UNIVERSITY OF CAPE COAST

EFFECT OF MEASUREMENT SCALES ON RESULTS OF ITEM RESPONSE THEORY MODELS AND MULTIVARIATE STATISTICAL TECHNIQUES

BY

ARIMIYAW ZAKARIA

Thesis submitted to the Department of Statistics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial fulfilment of the requirements for the award of Doctor of Philosophy degree in Statistics

JULY 2018

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature .... Date 27/02/2019

Name: Arimiyaw Zakaria

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature .... Date 28/02/2019

Name: Dr. Bismark Kwao Nkansah

Co-Supervisor's Signature .... Date 27/02/2019

Name: Dr. Nathaniel Kwamina Howard

ABSTRACT

The study investigates the effects of response scales of items on results of item response theory (IRT) models and multivariate statistical techniques. A total of sixty-four datasets have been simulated under various conditions such as item response format, number of dimensions underlying response scales, and sample size using R package MIRT command: *simdata* (*a, d, N, itemtype*). Two main statistical techniques – IRT models and Factor Analysis – are employed in analysing the simulated datasets using standard R 3.4.3 codes. We find that there is a direct relationship between parameters of IRT and those of factor models, particularly item discrimination and factor loadings. The results also show that the overall fitness of the item response model increases with increasing scale points for higher dimensionality and sample size 150 and higher. The fitness deteriorates over increasing scale points for small sample sizes for unidimensional IRT model. Again, the number of influential indicators on factors increases with increasing scale-points, which improves the fitness of the model. The results indicate that unrealistic factor solution may be obtained if we attempt to extract higher factor solution than the underlying dimensionality on few scale-points with higher sample sizes. The study suggests that a five-point response scale gives most reasonable results among various scales examined. IRT analysis is recommended as a preliminary process to ascertain the observed features of items. The study also finds a sample size of 150 as adequate for a most plausible factor solution, under various conditions.

KEY WORDS

Dimensionality

Factor model

Item response theory

Likert scale

Sample size

Scale points

ACKNOWLEDGEMENTS

My sincere gratitude and appreciation go to my Principal Supervisor, Dr. Bismark Kwao Nkansah of the Department of Statistics, University of Cape Coast (UCC), for his invaluable scholarly suggestions, comments and guidance throughout the preparation and writing of this thesis. I am indeed very grateful. I am also grateful to my Co-Supervisor, Dr. Nathaniel Kwamina Howard of the Department of Statistics, UCC, for his contribution towards the success of this work and ensuring the successful completion of my programme.

I wish to thank all lecturers and staff of the Department of Statistics, and the Department of Mathematics, UCC, for their immense contribution to the success of this work.

Finally, I would like to express my gratitude to my family and friends for their unflinching support throughout my programme.

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

In this chapter, the background of the study will be explored. It will highlight the main motivation of the study, and introduce various works on item response theory and factor analysis which are further examined in Chapter Two. The background study will guide the statement of the problem, which will in turn guide the objectives of the study. A number of datasets have been used in this study, which are mainly simulated. A brief description of these datasets will be given in this chapter. The organisation of the rest of the thesis is outlined as the last section of this chapter.

Background to the Study

Measurement can be defined in several different ways, depending on the context and the particular field of study. Measurement entails the assignment of numbers (or labels) to persons or objects in a systematic manner based on the degree to which they possess some characteristic (Blerkom, 2009; de Ayala, 2009). One approach to evaluating the quality of measurements is to use their reliability and validity. That is, the measurements that should be used are the ones that are most reliable and valid. Another approach to evaluating measurements involves specific properties such as distinctiveness, ordering, equal intervals and absolute zero (Allen & Yen, 1979). These properties relate to how well the measurements represent the characteristic being measured. They are used in determining the level of measurement – nominal, ordinal, interval, or ratio – and are contained in a framework of scaling theory. Scaling theory focuses on techniques for determining what numbers should be used to represent the degree of the characteristic being measured. A scale is an organised set of measurements, all of which measure one characteristic or ability. That is, scales yield numbers that represent the characteristics or abilities of the individuals they measure. The number assigned to a particular individual is the scale value.

Scaling theory describes the properties of the scales in terms of their levels of measurement. A scale's level of measurement is determined by the type of transformation that will maintain the scale's representation of the ability being measured. For instance, scales that reach ordinal level of measurement have large numbers assigned to objects with more of the characteristic being measured than to objects with less of that characteristic. Once the scale has been assigned, it can be transformed in any way as long as the correct ordering of the scale values is preserved. Such transformations under ordinal scale is monotonic as it does not affect the relative order of the scale values – for example, adding a constant or multiplying by a positive number. The transformation that maintains the correct representation of the ability defined by the scale is identified using a scaling model, which is a symbolic representation of the relationship between the ability being scaled and a set of observations, such as response scores. For a scaling model to be useful, it must fit a set of observations. When a model fits a set of observations, it will determine which scale value should be assigned to each observation. Most scaling models have been developed for obtaining interval and ratio scales (Allen & Yen, 1979).

Item response theory (IRT) provides mathematical techniques for performing measurement in which the ability being measured is considered to be continuous in nature (de Ayala, 2009). IRT models assume that the ability being scaled has a normal distribution and that the observed scores (e.g., item responses) are monotonically related to the ability being measured. IRT models express the association between an individual's response to an item and the underlying latent variable (ability) being measured by the instrument (questionnaire) (Reeve, 2002). IRT uses latent characterisations of individuals and items as predictors of observed responses. Thus, a person's response to an item is influenced by the characteristics of the individual and by the characteristics of the item. The IRT describes, in probabilistic terms, how a person with higher ability level is likely to provide a response in a different response category in relation to a person with a low ability level (Ostini & Nering, 2006; de Ayala, 2009). Each item is characterised by one or more model parameters: discrimination (α), location (δ), and guess (c) parameters.

