estimate of error due to simulation (Monte Carlo error) (Schafer, 1997). Running m chains also avoids examination of as many ACF charts since one does not have to assess autocorrelations as often.

2.9 The mi command in Stata

A number of software for statistical analysis offers MI, some of which are SAS, Stata, SPSS and R. In particular, Stata provides the SRMI library that offers MI under both FCS and MVNI using the mi suite of command. The command offers to perform the MI, analyse the imputed data sets and pool the results of the analysis.

The mi suite of commands deals with MI data. mi first sets the data and stores them in one of four formats. MI data contain m imputations numbered m = 1; 2; _ _ _ ;M; and contain m = 0; the original data with missing values. Each variable in MI data can be registered as imputed, passive, or regular. Variables are registered as imputed or regular according as they contain or do not contain missing observations, while passive variables are algebraic combinations of imputed, regular, or other passive variables. mi also allows the user to perform passive imputation when a transformation of one or many variables in the data is desired. For example, one may want to compute a log transformation or calculate a row total. To make sure that the log transformation is sustained throughout the data, mi allows the user to impute the log of the original variable instead of any regular imputation model.

Stata uses the mi impute command to fill in missing data on a single variable or multiple variables with plausible values, in which case imputation is done under the MAR assumption. The command can be used repeatedly to impute multiple variables only when the variables are independent and will be used in separate analyses. In practice, multiple variables usually must be imputed simultaneously, and that requires using a multivariate imputation method. The choice of an imputation method in this case also depends on the pattern of missing values. Variables that follow MMP can be imputed sequentially using univariate conditional distributions. A separate univariate imputation model can be specified for each imputation variable, which allows simultaneous imputation of variables of different types (Rubin 1987).

Stata also includes guidelines on choosing variables to include in the imputation model. One of which is that the analytic model and the imputation model should be congenial. When a pattern of missing values is arbitrary, iterative methods are used to fill in missing values. The mi impute mvn method uses multivariate normal data augmentation to impute missing values of continuous imputation variables (Schafer, 1997). FCS also accommodates arbitrary missing value patterns (van Buuren et al., 1999) using the mi impute chained command.This command uses a Gibbs-like algorithm to impute multiple variables sequentially using univariate FCS. The algorithm samples from the conditional distribution until finally its draws are made from the joint distribution of the variables. The uncertainty about the imputations is captured by both drawing imputations and the parameters of the conditional imputation model. It starts with a random draw from the observed values and cycles through the conditional distributions until convergence, or as long as is desired. The m Gibbs samplers are run in parallel and in the last iteration the imputed values are taken to create the m imputed data sets (van Buuren, 2012).

Stata also includes guidelines on choosing variables to include in the imputation model. One of which is that the analytic model and the imputation model should be congenial. When a pattern of missing values is arbitrary, iterative methods are used to fill in missing values. The mi impute mvn method uses multivariate normal data augmentation to impute missing values of continuous imputation variables (Schafer, 1997). FCS also accommodates arbitrary missing value patterns (van Buuren et al., 1999) using the mi impute chained command.This command uses a Gibbs-like algorithm to impute multiple variables sequentially using univariate FCS. The algorithm samples from the conditional distribution until finally its draws are made from the joint distribution of the variables. The uncertainty about the imputations is captured by both drawing imputations and the parameters of the conditional imputation model. It starts with a random draw from the observed values and cycles through the conditional distributions until convergence, or as long as is desired. The m Gibbs samplers are run in parallel and in the last iteration the imputed values are taken to create the m imputed data sets (van Buuren, 2012).

CHAPTER THREE

METHODOLOGY

3.1 Preamble

Fully conditional specification (FCS) is a practical approach for imputing missing datasets based on a set of imputation models, given that there is one model for each variable with missing values. It has been described in the context of medical research and recommended as a suitable approach for imputing incomplete (fairly) large datasets (Royston and White (2011), van Buuren et al. (1999), and White et al. (2011)). Because FCS involves a series of univariate models rather than a single large model, it imputes data on a variable by variable basis by specifying an imputation model per variable. Hence, the method used in this study is substantially dependent on the specification of the imputation model.

3.2 Assessing the MAR assumption

The methodology of MI depends on the assumption that missing data mechanism is MAR. Although, there is no formal procedure to test this assumption, we employ several tools based on the variable affected. One way is to compare respondents with and without response on the basis of some variables. Consequently, a t-test is used when the average of some continuous variable is compared, while a chi-square test is used when the marginal distributions of a categorical variable is compared. A further test of whether a given variable is MCAR or MAR is to fit a logistic regression model that predicts the probability of missingness given other, possibly complete, variables. The data is MAR rather than MCAR provided the variables significantly predicts this probability of missingness on the variable affected. In this study, we employ the socio-demographic variables as predictors in the logistic models and as variables on the basis of which comparison is made. All significance is declared at 5% level of significance.

CHAPTER THREE

METHODOLOGY

3.1 Preamble

Fully conditional specification (FCS) is a practical approach for imputing missing datasets based on a set of imputation models, given that there is one model for each variable with missing values. It has been described in the context of medical research and recommended as a suitable approach for imputing incomplete (fairly) large datasets (Royston and White (2011), van Buuren et al. (1999), and White et al. (2011)). Because FCS involves a series of univariate models rather than a single large model, it imputes data on a variable by variable basis by specifying an imputation model per variable. Hence, the method used in this study is substantially dependent on the specification of the imputation model.

3.2 Assessing the MAR assumption

The methodology of MI depends on the assumption that missing data mechanism is MAR. Although, there is no formal procedure to test this assumption, we employ several tools based on the variable affected. One way is to compare respondents with and without response on the basis of some variables. Consequently, a t-test is used when the average of some continuous variable is compared, while a chi-square test is used when the marginal distributions of a categorical variable is compared. A further test of whether a given variable is MCAR or MAR is to fit a logistic regression model that predicts the probability of missingness given other, possibly complete, variables. The data is MAR rather than MCAR provided the variables significantly predicts this probability of missingness on the variable affected. In this study, we employ the socio-demographic variables as predictors in the logistic models and as variables on the basis of which comparison is made. All significance is declared at 5% level of significance.

3.3 Choice of variables to be imputed

Before building an imputation model for missing data, an important step is the choice of Ymis, the set of p variables with missing values that are going to be imputed. Depending on one's imputation strategy, this set need not always be equivalent with the set of all variables with missing values in the dataset. For example, an imputation strategy might aim at reducing the size of imputation model by restricting imputations to a small subset of all the variables with missing values in the data set. This presents an important drawback because excluding other missing variables from the regression model ignores their correlations with the included (observed and missing) variables and thus violates the three general imputation requirements by Little and Rubin (2002) that association should be preserved by imputation models in both observed and missing variables, and even between missing variables.

For the above reasons our imputation strategy for the APF data is to impute the biggest possible set of variables with missing data such that the amount of missing data in a variable does not exceed 50%, which in our case consists of p = 72 variables out of all the 74 variables with missing values in the data set.

3.4 Types of models

In this section we define a regression model for each variable in Ymis that we want to impute. The choice of such a model determines the functional form of the conditional posterior distribution of the regression coefficients and residual variance and the conditional posterior predictive distribution of Yj from which we are going to draw the values used to impute the missing observations. For example, if we chose a linear regression model for Yj ; then Yj would follow a Normal distribution by assumption, and it can be shown that both its posterior predictive distribution and the distribution of j would be Normal.

We choose each regression model depending upon the variable type for Yj . There are three basic variable types in our data set: binary (e.g. sex), ordinal (e.g. father's highest level of education) and nominal (e.g. mother's occupation) variables. For the purpose of this study, the choice of the regression models is as follows: we use a logit model for the binary variables, an ordered logit model for the ordinal variables and a multinomial logit model for the nominal variables.

3.5 Predictor selection

As mentioned above about the choice of the variables to be imputed, one of the main goals of imputation is to preserve association between missing and observed variables, and also between missing variables. Therefore, when choosing predictors for the imputation model, it is not enough to select the most accurate predictors for each outcome variable as this approach may bias the correlation structure between the excluded variables variable and outcome variable. Also, ignoring variables that are determinants of non-response of the outcome variable makes the ignorability assumption on which our imputation model relies less plausible. Hence, we choose the number of predictors as large as possible (broad conditioning approach): the more predictors, the lower the bias and the higher the certainty of our imputations. However, there is a limit, of course. In such a large data set as in the APF data with several variables, it is not feasible to include all of them mainly because of multicollinearity and computational problems. Similar to van Buuren, Boshuizen, and Knook (1999), we adopt the following strategy for selecting predictor variables:

1. *Include the variables that are determinants of non-response.* These are necessary to satisfy the ignorability assumption, on which our imputation model relies. According to the ignorability assumption, the distribution of the complete data (including the unobserved values) only depends on the observed data, conditional on the determinants of item-nonresponse and other covariates. Determinants of nonresponse are found by inspecting their correlations with the response indicator of the variable to be imputed.
2. *In addition, include variables that are very good at predicting and explaining the variable of interest we want to impute.* This is the classical criterion for predictors and helps to reduce uncertainty of the imputations. These predictors are identified by their correlation with the target variable.

3. *In addition, remove the predictor variables from above that have too many missing values within the subsample of missing observations of the variable to be imputed and substitute them with more complete predictors of these predictors.* As a rule of thumb, predictors with percentages of observed cases within this subsample lower than 50% are removed and substituted by more complete predictors. This criterion contributes to make imputations more robust.
4. *In addition, include all variables that appear in the models that will be applied to the data after imputation.* In other words, one should envisage the several applications in which the data may be involved and include the variables as predictors that are expected to affect or explain according to these applications the variable to be imputed. Failure to do so will tend to bias results of potential users of the data.

3.6 Imputation order

One weakness of the FCS approach is that conditional densities may not converge to a stationary distribution. In practice, however, choosing a particular ordering of the variables often aid convergence. In the APF data we start imputation by the variables with the least missing values, and so on. Variables with the same amount of missingness are processed in an arbitrary order, but always in the same order.

3.7 Number of iterations

The number of iterations t determines how often the imputation procedure cycles through the variables to be imputed, replacing variables that are being conditioned in any regression by the observed or currently imputed values. As t tends to infinity, the sequence of parameters and predicted values should converge to a draw from the posterior distribution of _ and a draw from the posterior predictive distribution of Ymis. However, according to van Buuren, Boshuizen, and Knook (1999) during the first few iterations convergence in these models usually occurs very fast in practice because the posterior distributions of the regression coefficients already absorb a lot of uncertainty in the predictors and because the procedure creates imputations that are already statistically independent. Given the substantial computational effort required for the APF imputation model and following the number of iterations used in other similar surveys (like SCF (Kennickell (1991)) we set the iteration number for the APF imputation model to t = 8.

