USING ITEM RESPONSE THEORY TO UNDERSTAND ITEM-NONRESPONSE (MISSING DATA) IN GHANAIAN SURVEYS

BY

IDDRISU WAHAB ABDUL

(B.A. Mathematics and Statistics)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF PHILOSOPHY DEGREE IN APPLIED MATHEMATICS

April, 2012

DECLARATION

This thesis was written in the Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi from August, 2011 to June 2012 in partial fulfillment of the requirements for the award of Master of Philosophy degree in Applied Mathematics under the supervision of Nana Kena Frempong. All materials used in this work have been duly acknowledged in the References. I therefore declare that no part of this thesis has been presented in this or any other University.

IDDRISU WAHAB ABDUL 16/04/2012
(PG5070610) Signature Date

Certified by:
NANA KENA FREMPONG 16/04/2012
(Supervisor) Signature Date

Certified by:
MR K.F. DARKWAH KFDarkwah 23/5/2012
(Head of Department) Signature Date

DEDICATION

This piece of work is dedicated to my family and all those who do good things in society.

ACKNOWLEDGEMENT

I first of all express my sincere praise to the Almighty Allah for His Divine providence which has enabled me to come this far. I thank my supervisor Nana Kena Frempong for his able supervision, encouragement, useful comments and sense of good judgement which has led to the successful completion of this thesis and may Allah guide him, grant him success in all his endeavours.

I am highly indebted to my parents and siblings for their support, encouragement and prayer in all my endeavours especially in carrying out this thesis. Special thanks go to my senior brother, Mr Alhassan Iddrisu for his immense contribution in my education. My heartfelt gratitude goes to my colleagues for their support.

And to all whose names and contributions have not been respectively mentioned and acknowledged, I say am forever grateful, may Allah richly bless you all.

ABSTRACT

After reviewing the theoretical and empirical literature on Item Response Theory (IRT) and Item-Nonresponse, this study investigates three issues: Firstly, to identify the most appropriate IRT model for understanding item-nonresponse. Secondly, to find out the reason behind ‘don’t know’ responses and missing data; whether respondents don’t really know, don’t care, or don’t want to tell. Finally, to find out the characteristics of nonrespondents in Ghanaian surveys. Secondary analyses were done on questionnaire data collected in the 5th wave of the world values survey. All items were dichotomously scored. 0 was assigned to missing or ‘don’t know’ responses, and 1 was assigned to answered items. The data was analysed based on four item response theory models namely, the constrained Rasch model, the unconstrained Rasch model, the two parameter logistic model, and the three parameter logistic model. These models were explored to determine the most appropriate model for the data. The unconstrained Rasch model appeared as the best model for understanding item-nonresponse. It was found that, giving a ‘don’t know’ answer to an item is because of the item’s difficulty, which means that respondents don’t really know the answer to the item. The results also revealed that, item-nonresponse can be predicted by some item and respondent characteristics. In a typical application, politics and income related questions recorded the highest item-nonresponse rates. ~~Female~~ respondents and respondents with no formal education also recorded very high item-nonresponse rates.

Key Words: Item-Nonresponse, Item Response Theory (IRT), Respondent Characteristics, Unconstrained Rasch Model, World Values Survey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

1-PLM	One Parameter Logistic Model
2-PLM	Two Parameter Logistic Model
3-PLM	Three Parameter Logistic Model
AIC	Akaike Information Criterion
BHPS	British Household Panel Survey
BIC	Bayesian Information Criterion
DIF	Differential Item Functioning
DRQ	Democracy Related Question
EVS	European Values Study
GSOEP	German Socio-Economic Panel
HILDA	Household, Income and Labour Dynamics
ICC	Item Characteristic Curve
IRF	Item Response Function
IRQ	Income Related Question
IRT	Item Response Theory
ISR	Institute of Social Research
LLRA	Linear Logistic Regression with relaxed Assumptions
LRQ	Life Related Question

LRT	Likelihood Ratio Test
LTM	Latent Trait Modelling
MAR	Missing At Random
MCAR	Missing Completely At Random
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
NAEP	National Assessment of Educational Progress
PISA	Programme for International Student Assessment
PRQ	Politics Related Question
TIMSS	Trend in International Mathematics and Science Study
VRQ	Value Related Question
WVS	World Values Survey

CHAPTER 1

INTRODUCTION

1.0 Background

Survey research has been widely used in public opinion research in Ghana. While scholars in comparative studies are excited about data richness, they are also worried about data quality. Among all such concerns, item-nonresponse (or 'don't know' answers) has caught scholars' special attention. Comparative studies are seriously plagued by this problem: it is obviously not appropriate to ignore it and discard all 'don't know' answers, but what should we do when confronting a large amount of missing data while still purporting to draw valid inferences from the available data? Proper understanding of the missing data mechanism can be a huge step in dealing with item-nonresponse.

By definition, item-nonresponse is the failure to obtain information for a question within an interview or questionnaire (de Leeuw, 2001). It results in missing values to a particular question. However, it does not mean that item-nonresponse fails to contain any information. Although the information is not self-evident, it can be revealed by further analysis of the missingness. Rubin's framework (Rubin, 1976) differentiates among three kinds of missing data according to the underlying missing data mechanism: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR).

Data are called missing completely at random (MCAR) if the missingness of a response to a question is unrelated to its unknown value and also unrelated to the values of responses to other questions. For example, an interviewer during an interview, or a respondent in a mail questionnaire, may accidentally overlook a question. In the case of missing completely at random, the missing values are a random sample of all values and not related to any observed or unobserved variable. Thus results of data analyses will not be biased, because there are no systematic differences between respondents and nonrespondents and problems that arise are mainly a matter of reduced statistical power. It should be noted that the standard solutions in many statistical packages, those of listwise and pairwise deletion both assume that the data are missing completely at random (MCAR); a very strong and often unrealistic assumption (de Leeuw et al., 2003).

When the missingness is related to the observed data but not to the unknown value of the missing response to the question itself, it is said that data are missing at random (MAR). For example, an elderly respondent has a difficulty remembering an event because of a deficient memory. The resulting missingness is related to age, not to the event itself. When the data are missing at random, the missingness is a random process conditional on the observed data. In other words, the missing values are a random sample of all values within classes defined by observed values. In the example of the elderly respondent, the data are a random sample within subgroups formed by age. If the data are missing at random (MAR) and if the proper statistical model is used, the missing is said to be *ignorable* with respect to a particular type of inference. For example, in the case of the elderly respondent, the variable related to the missingness (age) is measured and available for inclusion in the proper analysis (de Leeuw et al., 2003).

Finally, when the missingness is related to the unknown answer to the question itself, the data are missing not at random (MNAR). For example, a respondent perceives her real answer as socially undesirable (e.g., drinking a lot) and evades responding by providing a 'don't know' or 'no answer'. If the data are not missing at random serious bias may occur. In that case, the missingness is said to be nonignorable and no simple solution for treating the missing data exists. A model for the missingness must be postulated and included in the analysis to prevent bias (de Leeuw et al., 2003).

The first intimations of item response theory (IRT) can perhaps be seen in (Thurstone's, 1925) paper, entitled 'A Method of Scaling Psychological and Educational Tests.' In it, he offers a solution to a problem of great interest at the time-how best to place the items of the (Binet and Simon, 1905) test of children's mental development on an age-graded scale. Using data from a large sample of London school children, he plots the proportions of children in successive age cross sections succeeding on successive Binet tasks. In an accompanying plot, he shows the age grading of the tasks resulting from his scaling procedure. These plots have many features suggestive of modern IRT (Bock, 1997).

Over the past three decades, since the publishing of Lord and Novick's Statistical Theories of Mental Test Scores in 1968 and Fischer's Einfuhrung in die Theorie psychologischer Tests in 1974, item response theory (IRT) has developed rapidly. This is demonstrated in the Handbook of Modern Item Response Theory (Van der Linden and Hambleton, 1997) with chapters on a wide range of topics in IRT. The study of individual responses to behavioural stimuli has clearly evolved into a major discipline of psychometric theory (Boomsma et al., 2000).

1.1 Study Area Profile

Modern Ghana was created from the British Gold Coast Colony, established in 1874, and the UK-administered Trusteeship Territory of Togoland, incorporated in 1956 following a plebiscite. Agitation for independence grew strongly after the

Second World War. From the early 1950s, self-government was introduced with elections in 1951, 1954 and 1956 to the legislative assembly. Kwame Nkrumah's party, the CPP, won all 3 elections and led the country to independence, as Ghana, in March 1957. Ghana was the first sub-Saharan country in colonial Africa to gain its independence. Nkrumah was the first Prime Minister, and in 1960 became President with the change of Ghana's status to a Republic within the Commonwealth.

The Republic of Ghana has a total border of 2,093 kilometres (1,300 miles), including 548 kilometres (341 miles) with Burkina Faso to the north, 688 kilometres (428 miles) with Côte d'Ivoire to the west, and 877 kilometres (545 miles) with Togo to the east. It has a coastline on the Gulf of Guinea, part of the Atlantic Ocean, measuring 539 kilometres (335 miles). It has an area of 239,540 square kilometres (92,486 square miles), making it about the size of the state of Oregon. Water occupies 8,520 square kilometres (3,290 square miles) of the country, primarily Lake Volta. The capital of Accra is located along the south-eastern coast. Ghana has a tropical climate, warm and comparatively dry along the southeast coast, hot and humid in the southwest, and hot and dry in the north. Its terrain is mostly low plains with a plateau in the south-central area. Its highest point is Mount Afadjato, which rises to 880 meters (2,~~887 feet)~~. Lake Volta, its largest lake, is the world's largest artificial lake. Ghana has 10 regions: the Northern, Upper West, Upper East, Volta, Ashanti, Western, Eastern, Central, Brong-Ahafo, and Greater Accra.

Ghana's economy has always been dependent on a number of key exports principally gold and cocoa, although more recently it has developed a burgeoning oil and gas sector. Gold dominates the mining sector and contributes about 30% of foreign exchange earnings. Ghana also produces diamonds, manganese and bauxite.

Ghana is also a major cocoa producer. In 2010, with an output of about 1,000,000 tonnes, it has retained its position as the second largest producer in the world, a position it had not held for three decades before 2003. Cocoa production is subject to volatile prices and the vagaries of the weather. This makes the economy vulnerable.

1.2 Problem Statement

Item-nonresponse has drawn scholars' interest from the very beginning of the development of scientific public opinion polls. A school of scholars believe that a simple 'don't know' response may reflect several possible mind states: no idea, no opinion, and refusal. Findings from empirical studies are consistent with their belief and confirm that the distinctions are evident in examination of the linkage of underlying attitude and opinion expression (Bogart, 1967).

However, item-nonresponse in Ghanaian surveys is often speculated as a problem but rarely researched. Even without sampling problems, this should be a great concern for those who might wish to do poll survey on political issues in Ghana. There may be more than just a few students in comparative studies who hold this opinion, and they usually consider two strategies that may be employed by the respondents to avoid the trouble: one is to tell a lie, another is to keep silence. Yet, relevant evidence from ~~empirical~~ research are rare. It is also suspected that the prevalence of MAR problems may also be caused by insufficient education resources in Ghana. Ordinary Ghanaian people may lack cognitive abilities to form concrete opinions to certain survey questions due to their low education level. These suspicions give rise to the controversial role of Ghanaian surveys in public opinion research.

Therefore, more and in-depth research with regards to measuring the abilities, attributes, interests, knowledge or proficiencies of respondents to address the problem of item-nonresponse is in great need.

This research shall be guided by the following separate but interrelated questions;

1. Which Item Response Theory (IRT) model is the most appropriate for understanding item-nonresponse?
2. Don't know answers and missing data; do respondents don't really know, don't care, or don't want to tell?
3. What are the characteristics of nonrespondents in Ghanaian surveys?

1.3 Objectives

1. To identify the most appropriate IRT model for understanding item-nonresponse.
2. To find out the reason behind don't know responses and missing data; whether respondents don't really know, don't care, or don't want to tell.
3. To find out the characteristics of nonrespondents in Ghanaian surveys.

1.4 Methodology

A summary of the research method is presented here. This includes a brief description of the problem model, the data set to be used for empirical analysis, and a description of the software to be used.

1.4.1 Problem Model

Item response theory (IRT) is an umbrella of statistical models that attempts to measure the abilities, attitudes, interests, knowledge or proficiencies of respondents as well as specific psychometric characteristics of test items.

Hambleton (2000) stated that item response theory places the ability of the respondent and the difficulty of the item on the same measurement scale so direct comparisons between respondents' abilities and items are possible. The ability or proficiency of the respondent is labelled theta (θ). The test item characteristics are described by the difficulty (b), discrimination (a), and pseudo-chance (c) parameters. Not all IRT models utilize all item parameters, and there is a continuing debate about the appropriateness of these parameters. For example, the Rasch model uses only the difficulty parameter and ignores the discrimination and pseudo-chance parameters completely. Because the Rasch model uses only the difficulty parameter as the only item parameter, it is called a 1-PL model (one parameter logistic).

Another model proposed by (Birnbaum, 1968) uses both the difficulty and discrimination item parameters and is called a 2-PL model. The model that utilizes all three parameters was proposed by (Lord, 1980) and it is called the 3-PL model.

1.4.2 Data

When choosing survey data for this project, we had several considerations in mind. First, we wished that the data are from a survey that is designed and conducted by serious and professional survey practitioners. Since this study focuses on the validity of survey responses, we would use data that is least contaminated by technical problems in the process of survey research operation. Second, we wished the data are from a study that is well-known and influential.

The World Values Survey (WVS) data stand out as an ideal source for this research purpose. The WVS originated from European Values Study (EVS) and extended to countries outside Europe in 1981, which constituted the first wave of the WVS. The surveys aim to be longitudinal as well as cross-cultural. The second wave of the WVS (1990) was conducted 10 years after the first and embraces 42 countries. The interval between the waves was shortened to 5 years for the third (1995), fourth (2000), and fifth (2005) waves, which includes 52 and 64 countries separately. In total, the WVS covers 81 societies containing about 85 percent of the world's population.

The WVS was conducted by the Institute of Social Research at the University of Michigan (ISR) in collaboration with leading survey research organizations in each country. Tracy Hammond & Mari Harris from Markinor Thinking are the principal investigators in this project. The survey covers a variety of research topics, such as socio-cultural, moral, religious, and political values and attitudes. It employs detailed questionnaires and face-to-face interview techniques in methodology. Representative samples were drawn from each country and the number varies from 1000 to 3500 per country. We selected the data from the fifth wave of the WVS for this project.

1.4.3 Software

An R Package for Latent Trait Modelling and Item Response Theory Analyses(R Development Core Team, 2010), shall be used. This package has been developed for the analysis of multivariate dichotomous and polytomous data using latent variable models, under the Item Response Theory approach. For dichotomous data the Rasch, the Two-Parameter Logistic, and Birnbaum's Three-Parameter models have been implemented, whereas for polytomous data Semejima's Graded Response model is available. Parameter estimates are obtained under marginal maximum likelihood using the Gauss-Hermite quadrature rule.

1.5 Justification

A survey is a potentially powerful assessment, monitoring, and evaluation tool available to researchers to effectively and efficiently assess stakeholder (employees, management, students, clients, etc.) perceptions and attitudes for a variety of purposes. According to (Kraut, 1996), survey purposes include the pinpointing of organisational concerns, observing long-term trends, monitoring program impact, providing input for future decisions, adding a communication channel, performing organisational behaviour research, assisting in organisational change and improvement, and providing symbolic communication.

Because the value of a survey in addressing these purposes is dependent on individuals participating in the research effort, item-nonresponse is a great concern among researchers and others, who conduct, analyse, interpret, and act on survey results. Low item response rates can cause smaller data samples. Smaller data samples decrease statistical power, increase the size of confidence intervals around sample statistics, and may limit the types of statistical techniques that can effectively be applied to the collected data.

Research on issues pertaining to item-nonresponse has a long history; the research continues today as evidenced by a quick look at the large number of recent publications and books referenced in this study. F. Stanton (1939) wrote one of the first empirical pieces on the topic in the *Journal of Applied Psychology* titled, Notes on the Validity of Mail Questionnaire Returns.

However, very little scholarly work is done on item-nonresponse in Ghana. The realization of how badly item-nonresponse bias can affect the interpretation of survey results can be a key driver towards the development of a variety of techniques to enhance item response rates. Therefore, it is crucial to have a better understanding of item-nonresponse in Ghanaian surveys.

1.6 Scope and Limitations

The study will basically be structured within the confines of the project subject matter. Typical of most scientific social researches, it is envisaged that, the proposed study will suffer the following set-backs;

- Resources inadequacy: finance and logistics since the entire project will be funded solely by me.
- Time constrain will also be a challenge.
- The unavailability of relevant literature on the study subject matter in the school library could be a threat to the project quality.

1.7 Thesis Organization

The study shall be organized in five related chapters. The opening chapter, chapter 1 will highlight the central theme of the research. Chapter 2 will be on the review of relevant literature, these will include summary of abstracts of contribution of variables with regard to the model used. Chapter 3 will consider some methodological approaches in arriving at reliable empirical findings (formulations and methods of solution). Chapter 4 will consider data analysis and discussion of results. Chapter 5 covers the conclusions and recommendations, the summary and suggested areas of possible future investigation.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

The purpose of this research is to understand how item and person characteristics affect item-nonresponse using data from a Ghanaian opinion survey. To provide the necessary background, this literature review will cover theories on item-nonresponse, types of item-nonresponse, factors affecting item-nonresponse, and previous work on item-nonresponse. Previous applications of IRT models will also be covered.

2.1 What is Item-Nonresponse?

We will start by first looking at unit nonresponse, this occurs when data for a whole unit of analysis are not available for statistical analysis. It may be because units could not be contacted or refused to cooperate, or because the questionnaire of a unit that did cooperate got lost during data editing or analysis (Lessler and Kalsbeck, 1992). Unit nonresponse falls outside the scope of this research. Unit nonresponse is often called first-level nonresponse.

Item-nonresponse is referred to as second-level nonresponse. Here, the unit has participated, but data on particular items are unavailable. The term unavailable is used on a purpose. Whether or not an answer is counted as missing depends on the goal of the study. For instance, 'don't know' can be viewed as a meaningful response to a question about voting intentions in an election poll. For other questions (e.g., income), 'don't know' has no informational value and is counted as missing. Therefore, an item is missing if the researcher interprets it as such, and decides that some kind of treatment is required. Thus, item-nonresponse is defined as the failure to obtain information for a question in an interview or questionnaire, so data are missing (see also Groves, 1989). In this research, 'don't know' responses are considered missing.

2.2 Types of Item-Nonresponse

A basic distinction concerning missing data is that data are *missing completely at random, missing at random, or not missing at random.* This distinction is important because it refers to quite different processes, requiring specific strategies in subsequent data analysis (Little and Rubin, 1987 cited from de Leeuw et al., 2003).