The discrimination parameter expresses an item's capacity to differentiate between persons who have high ability levels from persons who have low ability levels. This capacity to differentiate among people with different locations may be held constant or allowed to vary across items. The α value indicates the relevance of the item to the ability being measured by the questionnaire. An item with a positive α value is, at least somewhat, consistent with the underlying ability (trait) being measured, and a relatively large α value indicates a relatively strong consistency between the item and the underlying ability. In contrast, an item with a discriminating value of zero is unrelated to the underlying ability being measured, and an item with a negative α value is inversely related to the underlying ability. Thus, it is generally desirable for items to have a large positive discrimination value. Determining the item's discrimination value is particularly essential in identifying the group of individuals that are most typical to respond to items in a given study.

The item location parameter, δ commonly referred to as the item difficulty parameter, shows the position of an item on the ability scale. Item difficulty is an indication of the level of the underlying ability that is needed to respond in a certain way to the item (Osteen, 2010). An item with low (or negative) δ value is considered to be "easy", and persons with low ability levels have a tendency to respond positively (e.g., responding Yes on a dichotomous item) to it. Conversely, an item with a positive (and large) δ value is considered to be "difficult" and persons with high ability levels tend to respond favourably to it. Responding positively, favourably, or endorsing an item literally means that a person's response to the item is consistent with the direction of the item's expected response. In IRT, persons and items are located on the same continuum. That is, the item location and the person's ability level (θ) are indexed on the same metric. In this case, when a person's ability level is higher than the item location on the continuum, that person is more likely to provide a positive (favourable) response (Ostini & Nering, 2006).

The guess parameter represents the chance of persons with low ability responding favourably to an item. It is incorporated into an IRT model to account for responses at the lower end of the ability continuum. This applies to situations where guess is a factor in responses on selected (e.g., multiple choice) items (Hambleton, Swaminathan, & Rogers, 1991).

The measurement and analysis of dependence between variables, between sets of variables, and between variables and sets of variables are fundamental to multivariate statistical techniques (Anderson, 2003). Multivariate statistical techniques often involve modelling relationships among variables, and for exploring patterns that may exist in one or more dimensions of datasets (Timm, 2002). Factor analysis is a widely used multivariate statistical technique for measurement of unobservable constructs. It has been applied in this study as the main multivariate statistical technique due to its relevance. The technique is designed to determine the number of distinct constructs (abilities) needed to account for the pattern of correlations among a set of measures (indicator variables), for example, Likert-type responses. These unobservable abilities (common factors) are assumed to account for the structure of correlations among the indicator variables. The factor structure provides information about the number of common factors underlying a set of indicators. They also make available information to facilitate in interpreting the nature of these factors by providing estimates of the influence (factor loadings) each factor exerts on each of the indicators being assessed (Fabrigar & Wegener, 2012). The goal of factor analysis is to obtain a relatively parsimonious representation of the structure of correlations. In this case, the number of common factors needed to account for the correlations among the indicators is considerably less than the number of indicators. The factor model also assumes that each indicator variable is influenced by a unique factor, which represents that portion of the score on an indicator variable that is not accounted for by the common factors. These unique factors are restricted to only a single indicator in the model and cannot be used to explain the correlations among indicator variables.

In many instances, several challenges are faced in the application of factor analysis. Firstly, it is important to determine if the factor model is appropriate for the data. In this case, it is necessary to decide if the objectives of the study are adequately addressed by the model, and if the data satisfies the assumptions of the model. Secondly, it must be determined if the data is adequately represented by a single-factor, two-factor, or multiple-factor model. Other challenges include the procedure to use in estimating the parameters of the specified factor model, and interpretation of the results of the analysis.

Scales of measurement are quite useful in determining the appropriateness of use of certain statistical analyses. Scale of measurement can have implications for the meaningfulness of the analysis. That is, some standard statistical procedures should be used only with measurements that are interval or ratio, but not with nominal or ordinal (Furr & Bacharach, 2013). Parametric statistics are often valid only when interval or ratio data are used (Cohen, 2001).

Statement of the Problem

Modelling the relationship between item responses and the characteristics of persons falls under the realm of item response theory (IRT) models. The

IRT models are quite useful in the construction of scales (e.g., Likert scale) for measuring latent constructs of persons. The soundness of IRT results is often affected by several issues. An important issue to consider when designing Likert scale items is the optimal number of response categories. Considering reliability and validity, Jacoby and Matell (1971) attempted to determine the number of response alternatives to use in the construction of Likert-type scales. They indicated that both reliability and validity are independent of the number of scale points used for Likert-type items. They suggested that two or three-point Likert scales are good enough. Martin (1973) studied the effects of varying the number of scale points on the correlation coefficient using the bivariate normal distribution. Martin argued that the correlation coefficient generally decreases as the number of response categories becomes smaller, and suggested the use of ten to twenty points on a scale. Performances of IRT models have been studied only for specific scales. Results have rarely been compared on different scales. This study will examine the optimal number of scale points to consider when conducting IRT and factor analysis.