3.8 Number of imputations

Finally, we choose the number of realizations D that we want to have from the posterior predictive distribution P(Ymis Yobs) or, in other words, the number of multiply imputed data sets. Setting D too low leads to standard errors of the estimates that are too low and to p-values that are too low. Schafer and Olsen (1998) show that the gains of efficiency of an estimate rapidly diminish after the first few D imputations. They claim that good inferences can already be made with D = 3 to 5. However, Graham et al, (2007) show that another important quantity such as statistical power can vary more dramatically with D than is implied by efficiency. They claim that good inferences can be made with D = 20 to 40. It seems unlikely that a single correct value for D will be established in the literature because, like sample size, the number of imputation that are necessary depends on features of the individual data set and analysis model. In the APF imputation model, given the substantial increase in computational effort for every further imputation and following other similar surveys like the SCF we set the number of imputations to D = 5.

3.9 Method for combining analysis results

The multiple imputation methodology entails combining estimates from imputed datasets so as to produce one set of parameter estimates. For the APF dataset and in particular, the RSES a regression model is fitted to each imputed dataset and estimates are combined.

To combine the estimates across imputations, Rubin (1987) specifies that the average of individual estimates produced at each imputation be taken. The combined variance of this estimate consists of two parts: one accounts for natural variability. This part is often called the "within-imputation component", while the other accounts for "between-imputation" uncertainty introduced by missing data.

CHAPTER FOUR

RESULTS

4.1 Brief description of the APF data

The data used in this study was collected among adolescents in Ikere-Ekiti Local Government Area in Ekiti State of Nigeria to model predictors of Adolescent Psychosocial Functioning (APF). We shall refer to this data as the APF data. It consists of three psychosocial outcomes scales namely: the Rosenberg Self Esteem Scale (RSES), the Strength and Difficulty Questionnaire (SDQ), and the Center for Epidemiological Studies Depression Scale for Children (CES-DC). We refer to each item of these scales as r, s, and d, respectively. Each scale identifies variables that mostly measures the characteristics of interest. The data also consist of background information about students such as age, weight, height, as well as information about family type and status, parents' highest level of education and occupations. However, this study only considers imputing the RSES.

Table 4.1: Frequency distribution of observed responses on socio-demographic variables

Item	Frequency	Percent
Sex		
Male	201	41.0
Female	289	59.0
Class		
SSS 1	137	28.0
SSS 2	283	57.8
SSS 3	70	14.2
Religion		
Christianity	471	96.1
Islam	19	3.9
Area of Residence		
Rural	206	42.0
Urban	284	58.0
Ethnicity		
Yoruba	455	92.9
Hausa or Fulani	3	.6
Igbo	29	5.9
Others	3	.6
Family type		
Monogamy	378	77.1
Polygamy	112	22.9
Family status		
Parents are together	414	84.5
Parents are divorced	10	2.0
Parents are separated	34	6.9
Single mother	32	6.5

Of the 490 students recruited into the study Table 4.1 reveals that 201 (41.0%) were males while 289 (59.0%) were females. Majority (283, 57.8%) of the students were in the Senior Secondary School II (SSS 2) compare to 137 (28.0%) and 70 (14.3%) students who were in SSS 1 and SSS 3 respectively. Also, most of the students were Christians (471, 96.1%) as against 19 (3.9%) who were Muslims. Almost all the respondents were Yoruba (455, 92.9%) with 29 (5.9%) Igbo students, 3 (0.6%) Hausa or Fulani, and 3 (0.6%) students who were of other ethnic groups. There are 284 (58.0%) resided in the urban area of Ekiti State while 206 (42.0%) lives in the rural area.

Parents of the adolescent students interviewed had majorly a monogamy family type (378, 77.1%) and 112 (22.9%) families were of the polygamous family type. While most parents (414, 84.5%) lived in the same residence together, 34 (6.9%) parents were separated, 32 (6.5%) parents were single mother, and 10 (2.0%) parents were divorced.

Table 4.2: Frequency distribution of observed responses on socio-demographic variables

Item	Frequency	Percent
Father's highest level of education		
No formal education	21	4.3
Primary	32	6.5
Secondary	101	20.6
Tertiary	233	47.6
No idea	103	21.0
Father's occupation		
Farming	56	11.4
Trading	80	16.3
Civil servant	194	39.6
Employee of private organization	107	21.8
Others	53	10.8
Mother's highest level of education		
No formal education	21	4.3
Primary	36	7.3
Secondary	109	22.2
Tertiary	231	47.1
No idea	93	19.0
Mother's occupation		
Farming	12	2.4
Trading	224	45.7
Civil servant	178	36.3
Employee of private organization	43	8.8
Others	33	6.7

Table 4.2: Frequency distribution of observed responses on socio-demographic variables

Item	Frequency	Percent
Father's highest level of education		
No formal education	21	4.3
Primary	32	6.5
Secondary	101	20.6
Tertiary	233	47.6
No idea	103	21.0
Father's occupation		
Farming	56	11.4
Trading	80	16.3
Civil servant	194	39.6
Employee of private organization	107	21.8
Others	53	10.8
Mother's highest level of education		
No formal education	21	4.3
Primary	36	7.3
Secondary	109	22.2
Tertiary	231	47.1
No idea	93	19.0
Mother's occupation		
Farming	12	2.4
Trading	224	45.7
Civil servant	178	36.3
Employee of private organization	43	8.8
Others	33	6.7

The academic history of these parents is considerably fascinating as Table 4.2 reveals that while most fathers have attained a tertiary level of education (233, 47.6%), 101 (20.6%) fathers had at most a secondary education, 32 (6.5%) had at most a primary education, and 21 (4.3%) had no formal education. Nevertheless, 103 (21.0%) students reported that they had no idea of their fathers level of education. Similarly, for the students' mothers, most have attained a tertiary level of education (231, 47.1%), 109 (22.2%) had at most a secondary education, 36 (7.3%) had at most a primary education, and 21 (4.3%) mothers had no formal education.

In addition, more than a third of respondents' fathers were civil servants (194, 39.6%) while 107 (21.8%) students had fathers who were employee of private organizations, 80 (16.3%) fathers were traders, 56 (11.4%) farmers and 53 (10.8%) fathers were into other occupations. For mothers however, up to one half (224, 45.7%) were traders while 178 (36.3%) students had mothers who were civil servants, 80 (16.3%) mothers were employee of private organizations, 12 (2.4%) farmers and 33 (6.7%) mothers were into other occupations.

Table 4.3: Frequency distribution of observed responses on the RSES item

Item	Frequency	Percent
On the whole, I am satisfied with myself		
Strongly disagree	4	.8
Disagree	40	8.4
Agree	214	45.2
Strongly agree	216	45.6
At times I think I am not good at all		
Strongly disagree	68	14.1
Disagree	192	39.8
Agree	163	33.7
Strongly agree	60	12.4
I feel that I have a number of good qualities		
Strongly disagree	11	2.3
Disagree	34	7.0
Agree	249	51.6
Strongly agree	189	39.1
I am able to do things as well as most other people		
Strongly disagree	12	2.5
Disagree	57	12.0
Agree	227	47.8
Strongly agree	179	37.7
I feel I do not have much to be proud of		
Strongly disagree	92	19.5
Disagree	222	46.7
Agree	109	22.9
Strongly agree	52	10.9

In Table 4.3, about a half of the students (216, 45.6%) strongly agreed they were satisfied with themselves, while 214 (45.1%) simply agreed. However, 40 (8.4%) were not satisfied with themselves and 4 (0.8%) strongly declined they were satisfied with themselves. 163 (33.7%) students at times thought they were not good at all, while 192 (39.8%) declined. Also, 60 (12.4%) students strongly agreed and 68 (14.1%) students strongly disagreed that at times they thought they were not good at all. About half of the students (249, 51.6%) reported that they had a number of good qualities and another 189 (39.1%) in addition strongly agreed, remaining 34 (7.0%) and 11 (2.3%) who disagreed and strongly disagreed that they had a number of good qualities respectively.

Moreover, 227 (47.8%) students agreed and another 179 (37.7%) students strongly agreed that they were able to do things as well as most other people compare with 57 (12.0%) students and 12 (2.5%) students who disagree and strongly disagree, respectively, that they were able to do things as well as most other people. Most students declined they did feel they had too much to be proud of. In fact, 222 (46.7%) students disagree while another 92 (19.4%) students strongly disagreed. In contrast, only 109 (22.9%) agreed they did feel they had much to be proud of, while 52 (10.9%) students strongly agreed to this statement.

Table 4.4: Frequency distribution of observed responses on the RSES item

Item	Frequency	Percent
I certainly feel useless at times		
Strongly disagree	44	9.4
Disagree	109	23.2
Agree	194	41.3
Strongly agree	123	26.1
I feel that I'm a person of worth, at least on an equal plane with others		
Strongly disagree	19	4.1
Disagree	56	11.9
Agree	217	46.3
Strongly agree	177	37.7
I wish I could have more respect for myself		
Strongly disagree	205	42.5
Disagree	216	44.8
Agree	38	7.9
Strongly agree	23	4.8
All in all, I am inclined to feel that I am a failure		
Strongly disagree	24	5.2
Disagree	66	14.2
Agree	177	38.1
Strongly agree	198	42.5
I take a positive attitude toward myself		
Strongly disagree	24	5.1
Disagree	66	14.2
Agree	177	38.1
Strongly agree	198	42.6

In Table 4.4, 194 (41.3%) students agreed and 123 (26.2%) students strongly agreed that they certainly felt useless at times, while 109 (23.2%) students disagreed and 44 (9.4%) student strongly disagreed. Also, at least on an equal plane with others, about a half students (217, 46.3%) students agreed and another 117 (37.7%) students strongly agreed that they felt they were persons of worth. However, only 56 (11.9%) students disagreed and 19 (4.1%) student strongly disagreed with this claim. While 38 (7.9%) students agreed and 23 (4.8%) students strongly disagree with the claim that they wish they could have more respect for themselves, most of the students (216, 44.8%) merely declined and most of them (205, 42.5%) also strongly declined the claim.

Moreover, 177 (38.1%) students reported that they were inclined to feel like a failure in addition to 198 (42.6%) students who strongly agreed to the claim, remaining 66 (14.2%) and 24 (5.2%) who disagreed and strongly disagreed that they were inclined to feel like a failure respectively. Most students agreed they took positive attitude toward themselves. In fact, 222 (46.7%) students disagree while another 92 (19.4%) students strongly disagreed. In contrast, only 109 (22.9%) agreed they took positive attitude toward themselves, while 52 (10.9%) students strongly agreed to this statement.