In order to define the three kinds of missing data, (Rubin, 1976) distinguishes between the observed data $\boldsymbol{Y_{obs}}$ and the missing data $\boldsymbol{Y_{miss}}$. Together, these constitute the complete data matrix $\boldsymbol{Y} = (\boldsymbol{Y_{obs}}, \boldsymbol{Y_{miss}})$. Here, we adapted this notation to the latent variable framework. $\boldsymbol{Y}$ here is the complete data matrix that consists of the observed item responses Y_{obs} and the omitted responses $\boldsymbol{Y_{miss}}$ of the k items Y_1 to Y_k, indexed by i. The values of a latent variable ξ can also be considered to be missing data. In large-scale assessments, it is a common practice to include covariates such as gender and social economic status in the analyses. All covariates constitute the matrix $\boldsymbol{Z}$. *MCAR* denotes the case where the distribution of missing data is independent of $\boldsymbol{Y_{obs}}$, $\boldsymbol{Y_{miss}}$ and a given multivariate covariate $\boldsymbol{Z}$. That is, $P(\boldsymbol{D} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{miss}}, \boldsymbol{Z}, \xi) = P(\boldsymbol{D})$. The matrix $\boldsymbol{D}$ is an indicator matrix consisting of the indicator variables d_i that marks the occurrence of the values of Y_i,

$$d_i = \begin{cases} 1, & if\ Y_i\ is\ observed \\ 0, if\ Y_i\ is\ not\ observed \end{cases} \qquad \ldots\ldots\ldots\ldots\ldots (2.1)$$

MAR holds if the distribution of the missing data is only dependent on the observed variables $\boldsymbol{Y_{obs}}$ and $\boldsymbol{Z}$ but not dependent on the unobserved values of missing data $\boldsymbol{Y_{miss}}$ and ξ. This is equivalent to the expression $P(\boldsymbol{D} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{miss}}, \boldsymbol{Z}, \xi) = P(\boldsymbol{D} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Z})$.

The third type, called MNAR, can be written as $P(\boldsymbol{D} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{miss}}, \boldsymbol{Z}, \xi) \neq P(\boldsymbol{D} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Z})$. It is the opposite of MAR. That means the conditional distribution of the missing data given $\boldsymbol{Y_{obs}}$ and $\boldsymbol{Z}$ depends on the unobserved data $\boldsymbol{Y_{miss}}$ and possibly ξ.

2.3 Theories on Item-Nonresponse

Three theoretical frameworks that seem to dominate the literature on individual response behaviour are the cognitive model, the rational choice framework, and the satisficing theory. These models can be applied to the explanation of a variety of aspects of respondent behaviour, one of them being item-nonresponse.

2.3.1 Cognitive Model Theory

After early models developed in cognitive psychology (Lachman et al., 1979) focused exclusively on individual thought processes, the cognitive model of respondent behaviour extended this framework by also taking social aspects of the survey situation into consideration. This conceptualization of response behaviour separates several stages in the process of answering a question at which a respondent has to master distinct tasks: After having heard or read a question, the respondent must interpret the information. The issue addressed by the interviewer has to be recognized; the respondent must understand the content of the question and mentally connect it to familiar concepts. At this stage of the interview, problems may arise depending on the technicality of the question or the common understanding and definitions used by respondent and interviewer. Here face-to-face interviews bear the advantage of allowing for feedback between the interview partners regarding the intended content of the question (Riphahn and Serfling, 2002).

At the second cognitive stage after the content of a question has been communicated the respondent has to gather the required information. Here the familiarity of the issue matters: Being asked about ones' age imposes less of a cognitive effort than answering, say, about the amount of interest earned on building society accounts during the past calendar year. The more complex the issue the more is required of the respondent's knowledge, cognitive ability, and willingness to recollect the relevant information (Riphahn and Serfling, 2002).

After the respondent successfully gathered the required information, the questionnaire may impose a certain format on the answer to which the information has to be translated. Examples are categorical answers, or subjective intensity statements. Even when being aware of the correct answer, its format may require additional 'translation efforts' by the respondent. The final cognitive stage then consists of a possible adjustment of the information with the purpose of attaining objectives such as self-representation or social desirability of the answer. Only after the respondent refiltered the intended answer through these additional 'mental screens' is the answer provided (Riphahn and Serfling, 2002).

2.3.2 Rational Choice Theory

Rational choice theory was early on advanced through (Esser, 1984), and is still heavily relied upon. Esser (1984) views respondent behaviour as the outcome of an evaluation process: In any given situation the respondent evaluates behavioural alternatives based on their expected consequences and chooses the best maximizing subjectively expected utility. Within this framework the process of responding to a question consists of three stages: First a situation and the particular question at hand must be understood, then behavioural alternatives are to be evaluated, and finally the preferred behaviour is chosen (Riphahn and Serfling, 2002).

The rational choice approach predominates the analysis of survey response: Referring back to Dillman (1978), Hill and Willis (2001) state that an individual answers a survey if the act of participation is expected to bring rewards that exceed the cost of participation. The *rewards* to participation may include pecuniary and non-pecuniary rewards such as social acknowledgment by being positively regarded and appreciated by others. The *costs* considered by (Hill and Willis, 2001) consist of the length of time it takes to respond, but also of the emotional experience of going through potentially embarrassing, painful or cognitively difficult interviews. These cost considerations allow one to expect different response behaviours based on the perceived privacy of questions (Riphahn and Serfling, 2002).

Schrapler (2001) in his investigation of item-nonresponse similarly discusses costs and benefits connected to the decision of whether to provide the desired information or not: Relevant *benefits* of responding consist of supporting a potentially appreciated cause (e.g. scientific value, public interest) and of avoiding the negative effects of a refusal such as breaking social norms generated by the interview situation or violating courtesy towards the interviewer. The impact of the latter two factors distinguishes face-to-face from telephone interviews. Key *costs* of answering a survey consist of the potential negative consequence of providing private information (e.g. from tax authorities or through data abuse and breach of privacy) as well as of the necessary effort to recall the facts desired by the questionnaire. The cost of providing income information in particular is hypothesized by Schrapler to show a U-shaped pattern: If income is low, individuals may be too embarrassed to tell the truth, if it is very high, individuals may be reluctant to reveal this information to a stranger.

A separate aspect connected to the rational choice framework of participation decision is the relevance of trust in the interview situation. If a respondent is distrustful of the interviewer or his motivation, he is less ready to expend effort to provide information or even to reveal information at all. Hill and Willis (2001) first refer to Dillman (1978) who emphasized the relevance of trust, and then describe the steps taken in the Health and Retirement Survey to render the interviewer more trustworthy. Schrapler (2001) discusses the importance of a process he terms 'confidence building' which involves reducing the social distance between the interviewer and the respondent over time to increase trust and to reduce the fear of negative consequences of sensitive statements.

2.3.3 Satisficing Theory

Krosnick's satisficing theory (1991), posits three factors that affect the process of answering questions. The first being the motivation of the respondent to perform the task, the second is the difficulty of the task, and the last is the respondent's cognitive ability to perform the task. This theory elaborates on a standard question-answering process-model developed by (Tourangeau and Rasinski, 1988). This identifies two processes that explain differences in response quality, namely optimizing and satisficing. Optimizing means that the respondent goes through all the cognitive steps needed to answer a survey question, whiles satisficing means a respondent gives responses that appear reasonable or acceptable without going through all the steps involved in the question-answering process. Satisficing has a relationship with the motivation of the respondent, the difficulties of the task, and the cognitive abilities of the respondent. Difficult questions and low cognitive abilities may lead respondents to provide a satisfactory response instead of an optimal one. Optimizing strategy often provides more reliable responses than a satisficing strategy (Borgers and Hox, 2001).

In sum, the main theoretical approaches focus on the cognitive process of providing information, motivation, as well as on cost-benefit calculations of the individual. The latter may well be influenced by measures of confidence building on the part of the survey administration and the level of trust established in the personal relationship between interviewer and respondent.

2.4 Previous Research on Item-Nonresponse

Evidence on the determinants of item-nonresponse is relatively scarce and most relevant studies are rather recent. Among the earlier studies Lillard et al. (1986), motivated by rising nonresponse rates, investigate the distribution of item-nonresponse in the United States' Current Population Survey. The authors distinguish between individuals who refused to respond only to the income question and those who did not answer a number of questions. While the latter group represents the lower part of the income distribution, the probability of an exclusive income nonresponse increases with income. Earnings were more likely to be reported in personal compared to telephone interviews (Riphahn and Serfling, 2002).

Similarly, Sousa-Poza and Henneberger (2000) focus on the income question in Swiss telephone interviews and attempt to explain the determinants of item-nonresponse. They find that nonresponse probabilities are significantly lower for respondents with high education and higher among the self-employed and home owners. The authors investigate the relevance of matching the characteristics of interviewers and respondents and show that similarity in age increases the response probability, that education differences do not affect item-nonresponse, and that male interviewers are more successful in eliciting income information than female. Since the share of response behaviour that can be explained by characteristics of the interview partners or by their matching is limited, they conclude that the observed wage data are not biased by the large (wage) item-nonresponse encountered in telephone interviews.

This finding is basically confirmed by Biewen (2001), who compares alternative methods to address item-nonresponse based on GSOEP data, and shows that nonresponse of income is highest in the tails of the income distribution. However, he points out that nonresponse is only weakly associated with personal characteristics and mainly driven by unobservables. An earlier study of (Zweimuller, 1992) using Austrian data is similar to (Biewen, 2001). Estimating wage equations for women, (Zweimuller, 2001) however concludes that selection due to survey nonresponse is of larger importance than the usually addressed selectivity bias.

Schrapler's (2001) contribution focuses on the longitudinal development of item nonresponse for gross earnings and measures of individual concerns. Similar to (Lillard et al., 1986) and to (Biewen, 2001) he finds evidence that those in low social positions (also females and the young) tend to withhold income information. Again and just as in the data of (Sousa-Poza and Henneberger, 2000), respondents seem to be much more uncooperative in front of females than males. Schrapler concludes that the relationship between respondent and interviewer is of key significance, as with increasing trust the item-nonresponse rate falls off over time.

With a focus on improving survey administration, Hill and Willis (2001) evaluated the effectiveness of paying respondents for their time and of enhancing the psychic value of participation for unit nonresponse: Reassigning the same interviewer to a given respondent has powerful effects on the propensity to respond, as trust can be established between the interview partners over time. The authors emphasize the importance of the respondent's engagement and cognitive ease with the interview as predictors of survey participation (Riphahn and Serfling, 2002).

Loosveldt et al. (1999) used Belgian data to compare the correlation of item-nonresponse behaviour for a variety of questions with subsequent unit nonresponse. The differences in item-nonresponse across questions are interpreted as a consequence of the questions' varying cognitive difficulty and the sensitivity of the relevant issues. The authors find a positive correlation between item and subsequent unit nonresponse (Riphahn and Serfling, 2002).

In his investigation of item-nonresponse as a precursor of panel attrition, and whether or not the results in models of item nonresponse behaviour are affected by a selectivity bias due to panel attrition using data from the German socio-economic panel (GSOEP), Serfling (2004) found evidence for negative correlation of item-nonresponse and unit nonresponse. He also found by means of a bivariate probit model that attrition bias is item-specific.

Durrant (2005) reviewed and discussed the advantages and disadvantages of several imputation methods for handling item-nonresponse in the social sciences and found that, certain forms of fractional and multiple hot deck methods perform well with regards to bias and efficiency of a point estimator and robustness against model misspecifications but standard parametric imputation methods were not found adequate.

Borgers and Hox (2001) investigated into the effect of item and person characteristics on item nonresponse for written questionnaires used with school children. Secondary analyses were done on questionnaire data collected in five distinct studies. To analyse the data, logistic multilevel analysis was used with the items at the lowest and the children at the highest level. Item-nonresponse turned out to be relatively rare. They realized that item-nonresponse can be predicted by some of the item and person characteristics. They also found that there are interactions between item and person characteristics, especially with the number of years of education, which was used as a proxy indicator for cognitive skill. They finally concluded that young children do not perform as well as children with more years of education, by producing more item-nonresponse.

After reviewing the theoretical and empirical literature on item-nonresponse, Riphahn and Serfling (2002) in their study on item-nonresponse on income and wealth questions, focused on three issues: First, they found a significant heterogeneity in item-nonresponse across financial questions. Nonresponse rates varied widely, which showed up in significant question specific fixed effects in models of item-nonresponse outcomes. Second, they realized that there was not much to be gained for the informational value of surveys from matching interviewers and respondents based on their key characteristics, once age and gender effects are controlled for. Their third key result with respect to 'don't know' answer options in the questionnaire was that, the characteristics of don't know respondents are neither very close to those providing informative answers nor to those refusing to respond.

In their article on prevention and treatment of item-nonresponse, de Leeuw et al. (2003) stated that, understanding, prevention, and imputations are chained together in handling missing data successfully.

Using data on annual individual labour income from three representative panel datasets (GSOEP, BHPS, HILDA), Frick and Grabka (2007) investigated into the selectivity of item-nonresponse, and the impact of imputation as a prominent post-survey means to cope with this type of measurement error on prototypical analyses (earnings inequality, mobility and wage regressions) in a cross national setting. They found that all three panels made use of longitudinal information in their respective imputation procedures, however, there were marked differences in the implementation. Firstly, although the probability of item-nonresponse is quantitatively similar across countries, their empirical investigation identified cross-country differences with respect to the factors driving item-nonresponse. Secondly, longitudinal analyses yielded a positive correlation of item-nonresponse on labour income data over time and provided evidence of item-nonresponse being a predictor of subsequent unit nonresponse, thus supporting the 'cooperation continuum' hypothesis in all three panels. Regression results for wage equations based on observed (complete case analysis) vs. all cases and controlling for imputation status, indicated that individuals with imputed incomes, ceteris paribus, earn significantly above average in GSOEP and HILDA, while this relationship is negative using BHPS data.

The Gallup organisation has developed a series of 13 questions to gauge employee satisfaction with the workplace. The questions consist of an overall satisfaction item, followed by twelve agree-disagree statements regarding various aspects of the workplace. The twelve questions are referred to as the 'Q-12' and are administered both in interviewer-assisted formats and self-administered mode for a paper administration, there have been several different designs used, depending on the preferences of the client. Through these various designs, it was observed that item-nonresponse to the overall satisfaction question was often higher than the next question, which is the first of the 12 agree-disagree questions. In some cases, nonresponse to the satisfaction item had exceeded 25%, while nonresponse to the Q-12 agree/disagree items that followed was seldom more than 1 or 2%. Dillman et al. (2000) reviewed the various self-administered formats that had been used to ask this series of 13 questions in the prescribed order and their resulting item-nonresponse rates. Overall, they found that three of the eleven visual formats appeared to bring the item-nonresponse rates down to the point of exhibiting little if any differences between the satisfaction and agree/disagree items. They included; an extensive word simplification, changing the visual organisation of the first question in the sequence, and grouping of a second item with the satisfaction items in conjunction with the omission of the don't know categories for the first items.

Groves and Peytcheva (2006) designed fifty-nine methodological studies to estimate the magnitude of nonresponse bias in statistics of interest. These studies use a variety of designs: sampling frames with rich variables, data from administrative records matched to sample case, use of screening interview data to describe nonrespondents to main interviews, follow-up of nonrespondents to initial phases of field effort, and measures of behaviour intentions to respond to a survey. This permits exploration of what circumstances produce a relationship between nonresponse rates and nonresponse bias and what do not. The predictors are design features of the surveys, characteristics of the sample, and attributes of the survey statistics computed in the surveys.

Devey et al. (2006) advanced a theoretical model that explains organizations' nonresponse to surveys as a predictable aspect of organizational behaviour and structure. They argued that survey researchers must take into account the authority, capacity, and motive to respond of both the organizations sampled and the designated respondent within the organization. Their analysis identified a series of organizational sources of nonresponse that have clear consequences for final sample bias. These include resource independence from the environment, subsidiary status, information dispersal in large establishments, and lack of staff dedicated to information processing. They provided suggestions for future organizational survey design and for analysis strategies to cope with sample selection bias.

Immerwahr et al. (2008) evaluated techniques used in a telephone survey to 'convert' initial item-nonresponse in two questions: self-reported age and annual household income. The use of follow-up questions to initial 'don't know/not sure' and 'refused' responses substantially reduced overall item-nonresponse for both age and income. Initial nonresponse to self-reported age was 4.5%. A follow-up question to convert initial item-nonresponse brought the proportion of cases without age data down to 0.3%, a 93.6% reduction. Use of a 'critical value' follow-up reduced item-nonresponse to household income from an initial level of 13.7% to 8.7%, a reduction of 33.7%. These techniques can reduce item-nonresponse during data collection, reducing the potential for survey error.

Sijtsma and Van der Ark (2003) discussed a statistical test for investigating whether or not the pattern of missing scores in a respondent-by-item data matrix is random. Since this is an asymptotic test, they investigated whether it was useful in small but realistic sample sizes. They then discussed two known simple imputation methods, person mean and two-way imputation, and they proposed two new imputation methods, response-function and mean response-function imputation. These methods are based on few assumptions about the data structure. An empirical data example with simulated missing item scores showed that the new method

(Response-Function) was superior to the methods (Person Mean), (Two-Way), and (Mean Response-Function) in recovering from incomplete data several statistical properties of the original complete data. Methods Two-Way and Response-Function are useful both when item score missingness is ignorable and nonignorable.

2.5 History of Item Response Theory (Brief)

The first intimations of IRT can perhaps be seen in Thurstone's (1925) paper, entitled 'A Method of Scaling Psychological and Educational Tests.' In it, he offers a solution to a problem of great interest at the time-how best to place the items of the Binet and Simon (1905) test of children's mental development on an age-graded scale. Using data from a large sample of London school children, he plots the proportions of children in successive age cross sections succeeding on successive Binet tasks. In an accompanying plot, he shows the age grading of the tasks resulting from his scaling procedure. These plots have many features suggestive of modern IRT (Bock, 1997).

Thurstone dropped his work in measurement to pursue the development of multiple factor analysis, but his colleagues and students continued to refine the theoretical bases of IRT (Steinberg and Thissen, in draft). Richardson (1936) and Ferguson (1943) introduced the normal ogive model as a means to display the proportions correct for individual items as a function of normalized scores. Lawley (1943) extended the statistical analysis of the properties of the normal ogive curve and described maximum-likelihood estimation procedures for the item parameters and linear approximations to those estimates. Lord (1952) introduced the idea of a latent trait or ability and differentiated this construct from observed test score.

Lazarsfeld (1950) described the unobserved variable as accounting for the observed interrelationships among the item responses.

Considered a milestone in psychometrics, Embretson and Reise (2000), Lord and Novick's (1968) textbook entitled Statistical Theories of Mental Test Scores provides a rigorous and unified statistical treatment of classical test theory. The remaining half of the book, written by Allen Birnbaum, provides an equally solid description of the IRT models. Bock and several student collaborators at the University of Chicago, including David Thissen, Eiji Muraki, Richard Gibbons, and Robert Mislevy developed effective estimation methods and computer programs such as Bilog, Multilog, Parscale, and Testfact. Along with Aitken (Bock and Aitken, 1981), Bock developed the algorithm of marginal maximum likelihood method to estimate the item parameters that are used in many of these IRT programs.

In a separate line of development of IRT models, Rasch (1960) discussed the need for creating statistical models that maintain the property of specific objectivity, the idea that people and item parameters be estimated separately but comparable on a similar metric. Rasch inspired Fischer (1968) to extend the applicability of the Rasch models into psychological measurement and Ben Wright to teach these methods and help to inspire other students to the development of the Rasch models. These students, including David Andrich, Geoffrey Masters, Graham Douglas, and Mark Wilson, helped to push the methodology into education and behavioural medicine (Wright, 1997).

2.6 Previous Applications of IRT

Over the past three decades, since the publishing of Lord and Novick's Statistical Theories of Mental Test Scores in 1968 and Fischer's Einfuhrung in die

Theorie psychologischer Tests in 1974, item response theory (IRT) has developed rapidly. This is demonstrated in the Handbook of Modern Item Response Theory (Van der Linden and Hambleton, 1997) with chapters on a wide range of topics in IRT. The study of individual responses to behavioural stimuli has clearly evolved into a major discipline of psychometric theory (Boomsma et al., 2000).