IRT results has been found to be highly influenced by sample size. Notably, the problem of estimation of item parameters has a link with sample size. In other words, how large a sample to be used in IRT analysis will depend on how many item parameters to be estimated. For complex IRT models that requires estimation of more parameters, sample size should increase accordingly. The task of determining minimum sample size has been attempted by some researchers through simulation studies. Reise and Yu (1990) estimated the parameters of the graded response (GR) model, and recommended that a sample size of at least 500 is required to achieve adequate estimation under GR model. For Rasch item response model, useful information can be obtained from samples as small as 100 and sample sizes of 500 are more than adequate in estimating item parameters (de Ayala, 2009). Other varying opinions and findings have been observed (e.g., Stone, 1992; Osteen, 2010) regarding the suitability of the sample size for reasonable results in IRT models.

Factor analysis, undoubtedly, an important multivariate statistical technique, is also widely applied in analysing questionnaire items. Within the context of factor analysis, individual items typically represent indicator variables, and the latent abilities that the questionnaire seeks to measure represent the factors. The factor analysis model is based on three basic assumptions about the indicator variables – normality, constant variance and linearity. The indicator variables are also considered to be measured on at least the interval scale. When these assumptions are satisfied, the usual Pearson product-moment correlation coefficient provides a reliable measure of the extent of correlation between each pair of indicator variables, and the linear factor model reasonably fits the data.

However, a major concern in the literature (e.g., van der Eijk & Rose, 2015) has to do with the factor analysis of item responses from questionnaires. Item responses give categorical data, which suggest a violation of the continuous nature of the indicator variables. The implication is that the Pearson correlations between pairs of indicator variables in this case are less reliable and is a potential source of distortions in the factor structure. The severity of the distortions tend to increase as the number of response categories on the items decreases (Comrey & Lee, 1992). The unreliability of items may also contribute to difficulties with rotation of factors to obtain independent clusters, an incidence which is mostly due to the overlap in the content of items. As a remedy, Ferrando and Lorenzo-Seva (2013) recommended the use of tetrachoric correlations for factor analysis of dichotomous response data. For factor analysis of ordered polytomous data, it is recommended to use polychoric correlations.

Problems are also found to be connected to non-linear relations between items, which violates the assumption of linearity and normality underlying factor analysis. The non-linear relation leads to the problem of significant univariate skewness, univariate and multivariate kurtosis, and “difficult factors”, where items with similar distributions tend to form factors irrespective of their content.

This research attempts at examining the influence of the number of points on the response scales of items on the results of IRT and how it translates into suitable factor structure. Motivated by the literature in the area, the study is carried out using tetrachoric and polychoric correlations. Since results on optimal sample size for IRT has been inconsistent, the study will also investigate the effect of sample size on the factor structure.

Objectives of the Study

The main objective of the study is to examine the effect of measurement scales on the results of item response theory models and multivariate statistical techniques.

Specifically, the study seeks to:

1. examine the relationship between IRT and Factor Analysis models.
2. assess the effect of scale points on IRT results.
3. examine the effect of sample size on the results of IRT models.
4. investigate the effect of scale points on Factor Analysis results.
5. examine the effect of sample size on the results of Factor Analysis models.

Description of Datasets Used in the Study

Several datasets have been used in the thesis to study the effects of measurement scales on results of item response theory and factor analysis models. The first dataset, which is empirical and contains ten brooding items, is used in Chapter Three to study the graphical properties of IRT models. Other datasets have been simulated under various conditions and used in Chapter Four to address the objectives of the study. In this section, we provide a description of these datasets.

Brooding scale dataset

The dataset contains ten dichotomous items on brooding scale. It emanated from the responses of 2,569 females in a clinical group. Table 1 displays the estimated parameters for the ten items in the brooding scale.

Table 1: Estimated Parameters for Brooding Scale

		Parameters	
Item	Description	$\hat{\alpha}$	$\hat{\delta}$
1	Periods when I couldn't "get going"	1.95	-0.02
2	I wish I could be as happy as others	2.46	-0.15
3	I don't seem to care what happens to me	2.20	1.33
4	Criticism or scolding hurts me terribly	1.03	-0.26
5	I certainly feel useless at times	2.42	-0.03
6	I cry easily	1.11	-0.23
7	I am afraid of losing my mind	1.71	0.75
8	I brood a great deal	1.84	0.93
9	I usually feel that life is worthwhile	1.84	1.24
10	I am happy most of the time	2.83	0.25

Source: Reeve, 2002

Simulated datasets

These datasets consist of responses to twenty items of different response scales, namely two-point, three-point, five-point, and seven-point scales. They are generated using specified item parameter values of a given IRT model. Also, the datasets are simulated under various sample sizes such as 30, 100, 150, 200, 500, 800, and 1000. In addition, different dimensions of underlying person-ability are considered, particularly unidimensional, two-dimensional and three-dimensional. Further details of the description of simulated datasets are done in Chapter Four.

Organisation of the Thesis

This thesis is divided into five chapters under the headings: Introduction, Literature Review, Research Methods, Analysis and Results, and Summary, Conclusions and Recommendations.

The first chapter is the introduction of the thesis. It presents the background to the study, statement of the problem, objectives, and description of datasets used in the study. In the background, measurement scales, and the techniques of IRT and factor analysis are introduced. Next is the statement of the problem, where a number of problems associated with both techniques are highlighted. It is followed by the objectives of the study.

The literature review is presented in Chapter Two. It describes some studies already made in the application of IRT and factor analysis of items.

Chapter Three entails a review of key concepts and methods used in IRT and factor analysis. The chapter also presents two measures of correlation coefficients – tetrachoric and polychoric. The presentation of simulation, analysis of data, and results of the study are done in Chapter Four. The chapter describes in detail the simulation and analyses of datasets employed in the study. The chapter presents summaries of results in this study in the form of tables and figures. The major findings in this study are then discussed in relation to results from similar and related research. Chapter Five is the last chapter of this thesis. It encompasses the summary of all the major findings and presents them with reference to the objectives of the study. Conclusions emanating from the findings are outlined. Recommendations are also made based on the findings and on issues that require further study.