In Table 4.4, 194 (41.3%) students agreed and 123 (26.2%) students strongly agreed that they certainly felt useless at times, while 109 (23.2%) students disagreed and 44 (9.4%) student strongly disagreed. Also, at least on an equal plane with others, about a half students (217, 46.3%) students agreed and another 117 (37.7%) students strongly agreed that they felt they were persons of worth. However, only 56 (11.9%) students disagreed and 19 (4.1%) student strongly disagreed with this claim. While 38 (7.9%) students agreed and 23 (4.8%) students strongly disagree with the claim that they wish they could have more respect for themselves, most of the students (216, 44.8%) merely declined and most of them (205, 42.5%) also strongly declined the claim.

Moreover, 177 (38.1%) students reported that they were inclined to feel like a failure in addition to 198 (42.6%) students who strongly agreed to the claim, remaining 66 (14.2%) and 24 (5.2%) who disagreed and strongly disagreed that they were inclined to feel like a failure respectively. Most students agreed they took positive attitude toward themselves. In fact, 222 (46.7%) students disagree while another 92 (19.4%) students strongly disagreed. In contrast, only 109 (22.9%) agreed they took positive attitude toward themselves, while 52 (10.9%) students strongly agreed to this statement.

4.2 Assessing missing data

All the missing data in this study are unknown and not ignorable since they are due to nonresponse by the students.

Table 4.5: Overall summary of missing values

	Complete		Incomplete	
Missing	Number	Percentage	Number	Percentage
Variables	7	8.7	74	81.3
Cases	490	100		
Values	42456	95.2	2134	4.8

Table 4.5 summarizes the missing values present in the APF data. All the records had at least a value missing on some variables. Only seven (8.7%) variables provide complete data on all students. In all, the nonresponse rate is about 4.8%.

Table 4.6: Percentage of values missing on the socio-demographic variables.

Item	Missing Number	Missing Percent	Mean	Standard Deviation
Height	490	100		
Age	9	1.8	15.23	1.38
Weight	414	84.5	47.87	9.582
Sex	3	0.6		
Religion	1	0.2		
Area of Residence	60	12.2		
Ethnicity	1	0.2		
Family type	24	4.9		
Family status	4	0.8		
Father's highest level of education	37	7.6		
Father's occupation	17	3.5		
Mother's highest level of education	30	6.1		
Mother's occupation	10	2		

Table 4.6 reveals that no student gives information on height, while 76 (15.5%) students provided information on weight. Location is the next variable with highest missing values with 60 (12.2%) values missing. Nearly all students provided information on some variables, three (0.6%) on sex, while only one (0.2%) students failed to provide data on ethnicity and religion.

Table 4.7: A chi-square values comparing respondents with observed and missing responses.

Variable[a]	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
sex	0.207	0.031	0.011	0.378	0.042	0.763	1.586	0.137	0.010	0.204
sch	2.659	3.323	3.487	1.326	3.227	8.953	2.186	2.268	0.797	1.212
cls	2.057	0.576	2.283	0.573	1.458	1.579	0.108	0.725	0.001	0.033
rel	5.836*	0.885	1.063	0.323	0.328	0.287	9.816**	0.071	0.287	0.323
res	0.599	0.006	1.072	1.501	0.211	0.529	0.430	1.241	4.599*	0.482
fat	1.111	4.074*	0.703	0.796	2.410	0.296	2.587	1.955	0.296	2.300
fas	1.047	2.185	0.797	0.327	1.068	1.162	4.764	1.301	1.495	1.471
fed	8.723**	2.268	3.871	2.041	8.953	2.659	3.227	3.323	1.212	2.186
med	2.770	4.760	4.664*	8.739**	4.485	2.293	1.171	1.375	2.041	4.745*
foc	3.482	1.419	2.673	2.013	8.504*	3.612	4.862	3.612*	2.564	4.702
moc	2.489	2.710	3.985	3.029	4.351	1.218	2.846	1.377	1.562	4.303

* $p < 0.05$

[a] sex = Sex
sch = School
cls = Class
rel = Religion
res = Area of Residence
fat = Family type
fas = Family status
fed = Father's highest level of education
med = Mother's highest level of education
foc = Father's occupation
moc = Mother's occupation

Table 4.7 presents the result of a chi-square analysis to examine the comparability of respondents with observed and missing responses on each of sex, class, religion, area of residence, family type, family status, father's education, mother's education, father's occupation, and mother's occupation. Noticeable pattern of significant chi-square value occurs for r1 when comparison of respondents with observed and missing responses is made by religion (χ^2=5.836, $p < 0.05$), as well as when comparison is made by father's highest level of education (χ^2=8.732, $p < 0.01$). Also, for r2 there is significant difference in the groups of respondents when comparison is made with family type (χ^2=4.074, $p < 0.05$), for r3 when comparison is made by mother's highest level of education (χ^2=4.664, $p < 0.05$) with similar comparison for r4 (χ^2=8.739, $p < 0.01$). Furthermore, for r5 a significant difference exists when comparison is made by father's occupation, for r7 when comparison is made by religion (χ^2=9.816, $p < 0.01$), for r8 when comparison is made by father's occupation (χ^2=3.612, $p < 0.05$), for r9 when comparison is made by area of residence (χ^2= 4.559, $p < 0.05$), and for r10 when comparison is made by mother's education (χ^2= 4.745, $p < 0.05$).

Furthermore, Table 4.8 shows the results of t-test for the comparison of respondents with observed and missing responses on age. No noticeable significant t-value occurs when comparison of respondents' ages were made among those with observed and missing responses on each item of the RSES.

Table 4.8: A t-test analysis comparing mean ages of respondents with observed values and respondents with missing values

	Number		Mean			95% CI	
Item	Observed	Missing	Observed	Missing	t-value	Lower	Upper
R1	474	16	15.21	15.56	-1.00	-1.04	0.34
R2	482	8	15.23	15.00	0.43	-0.81	1.26
R3	483	7	15.22	15.43	-0.40	-1.24	0.82
R4	475	15	15.23	15.07	0.44	-0.55	0.87
R5	474	16	15.23	15.07	0.44	-0.55	0.87
R6	470	20	15.23	15.10	0.40	-0.49	0.75
R7	469	21	15.24	14.90	1.08	-0.27	0.94
R8	482	8	15.23	14.88	0.72	-0.61	1.32
R9	467	23	15.23	15.00	0.77	-0.36	0.83
R10	465	25	15.25	14.80	1.57	-0.11	1.00

4.3 Justification for Imputations

The results shown in Table 4.9 to Table 4.18 model the probability of missingness of items on the RSES as a function of the socio-demographic variables. This step becomes necessary as variables that significantly contribute to each model are deemed to affect these probabilities, providing partial evidence for the assumption of MAR, and thus these variables are incorporated into the imputation model for respective items.

Table 4.9: Logistic regression of nonresponse on item 1 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	-0.088	0.753	0.209	4.011
Sex (Female)				
Male	-0.640	0.957	0.081	3.444
School (Amoye)				
Comprehensive	0.670	1.132	0.213	17.961
Victory College	-1.340	1.298	0.021	3.337
St. Louis	0.983	0.963	0.404	17.654
Govt. College	-1.675	1.413	0.012	2.989
Class (SSS 2)				
SSS 1	-0.384	0.843	0.131	3.557
SSS 3	-0.818	0.960	0.067	2.898
Religion (Christianity)				
Islam	1.549*	0.764	1.053	21.021
Area of Residence (Urban)				
Rural	-0.595	1.114	0.062	4.894
Family type (Monogamy)				
Polygamy	-0.343	0.899	0.122	4.132
Family status (Parents are together)				
Parents are divorced	-0.585	1.431	0.138	2.248
Parents are separated	0.518	0.959	0.659	4.276
Single mother	-0.765	1.426	0.116	1.869
Father's education (Tertiary)				
No formal education	-1.125	0.677	0.168	0.628
Primary	2.912	1.779	3.246	38.222
Secondary	1.565	0.975	1.849	12.374
No idea	1.672*	0.678	1.411	20.097
Mother's education (Tertiary)				
No formal education	-0.797	1.210	0.139	1.466
Primary	0.524	0.714	0.842	3.388
Secondary	-1.021	1.162	0.116	1.118
No idea	1.607	1.084	1.733	14.352
Father's occupation (Civil servant)				
Farming	0.875	0.893	1.004	5.730
Trading	0.665	0.624	1.058	3.573
Employee of private organization	0.092	0.866	0.471	2.551
Others	1.532	0.800	2.121	10.095

Significance marker: * $p < 0.05$

4.3.1 Predictors of nonresponse on item 1

Table 4.9 above shows the result of a multiple logistic regression model estimation for nonresponse on item 1 of the RSES as a function of age, sex, school, class, religion, area of residence, family type, family status, father's education, mother's education, and father's occupation. We found that the effect of religion is significant ($\beta = 1.549$, $p < 0.05$), as well as the effect of father's education ($\beta = 1.672$, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 1 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 1 of the RSES only religion and father's education were used as predictors of the missing values on the item.

Table 4.10: Logistic regression of nonresponse on item 2 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	0.259	1.467	0.073	22.993
Sex (Female)				
Male	-0.667	0.633	0.148	1.775
School (Amoye)				
Comprehensive	-0.045	0.967	0.144	6.36
Victory College	-0.099	0.791	0.192	4.269
St. Louis	-0.477	0.841	0.119	3.227
Govt. College	-1.557	0.033	0.204	1.034
Class (SSS 2)				
SSS 1	1.065	0.681	0.764	11.014
SSS 3	0.127	0.925	0.185	6.961
Area of Residence (Urban)				
Rural	4.364*	1.621	3.277	182.71
Family type (Monogamy)				
Polygamy	-2.92*	1.306	0.004	0.697
Family status (Parents are together)				
Parents are divorced	-6.799	0.286	0.001	1.331
Parents are separated	1.286	0.792	0.767	17.067
Single mother	-5.597	1.027	0.001	2.793
Father's education (Tertiary)				
No formal education	1.683	1.384	0.357	81.104
Primary	-3.911	1.046	0.007	0.056
Secondary	0.836	0.822	0.46	11.556
No idea	-2.574*	1.205	0.007	0.808
Father's occupation (Civil servant)				
Farming	-0.124	1.048	0.113	6.882
Trading	0.045	0.775	0.229	4.775
Employee of private organization	0.247	0.683	0.336	4.888
Others	-0.103	0.859	0.390	2.361

Significance marker: * $p < 0.05$

4.3.2 Predictors of nonresponse on item 2

Table 4.10 above shows the result of a multiple logistic regression model estimation for nonresponse on item 2 of the RSES as a function of age, sex, school, class, area of residence, family type, family status, father's education, and father's occupation. We found that the effect of area of residence is significant ($\beta = 4.364$, $p < 0.05$), as well as the effect of family type ($\beta = -2.92$, $p < 0.05$) and father's education ($\beta = -2.547$, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 2 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 2 of the RSES only area of residence, family type, and father's education were used as predictors of the missing values on the item.