Vendramini and Silva (2001) conducted a study to analyse the validity evidences based on intern structure of the Attitudes toward Statistics scale by Item Response Theory. The Rasch polytomous models were shown as favourites, for their characteristics, for the application in the field of personality, attitudes and interests measurement. They analysed the responses of 693 undergraduate that had already studied the Statistical course in the higher education. The principal results of their study indicated validity evidences in the intern structure of the scale. They believed that the study of the attitudes in relation to the Statistics can contribute for the improvement of the teaching and of the learning of this course and of another that need statistical concepts.

Hedeker et al. (2001) applied item response theory models for intensive longitudinal data. They described IRT models and some key developments in the IRT literature, and illustrated their application for intensive longitudinal data. In particular, they related IRT models to mixed models and indicated how software for the latter could be used to estimate the IRT model parameters. Using Ecological Momentary Assessment data from a study of adolescent smoking, they described how IRT models can be used to address key questions in smoking research. They analysed whether or not a smoking report had been made in each of 35 time periods, defined by the crossing of seven days and five time intervals within each day. The IRT model was able to identify which time periods were most associated with smoking reports; and also which time periods were the most informative, in the sense of discriminating underlying levels of smoking behaviour. As indicated, weekend and evening hours yielded the most frequent smoking reports, however morning and, to some extent, mid-week reports were most discriminating in separating smoking levels.

Kirisci et al. (2001) applied item response theory to quantify substance use disorder severity. Their investigation had two main goals: (1) to determine whether binary substance use disorder diagnoses are indicators of a unidimensional trait indexing severity of disorder; and, (2) demonstrate the predictive, concurrent and construct validity of the substance use disorder severity scale. Boys and their biological parents were administered structured diagnostic interviews to diagnose substance use disorder. Item response theory (IRT) was applied to determine whether the diagnoses are indicators of a unidimensional trait. The score on this scale was correlated with substance use behaviour, violence, treatment history, risky sex, and social adjustment. They found that substance use disorder diagnoses are indicators of a unidimensional latent trait. Maternal and paternal substance use disorder severity predicted son's substance use disorder severity at age 19. The score on the substance use disorder severity scale correlated with drug use frequency, number of different drugs used in lifetime, treatment seeking, illegal behaviour, social maladjustment, and risky sex. They concluded that substance use disorder can be quantified on an interval scale indexing severity of disorder.

In their article on applications of item response theory to identify and account for suspect rater data, Zoanetti et al. (2001) described a plausible values imputation approach for deriving population scores on several language proficiency domains. The approach harnessed a multi-dimensional item response analysis combining student responses, rater judgements and student background variables. The target population was grade one and grade two primary school students enrolled in the Hong Kong schooling system. The raters were local teachers of English employed within the sampled schools. The primary objective of their research was to impute plausible values for data where no data was provided or where rater data was deemed suspect. By necessity, a secondary objective of this study was to establish rules for justly excluding particular data on the basis of questionable validity. Surveys such as TIMSS, PISA and NAEP have used such 'plausible value' methodologies to account for incomplete test designs and person nonresponse. The point of difference between their study and other similar studies was the use of item response theory (in particular plausible values) to replace and quantify the impact of potentially invalid rater judgements in a large-scale educational survey.

Klauer and Sydow (2001) proposed and applied a new model-based method for analysing learning processes in so-called short-term learning tests. Learning was measured by means of a two-dimensional item response theory model that incorporates two latent factors: ability level and learning ability. The model allows one to assess the impact of both variables and their correlation for given data sets in a manner that implicitly corrects for statistical artefacts that arise in conventional analyses. The model was applied to four short term learning tests administered to 434 children aged five to six years. Model tests and tests for empirical validity of the model parameters succeeded in establishing construct validity for one of these tests.

Verhelst and Verstralen (2001) presented an IRT model for multiple ratings. They assumed that the quality of a student performance has a stochastic relationship with the latent variable of interest. They showed that the ratings of several raters are not conditionally independent given the latent variable. The model gives a full account of this dependence. Several relationships with other models appear to exist. The proposed model is a special case of a nonlinear multilevel model with three levels, but it can also be seen as a linear logistic model with relaxed assumptions (LLRA). Moreover, a linearized version of the model turns out to be a special case of a generalizability model with two crossed measurement facets (items and raters) with a single first-order interaction term (persons and items). Using this linearized model, they showed how the estimated standard errors of the parameters are affected if the dependence between the ratings is ignored.

Jansen and Glas (2001) considered a latent trait model developed by Rasch for the response time on a set of pure speed tests, which is based on the assumption that the test response times are approximately gamma distributed with known index parameters and scale parameters depending on subject ability and test difficulty parameters. In their study, the principle of Lagrange multiplier tests was used to evaluate differential test functioning and subgroup invariance of the test parameters. Two numerical illustrations were given.

Confronted with incomplete data due to nonresponse, a researcher may want to impute missing values to estimate latent properties of respondents. In their study, Huisman and Molenaar (2001) presented the results of a simulation study. Investigating the performance of several imputation techniques, it appeared that some imputation procedures are based on item response theory (IRT) models, which can also be used to estimate latent abilities directly from the incomplete data by using incomplete testing designs when data are missing by design. This strategy had some serious disadvantages in the case of item-nonresponse, because nonresponse was assumed to be ignorable and computational problems arose in scales with many items. In a second simulation study, the performance of some imputation techniques was compared to the incomplete design strategy, in the case of item-nonresponse. The latter strategy resulted in slightly better ability estimates, but imputation is almost as good, especially when it is based on IRT models.

A nonparametric item response theory (IRT) model for the circumplex is introduced by Mokken et al. (2001), based on previous nonparametric IRT model for cumulative scaling, and nonparametric IRT model for unfolding. Some examples of circumplex representations in the social sciences were given. Model fit was based first on an extension of Loevinger's coefficient of homogeneity using quadruples as elementary units of analysis. Diagnostics for the probabilistic circumplex were suggested. Assignment of (ordinal) scale values was based on the notion of item steps, as developed earlier. The model they presented was based on dichotomous pick any/k data. Suggestions for extension to rank m=k data and to polytomous data were discussed.

Following in the nonparametric item response theory tradition, DIMTEST is an asymptotically justified nonparametric procedure that provides a test of hypothesis of unidimensionality of a test data set. Stout et al. (2001) introduced a new bias correction method for the DIMTEST procedure based on the statistical principle of resampling. A simulation study showed this new version of DIMTEST has a Type I error rate close to the nominal rate of α= 0.05 in most cases and very high power to detect multidimensionality in a variety of realistic multidimensional models. The result with this new bias correction method was seen as an improved DIMTEST procedure with much wider applicability and good statistical performance.

Item response theory (IRT) models are used to describe answering behaviour on tests and examinations. Although items may fit an IRT model, some persons may produce misfitting item score patterns, for example, as a result of cheating or lack of motivation. Several statistics have been proposed to detect deviant item score patterns. Misfitting item score patterns may be related to group characteristics such as gender or race. Investigating misfitting item score patterns across different groups is strongly related to differential item functioning (DIF). Meijer and Krimpen-Stoop (2001) studied the usefulness of person fit to compare item score patterns for different groups. In particular, the effect of misspecification of a model due to DIF on person fit was explored. Empirical data of a math test were analyzed with respect to misfitting item score patterns and DIF for men and women and blacks and whites. Results indicated that there were small differences between subgroups with respect to the number of misfitting item score patterns. Also, the influence of DIF on the fit of a score pattern was small for both gender and ethnic groups. The results imply that person-fit analysis is not very sensitive to model misspecification on the item level.

Post, Duijn, and Baarsen (2001) investigated how a motivated choice can be made for the analysis of an item set using either a cumulative item response model with monotone tracelines, modeling dominance relations, or a unimodal item response model with single-peaked tracelines, modeling proximity relations. The focus was on item sets consisting of positively and negatively formulated items (with respect to the latent trait to be measured), where the common practice is to reverse one type of item. The differences between the cumulative and unimodal model were studied theoretically, in terms of item location and item order, and empirically, in a reanalysis of a sample of De Jong Gierveld loneliness scale data. Item locations, and, in the case of the unimodal model, also subject locations were shown to be important determinants of the differences. For the loneliness scale data the analysis with the unimodal model is preferred over the cumulative model. An outline of a recommended strategy for an IRT analysis of scaling data was given.

Several large-scale educational surveys use item response theory (IRT) models to summarize complex cognitive responses and relate them to educational and demographic variables. The IRT models are often multidimensional with pre-specified traits, and maximum likelihood estimates (MLEs) are found using an EM algorithm, which is typically very slow to converge. Rubin and Thomas (2001) showed that the slow convergence is due primarily to missing information about the correlations between latent traits relative to the information that would be present if these traits were observed ('complete data'). They showed how to accelerate convergence using parameter extended EM (PX-EM), a recent modification of the EM algorithm. The PX-EM modification is simple to implement for the IRT survey model and reduces the number of required iterations by approximately one-half without adding any appreciable computation to each EM-iteration.

In item response theory, dominance relations are modelled by cumulative models and proximity relations by unfolding models. Usually cumulative models are used for the measurement of latent abilities and unfolding models for the measurement of latent attitudes and preferences. The distinction between both types of measurement models is best represented by the shape of the item characteristic curve which is monotone increasing in the cumulative model and single-peaked in the unfolding model. A boundary case is a situation in which some items are monotone decreasing and others are monotone increasing. In a study by Klinkenberg, (2001), an extension of the one-parameter logistic model was proposed for that situation. It was shown that the parameters of the proposed model could be estimated by a simple data transformation. The model was illustrated using an empirical data set.

Item response theory makes use of what are sometimes referred to as item response functions. Such a function is actually a probability density function for the response of an individual to a test item, given the values of certain parameters, classified as item parameters and person parameters (or abilities). In testing, there is typically a calibration phase, in which item parameters are estimated and abilities are ignored. This is followed by an application phase, in which abilities are estimated while conditioning on the estimated values of the item parameters. A Bayesian alternative to treating the item parameters as known quantities involves replacing item response functions with another class of probability density functions, referred to as expected response functions. The latter take the uncertainty regarding item parameters into account for purposes of estimating abilities. Lewis (2001) provided a formal description of expected response functions and briefly illustrated their application to the Rasch model for binary item responses.

Item response theory (IRT) is a powerful tool for the detection of differential item functioning (DIF). Glas (2001) showed that the class of IRT models with manifest predictors is a comprehensive framework for the detection of DIF. These models also supported the investigation of the causes of DIF. In principle, the responses to every item in a test can be subject to DIF, and traditional IRT-based detection methods require one or more estimation runs for every single item. Therefore, he proposed an alternative procedure that can be performed using only a single estimate of the item parameters. This procedure is based on the Lagrange multiplier test or the equivalent Rao efficient score test. Glas (2001) also generalized the procedure in various directions, the most important one being the possibility of conditioning on general covariates. A small simulation study was presented to give an impression of the power of the test. In an example using real data it is shown how the method can be applied to the identification of main and interaction effects in DIF.

CHAPTER 3

METHODOLOGY

3.0 Introduction

In this chapter, a new model for understanding item-nonresponse is introduced. We begin by formally describing the data set and the coding scheme used. To provide the necessary background, this chapter will also cover a detailed description of the problem model, the analysis plan, and the expected results. Under this model, we define quantities of interest and briefly explain the relevance of parameter estimates.

3.1 Data Description

Since this study focuses on respondents cognitive ability, we wish to use data that is least contaminated by technical problems in the process of survey research operation. We also wish that the data are from a survey that is designed and conducted by serious and professional survey practitioners. Finally, we wish the data are from a study that is well-known and influential. The World Values Survey (WVS) data stand out as an ideal source for this research purpose.

The World Values Survey is a global research project that explores people's values and beliefs, how they change over time and what social and political impact they have. It is carried out by a worldwide network of social scientists who, since 1981, have conducted representative national surveys in almost 100 countries. The WVS is the only source of empirical data on attitudes covering a majority of the world's population (nearly 90%).

The WVS measures, monitors and analyses: support for democracy, tolerance of foreigners and ethnic minorities, support for gender equality, the role of religion and changing levels of religiosity, the impact of globalization, attitudes toward the environment, work, family, politics, national identity, culture, diversity, insecurity, and subjective well-being.

The findings are valuable for policy makers seeking to build civil society and democratic institutions in developing countries. The work is also frequently used by governments around the world, scholars, students, journalists and international organizations and institutions such as the World Bank and the United Nations. Data from the World Values Survey have for example been used to better understand the motivations behind events such as the 2010-2011 Middle East and North Africa protests, the 2005 French civil unrest, the Rwandan genocide in 1994 and the Yugoslav wars and political upheaval in the 1990s.

The WVS originated from European Values Study (EVS) and extended to countries outside Europe in 1981, which constituted the first wave of the WVS. The surveys aim to be longitudinal as well as cross-cultural. The second wave of the WVS (1990) was conducted 10 years after the first and embraces 42 countries. The interval between the waves was shortened to 5 years for the third (1995), fourth (2000), and fifth (2005) waves, which includes 52 and 64 countries separately. In total, the WVS covers 81 societies.

We selected the data from the fifth wave of the WVS for this project since that included Ghana. The WVS was conducted by the Institute of Social Research at the University of Michigan (ISR) in collaboration with leading survey research organizations in each country. The Ghanaian survey in this wave was conducted by Markinor Thinking. Tracy Hammond and Mari Harris were the principal investigators in this project. The survey covers a variety of research topics, such as socio-cultural, moral, religious, and political values and attitudes. It employs detailed questionnaires and face-to-face interview techniques in methodology. Representative samples were drawn from each country and the number varies from 1000 to 3500 per country. The survey period for Ghana was from 19th February to 04th April 2007 which included a sample of about 1500 individuals.

3.2 Sample Selection

It is a fact that respondents are inclined to give 'don't know' answers to difficult questions due to limitation of cognitive ability, or to sensitive questions due to social desirability or political fear. The WVS covers a variety of topics that can enable us test topical effect on item-nonresponse. We grouped all the questions into six categories: life-related, value-related, politics-related, Income and wealth-related, democracy-related, and questions on socio-demographic features. Life-related questions include those on attitudes to life, confidence, marriage, religion, and morality whereas Value-related questions consist of those reflecting personal values on environment, country priority, and future changes. Politics-related questions include questions on institutional trust, political system, and international politics. Income and wealth related questions consist of those relating to family savings, and scale of income whiles democracy-related questions consist of those on governance and democracy. The socio-demographic features include age, sex, educational level, and employment status. Politics-related questions may be the most difficult to answer in terms of task difficulty while life-related questions are the easiest. However, questions in all six categories can be sensitive depending on social and political contexts. We finally selected five questions from each of the six categories randomly for analysis. We further selected one question randomly from each of the categories except those on socio-demographic features to be used for the IRT modelling and construction of item characteristics curves.

3.3 Coding Scheme

Based on literature study, we used manual coding scheme to code all items. All items were either assigned a value zero or one. Zero was used when an item was not answered or when a 'don't know' answer was provided, and one was assigned when an item was answered. This resulted in a matrix called an indicator matrix consisting of the indicator variables d_i that marks the occurrence of the values of item responses, i.e.

$$d_i = \begin{cases} 1, & if\ Item\ i\ is\ observed \\ 0, & if\ Item\ i\ is\ not\ observed \end{cases} \quad \dots\dots\dots\dots\dots (3.1)$$

3.4 Model

Item response theory (IRT) is an umbrella of statistical models that describe, in probabilistic terms, the ~~relationship~~ between a person's response to a survey question and his or her level of the 'latent variable' being measured by the scale. This latent variable is usually a hypothetical construct, trait, domain, or ability, which is postulated to exist but cannot be directly measured by a single observable variable or item. Instead, it is indirectly measured using multiple items or questions in a multi-item scale.

Hambleton (2000) stated that item response theory places the ability of the respondent and the difficulty of the item on the same measurement scale so direct comparisons between respondents' abilities and items are possible. The ability or proficiency of the respondent is labelled theta (θ). The test item characteristics are described by the difficulty (b), discrimination (a), and pseudo-chance (c) parameters. It is important to mention that not all IRT models utilize all item parameters, and there is a continuing debate about the appropriateness of these parameters. For example, the Rasch model uses only the difficulty parameter and ignores the discrimination and pseudo-chance parameters completely. Because the Rasch model uses only the difficulty parameter as the only item parameter, it is called a 1-PL model (one parameter logistic). Another model uses both the difficulty and discrimination item parameters and is called a 2-PL model. The model that utilizes all three parameters is called the 3-PL model.

The underlying latent variable, expressed mathematically by the Greek letter theta (θ), may be any measurable construct, such as mental health, fatigue, or physical functioning. The person's level on this construct is assumed to be the only factor that accounts for their response to each item in a scale. For example, a person who knows the answer to a survey question will have a high probability of responding. Someone who doesn't know anything about the question is more likely not to provide a response to that question or give a 'don't know' answer.

3.4.1 Rasch Model (1-PLM)

Rasch (1960) was the first to develop the one-parameter logistic model (sometimes referred to as the simple logistic model), however this model differed from models discussed below. In the Rasch Model, a person is characterized by a level on a latent trait ξ, and an item is characterized by a degree of difficulty δ. The probability of an item endorsement is a function of the ratio of a person's level on the trait to the item difficulty δ/ξ (Tinsley, 1992).

Given that the data adequately fit the Rasch model, one can make simple comparisons of the items and respondents according to the principles of specific objectivity. Specific objectivity means that comparison of two items' difficulty parameters are assumed to be independent of any group of subjects being surveyed, and the comparison of two subjects' trait levels does not depend on any subset of items being administered (Mellenbergh, 1994). The Rasch model assumes that the items are all equal in discrimination (weight equally on a factor) and that chance factors (e.g. guessing) do not influence the response.

For a particular item, Rasch proposed a simple probability function, which increases from zero to one with trait level, as:

$$P_i(\xi) = \frac{\xi}{\xi + \delta} \quad \ldots\ldots\ldots\ldots\ldots (3.2)$$

Model (3.2) has the interpretation of the probability of responding being equal to the value of the person parameter ξ relative to the value of the item parameter δ (Linden and Hambleton, 1997). If we use current item response theory notation, substituting e^{θ} for ξ and e^{b} for δ, we have:

$$P_i(\theta) = \frac{e^{\theta}}{e^{\theta} + e^{b}} = \frac{1}{1 + e^{-(\theta - b)}} = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad \ldots\ldots\ldots\ldots\ldots (3.3)$$

Where:

$P_i(\theta)$ is the probability of an individual with ability θ responding to item *i*.

b_i is the difficulty parameter for the i^{th} item.

e is a transcendental number (natural log constant) whose value to three decimal places is 2.718.

Model (3.3) shows the dependent variable, the probability of responding to an item, as a function of the difference between two independent variables, the person's level on the underlying trait θ and the item threshold b (difficulty). Rasch constrained the sum of the difficulty parameters for all scale items to be equal to zero (i.e $\Sigma b_i = 0$), thus setting the scale of the θ parameter. Given this constraint, the population distribution of θ is unspecified. The distribution has some mean, relative to the average item difficulty, and some variance, relative to the unit slope of the trace lines. The shape of the population distribution [of θ] is unknown; it is whatever shape it has to be to produce the observed score distribution (Thissen and Orlando, 2001).

3.4.2 Two-Parameter Logistic Regression Model (2-PLM)

The two-parameter logistic model (Birnbaum, 1968) allows the slope or discrimination parameter a, to vary across items instead of being constrained to be equal as in the one-parameter logistic or Rasch model. The relative importance of the difference between a person's trait level and item threshold is determined by the magnitude of the discriminating power of the item (Embretson and Reise, 2000). The two-parameter logistic model trace line for the probability of responding to item i for a person with latent trait level θ is:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \dots\dots\dots\dots\dots (3.4)$$

Where:

$P_i(\theta)$, b_i, and e have the same definition as in equation (3.3).