Chapter Summary

The chapter presents the background to the study, statement of the problem, objectives, outline of the thesis. The background of the study revealed that measurement scales determine what numbers should be used to represent the degree of the characteristic or ability being measured. Typically, responses to items on questionnaires can be classified under various measurement scales. For instance, Likert-type data constitute ordinal scale of measurement which are assumed to represent continuous unobservable characteristic or ability. It is noted that item response theory and factor analysis models are widely used statistical technique for measurement of continuous unobservable abilities. The statement of the problem indicated that results of these techniques are affected by various issues such as number of scale-points, sample size, dimensionality, number of items/indicators, and type of correlation matrix input. This study will examine the influence of the number of points on the response scales of items on the results of IRT and how it translates into suitable factor structure. It will also investigate the effect of sample size on the factor structure.

CHAPTER TWO

LITERATURE REVIEW

Introduction

The study investigates the effects of measurement scales on results of item response theory models and correlation-based multivariate techniques. This chapter presents a review of studies already made in the application of IRT and factor analysis of item responses. The chapter is structured into three main themes: (1) studies pertaining to IRT analysis of items, (2) studies relating to factor analysis of items, and (3) studies that compare the results of IRT and factor analyses of items. In what follows, we present a review of studies concerning IRT analysis of item responses. The next concentrates on factor analysis of items.

IRT Analysis of Items

Masters (1974) investigated the relationship between number of response categories employed and internal-consistency reliability of Likert-type questionnaires. The results indicated that in situations where low total score variability is achieved with a small number of categories, reliability can be increased through increasing the number of categories employed. In situations where opinion is widely divided toward the content being measured, reliability appeared to be independent of the number of response categories. Dodeen (2004) investigated the effect of item parameters on the item-fitness statistics using simulated data. Nine datasets were simulated using a sample size of 1000, 50 items, three levels of item discrimination, three levels of item difficulty and three levels of guess parameter. Results showed that item discrimination and guess parameters affected item-fitness. That is, as the level of item discrimination or guess parameter increased, item-fitness values increased, resulting in many items not fitting the model. The level of item difficulty did not affect the item-fitness statistic.

Koch (1983) applied two-parameter graded response latent trait model to data collected from a conventionally constructed Likert-type attitude scale. Comparisons were made of both the person latent trait estimates and the item parameter estimates with their counterparts from the conventional scaling method. Also studied were the goodness-of-fit of the graded response model and the information function feature of the model indicating the precision of measurement at each level of the attitude trait continuum. The results demonstrated that the graded response model could be successfully used to perform attitude measurement for Likert scales. Maydeu-Olivares, Drasgow, and Mead (1994) compared two models with the same number of parameters, graded response model (a difference model) and partial credit model (a divide-by-total model), with the aim of investigating whether difference models or divide-by-total models should be preferred for fitting Likert-type data. The models were found to be very similar under the conditions investigated, which included scale lengths from 5 to 25 items (five-option items were used) and samples of 250 to 3,000. The results suggested that both models fit approximately equally well in most practical applications. Under two-parameter logistic (2PL) model, Stone (1992) found that with sample size of 500 or more and 20 or more items, both item difficulty and discrimination parameters are generally stable and precise. Smith, Schumacker, and Bush (as cited in Osteen, 2010) examined the fitness of items using the mean square (MSQ) statistic and provided the following guidelines for sample size: misfit is evident when MSQ values are larger than 1.3 for samples less than 500, 1.2 for samples between 500 and 1,000, and 1.1 for samples larger than 1,000 respondents.

Fitzpatrick et al. (1996) compared the performances of one-parameter and two- parameter partial credit (1PPC and 2PPC) models using four real and four simulated datasets. The study included two sets of items: constructed-response (CR) items (i.e., open-ended questions), and multiple-choice (MC) items. In the study, where MC items were present, the partial credit models were combined with the one-parameter and three-parameter logistic (1PL and 3PL) models, respectively. Analyses of the real datasets showed that the 2PPC model alone or in combination with the 3PL model provided uniformly better fitness than did the 1PPC model used alone or in combination with the 1PL model. Also, IRT statistics for the real dataset indicated that the discriminations of MC and CR items differed substantially from one another, and that within item type they differed also. The authors noted that the poorer fit performance by the lPPC model alone or in combination with the 1PL model is likely produced by the considerable variability in item discrimination, as well as the guess on the MC items. In the simulation study, the percentages of items with good fitness tended to be larger when the 3PL-2PPC model combination was used. Also, this model combination tended to produce better item fitness across datasets with dissimilar properties.

Following the work of Fitzpatrick et al. (1996), Sykes and Yen (2000) conducted IRT scaling for six tests with mixed item formats. These tests differed in their proportions of constructed response (CR) and multiple choice (MC) items and in overall difficulty. One-parameter (1PPC) or two-parameter (2PPC) partial credit model was used for the CR items and the one-parameter logistic (1PL) or three-parameter logistic (3PL) model for the MC items. The study indicated that substantial number of items were not fitted by the 1PL/1PPC model as compared to the 3PL/2PPC model when item response data from six mixed-item-format tests, varying in difficulty, were analysed. The smallest percentage of items that were not fitted by the Rasch model was 33% compared to a maximum of 5% of the items that misfit the generalised model. The results also showed that the magnitude of 3PL/2PPC discrimination parameter estimates clearly decrease as the number of levels of the CR items increase. A 1PL/1PPC model constrains item discriminations to be equal. Sykes and Yen (2000) argued that by not allowing item discriminations to decrease with increasing numbers of score levels, the Rasch model can spuriously inflate its representation of the information contributed by CR items, with the magnitude of the inflation likely to increase with an increase in the number of item score levels. Again, items fitness was substantially worse with the combination IPLI/PPC model than the 3PL/2PPC model due to the former's restrictive assumptions that there would be no guess on the MC items, equal discrimination across items, and item types. Information for some items with summed ratings were usually over-estimated by 300% or more for the 1PL/1PPC model.