Table 4.11: Logistic regression of nonresponse on item 3 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	1.494	1.161	0.457	43.379
Sex (Female)				
Male	1.445*	0.674	1.132	15.896
School (Amoye)				
Comprehensive	-0.382	1.071	0.084	5.57
Victory College	-0.079	0.818	0.186	4.595
St. Louis	1.466	0.97	0.647	28.998
Govt. College	-0.408	0.98	0.097	4.535
Class (SSS 2)				
SSS 1	-0.322	0.757	0.164	3.193
SSS 3	0.362	0.755	0.327	6.312
Area of Residence (Urban)				
Rural	1.82*	0.819	1.24	30.706
Family type (Monogamy)				
Polygamy	-0.724	0.788	0.104	2.269
Family status (Parents are together)				
Parents are divorced	1.282	1.944	0.542	23.984
Parents are separated	1.672*	0.678	1.411	20.097
Single mother	-1.177	0.785	0.143	0.663
Father's education (Tertiary)				
No formal education	1.05	1.519	0.145	56.121
Primary	-2.201	1.964	0.016	0.751
Secondary	1.574	0.881	0.859	27.142
No idea	1.721	0.91	0.938	33.271
Mother's education (Tertiary)				
No formal education	-1.099	2.663	0.025	4.470
Primary	-1.335	1.248	0.078	0.889
Secondary	-1.443	0.966	0.036	1.567
No idea	-1.913	1.025	0.020	1.100

Significance marker: * $p < 0.05$

4.3.3 Predictors of nonresponse on item 3

Table 4.11 above shows the result of a multiple logistic regression model estimation for nonresponse on item 3 of the RSES as a function of age, sex, school, class, area of residence, family type, family status, father's education, and mother's education. We found that the effect of sex is significant (β = 1.445, $p < 0.05$), as well as the effect of area of residence (β = 1.82, $p < 0.05$) and family status (β = 1.672, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 3 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 3 of the RSES only sex, area of residence, and family status were used as predictors of the missing values on the item.

Table 4.12: Logistic regression of nonresponse on item 4 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio Lower	Upper
Age	1.649	1.227	0.469	57.632
Sex (Female)				
Male	0.908	1.366	0.17	3.059
School (Amoye)				
Comprehensive	1.158	1.499	0.169	16.087
Victory College	-1.006	2.094	0.047	24.817
St. Louis	2.541	1.594	0.558	89.051
Govt. College	2.028	1.293	0.603	95.841
Class (SSS 2)				
SSS 1	-1.306	2.717	0.019	3.831
SSS 3	-1.951	1.387	0.009	2.156
Area of Residence (Urban)				
Rural	1.289	1.19	0.353	37.358
Family type (Monogamy)				
Polygamy	-2.162	2.096	0.015	0.888
Family status (Parents are together)				
Parents are divorced	-1.727	0.832	0.079	0.400
Parents are separated	-0.937	1.893	0.010	16.028
Single mother	-5.351	4.867	0.000	0.546
Father's education (Tertiary)				
No formal education	1.286	0.792	0.767	17.067
Primary	-5.597	1.027	0.001	2.793
Secondary	3.486**	1.226	3.727	92.71
No idea	0.127	0.925	0.185	6.961
Mother's education (Tertiary)				
No formal education	-1.797	2.331	0.017	1.609
Primary	-2.661	1.588	0.015	0.329
Secondary	-1.112	2.015	0.046	2.346
No idea	1.929	2.048	0.124	30.948
Father's occupation (Civil servant)				
Farming	-1.568	2.028	0.029	1.506
Trading	0.259	1.467	0.073	22.993
Employee of private organization	0.296	1.321	0.101	17.894
Others	-0.299	1.616	0.031	17.611
Mother's occupation (Trading)				
Farming	-2.815	6.243	0.000	26.362
Civil servant	-2.639*	1.288	0.006	0.893
Employee of private organization	-2.962	2.021	0.001	2.718
Others	-0.676	1.642	0.02	12.703

Significance marker: * $p < 0.05$

4.3.4 Predictors of nonresponse on item 4

Table 4.12 above shows the result of a multiple logistic regression model estimation for nonresponse on item 4 of the RSES as a function of age, sex, school, class, area of residence, family type, family status, father's education, mother's education, father's occupation, and mother's occupation . We found that the effect of father's education is significant (β = 3.486, $p < 0.05$), as well as the effect of mother's occupation (β = -2.639, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 4 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 4 of the RSES only father's education and mother's occupation were used as predictors of the missing values on the item.

Table 4.13: Logistic regression of nonresponse on item 5 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	1.579	1.58	0.219	17.295
Sex (Female)				
Male	-0.47	0.591	0.196	1.991
School (Amoye)				
Comprehensive	0.461	0.73	0.379	6.635
Victory College	-0.088	0.753	0.209	4.011
St. Louis	0.125	0.73	0.271	4.738
Govt. College	0.196	1.542	0.059	24.97
Class (SSS 2)				
SSS 1	-0.817	0.721	0.108	1.814
SSS 3	-0.18	0.766	0.186	3.745
Area of Residence (Urban)				
Rural	1.179	1.527	0.163	64.79
Family type (Monogamy)				
Polygamy	0.129	0.604	0.348	3.715
Family status (Parents are together)				
Parents are divorced	0.039	1.672	0.039	27.542
Parents are separated	-0.883	1.113	0.047	3.665
Single mother	-0.595	1.114	0.062	4.894
Father's occupation (Civil servant)				
Farming	-0.797	1.21	0.042	4.834
Trading	1.549*	0.764	1.053	21.021
Employee of private organization	0.665	0.624	0.572	6.607
Others	0.875	0.893	0.417	13.795
Mother's occupation (Trading)				
Farming	-0.232	1.533	0.039	16.006
Civil servant	-0.008	0.578	0.32	3.077
Employee of private organization	0.11	0.783	0.241	5.177
Others	0.373	1.159	0.15	14.085

Significance marker: * $p < 0.05$

4.3.5 Predictors of nonresponse on item 5

Table 4.13 above shows the result of a multiple logistic regression model estimation for nonresponse on item 5 of the RSES as a function of age, sex, school, class, area of residence, family type, father's occupation, and mother's occupation . We found that the effect of father's occupation is significant (β = 1.549, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 5 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 5 of the RSES only father's occupation was used as predictors of the missing values on the item.

Table 4.14 Logistic regression of nonresponse on item 6 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	2.086	1.76	0.256	53.559
Sex (Female)				
Male	-0.025	0.51	0.359	2.652
School (Amoye)				
Comprehensive	0.025	0.756	0.233	4.514
Victory College	-1.137	0.8	0.067	1.54
St. Louis	-0.681	0.837	0.098	2.608
Govt. College	-0.609	0.749	0.125	2.36
Class (SSS 2)				
SSS 1	1.116	0.584	0.973	9.585
SSS 3	0.001	0.869	0.182	5.497
Area of Residence (Urban)				
Rural	-1.547	2.103	0.003	13.131
Family type (Monogamy)				
Polygamy	-0.542	0.607	0.177	1.911
Mother's education (Tertiary)				
No formal education	0.322	1.391	0.09	21.068
Primary	-2.199	1.832	0.003	4.023
Secondary	0.953	0.794	0.547	12.301
No idea	-0.584	1.076	0.068	4.598
Father's occupation (Civil servant)				
Farming	-0.079	0.788	0.197	4.329
Trading	-0.826	0.769	0.097	1.977
Employee of private organization	-0.242	0.646	0.221	2.785
Others	-1.380	1.172	0.025	2.501
Mother's occupation (Trading)				
Farming	0.843	1.383	0.155	34.926
Civil servant	0.461	0.613	0.477	5.279
Employee of private organization	1.134	0.719	0.759	12.734
Others	0.075	1.170	0.109	10.686

Significance marker: * $p < 0.05$

4.3.6 Predictors of nonresponse on item 6

Table 4.14 above shows the result of a multiple logistic regression model estimation for nonresponse on item 6 of the RSES as a function of age, sex, school, class, area of residence, family type, father's occupation, and mother's occupation . We found that none of these variables has significant effect on the pattern of missingness on this item. Hence there is no sufficient evidence that the missing data mechanism governing nonresponse on item 6 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 6 of the RSES only a nonzero regression parameter was used.

Table 4.15 Logistic regression of nonresponse on item 7 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	0.179	1.116	0.134	10.649
Sex (Female)				
Male	-1.339	1.243	0.023	2.997
School (Amoye)				
Comprehensive	1.579	1.58	0.219	17.295
Victory College	0.196	1.542	0.059	24.97
St. Louis	-0.232	1.533	0.039	16.006
Govt. College	0.039	1.672	0.039	27.542
Class (SSS 2)				
SSS 1	-0.376	1.396	0.044	10.597
SSS 3	0.373	1.159	0.15	14.085
Religion (Christianity)				
Islam	1.672*	0.678	1.411	20.097
Area of Residence (Urban)				
Rural	2.912	1.779	0.562	61.446
Family type (Monogamy)				
Polygamy	1.179	1.527	0.163	64.79
Father's occupation (Civil servant)				
Farming	1.565	0.975	0.708	32.298
Trading	-1.152	1.225	0.029	3.484
Employee of private organization	-1.021	1.162	0.037	3.514
Others	1.607	1.084	0.596	41.786
Mother's occupation (Trading)				
Farming	1.872	1.162	0.666	63.482
Civil servant	1.932	1.765	0.217	29.533
Employee of private organization	-0.053	1.082	0.114	7.912
Others	0.179	1.116	0.134	10.649

Significance marker: * $p < 0.05$

4.3.7 Predictors of nonresponse on item 7

Table 4.15 above shows the result of a multiple logistic regression model estimation for nonresponse on item 7 of the RSES as a function of age, sex, school, class, religion, area of residence, family type, father's occupation, and mother's occupation . We found that the effect of religion is significant (β = 1.672, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 7 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 7 of the RSES only religion was used as predictors of the missing values on the item.