D is a scaling factor equal to 1.7

a_i is the item discrimination parameter for the i^{th} item.

The scaling factor D in (3.4) is added to the model as an adjustment so that the logistic model approximates the normal ogive model. Approximately half of the literature includes the adjustment and half does not (Thissen and Steinberg, 1988).

3.4.3 Three-Parameter Logistic Regression Model (3-PLM)

The three-parameter logistic model (Lord, 1980) was developed in educational testing to extend the application of item response theory to multiple choice items that may elicit guessing. For item i, the three-parameter logistic trace line is:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \dots\dots\dots\dots\dots (3.5)$$

Where:

$P_i(\theta)$, b_i, a_i, D, and e, have the same definition as in equations (3.3) and (3.4).

c_i is the item pseudo-chance parameter for the i^{th} item.

The guessing parameter c is the probability of responding to an item i even if the person does not know the answer. When $c = 0$, the three-parameter model is equivalent to the 2-PL model. Including the guessing parameter changes the interpretation of other parameters in the model. The threshold parameter b is the value of theta at which respondents have a $(0.5 + 0.5c)$*(100) % chance of responding to an item.

The parameters (θ, b, a, c) are graphed in such a way as to yield important information about the test items themselves. Figure 3.1 shows an item characteristic curve for a hypothetical item with the identifying item parameters (Spencer, 2004). The x-axis represents the item's difficulty. Because this is on the same scale as the respondent's ability, we can quickly identify which items are appropriate or 'answerable' by a given respondent with a given ability or proficiency.

The item difficulty parameter (*b*) is plotted on the x-axis and is an indicator as to the difficulty of the item. Easier items have a lower value for *b* and the corresponding item traceline is shifted to the left. Harder items have a higher value for *b* and the corresponding item traceline is shifted to the right.

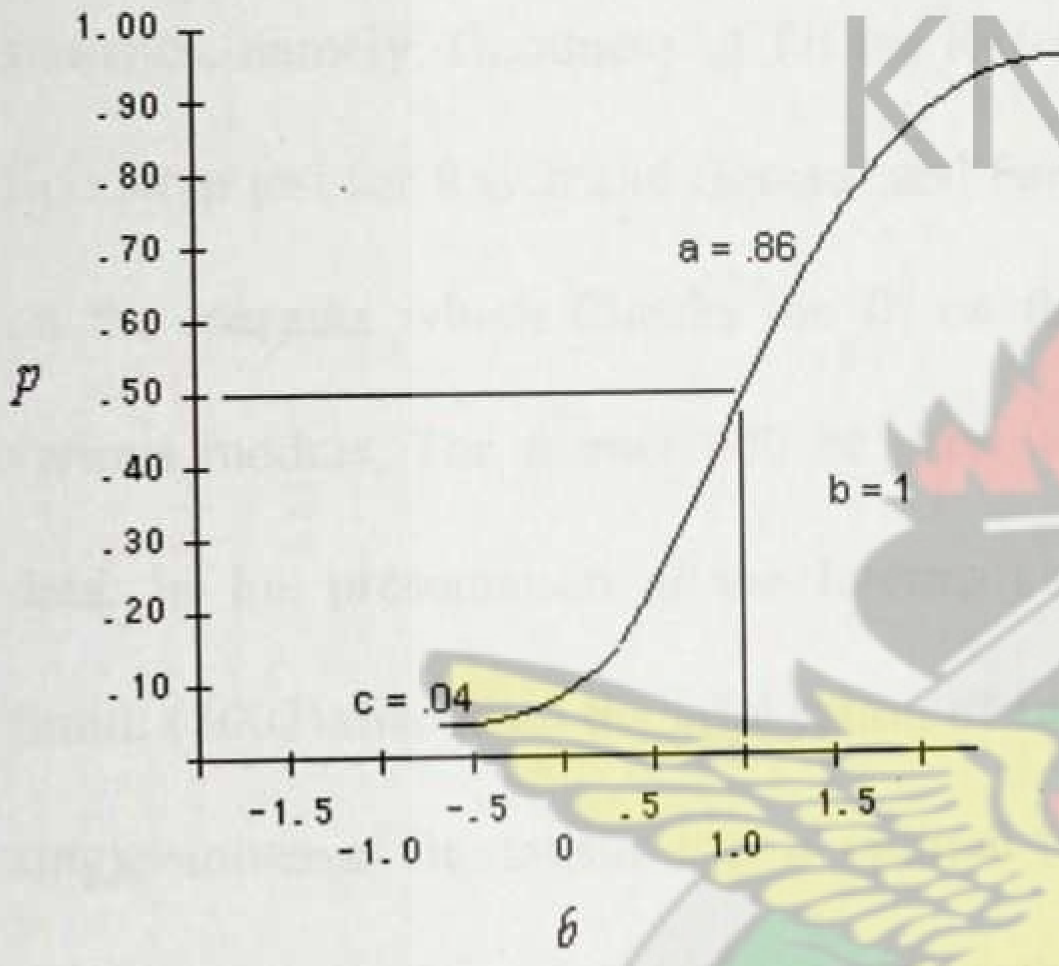

Figure 3.1: Item Characteristics Curve

The discrimination parameter (*a*) indicates the slope at the inflection point of the traceline. More discriminating items have a steeper slope. Less discriminating items have a much flatter slope, indicating that respondents of varying abilities have a similar probability in responding to the item. The pseudo-chance parameter (*c*) shifts the lower half of the traceline to a designated point above the x-axis. As the traceline shifts upward, students of lesser ability have a greater probability of responding. The pseudo-chance parameter can be construed as the probability of responding by an examinee of extremely low ability.

3.4.4 Goodness of Fit

A battery of fit statistics exists that indicate the degree to which a given IRT model adequately fits the empirical data. These are typically called Goodness of Fit Indices. A poorly fitting model cannot yield theoretically invariant item and ability parameters. Test for goodness of fit was performed to ensure that the appropriate model is applied.

The R Package for Latent Trait Modelling provides two goodness of fit statistics, namely, Goodness of Fit for Rasch Models which performs a parametric Bootstrap test for Rasch and Generalized Partial Credit models, and Fit of the model on the margins which Checks the fit on the two- and three-way margins for the various models. The former will be used when fitting the models to the empirical data. In his presentation at the International Objective Measurement Workshop, Smith (2002) discussed the application of fit statistics. His insight is that there is no single universal fit statistic that is optimal for detecting every type of measurement disturbance. Each statistic has its strengths and weaknesses. By identifying the different types of measurement disturbances, one can select the most appropriate fit statistic. This fit statistic can then be used to determine how adequately the selected IRT model fits the data.

Smith further classifies fit statistics into three categories: total fit, within fit, and between fit. These types differ in their purpose, and in the manner in which they summarize the squared standardized residuals. Another term, misfit, is used to identify when a model fails to adequately fit the data. The total fit statistic describes misfit due to the interactions of any item/person combination. This statistic works best in identifying random types of measurement disturbances between a target and focal group. The between fit statistic compares logical groups such as gender, ethnicity, or age to detect item bias and is best at identifying systematic measurement disturbances. The within fit statistic is similar to the between fit statistic. Whereas the between fit statistic sums over the entire respondent sample, the within fit statistic is summed over only the group of interest. Just as no single fit statistic functions optimally to describe the various types of misfit, no single fit statistic functions best for all conditions within these three categories.

3.4.6 Model Assumptions

IRT entails three assumptions:

1. A unidimensional trait denoted by θ ;
2. Local independence of items;
3. The response of a person to an item can be modelled by a mathematical *item response function* (IRF).

The trait is further assumed to be measurable on a scale (the mere existence of a test assumes this), typically set to a standard scale with a mean of zero and a standard deviation of one. 'Local independence' means that items are not related except for the fact that they measure the same trait, which equivalent to the assumption of unidimensionality, but presented separately because multidimensionality can be caused by other issues. The topic of dimensionality is often investigated with factor analysis, while the IRF is the basic building block of IRT and is the centre of much of the research and literature.

3.5 Methods of Estimation

The various methods of estimation used in this study are presented here. They include maximum likelihood estimation, estimation of model parameters, ability estimation, estimation of goodness of fit indices, and estimation of model selection indices.

3.5.1 Maximum Likelihood Estimation

IRT models are used to calculate a person's trait level by first estimating the likelihood of the pattern of responses to the items, given the level on the underlying trait being measured by the scale. Because the items are locally independent, the likelihood function L is,

$$L = \prod_{i=1}^{n} p_i(\theta) \quad \ldots\ldots\ldots\ldots\ldots (3.6)$$

Which is simply the product of the individual item probabilities $p_i(\theta)$. $p_i(\theta)$ models the probability of responding to the item i conditional on the underlying trait θ. Often, information about the population is included in the estimation process along with the information of the item response patterns. Therefore, the likelihood function is a product of the IRT probabilities for each individual item i multiplied by the population distribution of the latent construct $\phi(\theta)$:

$$L = \prod_{i=1}^{n} p_i(\theta)\phi(\theta) \quad \ldots\ldots\ldots\ldots\ldots (3.7)$$

Next, trait levels are estimated typically by a maximum likelihood method; specifically, the person's trait level maximizes the likelihood function given the item properties. Thus, a respondent's trait level is estimated by a process that, (1) calculates the likelihood of a response pattern across the continuous levels of the underlying trait θ, and (2) uses some search method to find the trait level at the maximum of the likelihood (Embretson and Reise, 2000). Often times, this search method uses some form of the estimate of the mode (highest peak) or the average of the likelihood function. These estimates can be linearly transformed to have any mean and standard deviation a researcher may desire.

3.5.2 Parameter Estimation

Because the actual values of the parameters of the items in a survey are unknown, one of the tasks performed when a survey is analysed under item response theory is to estimate these parameters. The obtained item parameter estimates then provide information as to the technical properties of the survey items. In reality, individuals ability scores are not known, but it is easier to explain how item parameter estimation is accomplished if this assumption is made.

In the case of a typical survey, a sample of M individuals responds to the N items in the questionnaire. The ability scores of these individuals will be distributed over a range of ability levels on the ability scale. These individuals may be divided into, say, J groups along the scale so that all the individuals within a given group have the same ability level θ_j and there will be m_j individuals within group j, where $j = 1,2,3, \dots, J$. Within a particular ability score group, r_j individuals respond to the given item. Thus, at an ability level of θ_j, the observed proportion of responses is $p(\theta_j) = {}^{r_j}/_{m_j}$, which is an estimate of the probability of responding at that ability level. Now the value of r_j can be obtained and $p(\theta_j)$ computed for each of the j ability levels established along the ability scale. If the observed proportions of responses in each ability group are plotted, the result will be something like that shown in Figure 3.2.

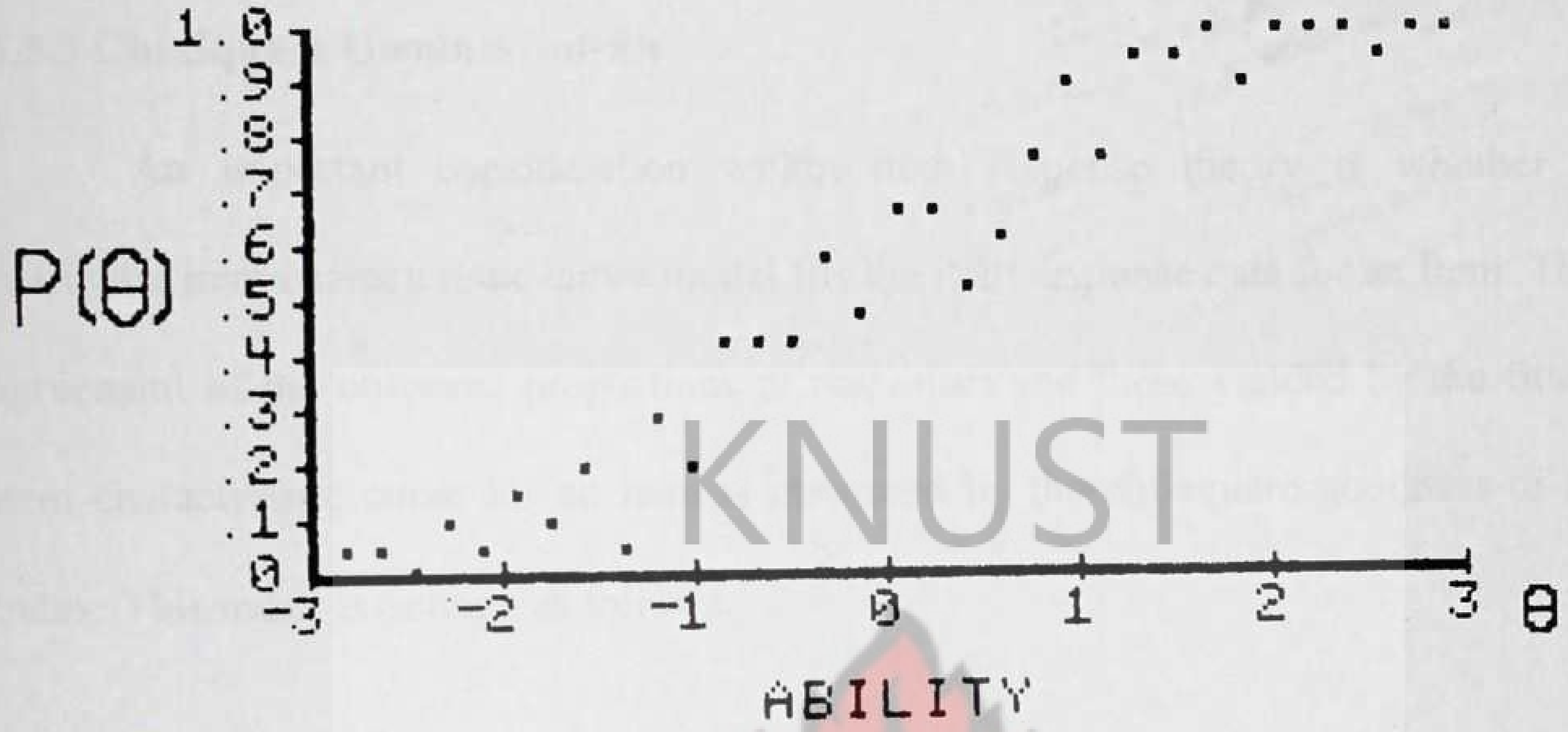

Figure 3.2: Observed proportion of responses as a function of ability

Source: (Baker, 2001)

The basic task now is to find the item characteristic curve that best fits the observed proportions of responses. To do so, one must first select a model for the curve to be fitted. Any of the three logistic models could be used. The procedure used to fit the curve is based upon maximum likelihood estimation. Under this approach, initial values for the item parameters, such as $b = 0$, $a = 1$, and $c = 0$ are established a *priori*. Then, using these estimates, the value of $P(\theta_j)$ is computed at each ability level via the equation for the item characteristic curve model. The agreement of the observed value of $p(\theta_j)$ and computed value $P(\theta_j)$ is determined across all ability groups. Then, adjustments to the estimated item parameters are found that result in better agreement between the item characteristic curve defined by the estimated values of the parameters and the observed proportions of responses. This process of adjusting the estimates is continued until the adjustments get so small that little improvement in the agreement is possible. At this point, the estimation procedure is terminated and the current values of a, b, and c are the item parameter estimates (Baker, 2001).

3.5.3 Chi-Square Goodness-of-Fit

An important consideration within item response theory is whether a particular item characteristic curve model fits the item response data for an item. The agreement of the observed proportions of responses and those yielded by the fitted item characteristic curve for an item is measured by the chi-square goodness-of-fit index. This index is defined as follows:

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{\left[p(\theta_j) - P(\theta_j)\right]^2}{P(\theta_j)Q(\theta_j)} \quad \dots\dots\dots\dots\dots (3.8)$$

where:

J is the number of ability groups.

θ_j is the ability level of group j.

m_j is the number of individuals having ability θ_j.

$p(\theta_j)$ is the observed proportion of responses for group j.

$P(\theta_j)$ is the probability of responding for group j computed from the item characteristic curve model using the item parameter estimates.

If the value of the obtained index is greater than a criterion value, the item characteristic curve specified by the values of the item parameter estimates does not fit the data. This can be caused by two things. First, the wrong item characteristic curve model may have been employed. Second, the values of the observed proportions of responses are so widely scattered that a good fit, regardless of model, cannot be obtained. In most tests, a few items will yield large values of the chi-square index due to the second reason. However, if many items fail to yield well-fitting item characteristic curves, there may be reason to suspect that the wrong model has been employed. In such cases, re-analysing the test under an alternative model may yield better results (Baker, 2001).

3.5.4 Ability Estimation Procedures

Under item response theory, maximum likelihood procedures are used to estimate an individual's ability. As was the case for item parameter estimation, this procedure is an iterative process. It begins with some a *priori* value for the ability of the individual and the known values of the item parameters. These are used to compute the probability of responding to each item for that individual. Then an adjustment to the ability estimate is obtained that improves the agreement of the computed probabilities with the individual's item response vector. The process is repeated until the adjustment becomes small enough that the change in the estimated ability is negligible. The result is an estimate of the individual's ability parameter. This process is then repeated separately for each individual in the survey. However, this procedure is based upon an approach that treats each individual separately. Hence, the basic issue is how the ability of a single individual can be estimated. The estimation equation used is shown below:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} -a_i[u_i - P_i(\hat{\theta}_s)]}{\sum_{i=1}^{N} a_i{}^2 P_i(\hat{\theta}_s) Q_i(\hat{\theta}_s)} \quad \dots\dots\dots\dots\dots (3.9)$$

where:

$\hat{\theta}_s$ is the estimated ability of the individual within iteration s

a_i is the discrimination parameter of item i, $i = 1,2,\ldots,N$

u_i is the response made by the individual to item i: u_i= 1 for a response, and u_i = 0 for a nonresponse

$P_i(\hat{\theta}_s)$ is the probability of responding to item i, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration s.

$Q_i(\hat{\theta}_s)$ = 1-$P_i(\hat{\theta}_s)$ is the probability of not responding to item i, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration s.

The equation has a rather simple explanation. Initially, the $\hat{\theta}_s$ on the right side of the equal sign is set to some arbitrary value, such as 1. The probability of responding to each of the N items in the questionnaire is calculated at this ability level using the known item parameters in the given item characteristic curve model. Then the second term to the right of the equal sign is evaluated. This is the adjustment term. The value of $\hat{\theta}_{s+1}$ on the left side of the equal sign is obtained by adding the adjustment term to $\hat{\theta}_s$. This value, $\hat{\theta}_{s+1}$, becomes $\hat{\theta}_s$ in the next iteration. The numerator of the adjustment term contains the essence of the procedure. Notice that $(u_i - P_i(\hat{\theta}_s)$) is the difference between the individual's item response and the probability of responding at an ability level of $\hat{\theta}_s$.

Now, as the ability estimate gets closer to the individual's ability, the sum of the differences between u_i and $P_i(\hat{\theta}_s)$ gets smaller. Thus, the goal is to find the ability estimate yielding values of $P_i(\hat{\theta}_s)$ for all items simultaneously that minimizes this sum. When this happens, the adjustment term becomes as small as possible and the value of $\hat{\theta}_{s+1}$ will not change from iteration to iteration. This final value of $\hat{\theta}_{s+1}$

is then used as the individual's estimated ability. The ability estimate will be in the same metric as the numerical values of the item parameters.