DeMars (2012) assessed how violations of the normality assumption impact the item parameter (i.e., discrimination and difficulty) estimates and factor correlations. For skewed and platykurtic latent variable distributions, three methods were compared in structural equation modelling package, Mplus – limited-information (LI), full-information (FI) integrating over a normal distribution, and FI integrating over the known underlying distribution. Dichotomous item responses were simulated to follow a two-parameter normal ogive MIRT model. Two factors were simulated with correlations of 0.5 or 0.8, and having the same distribution, either skewed negative or platykurtic. Responses to 44 items were simulated, each item measuring only one factor (22 items measured only Factor 1, and the other 22 items measured only Factor 2), and sample size of 300 or 3000 examinees. The results showed that for the platykurtic distribution, estimation method made little difference for item parameter estimates. When the latent variable was negatively skewed, for the most discriminating easy or difficult items, LI estimates of both parameters were considerably biased. Full-information estimates obtained by marginalising over a normal distribution were somewhat biased. Full-information estimates obtained by integrating over the true latent distribution were essentially unbiased. For the α parameters, standard errors were larger for the LI estimates when the bias was positive but smaller when the bias was negative. For the δ parameters, standard errors were larger for the LI estimates of the easiest, most discriminating items. Otherwise, they were generally similar for the LI and FI estimates. Sample size did not substantially impact the differences between the estimation methods.

Mount and Schumacker (1998) used simulated dichotomous data to determine the effects of guess on Rasch item fitness statistics (weighted total, unweighted total, and unweighted between fitness statistics) and the Logit Residual Index (LRI). The data were simulated using 100 items, 100 persons, three levels of guess (0%, 25%, and 50%), and two item difficulty distributions (normal and uniform). The results of the study indicated that no significant differences were found between the mean Rasch item fitness statistics for each distribution type as the probability of guessing the correct answer increased. The mean item scores differed significantly with uniformly distributed item difficulties, but not normally distributed item difficulties. The LRI was more sensitive to large positive item misfit values associated with the unweighted total fitness statistic than to similar values associated with the weighted total fitness or unweighted between fitness statistics. The greatest magnitude of change in LRI values (negative) was observed when the unweighted total fit statistic had large positive values greater than 2.4. The LRI statistic was most useful in identifying the linear trend in the residuals for each item, thereby indicating differences in ability groups (i.e., differential item functioning).

Rogers and Hattie (1987) investigated the behaviour of several person and item fitness statistics commonly used to test and obtain fitness to the one-parameter item response model. The sensitivity of the total-*t*, mean-square residual, and between-*t* fitness statistics to guess, heterogeneity in discrimination parameters, and multidimensionality was examined using simulated data for 500 persons and 15 items. Additionally, 25 misfit persons and a misfit item were generated to test the power of the three fit statistics to detect deviations in a subset of observations. Neither the total-*t* nor the mean-square residual were able to detect deviation from any of the models fitted. The use of these statistics appeared to be unwarranted. The between-*t* was a useful indicator of guess and heterogeneity in discrimination parameters, but was unable to detect multidimensionality. These results show that the use of person and item fitness statistics to test and obtain overall fitness to the one-parameter model can lead to acceptance of the model even when it is grossly inappropriate. Assessments of model fitness based on this strategy are inadequate.

Smith (1988) investigated the distributional properties of the standardised residuals used in estimating Rasch model's parameters when the data fit the model. The author also investigated the power of the standardised residual to detect measurement disturbances. The study was based on simulated data to control for the presence of confounding factors, such as multidimensionality, differences in the slopes of item characteristic curves, and guess. The results indicated that when the data fit the model, the distributional properties of the standardised residuals were close to hypothesised mean and standard deviation and that it is possible to construct reasonable Type I error rates that can be used as a frame of reference when investigating the fitness of actual data to the Rasch model. The analysis of the simulated measurement disturbance data indicated that although the shape of the standardised residual distribution reacts to the presence of the disturbance, the magnitude of the response is small and the residuals lack the power of the item or person fit statistics to detect measurement disturbances.

McKinley and Mills (1985) conducted a study to evaluate four goodness-of-fit procedures in item response theory using data simulation techniques. The procedures were evaluated using data generated according to three different item response theory models and a factor analytic model. Three different distributions of ability were used, as were three different sample sizes. It was concluded that the likelihood ratio Chi-square procedure yielded the fewest erroneous rejections of the hypothesis of fitness, whereas Bock's Chi- square procedure yielded the fewest erroneous acceptances of fitness. It was found that sample sizes between 500 and 1,000 were best. Shifts in the mean of the ability distribution were found to cause minor fluctuations, but they did not appear to be a major issue.