Table 4.16 Logistic regression of nonresponse on item 8 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio Lower	Upper
Age	-0.512	0.695	0.153	2.338
Sex (Female)				
Male	1.002	0.676	0.724	10.25
School (Amoye)				
Comprehensive	-1.138	1.03	0.043	2.412
Victory College	-1.085	1.015	0.046	2.471
St. Louis	0.264	1.076	0.158	10.731
Govt. College	-1.048	1.032	0.046	2.648
Class (SSS 2)				
SSS 1	0.715	0.815	0.414	10.095
SSS 3	0.616	1.008	0.257	13.347
Family type (Monogamy)				
Polygamy	-0.343	0.899	0.122	4.132
Family status (Parents are together)				
Parents are divorced	1.289	1.19	0.353	37.358
Parents are separated	-0.530	1.249	0.051	6.806
Single mother	0.457	1.156	0.164	15.238
Father's education (Tertiary)				
No formal education	1.128	1.61	0.132	72.511
Primary	1.649	1.227	0.469	57.632
Secondary	1.61	0.923	0.82	30.552
No idea	-0.199	1.189	0.08	8.423
Mother's education (Tertiary)				
No formal education	0.092	0.866	0.471	2.551
Primary	1.532	0.800	2.121	10.095
Secondary	-0.514	0.971	0.089	4.013
No idea	0.804	1.067	0.276	18.085
Father's occupation (Civil servant)				
Farming	-0.022	1.196	0.094	10.211
Trading	1.636*	0.72	1.253	21.046
Employee of private organization	-0.202	1.597	0.036	18.696
Others	-0.382	1.418	0.042	11.002

Significance marker: * $p < 0.05$

4.3.8 Predictors of nonresponse on item 8

Table 4.16 above shows the result of a multiple logistic regression model estimation for nonresponse on item 8 of the RSES as a function of age, sex, school, class, family type, family status, father's education, mother's education, and father's occupation . We found that the effect of father's occupation is significant (β = 1.636, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 8 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 8 of the RSES only father's occupation was used as predictors of the missing values on the item.

Table 4.17 Logistic regression of nonresponse on item 9 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	1.721	0.91	0.938	33.271
Sex (Female)				
Male	-0.36	0.692	0.18	2.711
School (Amoye)				
Comprehensive	0.68	0.895	0.342	11.399
Victory College	0.079	0.875	0.195	6.014
St. Louis	0.343	0.972	0.21	9.476
Govt. College	-0.382	1.071	0.084	5.57
Class (SSS 2)				
SSS 1	-1.438	1.135	0.026	2.196
SSS 3	1.445*	0.674	1.132	15.896
Area of Residence (Urban)				
Rural	-0.079	0.818	0.186	4.595
Family type (Monogamy)				
Polygamy	1.466	0.97	0.647	28.998
Father's education (Tertiary)				
No formal education	0.621	1.507	0.097	35.675
Primary	-0.154	1.488	0.046	15.851
Secondary	0.854	0.967	0.353	15.624
No idea	-1.246	1.132	0.031	2.645
Mother's education (Tertiary)				
No formal education	0.912	1.513	0.128	48.276
Primary	0.313	1.213	0.127	14.743
Secondary	-0.678	1.32	0.038	6.741
No idea	1.561	1.028	0.635	35.76
Father's occupation (Civil servant)				
Farming	0.34	1.079	0.169	11.648
Trading	1.672*	0.678	1.411	20.097
Employee of private organization	0.523	0.785	0.363	7.857
Others	0.029	1.052	0.131	8.093
Mother's occupation (Trading)				
Farming	2.209	1.295	0.719	115.356
Civil servant	0.767	0.857	0.402	11.541
Employee of private organization	1.82*	0.819	1.24	30.706
Others	1.289	1.129	0.397	33.152

Significance marker: * $p < 0.05$

4.3.9 Predictors of nonresponse on item 9

Table 4.17 above shows the result of a multiple logistic regression model estimation for nonresponse on item 9 of the RSES as a function of age, sex, school, class, area of residence, family type, father's education, mother's education, father's occupation and mother's occupation. We found that the effect of class is significant (β = 1.445, $p < 0.05$), as well as father's occupation (β = 1.672, $p < 0.05$), mother's occupation (β = 1.82, $p < 0.05$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 9 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 9 of the RSES only class, father's occupation, and mother's occupation were used as predictors of the missing values on the item.

Table 4.18 Logistic regression of nonresponse on item 10 of RSES on some selected variables

Variable	Coefficient	Standard Error	95% C.I. for odds ratio	
			Lower	Upper
Age	0.372	0.883	0.257	8.194
Sex (Female)				
Male	-0.216	1.476	0.045	14.542
School (Amoye)				
Comprehensive	2.086	1.76	0.256	53.559
Victory College	-1.547	2.103	0.003	13.131
St. Louis	-2.199	1.832	0.003	4.023
Govt. College				
Class (SSS 2)				
SSS 1	0.244	1.882	0.032	50.999
SSS 3	1.97	1.426	0.438	117.382
Area of Residence (Urban)				
Rural	2.293	2.161	0.144	684.206
Family type (Monogamy)				
Polygamy	1.796	1.366	0.415	87.583
Father's education (Tertiary)				
No formal education	1.286	0.792	0.767	17.067
Primary	0.836	0.822	0.46	11.556
Secondary	4.305**	1.516	3.792	145.853
No idea	3.358	1.993	0.578	128.591
Father's occupation (Civil servant)				
Farming	0.045	0.775	0.229	4.775
Trading	0.247	0.683	0.336	4.888
Employee of private organization	0.126	1.541	0.055	23.234
Others	-0.542	2.795	0.002	89.282
Mother's occupation (Trading)				
Farming	-0.124	1.048	0.113	6.882
Civil servant	0.837	1.484	0.126	42.367
Employee of private organization	-2.231	2.544	0.001	15.735
Others	1.774	2.564	0.039	97.133

Significance marker: * $p < 0.05$

Predictors of nonresponse on item 10Table 4.18 above shows the result of a multiple logistic regression model estimation for nonresponse on item 10 of the RSES as a function of age, sex, school, class, area of residence, family type, father's education, father's occupation and mother's occupation. We found that the effect of father's education is significant (β = 4.305, $p < 0.01$). Hence the data provide sufficient evidence that the missing data mechanism governing nonresponse on item 10 of the RSES is not MCAR. Consequently, in the ordinal logistic imputation model specified for item 10 of the RSES only father's education was used as predictors of the missing values on the item.

4.4 Modelling self-esteem before and after imputation

Table 4.19 shows the distribution of missingness on the RSES alongside some descriptive statistics.

Table 4.19: Summary statistics for the RSES items prior to imputation

Item	Number of responses Observed	Missing	Percent Missing	Median	Minimum	Maximum
R1	474	16	3.3	2	0	3
R2	482	8	1.6	1	0	3
R3	483	7	1.4	2	0	3
R4	475	15	3.1	2	0	3
R5	474	16	3.3	1	0	3
R6	470	20	4.1	2	0	3
R7	469	21	4.3	2	0	3
R8	481	9	1.8	1	0	3
R9	467	23	4.7	2	0	3
R10	465	25	5.1	2	0	3

There were 25 (5.1%) nonresponses on the tenth item, r10, of the scale, thus the item presents the largest percentage nonresponse across the items of the scale, while r3 has the lowest amount of missing data (7, 1.4%). Overall, in addition, more than half of the items had percentage nonresponse of at least 3.3. Furthermore, Table A.1 shows the missing data patterns for all the cases with missing data on the RSES. This shows that the pattern of missingness on this scale is arbitrary.

Table 4.20 shows that after imputation no nonresponse exists in the dataset. Furthermore, the summary statistics did not differ much from that in Table 4.19.

Table 4.20: Summary statistics for the RSES items after imputation

	Number of responses		Percent			
Item	Observed	Missing	Missing	Median	Minimum	Maximum
R1	490	0	0	1	0	3
R2	490	0	0	1	0	3
R3	490	0	0	2	0	3
R4	490	0	0	2	0	3
R5	490	0	0	1	0	3
R6	490	0	0	2	0	3
R7	490	0	0	2	0	3
R8	490	0	0	1	0	3
R9	490	0	0	1	0	3
R10	490	0	0	2	0	3

Table 4.21: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Sex (Female)					
Male					
		4.486***	1.114	2.296	6.676
	1	3.460	0.540	2.398	4.522
	2	4.060	0.583	2.915	5.206
	3	4.388	0.632	3.147	5.630
	4	4.444	0.687	3.095	5.793
	5	4.227	0.748	2.758	5.696
		4.116***	0.826	2.959	5.273
Class (SSS 2)					
SSS 1					
		6.93**	1.217	4.537	9.323
	1	6.738	0.815	5.137	8.339
	2	6.977	0.888	5.232	8.722
	3	6.943	0.968	5.042	8.845
	4	6.637	1.053	4.568	8.707
	5	6.059	1.145	3.809	8.309
		6.671***	1.138	5.327	8.014
SSS 3					
		2.696	1.651	-0.55	5.942
	1	2.491	1.370	-0.201	5.184
	2	2.548	1.442	-0.286	5.382
	3	2.512	1.489	-0.414	5.438
	4	2.383	1.510	-0.585	5.350
	5	2.161	1.506	-0.798	5.119
		2.419	1.493	-0.515	5.354

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,...,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.1 Description of Table 4.21

Table 4.21 presents the effects of sex and class of adolescent students on their level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.21, it was observed that male respondents had a significantly higher self-esteem score before imputation (β = 4.116, $p < 0.001$) and after imputation (β = 4.486, $p < 0.001$) than their female counterpart. Examination of the raw and pooled standard errors of the regression estimate reveals about 26% relative reduction and hence, a more precise estimate with narrower 95% confidence interval.

Similarly, adolescent students in the SSS 1 class had significantly higher self-esteem score before imputation (β = 6.930, $p < 0.01$) and after imputation (β = 6.671, $p < 0.001$) than the SSS 2 students with a relative reduction in standard error of about 6% and hence, a more precise with narrower 95% confidence interval. Also, adolescent students in the SSS 3 class had higher self-esteem score before imputation (β = 2.696) and after imputation (β = 2.419) than the SSS 2 students, however, this result is found not to be significant at each imputation step.

Table 4.22: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Family status (Parents are together)					
Parents are divorced					
		-4.464*	2.084	-7.494	-1.434
	1	-3.079	1.403	-5.835	-0.322
	2	-4.006	1.365	-6.688	-1.323
	3	-4.544	1.320	-7.138	-1.950
	4	-4.695	1.268	-7.186	-2.204
	5	-4.457	1.208	-6.830	-2.084
		-4.156*	1.827	-6.576	-1.736
Parents are separated					
		-1.282	2.118	-5.447	2.883
	1	-2.198*	0.906	-3.979	-0.418
	2	-1.592	0.890	-3.341	0.156
	3	-1.006	0.924	-2.823	0.811
	4	-0.440	1.010	-2.425	1.545
	5	0.106	1.147	-2.148	2.361
		-1.026	1.972	-5.848	3.796
Single mother					
		2.300	2.256	-2.135	6.735
	1	1.536	1.485	-1.381	4.453
	2	1.817	1.686	-1.497	5.130
	3	1.852	1.902	-1.885	5.588
	4	1.641	2.131	-2.546	5.828
	5	1.184	2.374	-3.480	5.849
		2.419	1.493	-0.515	5.354

Significance marker: * p < 0.05

m: Regression model estimates using imputed dataset m = 1,...,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.2 Description of Table 4.22

Table 4.22 presents the effects of family status of adolescent students on their level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.22, it was observed that after imputation adolescent students whose parents are divorced had a significantly lower self-esteem score before imputation (β = -4.464, $p < 0.05$) and after imputation (β = -4.156, $p < 0.05$) than students whose parents are together. Examination of the raw and pooled standard errors of the regression estimate reveals about 12% relative reduction and hence, a more precise estimate and narrower 95% confidence interval.