Unfortunately, there is no way to know the individual's actual ability parameter. The best one can do is estimate it. However, this does not prevent us from conceptualizing such a parameter. Fortunately, one can obtain a standard error of the estimated ability that provides some indication of the precision of the estimate. The underlying principle is that an individual, hypothetically, could answer the questionnaire a large number of times, assuming no recall of how the previous questionnaire items were answered. An ability estimate $\hat{\theta}$ would be obtained from each testing. The standard error is a measure of the variability of the values of $\hat{\theta}$ around the individual's unknown parameter value θ. In the present case, an estimated standard error can be computed using the equation given below:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{N} a_i{}^2 P(\hat{\theta})Q(\hat{\theta})}} \quad \ldots\ldots\ldots\ldots\ldots (3.10)$$

It is of interest to note that the term under the square root sign is exactly the denominator of equation (3.9). As a result, the estimated standard error can be obtained as a side product of estimating the individual's ability (Baker, 2001).

3.5.5 Model Selection

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) values were used to select the most appropriate model. The Bayesian information criterion was introduced by (Schwarz, 1978) as a competitor to the Akaike (1973, 1974) information criterion. Schwarz derived BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In large-sample settings, the fitted model favoured by BIC ideally corresponds to the candidate model which is a posteriori most probable; i.e., the model which is rendered most plausible by the data at hand. The computation of BIC is based on the empirical log-likelihood and does not require the specification of priors. Assuming two candidate models are regarded as equally probable a priori, a Bayes factor represents the ratio of the posterior probabilities of the models. The model which is a posteriori most probable is determined by whether the Bayes factor is less than or greater than one. The definitions of AIC and BIC are given as,

$$AIC = -2(\ln(likelihood)) + 2k \quad \ldots\ldots\ldots\ldots\ldots (3.11)$$

$$BIC = -2(\ln(likelihood)) + kln(n) \quad \ldots\ldots\ldots\ldots\ldots (3.12)$$

Where, *likelihood* is the probability of the data given a model and k, is the number of free parameters in the model. AIC and BIC feature the same goodness-of-fit term but the penalty term of BIC is more stringent than the penalty term of AIC. ($for\, n > 8$, $kln(n)$ exceeds $2k$). Consequently, BIC tends to favour smaller models than AIC.

3.5.6 Other Formulae Used

1. Cronbach's alpha is defined as;

$$\alpha = \frac{p}{p-1}(1 - \frac{\sum_{i=1}^{p} \sigma_{yi}{}^{2}}{\sigma_{x}{}^{2}})$$

Where p is the number of items, $\sigma_x{}^2$ is the variance of the observed total test scores, and $\sigma_{yi}{}^2$ is the variance of the i^{th} item.

2. The point biserial correlation is defined as;

$$r = \frac{(\bar{X}_1 - \bar{X}_0)\sqrt{\pi(1-\pi)}}{S_x}$$

Where $\bar{X}_1$ and $\bar{X}_0$ denote the sample means of the *X*-values corresponding to the first and second level of Y respectively, S_x is the sample standard deviation of *X*, and π is the sample proportion for *Y=1*.

3. The kernel density estimator is defined as;

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

Where *X*$X_1, X_2, X_3, \ldots, X_n$ is an independent and identically distributed sample drawn from some distribution with unknown density *f* , K(.) is the kernel – a symmetric but not necessarily positive function that integrates to one, and *h>0* is a smoothing parameter called the bandwidth.

4. The test information function is defined as;

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta)$$

Where, $I(\theta)$ is the amount of test information at an ability level of θ, $I_i(\theta)$ is the amount of information for item i at ability level θ, and N is the number of items in the test.

5. Under a one-parameter (Rasch) model, the item information is defined as;

$$I_i(\theta) = P_i(\theta)Q_i(\theta)$$

Where, $P_i(\theta) = \frac{1}{1+e^{-(\theta-b_i)}}$, b_i is the difficulty parameter for item i, $Q_i(\theta) = 1 - P_i(\theta)$, and θ is the ability level of interest.

3.6 Analysis Plan

The dependent variable in this study, item-nonresponse, is measured as whether or not an item is answered by an individual. This is a dichotomously scored variable: 0 = no answer or 'don't know' answer to the item, 1 = answer to the item. The item-nonresponse variable is extremely skewed; most people answered most items. Such data violate several assumptions of the normal regression method, and the appropriate analysis model is an IRT model. The degree to which these IRT models adequately fit the empirical data shall be indicated using various goodness-of-fit indices. Likelihood ratio test will then be conducted to select the most suitable model for the data. Descriptive statistics, tables, and graphs would be used to classify and describe variables.

3.6 Software

An R Package for Latent Trait Modelling and Item Response Theory Analyses (R Development Core Team, 2010) was used. This package has been developed for the analysis of multivariate dichotomous and polytomous data using latent variable models, under the Item Response Theory approach. For dichotomous data the Rasch, the Two-Parameter Logistic, and Birnbaum's Three-Parameter models have been implemented, whereas for polytomous data Semejima's Graded Response model is available. Parameter estimates are obtained under marginal maximum likelihood using the Gauss-Hermite quadrature rule. SPSS version 19 was also used to generate some frequencies and cross tabulations.

3.7 Expected Results

On the reason behind 'don't know' responses, whether respondents actually don't know or don't care, or don't want to tell, we expect to see 'don't know' responses having a positive relationship with the difficulties of questions. We also expect respondents with high ability (i.e. respondents who have high question knowledge) to have high probability of responding to such questions and vice-versa. There may also be some few respondents with low ability providing meaningful responses which will mean that they are guessing and for that matter 'don't care'.

On the characteristics of nonrespondents in Ghanaian surveys, it is commonly believed that females, less-educated, or older people are associated with lower cognitive ability, and thus give more 'don't know' answers in public opinion survey. Age may have a complex effect that not only involves cognitive ability but also relates to other effects. Therefore, we divided the respondents into three age groups: the first group includes people from 15 to 29 years old, the second from 30 to 49 years old, and the third 50 years or older. We expect that the effects are most prominent on responses to politics-related questions.

Females may focus more on life issues but have less interest in politics-related questions. Compared with young adults, the older generation may care more about social policies and thus tend to express concrete opinions on politics-related questions. We expect those with higher degree of political interest tend to give more substantive answers to survey questions rather than ignoring them by saying 'don't know.' Moreover, the effect would be more notable on politics-related questions. Finally, we expect to see that ignoring item-nonresponse as it is commonly done in most Ghanaian surveys results in misleading, biased, and untrue results.

CHAPTER 4

RESULTS AND DISCUSSION

4.0 Introduction

In this chapter, we applied the proposed methodology to the data from the Ghanaian survey in the 5th wave of the world values survey discussed in the previous chapter. For model fitting, we explored the Rasch, the two-parameter logistic, and the three-parameter logistic models. The degree to which these IRT models adequately fit the empirical data was indicated using the chi-square goodness-of-fit tests. Likelihood ratio test was then conducted to select the most suitable model for the data. We then proceeded with the appropriate model to investigate and discuss the effects of item and person characteristics on item-nonresponse.

4.1 Descriptive Statistics

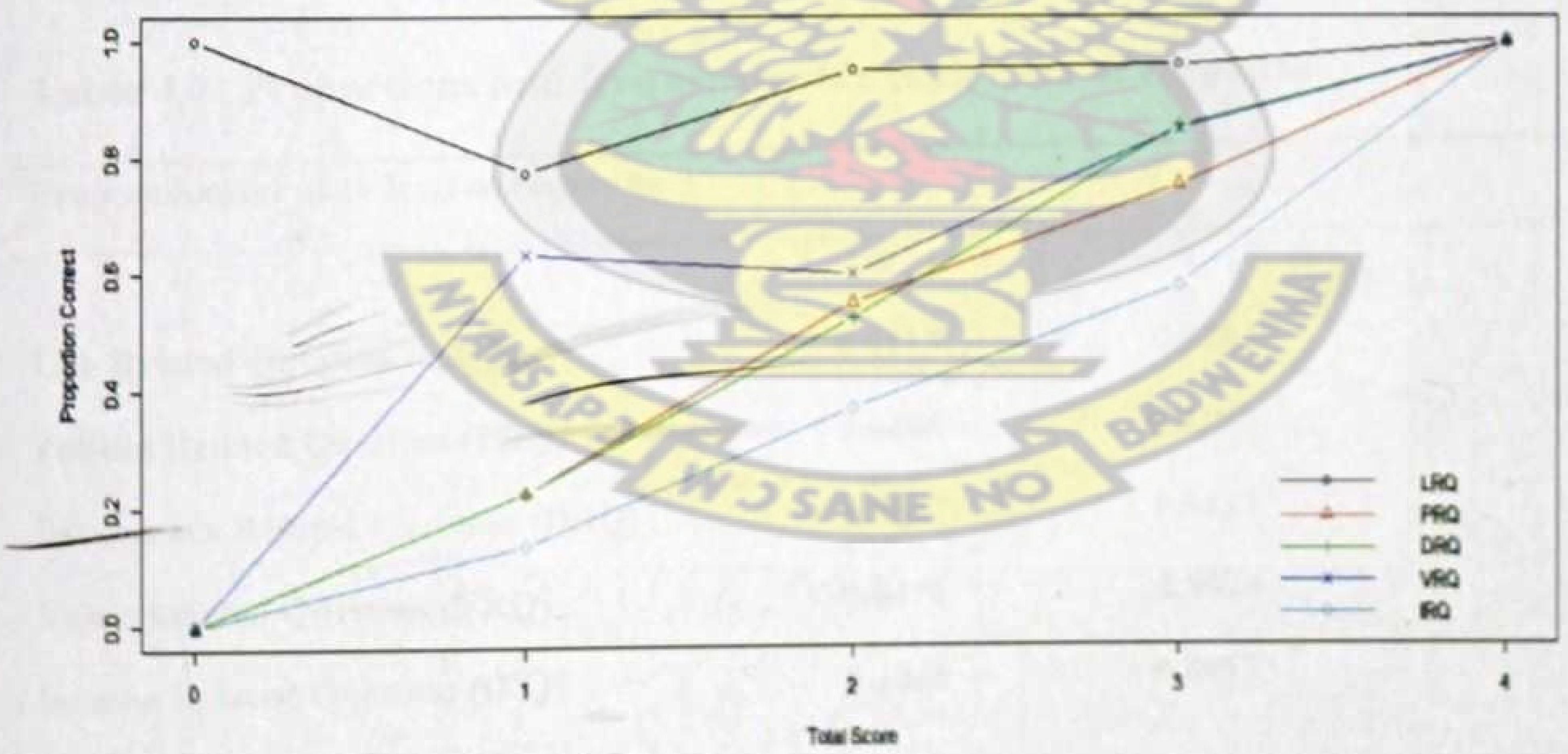

Figure 4.3: Descriptive Statistics Plot

At an initial step, descriptive statistics for our data was produced using the *descript ()* function in the LTM package in R. The output contains among others the proportions for all the possible response categories for each item. We observe from

Figure 4.3 and Table 4.1 that the life related question seems to be the easiest one having the highest proportion (0.9883) of responses, while the income related question seems to be the most difficult one having the lowest proportion (0.8957) of responses. The proportion of responses for the politics related question, democracy related question, and value related question are 0.9302, 0.9433, and 0.9524 respectively.

Frequencies of all possible total scores are also presented. The total score of a response pattern is simply its sum. For dichotomous items this is the number of positive responses. We observe from table 4.2 that 1217 out of the 1534 respondents responded to all questions on all 5 categories, 215 respondents responded to all questions on 4 categories, 78 respondents responded to all questions on 3 categories, 22 respondents responded to all questions on only 2 categories, and 2 respondents responded to all questions on only 1 category. This indicates that our data is extremely skewed since most of the respondents responded to most items.

Table 4 1: Proportions and frequencies for each level of response

Proportions for each level of response:

	0	**1**
Life Related Question (LRQ)	**0.0117**	**0.9883**
Politics Related Question (PRQ)	**0.0698**	**0.9302**
Democracy Related Question (DRQ)	**0.0567**	**0.9433**
Value Related Question (VRQ)	**0.0476**	**0.9524**
Income Related Question (IRQ)	**0.1043**	**0.8957**

Frequencies of total scores:

	0	**1**	**2**	**3**	**4**	**5**
Freq	**0**	**2**	**22**	**78**	**215**	**1217**

Source: WVS (Ghana, 5th wave)

Again since our data is dichotomous response matrices, two versions of point-biserial correlation of each item with the total score are returned. We observe from Table 4.2 that each item is included in the computation of the total score in the first instance, and in the second instance is excluded. The point-biserial correlation is mathematically equivalent to the Pearson (product moment) correlation except that it uses dichotomous variables. We observe that correlations are actually higher when an item is included than when it is excluded. It is also observed that apart from the life related item, all the other items have medium point biserial correlation with the total score when they are included.

We also have Cronbach's alpha, for all items and excluding each time one of the items. It is a coefficient of reliability. It is commonly used as a measure of the internal consistency or reliability of a psychometric test score for a sample of examinees. The higher the Cronbach alpha, the more reliable the test results will be. Theoretically, alpha varies from zero to one, since it is the ratio of two variances. We observe from Table 4.2 that, the cronbach alpha for all the items (0.4294) is acceptable for our purpose.

Table 4.2: Point Biserial Correlations and Cronbach's Alpha

Point Biserial correlation with Total Score:

	Included	Excluded
LRQ	0.2533	0.0876
PRQ	0.6036	0.2498
DRQ	0.6111	0.3009
VRQ	0.5015	0.1916
IRQ	0.6884	0.2792

Cronbach's alpha:

	Value
All Items	0.4294
Excluding LRQ	0.4452
Excluding PRQ	0.3535
Excluding DRQ	0.3154
Excluding VRQ	0.3966
Excluding IRQ	0.3319

Source: WVS (Ghana, 5th wave)

Finally, we have the χ^2 p-values for pairwise associations between the five items, corresponding to the 2×2 contingency tables for all possible pairs. Before an analysis with latent variable models, it is useful to inspect the data for evidence of positive correlations. In this case, the ad hoc checks are performed by constructing the 2×2 contingency tables for all possible pairs of items and examine the chi-squared p-values. Inspection of non-significant results can be used to reveal 'problematic' items. We observe from table 4.3 that only three pairs of items seem to have low degree of association, and the life related item is included in all three pairs. This means the data is okay for IRT modelling.

Table 4.3: Pairwise Associations

Pairwise Associations:

	Item i	Item j	p.value
1	LRQ	VRQ	1.000
2	LRQ	PRQ	0.638
3	LRQ	DRQ	0.071
4	LRQ	IRQ	0.002
5	DRQ	VRQ	1e-03
6	VRQ	IRQ	1e-04
7	PRQ	VRQ	1e-06
8	PRQ	DRQ	1e-12
9	PRQ	IRQ	2e-13
10	DRQ	IRQ	<2e-16

Source: WVS (Ghana, 5th wave)

4.2 The Constrained Rasch Model

We start by fitting the original form of the Rasch model that assumes known discrimination parameter ~~fixed at one~~. In this Model, a person is characterized by a level on a latent trait (Question knowledge), and an item is characterized by a degree of difficulty. Given that the data adequately fit this model, one can make simple comparisons of the items and respondents since comparison of two items' difficulty parameters are assumed to be independent of any group of subjects being surveyed, and the comparison of two subjects' trait levels does not depend on any subset of items being administered. The higher the difficulty parameter, the more difficult the question. We observe from Table 4.4 that the results of the descriptive analysis are also validated by the model fit, where the income related question and the life related question are the most difficult and the easiest, respectively.

Table 4.4: Results for the constrained Rasch model

Coefficients:			
	value	std.err	z.vals
Difficulty. LRQ	-4.9433	0.2417	-20.4494
Difficulty. PRQ	-3.0124	0.1085	-27.7528
Difficulty. DRQ	-3.2511	0.1183	-27.4878
Difficulty. VRQ	-3.4492	0.1274	-27.0635
Difficulty. IRQ	-2.5282	0.0926	-27.2956

Source: WVS (Ghana, 5th wave)

A transformation of the parameter estimates into probability estimates was done. The probability of responding to an item is seen as a function of the ratio of a person's level on the trait (Question Knowledge) to the item difficulty. The column $P(X = 1|Z = 0)$ denotes the probability of responding to the ith item for the average individual. These probabilities were sorted according to the difficulty estimates. In Table 4.5, we observe that the probability of the average individual responding to the life related question is higher ~~than responding~~ to the Politics related question.

Table 4.5: Probability estimates under the constrained Rasch model

	Difficulty	Discrimination	P(x=1|z=0)
LRQ	-4.9433	1	0.9929
PRQ	-3.4492	1	0.9692
DRQ	-3.2511	1	0.9627
VRQ	-3.0124	1	0.9531
IRQ	-2.5282	1	0.9261

Source: WVS (Ghana, 5th wave)

4.3 The Unconstrained Rasch Model (1-PLM)

Both the constrained and the unconstrained Rasch models have similar features and are mathematically equivalent except that, where the constrained Rasch model had a fixed slope of one for all items, the unconstrained Rasch model only requires the slope to be equal for all items. The results in Table 4.6 suggest that the discrimination parameter is actually different from one. Comparing the difficulty parameters under this model, one will observe again that the income related question and the life related question are most difficult and easiest respectively.

Table 4.6: Results for the unconstrained Rasch model

Coefficients:			
	Value	**std.err**	**z.vals**
Difficulty. LRQ	**-3.4995**	**0.2289**	**-15.2912**
Difficulty. PRQ	**-2.2046**	**0.1188**	**-18.5567**
Difficulty. DRQ	**-2.3699**	**0.1293**	**-18.3324**
Difficulty. VRQ	**-2.5059**	**0.1385**	**-18.0929**
Difficulty. IRQ	**-1.8644**	**0.0996**	**-18.7238**
Discrimination	**1.6016**	**0.1141**	**14.0351**

Source: WVS (Ghana, 5th wave)

We observe from Table 4.7 that, the probability of an average individual under the unconstrained Rasch model responding to the life related question is higher than responding to the value related question. Similarly, the probability of responding to the democracy related question is higher than responding to the politics related question for the average individual under the unconstrained Rasch model.

Table 4.7: Probability estimates under the unconstrained Rasch model

	Difficulty	Discrimination	P(x=1|z=0)
LRQ	-3.4995	1.6016	0.9963
VRQ	-2.5059	1.6016	0.9822
DRQ	-2.3699	1.6016	0.9780
PRQ	-2.2046	1.6016	0.9716
IRQ	-1.8644	1.6016	0.9519

Source: WVS (Ghana, 5th wave)

4.4 The Two-Parameter Logistic Model (2-PLM)

Here, we explore how the two-parameter logistic fits the data. Whereas the Rasch models constrain the discrimination parameter to be equal, the two-parameter logistic model allows the slope or discrimination parameter to vary across items. Discrimination is deemed high if its value is greater than 1.35. The parameters are estimated by maximizing the approximate marginal log-likelihood under the conditional independence assumption, that is, conditionally on the latent structure the items are independent Bernoulli variates under the logit link. The results in Table 4.8 suggest that the discrimination parameter is actually not the same for all items. Comparing the difficulty parameters under this model, one will observe again that the income related question and the life related question are most difficult and easiest respectively. In terms of discrimination, we observe from Table 4.8 that, all the questions have high discrimination, especially the democracy related question.