Factor Analysis of Items

An issue to consider when conducting factor analysis is the characteristics of the sample from which the measurements of the indicator variables are taken. Obviously, an aspect of the sample that is worth considering is how large the sample should be in order to perform factor analysis. Correlations are less reliable when estimated from small samples (Tabachnick & Fidell, 2013). Gorsuch (1974) puts it bluntly that " no one seems to know exactly where a large *n* begins and a small *n* leaves off". Comrey and Lee (1992) noted that as the sample size increases, the reliability of the obtained correlations increases. They found that samples of size 50 give very inadequate reliability of correlation coefficients, while samples of size 1000 are more than adequate for factor analysis. With regards to evaluating the adequacy of the sample size, Comrey and Lee (1992) provided some guidelines: 50 is very poor, 100 is poor, 200 is fair, 300 is good, 500 is very good, and 1000 or greater is excellent. Other researchers are of the view that under optimal conditions (communalities of 0.70 or greater and 3 to 5 indicator variables loading on each factor), a sample of size 100 can be adequate; under moderately good conditions (communalities of 0.40 to 0.70 and at least 3 indicators loading on each factor), a sample of at least 200 should suffice; and under poor conditions (communalities lower than 0.40 and some factors with only two indicator variables on them), samples of at least 400 might be necessary

(Fabrigar & Wegener, 2012; Tabachnick & Fidell, 2013; MacCallum, Browne, & Sugawara, 1996).

Muthén and Kaplan (1985) considered the problem of applying factor analysis to non-normal categorical variables. A Monte Carlo study is conducted where five prototypical cases of non-normal variables are generated. Two normal theory estimators, maximum likelihood (ML) and generalised least squares (GLS), were compared to the asymptotically distribution-free (ADF) estimator. A categorical variable methodology (CVM) estimator was also considered for the most severely skewed case. Results showed that ML and GLS Chi-square tests were quite robust but obtain too large values for variables that were severely skewed and kurtotic. ADF, however, performed well. Parameter estimate bias appeared non-existent for all estimators. Results also showed that ML and GLS estimated standard errors were biased downward. For ADF, no such standard error bias was found. The CVM estimator appeared to work well when applied to severely skewed variables that had been dichotomised. ML and GLS results for kurtosis-only showed no distortion of Chi-square or parameter estimates and only a slight downward bias in estimated standard errors.

Babakus, Ferguson, and Jöreskog (1987) used a simulation design to study the sensitivity of maximum likelihood (ML) factor analysis to violations of measurement scale and distributional assumptions in the input data. Product-moment, polychoric, Spearman's rho, and Kendall's tau correlations computed from ordinal data were used to estimate a single-factor model. The resulting ML estimates were compared on the bases of convergence rates and improper solutions, accuracy of the loading estimates, fitness statistics, and estimated standard errors. Results showed that, for large samples (n = 500), all replications converged and the solutions were proper with both continuous and discrete data. In small samples (n = 100) with the larger loading vector (0.8, 0.8, 0.8, 0.8), all continuous cases converged and the solutions were proper. Though all small sample with large loading cases converged, there were three improper solutions. All three occurred with the polychoric correlation. Non-convergence and improper solutions occurred with small samples ($n = 100$) and smaller loading vector (0.4, 0.6, 0.6, 0.8) for both continuous and discrete cases. For continuous replications, there were four non-convergent cases and a total of 124 improper solutions (2%). When the same data were categorised, 43 non-convergent cases and 239 improper solutions were obtained (4%). Most of the non-convergent (44%) and improper solutions (60%) occurred when polychoric correlations were used as input. Generally, on the basis of convergence rates and improper solutions, Kendall's tau out-performed the other three measures, followed by the product-moment and Spearman's rho which produced similar results. The study revealed that, the polychoric correlation out-performed other measures on both the categorisation bias and squared error criteria. The product-moment correlation produced the second best overall results, followed by Spearman's rho and Kendall's tau. On the basis of estimated pairwise correlations, factor loadings and standard errors, the polychoric correlation gave consistently better estimates, but performed worst on all goodness-of-fit criteria.

Finch (2006) compared the ability of two commonly used methods of rotation in factor analysis, Varimax and Promax, to correctly link items to factors and to identify the presence of simple structure. Results suggested that the two approaches are equally able to recover the underlying factor structure, regardless of the correlations among the factors, though the Promax method is better able to identify the presence of a simple structure. The results further suggested that for identifying which items are associated with which factors, either approach is effective, but that for identifying simple structure when it is present, the Promax method is preferable.

Tate (2003) compared a number of common methods for assessing dimensionality in item response data, including the unweighted least squares (ULS), robust weighted least squares (RWLS), and a full information method using the TESTFACT software. Tate simulated all items with guess parameter values of 0.2, samples of 2,000 examinees, and 60 items. The author found that exploratory factor analysis (EFA) with the oblique PROMAX rotation, using both TESTFACT and NOHARM, was able to recover item parameters under a variety of multidimensional structures. On the other hand, confirmatory factor analysis (CFA) using RWLS in Mplus demonstrated less than optimal item parameter recovery in all cases where guess was present in the data.

Dolan (1994) studied two estimators in the factor analysis of categorical items, the weighted least squares function implemented in LISREL 7 and a generalised least squares function implemented in LISCOMP. Dolan's main interest was the performance of these estimators in relatively small samples (200 to 400) and the comparison of their performance with the normal theory maximum likelihood estimator given an increasing number of response categories. The author evaluated the performance of these estimators based on the variability of the parameter estimates, the bias of the parameter estimates, the distribution of the parameter estimates and the χ^2 goodness-of-fit statistics. The results indicated that in the ideal circumstances, 200 is too small a sample size to justify the use of large sample statistics associated with these estimators.