Students whose parents are separated had a lower self-esteem score before imputation (β = -1.282) and after imputation (β = -1.026) when compared with students whose parents are together. Although, a relative reduction of about 7% is found in its standard error, this result is not significant. Also, students with a single mother had a higher self-esteem score before imputation (β = 2.300) and after imputation (β = 2.419) when compared with students whose parents are together. This result is also not significant. However, examination of the raw and pooled standard errors of each regression estimate reveals a relative reduction in standard error of about 34% and hence, a narrower 95% confidence interval.

Table 4.23: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Father's education (Tertiary)					
No formal education					
		1.677	3.726	-5.65	9.004
	1	-1.573	3.029	-7.526	4.380
	2	-2.007	3.200	-8.295	4.280
	3	-2.174	3.284	-8.628	4.280
	4	-2.072	3.283	-8.524	4.379
	5	-1.703	3.196	-7.982	4.576
		-1.906	3.277	-8.346	4.535
Primary					
		-0.651	2.93	-6.412	5.111
	1	0.642	2.638	-4.541	5.826
	2	1.121	2.474	-3.742	5.983
	3	1.440	2.321	-3.121	6.001
	4	1.600	2.177	-2.678	5.879
	5	1.602	2.044	-2.415	5.618
		1.281	2.531	-3.690	6.252
Secondary					
		-0.219	1.95	-4.052	3.614
	1	0.572	1.885	-3.132	4.277
	2	0.548	1.780	-2.950	4.047
	3	0.458	1.695	-2.872	3.788
	4	0.402	1.627	-2.796	3.600
	5	0.480	1.579	-2.623	3.582
		0.492	1.719	-2.885	3.87
No idea					
		3.534	2.111	-0.618	7.685
	1	1.318	1.343	-1.321	3.957
	2	0.132	1.384	-2.587	2.851
	3	0.748	1.495	-2.190	3.686
	4	0.166	1.677	-3.129	3.460
	5	0.386	1.929	-3.404	4.176
		0.550	1.565	-2.525	3.625

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,...,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.3 Description of Table 4.23

Table 4.23 presents the effects of father's education on student's level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.23, it was observed that after imputation student whose father had no formal education had a lower self-esteem score (β = -1.906) as opposed to a higher pre-imputation self-esteem score (β = 1.677) compared to students whose father had a tertiary education. Even though these results are not significant, examination of the raw and pooled standard errors of the regression estimate reveals about 12% relative reduction and hence, a precise estimate and narrower 95% confidence interval.

Similarly, adolescent students who had no idea about their fathers' highest level of education had higher self-esteem score before imputation (β = 3.534) and after imputation (β = 0.550) than students whose father had a tertiary education with a relative reduction in standard error of about 25% and hence, a narrower 95% confidence interval.

Although, the data failed to provide sufficient evidence that students whose parents had primary and secondary education had a higher self-esteem score before imputation and after imputation (β = 1.281 and β = 0.492) than students whose father had a tertiary education, we observe a relative reduction in standard error of about 14% and 12% respectively.

Table 4.24: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Mother's education (Tertiary)					
No formal education					
		1.612	3.498	-5.265	8.49
	1	-4.073	2.942	-9.855	1.709
	2	-4.178	3.162	-10.392	2.036
	3	-4.311	3.280	-10.756	2.134
	4	-4.472	3.295	-10.947	2.003
	5	-4.660	3.208	-10.964	1.643
		-4.339	3.243	-10.711	2.032
Primary					
		-2.772	2.201	-5.607	5.394
	1	-2.202	2.656	-7.422	3.018
	2	-2.814	2.455	-7.637	2.010
	3	-2.423	2.241	-6.826	1.981
	4	-2.028	2.015	-5.987	1.932
	5	-2.629	1.777	-6.121	0.863
		-2.419	2.349	-7.034	2.196
Secondary					
		-0.512	1.563	-3.583	2.559
	1	-0.438	1.331	-3.053	2.177
	2	-0.407	1.118	-2.604	1.790
	3	-0.417	0.942	-2.268	1.434
	4	-0.408	0.804	-1.987	1.171
	5	-0.370	0.702	-1.749	1.010
		-0.408	0.98	-2.334	1.518
No idea					
		3.852*	0.819	1.22	6.484
	1	4.048	0.576	2.917	5.179
	2	4.017	0.564	2.909	5.124
	3	3.906	0.604	2.718	5.093
	4	3.716	0.698	2.346	5.087
	5	3.448	0.843	1.791	5.105
		3.827*	0.731	1.683	5.971

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,...,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.4 Description of Table 4.24

Table 4.24 presents the effects of mother's education on student's level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.24, it was observed that after imputation student whose mother had no formal education had a lower self-esteem score (β = -4.339) as opposed to a higher pre-imputation self-esteem score (β = 1.612) compared to students whose mother had a tertiary education. Even though these results are not significant, examination of the raw and pooled standard errors of the regression estimate reveals about 7% relative reduction and hence, a precise estimate and narrower 95% confidence interval.

Similarly, the data failed to provide sufficient evidence that students whose mother had primary and secondary education had a lower self-esteem score before imputation (β = -2.772 and β = -0.512) and after imputation (β = -2.419 and β = -0.408) than students whose mother had a tertiary education, we observe a relative reduction in standard error of about 7% and 37% respectively.

Meanwhile, adolescent students who had no idea about their mothers' highest level of education had significantly higher self-esteem score before imputation (β = 3.852, $p < 0.05$) and after imputation (β = 3.827, $p < 0.05$) than students whose father had a tertiary education with a relative reduction in standard error of about 11% and hence, a more precise estimate and narrower 95% confidence interval.

4.4.4 Description of Table 4.24

Table 4.24 presents the effects of mother's education on student's level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.24, it was observed that after imputation student whose mother had no formal education had a lower self-esteem score (β = -4.339) as opposed to a higher pre-imputation self-esteem score (β = 1.612) compared to students whose mother had a tertiary education. Even though these results are not significant, examination of the raw and pooled standard errors of the regression estimate reveals about 7% relative reduction and hence, a precise estimate and narrower 95% confidence interval.

Similarly, the data failed to provide sufficient evidence that students whose mother had primary and secondary education had a lower self-esteem score before imputation (β = -2.772 and β = -0.512) and after imputation (β = -2.419 and β = -0.408) than students whose mother had a tertiary education, we observe a relative reduction in standard error of about 7% and 37% respectively.

Meanwhile, adolescent students who had no idea about their mothers' highest level of education had significantly higher self-esteem score before imputation (β = 3.852, $p < 0.05$) and after imputation (β = 3.827, $p < 0.05$) than students whose father had a tertiary education with a relative reduction in standard error of about 11% and hence, a more precise estimate and narrower 95% confidence interval.

Table 4.25: **A regression model for determinants of self-esteem before and after imputation**

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Father's occupation (Civil servant)					
Farming					
		0.566	2.046	-3.458	4.59
	1	-4.138	1.484	-7.053	-1.222
	2	-3.354	1.624	-6.546	-0.163
	3	-4.240	1.707	-7.594	-0.886
	4	-4.095	1.732	-7.497	-0.692
	5	-3.719	1.698	-7.056	-0.381
		-3.909*	1.811	-7.469	-0.349
Trading					
		-0.822	1.664	-4.095	2.451
	1	-3.940	1.290	-6.476	-1.405
	2	-3.746	1.221	-6.144	-1.347
	3	-3.463	1.172	-5.767	-1.159
	4	-3.092	1.146	-5.344	-0.841
	5	-2.634	1.140	-4.875	-0.393
		-3.375*	1.522	-6.365	-0.384
Employee of private organization					
		1.639	1.41	-1.134	4.412
	1	-0.913	1.196	-3.263	1.436
	2	-0.572	1.223	-2.976	1.832
	3	-0.437	1.263	-2.919	2.044
	4	-0.508	1.314	-3.090	2.074
	5	-0.784	1.377	-3.489	1.921
		-0.643	1.322	-3.24	1.954
Others					
		1.416	1.912	-2.344	5.176
	1	-3.006	1.268	-5.498	-0.513
	2	-3.259	1.400	-6.010	-0.508
	3	-3.282	1.589	-6.404	-0.160
	4	-3.077	1.835	-6.682	0.529
	5	-2.642	2.138	-6.843	1.560
		-3.053	1.726	-6.445	0.338

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,...,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.5 Description of Table 4.25

Table 4.25 presents the effects of father's occupation on student's level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

In Table 4.25, it was observed that after imputation student whose father is a farmer had a significantly lower self-esteem score (β = -3.909, $p < 0.05$) as opposed to a higher but insignificant pre-imputation self-esteem score (β = 0.566) compared to students whose father is a civil servant. Examination of the raw and pooled standard errors of the regression estimate reveals about 11% relative reduction and hence, a more precise estimate and narrower 95% confidence interval.

Similarly, student whose father is a trader had a significantly lower self-esteem score (β = -3.375, $p < 0.05$) after imputation as opposed to a lower but insignificant pre-imputation self-esteem score (β = -0.822) compared to students whose father is a civil servant. We also observe about 11% relative reduction in the raw and pooled standard errors of the regression estimates, and hence, a more precise estimate and narrower 95% confidence interval.

However, the data failed to provide sufficient evidence that students whose father is an employee of private organization and those whose father engages in other occupation had a higher self-esteem score (β = 1.639 and β = 1.416) before imputation and a lower self-esteem score (β = -0.643 and β = -3.053) after imputation than students whose father is a civil servant. We observe a relative reduction in standard error of about 6% and 10% respectively.