Table 4.8: Results for the two-parameter logistic model

Coefficients:			
	value	std.err	z.vals
Difficulty. LRQ	-4.7096	1.3336	-3.5314
Difficulty. PRQ	-2.2738	0.2271	-10.0143
Difficulty. DRQ	-2.0177	0.1678	-12.0229
Difficulty. VRQ	-2.7544	0.3283	-8.3888
Difficulty. IRQ	-1.8684	0.1751	-10.6728
Discrimination. LRQ	1.0534	0.3700	2.8469
Discrimination. PRQ	1.5143	0.2469	6.1346
Discrimination. DRQ	2.2710	0.4487	5.0617
Discrimination. VRQ	1.3655	0.2437	5.6032
Discrimination. IRQ	1.5945	0.2583	6.1739

Source: WVS (Ghana, 5th wave)

We observe from Table 4.9 that, the probability of an average individual under the two-parameter logistic model responding to the life related question is higher than responding to the value related question. Similarly, the probability of responding to the democracy related question is higher than responding to the politics related question for the average individual under the two-parameter logistic model.

Table 4.9: Probability estimates under the two-parameter logistic model

	Difficulty	Discrimination	P(x=1|z=0)
LRQ	-4.7096	1.0534	0.9930
VRQ	-2.7544	1.3655	0.9773
PRQ	-2.2738	1.5143	0.9690
DRQ	-2.0177	2.2710	0.9899
IRQ	-1.8684	1.5945	0.9516

Source: WVS (Ghana, 5th wave)

4.5 The Three-Parameter Logistic Regression Model (3-PLM)

Here, we explore how the three-parameter logistic model fits the data. Whereas the Rasch models constrain the discrimination parameter to be equal, the three-parameter logistic model allows the slope or discrimination parameter to vary across items and also incorporates a guessing parameter. This model is usually employed to handle the phenomenon of non-random guessing in the case of difficult items. The parameters are estimated by maximizing the approximate marginal log-likelihood under the conditional independence assumption, that is, conditionally on the latent structure the items are independent Bernoulli variates under the logit link. The results in Table 4.10 confirm that the discrimination parameter is actually not the same for all items. Comparing the difficulty parameters under this model, one will observe this time that the value related question and the life related question are most difficult and easiest respectively. In terms of discrimination, we observe from Table 4.10 that, all the questions have very high discrimination. It is important to mention that under the three-parameter model, the values of *the guessing parameter* are not apparent since difficulty values are less than 0 and discrimination values are greater than 1.

Table 4.10: Results for the three-parameter logistic model

Coefficients:

	value	std.err	z.vals
Guessing. LRQ	0.0547	1.6634	0.0329
Guessing. PRQ	0.7869	0.0209	37.6242
Guessing. DRQ	0.7933	0.0374	21.2035
Guessing. VRQ	0.8678	0.0152	57.1432
Guessing. IRQ	0.2908	0.0365	7.9663
Difficulty. LRQ	-4.1906	1.9975	-2.0979
Difficulty. PRQ	-0.6422	0.0001	-6422.0
Difficulty. DRQ	-0.6688	0.0001	-6688.0
Difficulty. VRQ	-0.3507	4.8980	-0.0716
Difficulty. IRQ	-1.1578	5181.1996	-0.0002
Discrimination. LRQ	1.2071	0.4161	2.9009
Discrimination. PRQ	47.5954	0.0001	475954
Discrimination. DRQ	42.6518	0.0001	426518
Discrimination. VRQ	36.9218	531.5050	0.0695
Discrimination. IRQ	56.9407	1459291.0627	0.0000

Source: WVS (Ghana, 5th wave)

We observe from Table 4.11 that, unlike the Rasch and the two-parameter logistic models, the probability of an average individual under the three-parameter logistic model responding to the life related question is lower than responding to the value related question. Also, the probabilities of responding to the democracy related question, the politics related question, and the income related question for the average individual under the three-parameter logistic model is certain.

Table 4.11: Probability estimates under the three-parameter logistic model

	Guessing	Difficulty	Discrimination	P(x=1|z=0)
LRQ	**0.0547**	**-4.1906**	**1.2071**	**0.9940**
IRQ	**0.2908**	**-1.1578**	**56.9407**	**1.0000**
DRQ	**0.7933**	**-0.6688**	**42.6518**	**1.0000**
PRQ	**0.7869**	**-0.6422**	**47.5954**	**1.0000**
VRQ	**0.8678**	**-0.3507**	**36.9218**	**0.9999**

Source: WVS (Ghana, 5th wave)

To determine which of the four IRT models fitted above is the most appropriate for our data, Table 4.12 is a likelihood ratio table which compares the unconstrained version of the Rasch model, the constrained Rasch model, the two-parameter logistic model, and the three parameter logistic model. The likelihood ratio test (LRT) verifies that the unconstrained version of the Rasch model is more suitable for the data. We also have the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) both of which feature the same goodness-of-fit term. However, BIC tends to favour smaller models than AIC. AIC provides an asymptotically unbiased estimator of the expected discrepancy between the generating model and the fitted approximating model. BIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model. By choosing the fitted candidate model corresponding to the minimum value of BIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability. Therefore the smaller the AIC and BIC value, the better the model. We observe from Table 4.12 that the unconstrained Rasch model has the smallest AIC and BIC values, hence the more suitable for the data.

Table 4.12: Likelihood ratio test for the unconstrained Rasch, Constrained Rasch, 2-PLM and the 3-PLM

Likelihood Ratio Table

	AIC	BIC	log.Lik	LRT	df	p.value
Unconstrained Rasch	**3104.63**	**3136.64**	**-1546.31**			
Constrained Rasch	**3136.83**	**3163.51**	**-1563.41**	**34.2**	**1**	**<0.001**
2-PLM	**3106.11**	**3159.46**	**-1543.05**	**6.52**	**4**	**0.163**
3-PLM	**3109.89**	**3189.92**	**-1539.95**	**12.74**	**9**	**0.175**

Source: WVS (Ghana, 5th wave)

In order to check the fit of the most suitable model (the unconstrained Rasch model) to the data, we can look at the two-way margins. We construct the 2 × 2 contingency tables obtained by taking the variables two at a time. Comparing the observed and expected two-way margins is analogous to comparing the observed and expected correlations when judging the fit of a factor analysis model. The comparison is made using the Chi-squared residuals. As rule of thumb residuals greater than 3.5 are indicative of poor fit. We observe from Table 4.13 that none of the chi-squared values are more than 3.5 indicating that the unconstrained Rasch model fits the data well.

Table 4.13: Goodness-of-Fit for the unconstrained Rasch model

Fit on the Two-Way Margins using Pearson chi-squared:

Response: (0,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	1	4	1	4.29	2.52
2	1	2	2	5.57	2.29
3	4	5	18	24.97	1.94

Response: (1,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	4	5	142	134.68	0.40
2	1	4	72	68.89	0.14
3	1	2	105	101.38	0.13

Response: (0,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	4	5	55	48.21	0.96
2	1	2	16	12.74	0.83
3	1	4	17	14.03	0.63

Response: (1,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	4	5	1319	1326.14	0.04
2	3	5	1320	1316.17	0.01
3	1	2	1411	1414.30	0.01

Source: WVS (Ghana, 5th wave)

Now, adopting the unconstrained Rasch model as the most appropriate for our data, we produce the Item Characteristic, the Item Information and the Test Information Curves. The item characteristic curve (ICC) is the basic building block in IRT. The ICC models the relationship between a person's probability of responding to an item category and the level on the construct measured by the scale. The properties of the ICC needed to describe the item's characteristics are its location and the steepness. The steepness of the ICC reflects the discrimination property of an item whereas the difficulty parameter which is represented by location is the point on the ability scale at which the probability of responding to the item is 0.5. We observe from Figure 4.4 that the life related question and the income related question are the easiest and the most difficult respectively.

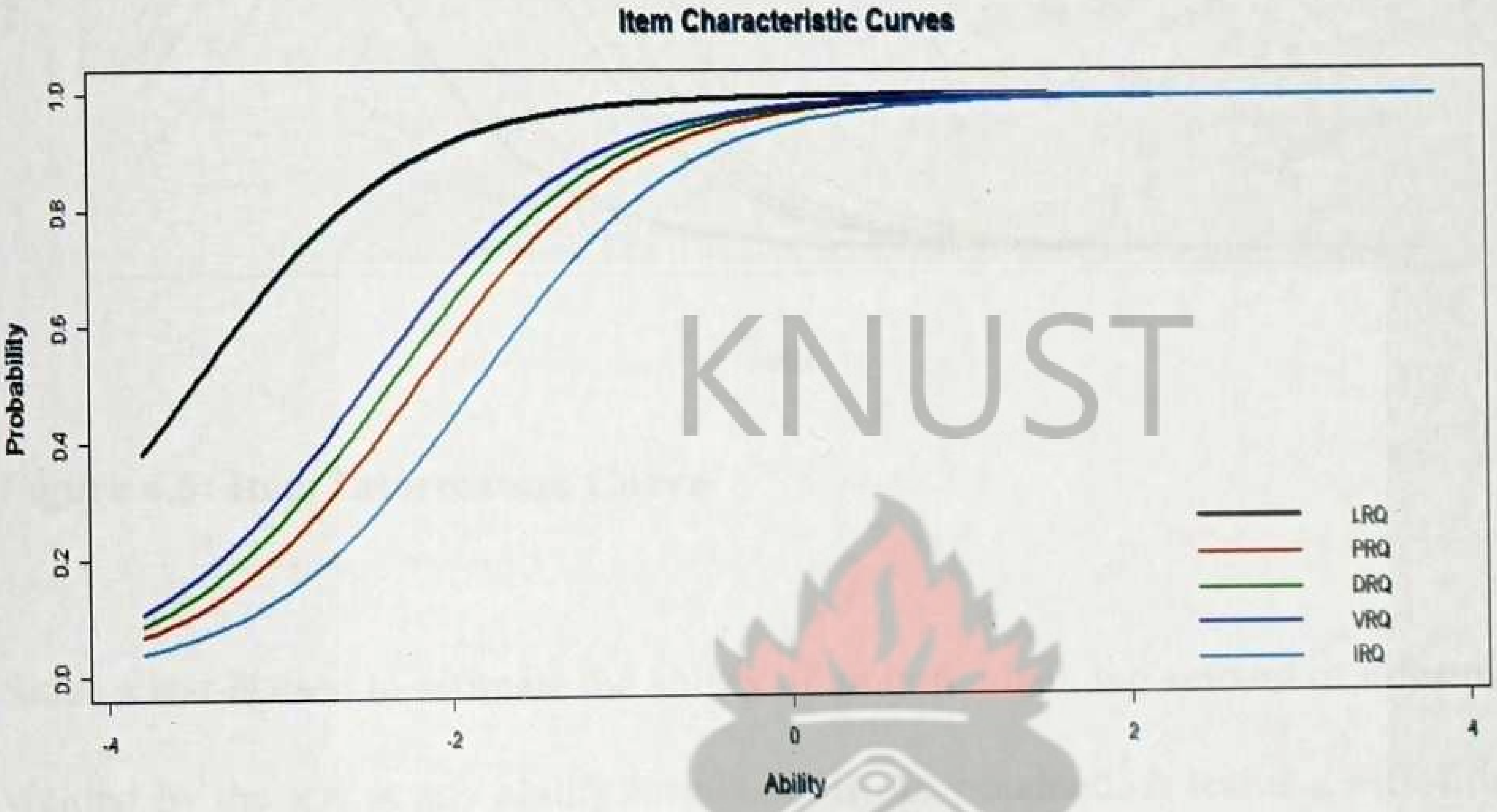

Figure 4.4: Item Characteristic Curve

Item information is the amount of information based upon a single item. It can be computed at any ability level. Because only a single item is involved, the amount of information at any point on the ability scale is going to be rather small. An item measures ability with greatest precision at the ability level corresponding to the item's difficulty parameter. We observe from Figure 4.5 below that the amount of item information for each item decreases as the ability level departs from the item difficulty and approaches zero at the extremes of the ability scale.

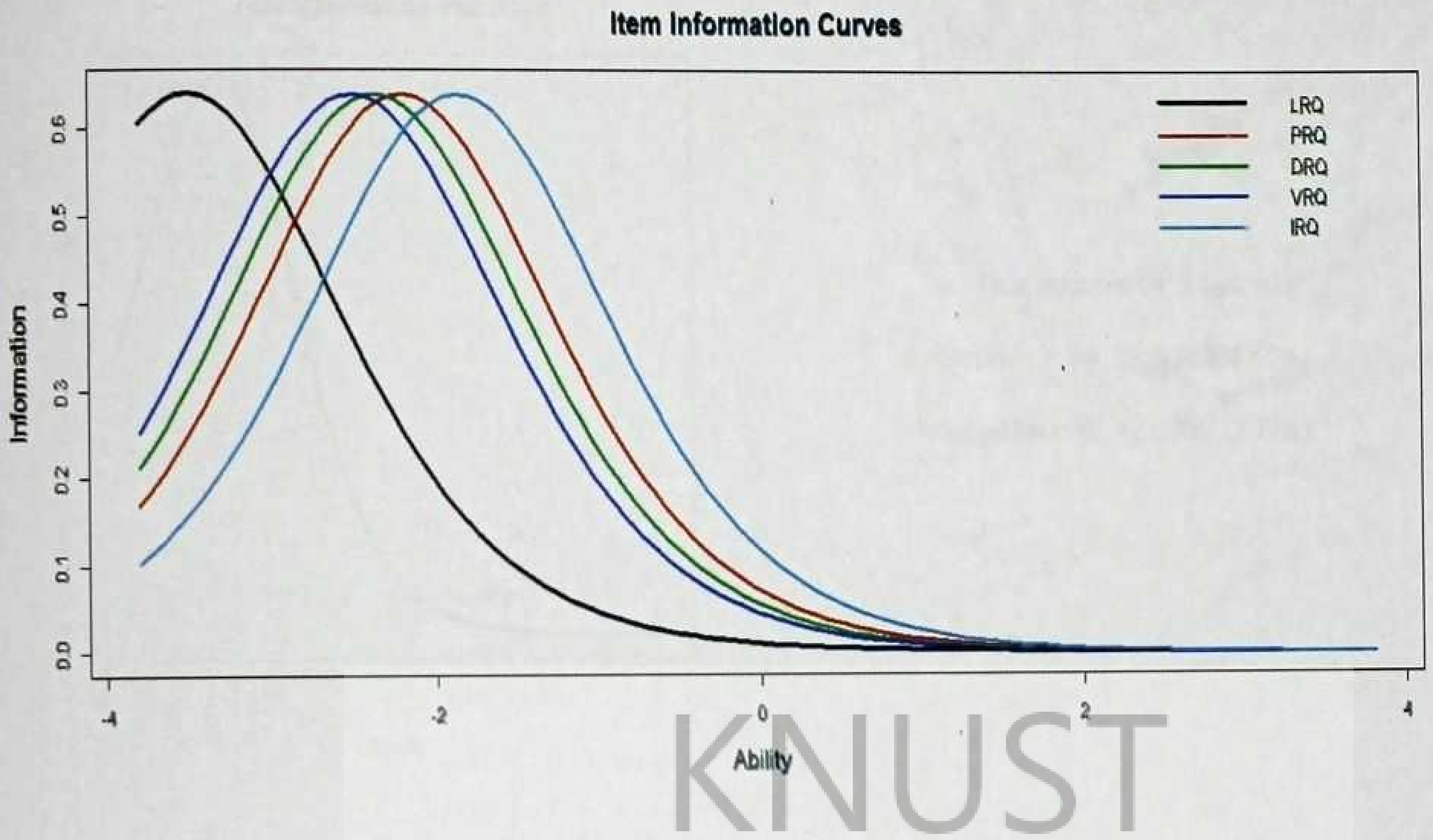

Figure 4.5: Item Information Curve

Since a test is used to estimate the ability of an individual, the amount of information yielded by the test at any ability level can also be obtained. A test is a set of items; therefore, the test information at a given ability level is simply the sum of the item informations at that level. The general level of the test information function will be much higher than that for a single item information function. Thus, a test measures ability more precisely than does a single item. We observe from Figure 4.6 below that the maximum value of the test information function is at ability level -2. However, as the ability level increases, the amount of test information decreases significantly. This indicates that the items asked in our data mainly provide information for respondents with low ability. In particular, the amount of test information for ability levels in the interval (−4, 0) is almost 90%.

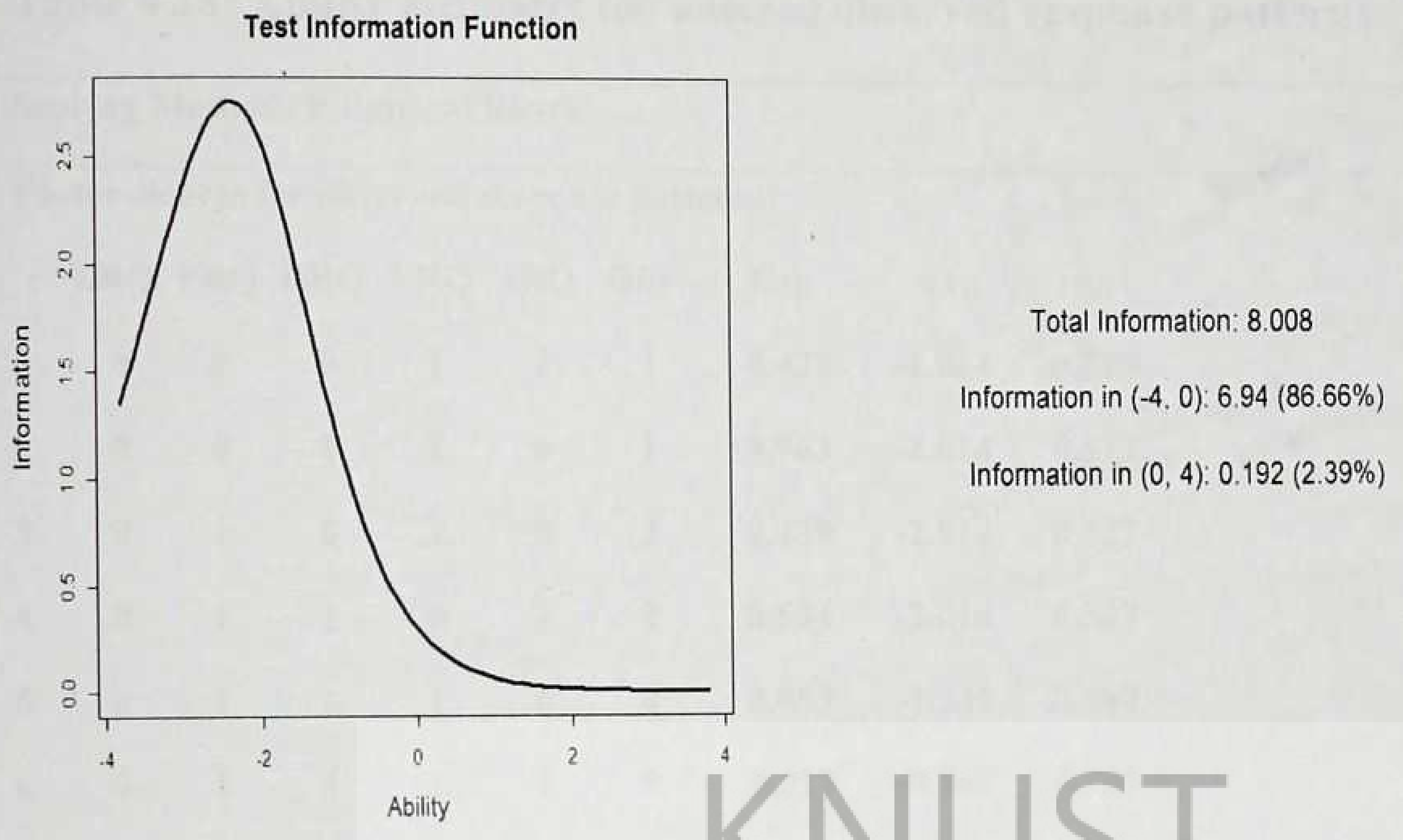

Figure 4.6: Test Information Curve

Source: WVS (Ghana, 5th wave)

4.6 Ability Estimates

Finally, the ability estimates for respondents are obtained. The primary purpose for using IRT in this study is to locate respondents on the ability scale. Since this will help us evaluate respondents in terms of how much underlying ability (Question knowledge) they possess. Factor scores or ability estimates are summary measures of the posterior distribution $P(Z/X)$, where Z denotes the vector of latent variables and X the vector of manifest variables. By default factor scores produces ~~ability~~ estimates for the observed response patterns. We observe from Table 4.14 that our data contains 22 response patterns. Also, the items asked in our data mainly provide information for respondents with low ability (i.e., below 0). That is, most of the items in our dataset are relatively easy for the average respondent to answer.