Potthast (1993) examined the utility of a categorical variable methodology (CVM) for confirmatory factor analysis of ordinal variables. Multivariate normal data were generated according to four different factor models (4, 9, 15 and 22 parameters) for samples of 500 and 1000. Indicators were classified into five categories so that manifest variables displayed negative, zero, positive or highly positive kurtosis. Each of the 32 design cells was replicated 100 times. Parameter estimates exhibited little or no bias under any condition. Standard errors were under-estimated with respect to the standard deviation of the parameter estimates. This negative bias worsened as model size grew or as positive kurtosis increased; it was more severe for factor correlations than indicator loadings. Chi-square fitness statistics rejected the true model more often than expected for nine-parameter and larger models. Although variables with high positive kurtosis led to the greatest misfit in large models, fitness was poor even with variables of zero kurtosis. As expected, larger samples always yielded more accurate results.

Yang-Wallentin, Jöreskog, and Luo (2010) studied the behaviour of maximum likelihood methods such as unweighted least squares (ULS), maximum likelihood (ML), weighted least squares (WLS), or diagonally weighted least squares (DWLS) in combination with polychoric correlations when the models are misspecified. Yang-Wallentin et al. also studied the effect of model size and number of categories on the parameter estimates, their standard errors, and the common Chi-square measures of fit when the models are both correct and misspecified. Results showed that when used routinely, these methods give consistent parameter estimates, but ULS, ML, and DWLS give incorrect standard errors. The authors noted that correct standard errors can be obtained for these methods by robustification using an estimate of the asymptotic covariance matrix (W) of the polychoric correlations.

Parry and McArdle (1991) provided a comparison of four selected least-squares methods of factor analysis of binary data: (1) calculation of a matrix of phi coefficients, followed by fitting of a factor model using a minimum unweighted least-squares (ULS) procedure (ULS-PHI); (2) calculation of a matrix of tetrachoric correlations, followed by fitting of a factor model using a minimum ULS procedure (ULS-TC); (3) calculation of a matrix of tetrachoric correlations, followed by fitting of a factor model based on a weighted least-squares (WLS) factor extraction (LISCOMP); and (4) calculation of a product-moment correlation matrix using phi coefficients means, followed by fitting of a factor model using an approximation to a ULS (NORHAM). The study was done using simulated data, generated under varying sample sizes, threshold values, and population loadings of a factor model. The results showed that, the advantage of one method over another depends on the sample size, as well as on the combination of magnitude of the loading and the skewness of the data (threshold). Parry and McArdle noted that LISCOMP does not appear to work well for datasets of small sample size, and differences among the three remaining methods appear to be smallest when the data is not highly skewed and when loadings are of moderate size (0.7). The study further revealed that the estimates of population loadings using NOHARM and LISCOMP procedures were not markedly superior to those obtained from ULS-PHI, except when population loadings were high (0.9). Again, NOHARM did not perform better than ULS-TC, even when the data was more highly skewed. Parry and McArdle concluded that NOHARM and LISCOMP did not out-perform factor analysis using the tetrachoric and Phi correlation coefficients estimated from bivariate tables of the observed variables as input to the analysis.

Muthén (1984) proposed a structural equation model with a generalised measurement part, allowing for dichotomous, ordered categorical, and continuous indicator variables. A computationally feasible three-stage estimator is proposed for any combination of observed variable types. The author noted that, the proposed model is a three-stage, limited information, generalized least-squares (GLS) estimator, which gives large-sample Chi-square tests of model fit and large-sample standard errors of estimates. Muthén outlined that, the techniques makes it possible for GLS factor analysis with (mixtures of continuous and) ordered polytomous indicators, testing hypotheses of both correlation and level structures in multiple-group structural equation models, and multivariate structural regression with ordered categorical response variables.

Flora and Curran (2004) used Monte Carlo simulation methodology to empirically study the effects of varying latent response variable (y^*) distribution, sample size (n), and model size on the computation of Chi-square model test statistics, parameter estimates, and associated standard errors pertaining to CFAs fitted to ordinal data. The y^* distributions considered include a multivariate normal distribution and four non-normal distributions with varying skewness and kurtosis. Each dataset generated conformed to four model specifications that hold fo y^*: Model 1 consisted of a single factor measured by five ordinal indicators; Model 2 consisted of a single factor measured by ten indicators; Model 3 consisted of two correlated factors each measured by five indicators; and Model 4 consisted of two correlated factors each measured by ten indicators. After sampling continuous multivariate data from various distributions, the samples were transformed into two-category and five-category ordinal data. For each combination of y^* distribution and model specification, Flora and Curran generated random samples of four different sizes: 100, 200, 500, and 1,000. For each simulated sample of ordinal data, the authors calculated the corresponding polychoric correlation matrix and fit the relevant population model using both full and robust WLS estimation. The study showed that the polychoric correlation estimates tended to become positively biased as a function of increasing non-normality in the y^* distributions; however, mean relative bias (RB) remained under 10% in almost all cases. Although the correlation estimates were frequently positively biased, the centre of these distributions did not depart substantially from the population correlation value, even with y^* non-normality. Also, sample size did not have any apparent effect on the accuracy of the polychoric correlations, although there was a tendency for correlations calculated from two-category data to be slightly more biased than those calculated from five-category data. With sample size of 100, full WLS did not produce any solutions for Model 4 (due to non-invertible weight matrices). In general, the rates of improper solutions were greater in the two-category versus five-category condition. For Models 2 and 3, two-category data produced high rates of improper solutions with sample size of 100, whereas the rates were near zero in the five-category condition. Also, nearly 100% of replications of Model 4 were improper in the two-category condition where $n = 200$, whereas the corresponding rates in the five-category condition were only around 30%. Although the rates of improper solution obtained with full WLS varied somewhat across different y^* distributions, this variation did not appear to be systematically associated with degree of non-normality in y^*. At the two largest sample sizes ($n = 500$ and $n = 1{,}000$), full WLS estimation converged to proper solutions of all four models across 100% of replications. Both the Chi-square test statistics and their standard deviations tend to be positively biased across all cases of the study, particularly with full WLS estimation. This bias increases as a function of increasing number of indicators for a model and by model complexity. The effect of sample size on the inflation in Chi-square test values varies substantially with model specification. Within each of the four models, the Chi-square RB decreases as sample size increases, but this effect is more pronounced for larger models. In addition, there appears to be some indication that the Chi-square statistics are affected by non-normality in y^*.