Table 4.26: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Mother's occupation (Trading)					
Farming					
		2.081	3.948	-5.682	9.844
	1	0.262	3.302	-6.227	6.752
	2	0.370	3.519	-6.545	7.286
	3	0.223	3.592	-6.836	7.282
	4	0.219	3.522	-6.701	7.139
	5	0.260	3.307	-6.238	6.759
		0.267	3.453	-6.518	7.053
Civil servant					
		2.87	1.939	-1.143	6.484
	1	-2.789	2.161	-7.035	1.458
	2	-3.101	1.856	-6.748	0.546
	3	-2.811	1.604	-5.963	0.342
	4	-2.818	1.406	-5.580	-0.055
	5	-2.522	1.261	-4.999	-0.045
		-2.808	1.708	-6.164	0.549
Employee of private organization					
		2.937	1.836	-0.671	6.545
	1	1.031	1.199	-1.325	3.387
	2	1.643	1.153	-0.622	3.908
	3	1.969	1.152	-0.295	4.233
	4	2.009	1.197	-0.343	4.361
	5	1.763	1.288	-0.768	4.293
		1.683	1.384	-1.037	4.402
Others					
		-2.772	2.201	-7.099	1.555
	1	-0.694	1.819	-4.268	2.880
	2	-1.032	1.902	-4.769	2.706
	3	-1.174	1.932	-4.971	2.623
	4	-1.121	1.909	-4.873	2.630
	5	-0.874	1.833	-4.476	2.729
		-0.979	1.925	-4.761	2.803

Significance marker: * $p < 0.05$

m. Regression model estimates using imputed dataset m = 1,...,5

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

Table 4.26: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Mother's occupation (Trading)					
Farming					
		2.081	3.948	-5.682	9.844
	1	0.262	3.302	-6.227	6.752
	2	0.370	3.519	-6.545	7.286
	3	0.223	3.592	-6.836	7.282
	4	0.219	3.522	-6.701	7.139
	5	0.260	3.307	-6.238	6.759
		0.267	3.453	-6.518	7.053
Civil servant					
		2.87	1.939	-1.143	6.484
	1	-2.789	2.161	-7.035	1.458
	2	-3.101	1.856	-6.748	0.546
	3	-2.811	1.604	-5.963	0.342
	4	-2.818	1.406	-5.580	-0.055
	5	-2.522	1.261	-4.999	-0.045
		-2.808	1.708	-6.164	0.549
Employee of private organization					
		2.937	1.836	-0.671	6.545
	1	1.031	1.199	-1.325	3.387
	2	1.643	1.153	-0.622	3.908
	3	1.969	1.152	-0.295	4.233
	4	2.009	1.197	-0.343	4.361
	5	1.763	1.288	-0.768	4.293
		1.683	1.384	-1.037	4.402
Others					
		-2.772	2.201	-7.099	1.555
	1	-0.694	1.819	-4.268	2.880
	2	-1.032	1.902	-4.769	2.706
	3	-1.174	1.932	-4.971	2.623
	4	-1.121	1.909	-4.873	2.630
	5	-0.874	1.833	-4.476	2.729
		-0.979	1.925	-4.761	2.803

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,...,5

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates

Table 4.26: A regression model for determinants of self-esteem before and after imputation

Variable	m	Coefficient	Standard Error	95% C.I. Lower	95% C.I. Upper
Mother's occupation	**(Trading)**				
Farming					
		2.081	3.948	-5.682	9.844
	1	0.262	3.302	-6.227	6.752
	2	0.370	3.519	-6.545	7.286
	3	0.223	3.592	-6.836	7.282
	4	0.219	3.522	-6.701	7.139
	5	0.260	3.307	-6.238	6.759
		0.267	3.453	-6.518	7.053
Civil servant					
		2.87	1.939	-1.143	6.484
	1	-2.789	2.161	-7.035	1.458
	2	-3.101	1.856	-6.748	0.546
	3	-2.811	1.604	-5.963	0.342
	4	-2.818	1.406	-5.580	-0.055
	5	-2.522	1.261	-4.999	-0.045
		-2.808	1.708	-6.164	0.549
Employee of private organization					
		2.937	1.836	-0.671	6.545
	1	1.031	1.199	-1.325	3.387
	2	1.643	1.153	-0.622	3.908
	3	1.969	1.152	-0.295	4.233
	4	2.009	1.197	-0.343	4.361
	5	1.763	1.288	-0.768	4.293
		1.683	1.384	-1.037	4.402
Others					
		-2.772	2.201	-7.099	1.555
	1	-0.694	1.819	-4.268	2.880
	2	-1.032	1.902	-4.769	2.706
	3	-1.174	1.932	-4.971	2.623
	4	-1.121	1.909	-4.873	2.630
	5	-0.874	1.833	-4.476	2.729
		-0.979	1.925	-4.761	2.803

Significance marker: * $p < 0.05$

m: Regression model estimates using imputed dataset m = 1,…,5.

The first unlabelled row presents estimates before imputation.

The last unlabelled row presents after-imputation pooled estimates.

4.4.6 Description of Table 4.26

Table 4.26 presents the effects of mother's occupation on student's level of self-esteem before imputation and at each imputation step. The table also shows the pooled effects, standard errors and 95% confidence interval.

Although we observe no significant effects in Table 4.26, student whose mother is a farmer had a higher self-esteem score before imputation (β = 2.081) as opposed to a higher but reduced pre-imputation self-esteem score (β = 0.267) compared to students whose mother is a trader. Examination of the raw and pooled standard errors of the regression reveals 12.5% relative reduction and hence, a more precise estimate and narrower 95% confidence interval.

Student whose mother is a civil servant had a higher self-esteem score (β = 2.808) after imputation as opposed to a lower pre-imputation self-esteem score (β = -2.87) compared to students whose mother is a trader. We also observe about 12% relative reduction in the raw and pooled standard errors of the regression estimates, and hence, a more precise estimate and narrower 95% confidence interval.

Also, students whose mother is an employee of private organization and those whose mother engages in other occupation had a higher self-esteem score before and after imputation (β = 2.937, β = 1.683) and a lower self-esteem score before and after imputation (β = -2.772, β = -0.979), respectively. We observe a relative reduction in standard error of about 25% and 13% respectively.

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATION

5.1 Discussion

Health researchers who carry out surveys, particularly those who collect data from self-reported scales will almost certainly be faced with the problem of missing data frequently. In this study, we have presented a missing data analysis for the APF dataset that was collected so as to model psychosocial disorder among adolescents in some selected secondary schools in Ekiti State. While it was recognized that imputing items on Strength and Difficulty Questionnaire and Centre for Epidemiological Studies Depression Scale for Children would have constituted a more complete study, we have however limited this analysis to the RSES. Hence, the report presented in this study is based on imputing the RSES only.

We found that significant estimates of the multiple linear regression parameters were given with relatively low standard errors. For example, male respondents had a significantly higher self-esteem score estimated with relatively high precision, while adolescent students in the SSS 1 class also scored significantly high on the self-esteem scale. Also, the estimated coefficient for students whose parents were divorced was significantly lower score and with low standard error.

Moreover, after accounting for missing data mechanism and employing imputation models that fill in missing observations with plausible values from the conditional distribution of the missing variable in concern, estimates that were not significant became significant. This is true of father's occupation and mother's education, so that students whose parents are farmers and traders had significantly lower score on RSES, while students who had no idea of their mother's occupation had significantly higher self-esteem score.

In this regard, MI almost always provides estimates that are more representative of the population parameter than popular missing data techniques implemented in most statistical software do, in particular, listwise deletion.

Apart from low statistical power and inflated standard errors, researchers who criticize listwise deletion (e.g. Lee and Carlin, 2010; Schafer and Graham, 2002) often based their arguments on its production of biased point estimates due to the assumption that set of observations with missing values do not differ from set of observations with valid values. Since, for example, students who had no idea of their father's education were more likely to miss item 1 of RSES, that assumption is suspect. similar conclusions were also made for items 2 through 10 of RSES. With this bias in mind and given listwise deletion approach to missing values, effects of socio-demographic variables on self-esteem were either underestimated or overestimated with low precision. This agrees with the submission of Leeaw et al (2003) and Jeffrey (2003).

5.2 Conclusion

This study presents the APF multiple imputation models and its implementation using FCS. After showing that missing values in the APF dataset do not follow the Missing Completely at Random assumptions, we also justify the choice of MI approach in the context of several other missing data methods.

Also, we summarize the resulting parameter estimates of a linear regression model describing the effect of some socio-demographic variables and self-esteem from both dataset with missing values and the imputed datasets obtained from the mi STATA command. We observe that properly accounting for missing values with multiple imputations provides a useful and more reliable approach than listwise deletion method.

5.3 Recommendations

Consequent upon the observation that multiple imputation provides a more precise parameter estimates, we recommend MI and hope to see researchers properly accounting for missing values using MI technique in their analysis and methods in future health studies so as to achieve substantial inference.

REFERENCES

1. Allison, P.D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research, 28, 301-309.

2. Allison, P.D. (2001). Missing Data: Quantitative Applications in the Social Sciences. Sage publications, Inc. Thousand Oaks, CA.

3. Allison, P.D. (2006). Multiple Imputation of Categorical Variables Under the Multivariate Normal Model. This paper was presented at the Annual Meeting of the American Sociological Association, Montreal, August 2006. An earlier version was presented at the Annual Meeting of SUGI (SAS Users Group International), Philadelphia, PA, April 2005.

4. Allison, P.D. (2012). Handling missing data by maximum likelihood. SAS Global Forum. Statistical Horizons, Haverford, PA, USA.

5. Chen, H.Y., Little, R.J.A. (1999). A Test of Missing Completely at Random for Generalized Estimating Equations with Missing Data. Biometrika, 94, 896-908.

6. Collins, L.M., Schafer, J.L., Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6, 330-351.

7. Crawford, T.N., Cohen P., Johnson, J.G., Sneed, J.R., and Brook, J.S. (2004). The Course and Psychosocial Correlates of Personality Disorder Symptoms in Adolescence: Eriksons Developmental Theory Revisited Journal of Youth and Adolescence 33(5), 373-387

8. de Leeuw, E.D., Hox, J., Husman, M. (2008). Prevention and treatment of item nonresponse. Journal of Official Statistics, 19, 153{176.

9. Dempster, A.P., Laird, N.M. Rubin, D.B. (1977). Maximum Likelihood from In complete Data via the EM Algorithm. Journal of Royal Statistical Society, 39, 1-38.

10. Fichman, M. and Cummings, J.M. (2003). Multiple Imputation for Missing Data: Making the Most of What you Know. Tepper School of Business Paper 113,

11. Gelman, A. and Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge.

12. Graham, J.W. (2009). Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 171(60), 549-576.

13. Graham J.W., Hofer, S.M., Donaldson, S.I., MacKinnon, D.P., and Schafer, J.L. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windles, and S. West (Eds.) "The science of prevention: methodological advances from alcohol and substance abuse research." American Psychological Association, Washington, D.C. 325-366.

14. Graham, J., A. Olchowski, and T. Gilreath (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory Prevention Science 8 206-213

15. Graham J.W. and Schafer, J.L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.) Statistical Strategies for Small Sample Research Thousand Oaks, CA: Sage

16. Graham J.W. and Hofer, S.M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, and J. Baumert, (Eds.) "Modelling Longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples." Erlbaum, Hillsdale, NJ. 201-218.