Table 4.14: Ability estimates for selected observed response patterns

Scoring Method: Empirical Bayes

Factor-Scores for observed response patterns:

	LRQ	PRQ	DRQ	VRQ	IRQ	Obs	Exp	z1	se.z1
1	0	0	0	1	1	1	0.428	-2.014	0.527
2	0	0	1	1	0	1	0.963	-2.014	0.527
3	0	1	0	1	0	2	0.739	-2.014	0.527
4	0	1	1	0	0	1	0.594	-2.014	0.527
5	0	1	1	1	0	4	2.089	-1.531	0.580
6	0	1	1	1	1	9	6.917	-0.883	0.705
7	1	0	0	0	0	2	3.694	-2.445	0.516
8	1	0	0	0	1	2	2.103	-2.014	0.527
9	1	0	0	1	0	10	5.876	-2.014	0.527
10	1	0	0	1	1	8	7.393	-1.531	0.580
11	1	0	1	0	0	3	4.727	-2.014	0.527
12	1	0	1	0	1	9	5.947	-1.531	0.580
13	1	0	1	1	0	18	16.615	-1.531	0.580
14	1	0	1	1	1	53	55.028	-0.883	0.705
15	1	1	0	0	0	2	3.627	-2.014	0.527
16	1	1	0	0	1	12	4.564	-1.531	0.580
17	1	1	0	1	0	17	12.750	-1.531	0.580
18	1	1	0	1	1	31	42.230	-0.883	0.705
19	1	1	1	0	0	10	10.256	-1.531	0.580
20	1	1	1	0	1	32	33.968	-0.883	0.705
21	1	1	1	1	0	90	94.899	-0.883	0.705
22	1	1	1	1	1	1217	1212.004	0.152	0.899

Source: WVS (Ghana, 5th wave)

Figure 4.7 is a Plot of a Kernel Density Estimation of the distribution of the factor scores (i.e., person parameters). Kernel density estimation is a non-parametric way of estimating the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. It includes in the plot the item difficulty parameters (similar to the Item Person Maps). The plot confirms the fact that our data is extremely skewed.

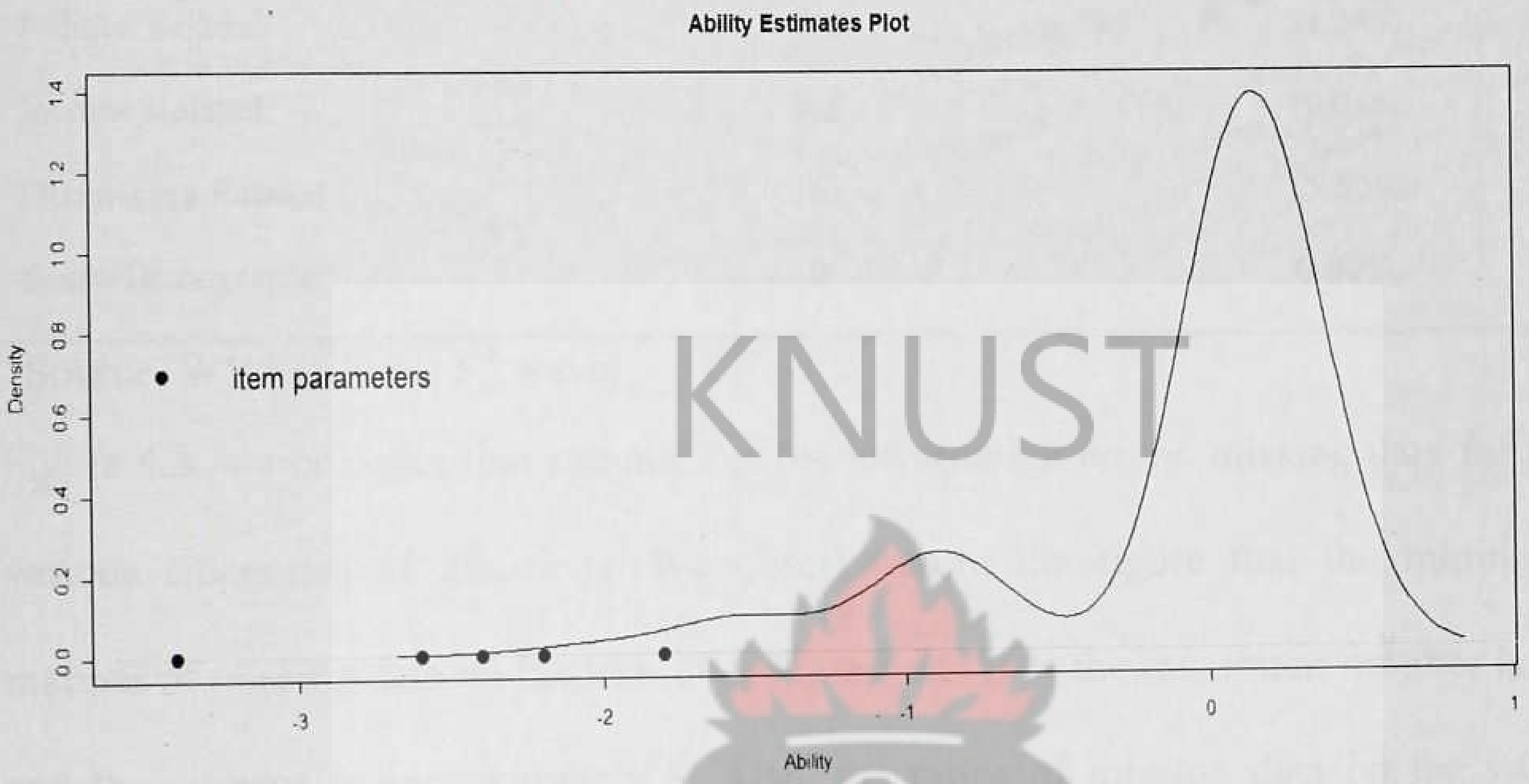

Figure 4.7: Ability Estimates Plot

Source: WVS (Ghana, 5th wave)

4.7 Item and Person Characteristics

We observe from Table 4.15 that the total number of missing data on the randomly selected thirty (30) ~~variables~~ used in this study is 1,050. Only forty-one out of this total representing (3.91%) is on life related questions and none of the missing data is on socio-demographic questions. Whiles two-hundred and thirteen of the total missing data representing (20.29%) is on value related questions, one-hundred and sixty-three representing (15.52%) is on democracy related questions. Furthermore, whereas majority (31.24%) of the missing data is on politics related questions, an equally significant proportion (29.05%) is on income related questions.

Table 4. 15: Item response rates for the various categories of questions

Type of Question	Number of Missing	Item Nonresponse Rate
Life Related	41	3.91%
Value Related	213	20.29%
Politics Related	328	31.24%
Income Related	305	29.05%
Democracy Related	163	15.52%
Socio-Demographic	0	0.00%

Source: WVS (Ghana, 5th wave)

Figure 4.8 is a box-plot that summarises the information on the missing data for the various categories of questions. We observe from the figure that the minimum number of missing data on the life related questions is 1, the maximum number is 18 and the average is approximately 8. Also, the range of missing data on the value related questions is 41 with an average of about 42 whiles the range on the democracy related questions is 75 with an average of about 33. The missing data on income related questions have the highest spread, a minimum of 10 and a maximum of 160 with an average of 61. The questions on politics recorded the highest average (65.6) number of missing data.

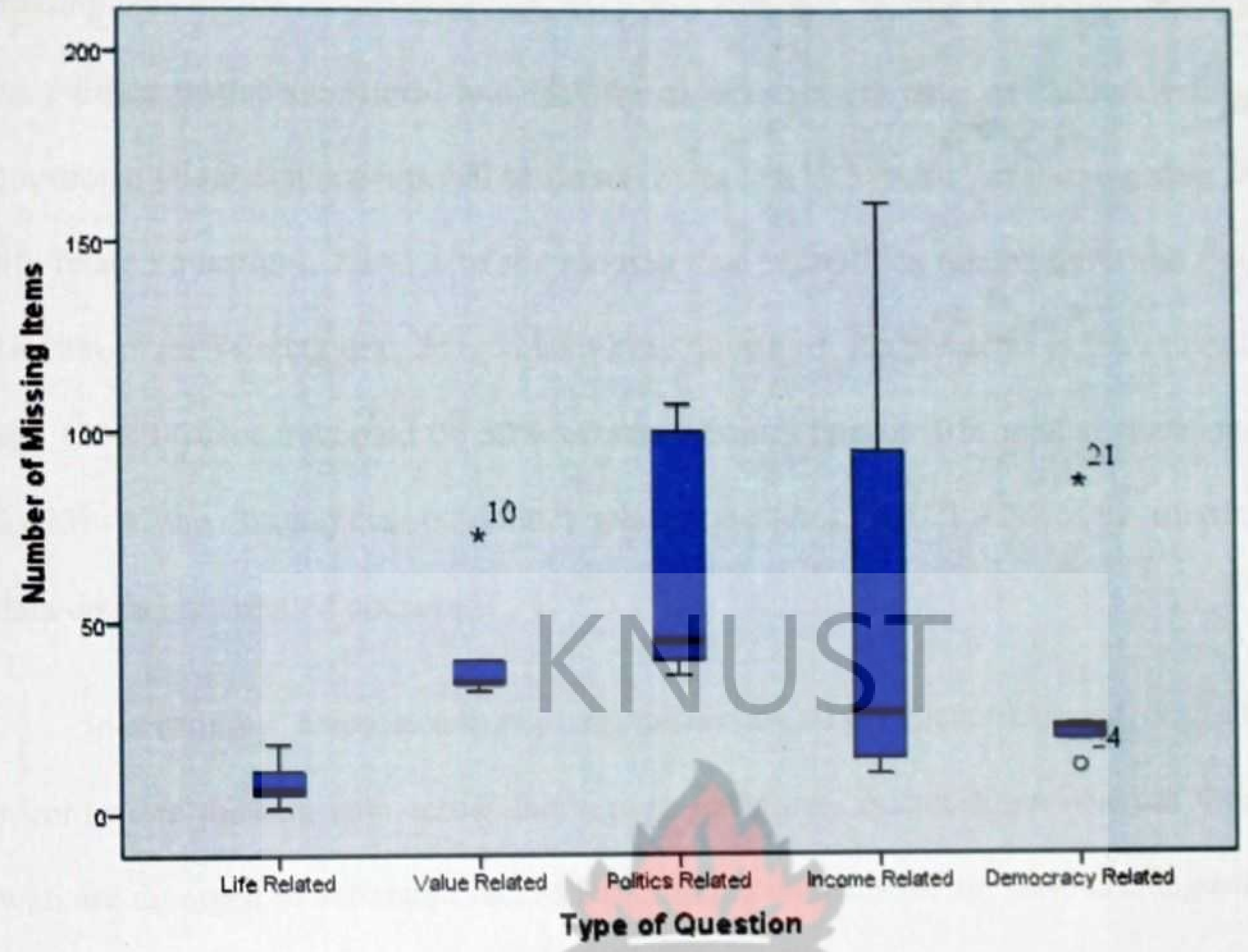

Figure 4.8: Box Plots for Number of Missing Items in the Dataset

Source: WVS (Ghana, 5th wave)

Table 4.16 shows some characteristics of the nonrespondents. We observe that female respondents alone recorded 60.98% of the missing data on life related questions, 56.34% of the missing data on value related questions, 60.37% of the missing data on politics related questions, 55.74% of the missing data on income related questions, and 57.06% of the missing data on democracy related questions. On the other hand, male respondents recorded 39.02% of the missing data on life related questions, 43.66% of the missing data on value related questions, 39.63% of the missing data on politics related questions, 44.26% of the missing data on income related questions, and 42.94% of the missing data on democracy related questions. Furthermore, whiles the younger (15-29 Years) respondents recorded 43.90% of the missing data on life related questions, they also recorded 46.34% of the missing data on politics related questions, and 60.00% of the missing data on income related questions. Older respondents (50-98 years) recorded 19.51% of the missing data on life related questions, 20.43% of the missing data on politics related questions, and 18.03% of the missing data on income related questions. Respondents in their middle age (30-49 years) recorded 36.59% of the missing data on life related questions, 33.23% of the missing data on politics related questions, and 21.97% of the missing data on income related questions.

Interestingly, respondents who are married or single (never married) seem to record more missing data across the various categories of questions whereas those who are divorced or separated record less missing data across the various categories of questions. Married respondents recorded 48.78% of the missing data on life related questions, 69.49% of the missing data on value related questions, 50.30% of the missing data on politics related questions, 37.70% of the missing data on income related questions, and 51.53% of the missing data on democracy related questions. Single respondents recorded 31.71% of the missing data on life related questions, 18.31% of the missing data on value related questions, 32.62% of the missing data on politics related questions, 50.49% of the missing data on income related questions, and 34.36% of the missing data on democracy related questions. Appendix A provides further characteristics of nonrespondents.

Table 4. 16: Characteristics of Nonrespondents

background Features	Item Nonresponse Rates					
	LRQ	VRQ	PRQ	IRQ	DRQ	Total
SEX						
Male	16 (39.02%)	93 (43.66%)	130 (39.63%)	135 (44.26%)	70 (42.94%)	444 (42.29%)
Female	25 (60.98%)	120 (56.34%)	198 (60.37%)	170 (55.74%)	93 (57.06%)	606 (57.71%)
AGE						
15-29	18 (43.90%)	50 (23.47%)	152 (46.34%)	183 (60.00%)	77 (47.24%)	480 (45.71%)
30-49	15 (36.59%)	86 (40.38%)	109 (33.23%)	67 (21.97%)	47 (28.83%)	324 (30.86%)
50+	8 (19.51%)	77 (36.15%)	67 (20.43%)	55 (18.03%)	39 (23.93%)	246 (23.43%)
MARITAL STATUS						
Married	20 (48.78%)	148 (69.49%)	165 (50.30%)	115 (37.70%)	84 (51.53%)	532 (50.67%)
Living together as married	1 (2.44%)	3 (1.41%)	15 (4.57%)	6 (1.97%)	3 (1.84%)	28 (2.67%)
Divorced	4 (9.76%)	1 (0.47%)	11 (3.35%)	13 (4.26%)	2 (1.23%)	31 (2.95%)
Separated	2 (4.88%)	0 (0.00%)	10 (3.05%)	4 (1.31%)	4 (2.45%)	20 (1.90%)
Widowed	1 (2.44%)	22 (10.33%)	20 (6.10%)	13 (4.26%)	14 (8.59%)	70 (6.67%)
Single/Never married	13 (31.71%)	39 (18.31%)	107 (32.62%)	154 (50.49%)	56 (34.36%)	369 (35.14%)

Source: WVS (Ghana, 5th wave)

4.8 Discussion

To answer the first research question "which IRT model is the most appropriate in understanding item-nonresponse", we explored the four IRT models for dichotomous data which include the constrained Rasch model, the unconstrained Rasch model, the two-parameter logistic model, and the three-parameter logistic model. From the likelihood ratio test in table 4.12, we observe by inspecting the AIC and BIC values that, the unconstrained Rasch model has the smallest AIC and BIC values. The AIC and BIC values are defined in a way that, the smaller the better. We also observe from table 4.13 that the unconstrained Rasch model has a very good fit to the data. Hence, the most appropriate model for the data.

To answer the second research question "Missing or don't know answers; Do respondents don't really know, don't care, or don't want to tell?" a school of scholars believe that a simple 'don't know' response may reflect several possible mind states: no idea, don't want to tell, and don't care. Findings from empirical studies are consistent with their belief and confirm that the distinctions are evident in examination of the linkage of underlying attitude and opinion expression (e.g., Bogart, 1967). When respondents during an interview begin guessing answers to questions (i.e., providing answers without going through the various cognitive procedures described in chapter one), it is an indication that they have lost interest in the interview and so 'don't care' about the outcome of it. Therefore it makes sense to associate 'don't care' with guessing.

From the model fitting described above, we realized that the unconstrained Rasch model which uses only the difficulty parameter appeared as the best model for the data. This means that ~~it is only~~ the difficulty parameter of an item that explains whether or not an individual will respond to that item or not. The guessing parameter does not play out in the data. Therefore if an individual does not respond to a survey question or give a 'don't know answer', it is because of the item's difficulty which means that he/she doesn't really know the answer to the question.

On the characteristics of nonrespondents, we observe from table 4.16 that female respondents record higher item-nonresponse rates than their male counterparts across all categories of questions. That is, (60.98%) against (39.02%) on life related questions, (56.34%) against (43.66%) on value related questions, (60.37%) against (39.63%) on politics related questions, (55.74%) against (44.26%) on income related questions, and (57.06%) against (42.94%) on democracy related questions. This is consistent with Ren (2009) who found that being a female respondent decreases the odds of never giving ‘don’t know’ answers by a factor of 0.39 (or 61%).

Also, we observe from table 4.16 that younger respondents (15-29 years) record higher item-nonresponse rates than respondents in the middle age group (30-49 years), who intern record higher rates than older respondents (50-98 years) across all categories of questions except questions on values. Furthermore, we observe again that, respondents who are married seem to have higher item-nonresponse rates than those who are single and never married across all categories of questions.

On education, it is observed from appendix A that respondents without any formal education record higher item-nonresponse rates than those who are educated across all categories of questions. We also observe that educational level has an inverse relationship with item-nonresponse rate. That is, the higher a person’s educational level, the less missing data he/she provides. This is also consistent with (Ren, 2009) who found that an additional unit increase in education increases the odds of never giving ‘don’t know’ answers by 80%, holding all other variables constant. On employment status, it is again observed from appendix A that, respondents who are self-employed record the highest item-nonresponse rates across all the categories of questions whereas respondents who are retire/pensioned record the lowest item-nonresponse rates across the various categories of questions.

CHAPTER 5

CONCLUSION AND RECCOMENDATIONS

5.0 Introduction

This chapter is devoted to drawing a fair conclusion to the study. A summary of the major findings based on the research questions and objectives is also presented. Finally, suggested directions for further research are offered.

5.1 Conclusion

IRT models have been extensively developed and used in educational and psychological measurement. However, use of IRT models outside of these areas is rather limited. Part of the reason for this is that these methods have not been well understood by non psychometricians. With the increasing rates of item-nonresponse in surveys, formulation of IRT models to provide in-depth understanding of this problem can help researchers overcome these obstacles. IRT modelling provides a useful method for addressing questions about patterning of behaviour beyond mere frequency reports.

The implications from this study are quite clear. First, we investigated to identify the most appropriate IRT model for understanding item-nonresponse. The results revealed clearly that, the unconstrained Rasch model is the best IRT model for understanding item-nonresponse.

Furthermore, we analysed the reason behind don't know responses and missing data; whether respondents don't really know, don't care, or don't want to tell. The IRT model was able to provide the reason. As indicated, if an individual does not respond to a survey question or give a 'don't know answer' to the question, it is because of the question's difficulty which means that he/she doesn't really know the answer to the question.