17. Graham J.W., Cumsille, P.E., and Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka and W. F. Velicer (Eds). Research Methods in Psychology. New York: John Wiley & Sons Handbook of Psychology. 2, 87-114.

18. Horton, N.J. and Lipsitz, S.R. (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. Journal of the American Statistical Association. 55, 244-254.

19. Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation mimeo, Board of Governors of the Federal Reserve System, in 1991 Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association.

20. Lavori P., R. Dawson and D. Shera. (1995). A multiple imputation strategy for clinical trial with truncation of patient data. Statistics in Medicine, 14, 1913-1925

21. Lee, K.J. and Carlin, J.B. (2010). Multiple Imputation for Missing Data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology, 171(5), 624-632.

22. Little, R.J.A. and Rubin, D.B. (1987). Statistical analysis with missing data. JohnWiley & Sons, New York.

23. Little, R.J.A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of American Statistical Association, 83(5), 1198{1202

24. Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. 2nd ed Wiley, New York

25. Park, T., Lee, S-Y. (1997). A test of missing completely at random for longitudinal data with missing observations. Statistics in Medicine, 16, 1859-1871.

26. Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-45.

27. Royston, P., and White, I.R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. Journal of Statistical Software, 454, 1-20.

28. Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc., 81, 366-374.

29. Rubin, D.B. (1976). Inference and missing data (with discussion). Biometrika, 63, 581-592.

30. Rubin, D.B. (1977). Formalizing subjective notion about the effect of nonrespondents in sample surveys. Journal of American Statistical Association, 72, 538-543.

31. Rubin, D.B. (1987). Multiple imputation for nonresponse in survey. John Wiley, New York.

32. Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). Journal of American Statistical Association, 91, 473-489.

33. Rubin, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC Statistica Neerlandica, 57(1), 3-18.

34. Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall, London. Schafer, J.L. (1999). Multiple imputation: a primer. Statistical Method in Medical Research, 8, 3-15.

35. Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our view of the state of the art. Psychological Methods, 7(2), 147-177.

36. Schafer, J. L., and M. K. Olsen (). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research 33, 545-571

37. Shrive F.M., Stuart, H, Quan H, William A Ghali W.A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Medical Research Methodology, 657,

38. Stata Corporation. (2011). Stata Statistical Software: Release 12 Software Stata Corporation, College Station, Texas.

39. Stuart, E.A, Azur, M., Frangais, C., Leaf, P. (2009). Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative. American Journal of Epidemiology, 169, 1133-1139.

40. van Buuren, S. (2012). Flexible Imputation of Missing Data. CRC, Chapman & Hall.

41. van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Method in Medical Research, 16(3). 219-242.

42. van Buuren S., Boshuizen, H.C., Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statist. Med., 18, 681-694.

43. van Buuren, S., and K.C. Oudshoorn (2000). Multivariate Imputation by Chained Equations. MICE V1.0 User's manual. TNO Prevention and Health, http://web. inter.nl.net/users/S.van.Buuren/mi/docs/Manual.pdf (07.10.2005).

44. van Buuren S., Brands, J.P.L., Groothuis-Oudshoorn, C.G.M, et al. (2006). Fully conditional specification in multivariate imputation. J Stat. Comput. Simul., 76(12), 1049-1064.

45. van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45, 3

46. White, I.R., Royston, P., Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30, 377-399

47. Yu L.M., Burton A, Rivero-Arias O. (2007). Evaluation of software for multiple imputation of semi-continuous data. Statistical Method in Medical Research, 16(3), 243-258.

APPENDIX

Table A.1: Missing data pattern on the RSES

s/n	*n*	R3	R2	R8	R4	R5	R6	R1	R7	R9	R10	*m*
1	398											398
2	7							o				405
3	1				o			o				409
4	3				o							401
5	1				o				o			409
6	7								o			405
7	1								o	o		417
8	11									o		409
9	1			o						o		413
10	3			o								401
11	17										o	415
12	2									o	o	428
13	1								o	o	o	437
14	1				o				o	o		422
15	1							o	o			413
16	1						o		o			415
17	9						o					407
18	2						o	o				416
19	1					o	o					414
20	6					o						404
21	1				o	o						408
22	1					o			o			412
23	1	o				o						406
24	1	o										399
25	1	o								o		411
26	3		o									401
27	1				o						o	419
28	1				o		o	o	o			431
29	1		o		o	o	o		o			432
30	1			o	o	o	o		o	o	o	467
31	3	o	o	o	o	o	o	o	o	o	o	490
32	1	o	o	o	o	o	o	o	o	o		465

Sample Questionnaire

MODELLING PREDICTORS OF ADOLESCENT PSYCHOSOCIAL FUNCTIONING IN SECONDARY SCHOOLS IN IKERE-EKITI LOCAL GOVERNMENT AREA, EKITI STATE, NIGERIA

SECTION A: BACKGROUND INFORMATION ***(Tick the code as appropriate)***

1. What is your sex ☐ Male ☐ Female
2. What is your current age *(fill the exact height)* ______________
3. What is your height *(fill the exact height)* ______________
4. What is your height *(fill the exact height)* ______________
5. What is the name of your school ______________
6. What class are you ______________
7. What is your religion ☐ Christianity ☐ Islam ☐ Others *(please specify)* ______________
8. Area of residence ☐ Rural area ☐ Urban area
9. Ethnicity ☐ Yoruba ☐ Hausa/Fulani ☐ Igbo ☐ Others *(please specify)* ______________
10. Family type ☐ Monogamy ☐ Polygamy
11. Family status ☐ Parents are together ☐ Parents are divorced ☐ Parents are separated ☐ Single mother
12. Father's highest level of education ☐ No formal education ☐ Primary ☐ Secondary ☐ Tertiary ☐ No idea
13. Father's occupation ☐ Farming ☐ Trading ☐ Farming ☐ Trading ☐ Others *(please specify)* ______________
14. Mother's highest level of education ☐ No formal education ☐ Primary ☐ Secondary ☐ Tertiary ☐ No idea
15. Mother's occupation ☐ Farming ☐ Trading ☐ Civil servant ☐ Employee of private organisation ☐ Others *(please specify)* ______________
16. Do you have friends of the opposite sex ☐ Yes ☐ No
17. Have you felt disappointed / jilted by a friend who is an opposite sex ☐ Yes ☐ No

18a. Which of the following have you ever done with an opposite sex *(You can tick more than one)*
☐ Kissing/Caressing ☐ Sex ☐ Petting

18b. Which of the following have you ever done with a person of the same sex *(You can tick more than one)*
☐ Kissing/Caressing ☐ Sex ☐ Petting

SECTION B: PSYCHOSOCIAL OUTCOMES

A. ROSENBERG SELF ESTEEM SCALE (RSES)

Below is a list of statements dealing with your general feelings about yourself. Please indicate how strongly you agree or disagree with each statement.

		Strongly Agree	Agree	Disagree	Strongly Disagree
1	On the whole, I am satisfied with myself	3	2	1	0
2	At times I think I am no good at all	0	1	2	3
3	I feel that I have a number of good qualities.	3	2	1	0
4	I am able to do things as well as most other people	3	2	1	0
5	I feel I do not have much to be proud of	0	1	2	3
6	I certainly feel useless at times	0	1	2	3
7	I feel that I'm a person of worth, at least on an equal plane with others	3	2	1	0
8	I wish I could have more respect for myself	0	1	2	3
9	All in all, I am inclined to feel that I am a failure	0	1	2	3
10	I take a positive attitude toward myself	3	2	1	0

Note: The filling of this questionnaire is voluntary

B. STRENGTH AND DIFFICULTY QUESTIONNAIRE (SELF RATED) *(cycle the code as appropriate)*

- For each item, please mark the box for **Not True, Somewhat True or Certainly True**.
- It would help us if you answered all items as best you can even if you are not absolutely certain or the item seems daft! Please give your answers on the basis of how things have been for you over the **last six months.**

Code	Questions	Not True	Somewhat True	Certainly True
Se1	I try to be nice to other people. I care about their feelings	0	1	2
Sc1	I am restless, I cannot stay still for long	0	1	2
Sa1	I get a lot of headaches, stomach-aches or sickness	0	1	2
Se2	I usually share with others (food, games, pens etc.)	0	1	2
Sb1	I get very angry and often lose my temper	0	1	2
Sd1	I am usually on my own. I generally play alone or keep to myself	0	1	2
Sb2	I usually do as I am told*	2	1	0
Sa2	I worry a lot	0	1	2
Se3	I am helpful if someone is hurt, upset or feeling ill	0	1	2
Sc2	I am constantly fidgeting or squirming	0	1	2
Sd2	I have one good friend or more*	2	1	0
Sb3	I fight a lot. I can make other people do what I want	0	1	2
Sa3	I am often unhappy, down-hearted or tearful	0	1	2
Sd3	Other people of my age generally like me*	2	1	0
Sc3	I am easily distracted, I find it difficult to concentrate	0	1	2
Sa4	I am nervous in new situations. I easily lose confidence	0	1	2
Se4	I am kind to younger children	0	1	2
Sb4	I am often accused of lying or cheating	0	1	2

Sd4	Other children or young people pick on me or bully me	0	1	2
Se5	I often volunteer to help others (parents, teachers, children)	0	1	2
Sc4	I think before I do things*	2	1	0
Sb5	I take things that are not mine from home, school or elsewhere	0	1	2
Sd5	I get on better with adults than with people my own age	0	1	2
Sa5	I have many fears, I am easily scared	0	1	2
Sc5	I finish the work I'm doing. My attention is good*	0	1	2

C. CENTER FOR EPIDEMIOLOGICAL STUDIES DEPRESSION SCALE FOR CHILDREN (CES-DC)

Below is a list of the ways you might have felt or acted.
Please check how *much* you have felt this way **during the *past week***.

Code	Questions	Not At All	A Little	Some	A Lot
1	I was bothered by things that usually don't bother me				
2	I did not feel like eating, I wasn't very hungry				
3	I wasn't able to feel happy, even when my family or friends tried to help me feel better				
4	I felt like I was just as good as other kids				
5	I felt like I couldn't pay attention to what I was doing				
6	I felt down and unhappy				
7	I felt like I was too tired to do things				
8	I felt like something good was going to happen				
9	I felt like things I did before didn't work out right				
10	I felt scared				
11	I didn't sleep as well as I usually sleep				
12	I was happy				
13	I was more quiet than usual				
14	I felt lonely, like I didn't have any friends				
15	I felt like kids I know were not friendly or that they didn't want to be with me				
16	I had a good time				
17	I felt like crying				
18	I felt sad				
19	I felt people didn't like me				
20	It was hard to get started doing things				