Finally, on the characteristics of nonrespondents, we found that female respondents record more item-nonresponse than male respondents. Also, younger respondents record more item-nonresponse than older respondents. Respondents with no formal education also record higher item-nonresponse than educated respondents. We also found that, respondents who are self-employed record higher item-nonresponse than respondents who are unemployed, who intern record higher item-nonresponse than respondents who work full-time, part-time, or retired. Interestingly, we realized again that married or single respondents record higher item-nonresponse than respondents who are separated or widowed.

To sum up, we found that the propensity to give 'don't know' responses is related to gender, formal education level, age, employment status, and marital status.

5.2 Recommendations

Other applications of IRT modelling of item-nonresponse are clearly possible. For example, researchers can begin looking at IRT model-based imputation for missing data since it has the theoretical advantage of using both respondent and item-specific information.

Also, because of the unique features of IRT models, it has helped health professionals immensely in various parts of the world in evaluating health outcomes. I therefore recommend to Ghanaian health professionals the use of IRT models for the evaluation of health outcomes.

Finally, we wish to end this study by advising researchers in Ghana that, when reporting results of analyses based on items with high nonresponse rates, they should point out obvious pitfalls in using these data and illustrate the kinds of conclusions that might be appropriate.

REFERENCES

Baker, F.B. (2001). The Basics of Item Response Theory. USA: ERIC Clearinghouse on Assessment and Evaluation.

Beatty, P. & Douglas H. (2002). To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse. In *Survey Nonresponse*, edited by R. M. Groves, et al. New York: John Wiley & Sons, Inc.

Berinsky, A. J. (2004). *Silent voices: public opinion and political participation in America*. Princeton, N.J.; Oxford: Princeton University Press.

Biewen, M. (2001). Item non-response and inequality measurement: Evidence from the German earnings distribution, *AllgemeinesStatistischesArchiv*85, 409-425.

Binet, A. & Simon, T. (1905). Methodesnouvelles pour le diagnostic du nieveauintellectuelanormoux {New methods for the diagnosis of levels of intellectual abnormality]. AnneePsychologique, 11, 191-244.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D. (1997). A brief history of item response theory. Educational Measurement: Issues and Practice, 16, 21-33.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

Bogart, L. (1967). No Opinion, Don't Know, and Maybe No Answer. *Public Opinion Quarterly* 31 (3):331-345.

Boomsma, J.J. & Duijn, S. (2000). Lecture Notes in Statistics. The Netherlands: University of Groningen. Book.

Borgers, N. & Hox, J. (2001). Item Nonresponse in Questionnaire Research with Children. Journal of Official Statistics, Vol. 17, No. 2, pp. 321-335.

Crossley, H. M. & Fink, R. (1951). Response and non-response in a probability sample. *International Journal of Opinion & Attitude Research*, *5*, 1-19.

de Leeuw, E.D. (2001). Reducing Missing Data in Surveys: An Overview of Methods. *Quality and Quantity* 35 (2):147-160.

de Leeuw, E.D. & Hox, H. (2003) Prevention and Treatment of Item Nonresponse. Journal of Official Statistics, Vol. 19, No. 2, pp. 153-176.

Deming, W. E. (1953). On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse.*Journal of the American Statistical Association*, *48*, 743-772.

Devey, D.T., Leiter, J. & Thompson, S. (2006). Organizational Survey Nonresponse. USA: North Carolina State University.

Dillman, D.A, Caldwell, S. & Gansemer, M. (2000). Visual Design Effects on Item Nonresponse to a Question about Work Satisfaction that Precedes the Q-12 Agree-Disagree Items. Lincoln, Nebraska: Research paper.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*, John Wiley and Sons, New York.

Durrant, G.B. (2005). Imputation Methods for Handling Item-Nonresponse in the social sciences: A Methodological Review. Southampton: University of Southampton.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Esser, H. (1984). Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischenErklärung und empirischenUntersuchung von Interviewereffekten, in: K.Mayer, P. Schmidt (Hrsg.) *AllgemeineBevölkerungsumfrage der Sozialwissenschaften*, 26-71, Frankfurt.

Esser, H. (1986). KönnenBefragtelügen? ZumKonzept des wahrenWertesimRahmen der handlungstheoretischenErklärung von Situationseinflüssenbei der Befragung, *KölnerZeitschriftfürSoziologie und Sozialpsychologie*38, 314-336.

Ferguson, G. A. (1943). Item selection by the constant process. Psychometrika, 7, 19-29.

Fischer, G. (1968). Psychologische test theorie [Psychological test theory]. Bern: Huber.

Frick, J.R. & Grabka, M.M. (2007). Item Nonresponse and Imputation of Annual Labor Income in Panel Surveys from a Cross National Perspective. IZA, DP No. 3043.

Glas, C. A. W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory* (pp.131-148). New York, NJ: Springer.

Groves, M.R. & Peytcheva, E. (2006). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. Unpublished ms.

Groves, R.M (1989). Survey Errors and Survey Costs. New York: Wiley

Hambleton, R.K. (2000). *Introduction to Item Response Theory*. Breakout session presented at the Sylvan Prometric Results 2000 Conference, Tucson AZ.

Harmon, M.D. (2001). Poll Question Readability and 'Don't Know' Replies. *International Journal of Public Opinion Research* 13 (1):72-79.

Hedeker, D., Mermelstein, R.J. & Flay, B.R. (2001). Application of Item Response Theory Models for Intensive Longitudinal data. Oxford University Press, New York.

Hill, D. & Robert J.W. (2001). Reducing Panel Attrition: A Search for Effective Policy Instruments, *Journal of Human Resources* 36(3), 416-438.

Huisman, M. & Molenaar, I.W. (2001). Imputation of Missing Scale Data with Item Response Models. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Immerwahr, S., Eisenhower, D. & Konty, K. (2008). Effective Conversion of Initial Item Nonresponse to Questions on Age and Income in a Telephone Survey. New York City Department of Health and Mental Hygiene.

Item Response Theory (2011, September 02nd) [online].-URL: http://en.wikipedia.org/wiki/Item_Response_Theory history.htm

Jansen, M. G. H., & Glas, C. A. W. (2001). Statistical tests for differential test functioning in Rasch's model for speed tests. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory* (pp.149-162). New York, NJ: Springer.

Kampen, J.K. (2006). The Impact of Survey Methodology and Context on Central Tendency, Nonresponse and Association of Subjective Indicators of Government Performance. *Quality and Quantity*. 41: 793-813

Kirisci, L., Tarter, R.E., Vanyukor, M., Martin, C., Mezzich, A. & Brown, S. (2001). Application of Item Response Theory to quantify substance use disorder severity. Center for Education and Drug Abuse Research, University of Pittsburgh, 711 Salk Hall, Pittsburgh, PA 15261, USA.

Klauer, K.C. & Sydow, H. (2001). Modeling Learning in Short-term. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Klinkenberg, E.L. (2001). A Logistic IRT Model for Decreasing and Increasing Item Characteristic Curves. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Know More About Ghana (2011, August 24th) [online]. URL: http://www.ghana.gov.gh/index.php?

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213-236.

Krosnick, J.A. (1999). Survey Research. *Annual Review of Psychology* 50:537-567.

Lachman, R., Lachman J.T., & E.C. Butterfield (1979). *Cognitive Psychology and Information Processing*, Hillsdale, NJ: Erlbaum.

Lawley, D. N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 62-A, Part 1, 74-82.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, Measurement and Prediction (pps. 362-412). New York: Wiley.

Lessler, J.T. & Kalsbeek, W.D (1992). Nonsampling Error in Surveys. New York: Wiley

Lewis, C. (2001). Expected Response Functions. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Lillard, L., James P.S., & Finis, W. (1986). What do we really know about wages? The Importance of Nonreporting and Census Imputation, *Journal of Political Economy* 94(3), 489-506.

Linden, W. J., & Hambleton, R. K. (Eds.) (1997). Handbook of modern item response theory. New York, NY: Springer-Verlag.

Little, R.J.A. & Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley

Loosveldt, G., Jan, P., & Jaak, B. (1999). Item non-response as a predictor of unit non-response in a panel survey, Paper presented at the International Conference on Survey Non-response, Portland Oregon, USA.

Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, Whole No. 7.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Luong, A. & Rogelberg, S. G. (1998). How to increase your survey response rate. *The IndustrialOrganizational Psychologist*, *36*, 61-65.

Meijer, R.R. & Krimpen-Stoop, E.M.L.A. (2001). Person Fit Across Subgroups: An Achievement Testing Example. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research, 29, 223-236.

Mokken, R.J., Schuur, W.H. & Leeferink, A.J. (2001). The Circles of our Minds: A Nonparametric IRT Model for the Circumplex. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Post, W.J., Duijn, M.A.J. & Baarsen, B. (2001). Single-Peaked or Monotone Tracelines? On the choice of an IRT Model for Scaling Data. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

'R' Development Core Team (2010). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Rapoport, R.B. (1979). What they don't know can hurt you. *American Journal of Political Science* 23 (4):805-815.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: MESA.

Reeve, B.B. & Fayers, P. (2003). Applying Item Response Theory Modelling for evaluating questionnaire item and scale properties.

Ren, L. (2009). Surveying Public Opinion in Transitional China: An Examination of Survey Response. Pittsburgh: University of Pittsburgh. PhD Thesis

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test.Psychometrika, 1, 33-49.

Riphahn, R.T. & Serfling, O. (2002). Item Nonresponse on Income and Wealth Questions. Switzerland: University of Basel, IZA, CEPR, DIW. Research paper.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. Catholic University of Leuven: Journal of Statistical Software, Vol. 17, Issue 5. http://www.jstatsoft.org

Rose, N., Davier, M.V. & Xu, X. (2010). Modeling Nonignorable Missing Data with Item Response Theory. ETS, Princeton, New Jersey.http://www.ets.org/research/contact.html

Rubin, D. B. (1976). Inference and missing data.*Biometrika*, *63*(3), 581–592.

Rubin, D.B. & Thomas, N. (2001). Using Parameter Expansion to Improve the Performance of the EM Algorithm for Multidimensional IRT Population-Survey Models. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

Schrapler, J.P. (2001). Respondent Behavior in Panel Studies. A Case Study of the German Socio-Economic Panel (GSOEP), *DIW Discussion Papers* No. 244, DIW - Berlin.

Serfling, R.J. (2004). The interaction between unit and item nonresponse in view of the reverse cooperation continuum. Switzerland: University of Basel. Research paper.

Shoemaker, P.J., Martin E., & Elizabeth, A.S. (2002). Item Nonresponse: Distinguishing between Don't Know and Refuse. *International Journal of Public Opinion* 14 (2):193-201.

Sijtsma, K. & Van der Ark, L.A. (2003). Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data. Tilburg: Tilburg University.

Sousa-Poza, A. & Henneberger, F. (2000). Wage data collected by telephone interviews: an empirical analysis of the item nonresponse problem and its implications for the estimation of wage functions, *SchweizerischeZeitschriftfürVolkswirtschaft und Statistik*136(1), 79-98.

Spencer, S.G. (2004). The Strength of Multidimensional Item Response Theory in Exploring Construct Space that is Multidimensional and Correlated. Brigham Young University.PhD Thesis.

Stanton, F. (1939). Notes on the validity of mail questionnaire returns. *Journal of Applied Psychology*, *23*, 170-187.

Steinberg, L., & Thissen, D. (in draft). Chapter 1 – An intellectual history of item response theory: Models for binary responses. From a draft of Item response theory for psychological research.

Stocke, V. (2006). Attitudes toward Surveys, Attitude Accessibility and the Effect on Respondents' Susceptibility to Nonresponse. *Quality and Quantity* 40:259-288.

Stout, W., Froelich, A.G. & Gao, F. (2001). Using Resampling Methods to Produce an Improved DIMTEST Procedure. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory*. New York, NJ: Springer.

Suchman, E. A., & McCandless, B. (1940). Who answers questionnaires? *Journal of Applied Psychology*, *24*, 758-769.

Thissen, D. & Orlando, M. (2001). Chapter 3-Item response theory for items scored in two categories. In D. Thissen& H. Wainer (Eds.), Test Scoring. Hillsdale, NJ: Erlbaum.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. Psychological Bulletin, 104, 385-395.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

Tinsley, H. E. A. (1992). Psychometric theory and counseling psychology research. In S. D. Brown & R. W. Lent (Eds.), Handbook of counseling psychology (2nd ed., pp. 37-70). New York, NY: Wiley.

Tourangeau, R. & Rasinski, K.A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. Psychological Bulletin, 103, 299-314.

Vendramini, C.M.M. & Silva, M.C.R. (2001). Application of Item Response Theory in the Attitudes Evaluation. *Universidade São Francisco, Brazil*

Verhelst, N.D. & Verstralen, H.H.F.M. (2001). An IRT Model for Multiple Raters. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory.* New York, NJ: Springer.

World Values Survey (2011, August 29th) [online].-URL: http://en.wikipedia.org/wiki/World_Values_Survey

World Values Survey Data (2010, October 25th) [online] –URL: http://www.worldvaluessurvey.org

Wright, B. D. (1997). A history of social science measurement. Educational Measurement: Issues and Practice, 16, 21-33.

Zoanetti, N., Griffin, P. & Adams, R. (2001). Applications of Item Response Theory to identify and account for suspect rater data. University of Melbourne

Zweimuller, J. (1992). Survey non-response and biases in wage regressions, *Economics Letters* 39, 105-109.

APPENDICES

Appendix A: Further Characteristics of Nonrespondents

background Features	Item-Nonresponse Rates					
	LRQ	VRQ	PRQ	IRQ	DRQ	Total
Highest Educational Level Attained						
No formal education	11 (26.83%)	131 (61.50%)	145 (44.21%)	70 (22.95%)	51 (31.29%)	408(38.86%)
Complete primary school	8 (19.51%)	43 (20.19%)	79 (24.09%)	67 (21.97%)	42 (25.77%)	239(22.76%)
Complete secondary: tech/ voc. Type	8 (19.51%)	14 (6.57%)	34 (10.37%)	64 (20.98%)	25 (15.34%)	145(13.81%)
Incomplete university-preparatory type	0 (0.00%)	0 (0.00%)	4 (1.22%)	6 (1.97%)	1 (0.61%)	11 (1.05%)
Complete university-preparatory type	0 (0.00%)	2 (0.94%)	3 (0.91%)	6 (1.97%)	1 (0.61%)	12 (1.14%)
University education, without degree	0 (0.00%)	0 (0.00%)	5 (1.52%)	8 (2.62%)	1 (0.61%)	14 (1.33%)
University education, with degree	0 (0.00%)	2 (0.94%)	1 (0.30%)	5 (1.64%)	1 (0.61%)	9 (0.86%)
Employment Status						
Full time employee	1 (2.44%)	14 (6.57%)	24 (7.32%)	27 (8.85%)	8 (4.91%)	74 (7.05%)
Part time employee	5 (12.20%)	1 (0.47%)	5 (1.52%)	5 (1.64%)	4 (2.45%)	20 (1.90%)
Self employed	17 (41.46%)	135 (63.38%)	153 (46.65%)	96 (31.48%)	77 (47.24%)	478(45.52%)
Retired/ pensioned	0 (0.00%)	1 (0.47%)	4 (1.22%)	9 (2.95%)	3 (1.84%)	17 (1.62%)
Housewife	0 (0.00%)	19 (8.92%)	9 (2.74%)	7 (2.30%)	8 (4.91%)	43 (4.10%)
Student	6 (14.63%)	17 (7.98%)	46 (14.02%)	71 (23.28%)	30 (18.40%)	170(16.19%)
Unemployed	12 (29.27%)	26 (12.21%)	87 (26.52%)	90 (29.51%)	33 (20.25%)	248(23.62%)

Source: WVS (Ghana, 5th wave)

Appendix B: Goodness-of-Fit Indices

Table 4.17: Goodness-of-Fit Indices for the constrained Rasch model

Bootstrap Goodness-of-Fit using Pearson chi-squared:

Tobs: 84.45

data-sets: 200

p-value: 0.005

Fit on the Two-Way Margins:

Response: (0,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	5	33	16.92	15.27 ***
2	3	4	18	8.33	11.23 ***
3	2	3	23	11.83	10.56 ***

Response: (1,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	5	127	141.04	1.40
2	2	3	64	73.55	1.24
3	3	4	55	63.21	1.07

Response: (0,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	5	54	68.46	3.05
2	2	5	73	84.70	1.62
3	1	5	10	13.78	1.04

Response: (1,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	5	1320	1307.58	0.12
2	2	5	1301	1291.33	0.07
3	2	3	1363	1355.24	0.04

'***' denotes a chi-squared residual greater than 3.5

Source: WVS (Ghana, 5th wave)

Table 4. 18: Goodness-of-Fit for the two-parameter logistic model

Fit on the Two-Way Margins

Response: (0,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	1	5	8	5.06	1.71
2	1	4	1	2.52	0.92
3	4	5	18	22.50	0.90

Response: (1,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	4	5	142	137.21	0.17
2	2	3	64	61.67	0.09
3	1	5	152	154.65	0.05

Response: (0,1)

	Item I	Item j	Obs	Exp	(O-E)^2/E
1	1	5	10	13.02	0.70
2	4	5	55	50.65	0.37
3	1	2	16	14.44	0.17

Response: (1,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	4	5	1319	1323.63	0.02
2	1	5	1364	1361.26	0.01
3	2	3	1363	1365.33	0.00

Source: WVS (Ghana, 5th wave)

Table 4.19: Goodness-of-Fit for the three-parameter logistic model

Fit on the Two-Way Margins

Response: (0,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	4	18	11.66	3.55
2	2	3	23	17.46	1.76
3	4	5	18	21.50	0.57

Response: (1,0)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	4	55	62.31	0.86
2	2	3	64	70.76	0.65
3	2	4	57	59.64	0.12

Response: (0,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	4	69	76.56	0.75
2	2	3	84	90.99	0.54
3	4	5	55	52.47	0.12

Response: (1,1)

	Item i	Item j	Obs	Exp	(O-E)^2/E
1	3	4	1392	1383.47	0.05
2	2	3	1363	1354.79	0.05
3	2	4	1370	1365.91	0.01

Source: WVS (Ghana, 5th wave)

Appendix C: Questions Selected for Analysis

Life Related Questions

V4. Family important

V5. Friends important

V6. Leisure time

V8. Work important

V22. How satisfied are you with your life

Value Related Questions

V202. Justifiable: homosexuality

V203. Justifiable: prostitution

V204. Justifiable: abortion

V205. Justifiable: divorce

V207. Justifiable: suicide

Politics Related Questions

V96. Political action: signing a petition

V97. Political action: joining in boycotts

V148. Having a strong leader

V149. Having experts make decisions

V150. Having the army rule

Income Related Questions

V51. It´s humiliating to receive money without having to work for

V68. Satisfaction with the financial situation of household

V106. Increase in taxes if extra money used to prevent environment

V251. Family savings during past year

V253. Scale of incomes

Democracy Related Questions

V137. Confidence: Justice System

V154. Democracy: People choose their leaders in free elections.

V160. Democracy: People can change the laws in referendums.

V161. Democracy: Women have the same rights as men.

V162. Importance of democracy

Socio-Demographic Features

V55. Marital status

V235. Sex

V237R3. Age of respondent - 3 intervals

V238. Highest educational level attained

V241. Employment status

Appendix D: Data Used for Analysis (Part)

v4	v5	v6	v8	v22	v51	v68	v96	v97	v106	v137	v148	v149	v150	v154	v160	v161	v162	v202	v203	v204	v205	v207	v251	v253
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1

Source: WVS (Ghana, 5th wave)