Forero, Maydeu-Olivares, and Gallardo-Pujol (2009) conducted a simulated study to compare DWLS and ULS in estimating a factor analysis model with categorical ordered indicators under different settings of dimensionality, factor loading, sample size, number of items per factor, number of response alternatives per item, and item skewness. A total of 324 conditions per estimation method were investigated, using 1,000 replications for each setting. A full factorial design was used by crossing three sample sizes (200, 500, and 2,000 respondents); two levels of factor dimensionality (one and three factors); three test lengths (9, 21, and 42 items); three levels of factor loadings λ: low ($\lambda = 0.4$), medium ($\lambda = 0.6$), and high ($\lambda = 0.8$); and six item types (three types consist of items with two categories, and another three of items with five categories) that varied in skewness, kurtosis, or both. Results indicated that, on average, convergence rates (i.e. rates of plausible solutions) across the 324 conditions were 97.4% for DWLS and 96.4% for ULS. However, convergence rates differed depending on the number of indicators per dimension, item skewness, and sample size. Both estimators showed smaller convergence rates for models with only three indicators per dimension. In this setting, convergence rates were better for DWLS: Average convergence was 90.6% for DWLS versus 85.4% for ULS. When the number of indicators per dimension was seven or more, average convergence rates were similar (roughly 99%). Increasing skewness worsened convergence: When item skewness was greater than or equal to 1.5, average convergence was 96.4% for DWLS and 94.7% for ULS. When item skewness was below 1.5, convergence performance was, on average, similar across the methods (98%). Finally, sample size improved convergence rates.

Morata-Ramirez and Holgado-Tello (2013) compared four estimation methods: maximum likelihood (ML), robust maximum likelihood (RML), unweighted least squares (ULS), and robust unweighted least squares (RULS) according to two of the assumptions CFA is supposed to fulfil – multivariate normality, and the continuous measurement nature of both latent and observed variables. In the study, three conditions were manipulated: hypothesized model dimensions (3, 5 and 7 uncorrelated factors), sample size (250, 450, 650, 850), and items skewness (all items symmetric, all items asymmetric). Each sample of continuous and normally was generated with 9, 15 or 21 items (3, 5 and 7 dimensions, respectively) were categorised to a five-point scale. Results showed that when ULS or RULS methods were applied to symmetrical item distributions, Chi-square statistics for three-factor models were high for samples of 250 subjects, but not for the remaining sample sizes. In respect of ML and RML estimators, Chi-square statistics showed high values which were greater than the ones reported for RULS method. Chi-square value for three-factor models were high along the different sample sizes, while they are pretty high for five-factor models with 450 or 650 subjects and high for 850 subjects. For asymmetrical item distributions, when ULS and RULS estimators were considered, five and seven-factor models had highest Chi-square values for samples of 850 subjects. Concerning ML and RML estimation methods, Chi-square values were higher for five-factor models compared to three and seven-factor models regardless of the sample size. Morata-Ramirez and Holgado-Tello suggested that ULS and RULS are preferable as polychoric correlations help to overcome grouping and transformation errors produced when using Pearson correlations for ordinal observed variables.

Li (2016) carried out a Monte Carlo simulation study to compare the effects of different configurations of latent response distributions, numbers of categories, and sample sizes on model parameter estimates, standard errors, and Chi-square test statistics in a correlated two-factor model. Two estimation procedures, robust maximum likelihood (RML) and diagonally weighted least squares (DWLS), were used in the study. Factor loading was held constant at 0.7, with its corresponding uniqueness automatically set to 0.51, inter-factor correlation was set to 0.3 , and factor variances were all set equal to 1. Two latent distributions that varied in skewness and kurtosis were employed: (1) a slightly non-normal latent distribution with skewness = 0.5 and kurtosis = 1.5, and (2) a moderately non-normal latent distribution with skewness = 1.5 and kurtosis = 3.0. Four, six, eight, and ten categories were generated for each ordinal indicator within both the slightly and moderately non-normal latent distributions. Three different empirical sample sizes, 200, 500, and 1,000 were employed in this study. The study found that, the problems of improper solutions or non-convergence did not occur for both RML and DWLS, irrespective of the number of categories, level of latent distribution violations (slightly and moderately non-normal), and sample sizes. Factor loadings were, on average, underestimated by RML when ordinal data had only four response categories. Conversely, the factor loadings were slightly overestimated, on average by DWLS, and considered essentially unbiased, especially when the latent distribution is only slightly non-normal. Regardless of the number of categories, DWLS was consistently superior to RML for factor loading estimates. Generally, the discrepancy in overall performance between DWLS and RML became larger as the sample size increased. DWLS was better than RML in the overall quality of factor loading estimates from four to ten categories across different sample sizes, even when ordinal observed data were generated from a moderately non-normal latent distribution.

Rhemtulla, Brosseau-Liard, and Savalei (2012) compared the performances of robust normal theory maximum likelihood (ML) and robust categorical least squares (cat-LS) methodology for estimating confirmatory factor analysis models with ordinal variables. Data were generated from two models with two to seven categories, four sample sizes, two latent distributions, and five patterns of category thresholds. Results revealed that factor loadings and robust standard errors were generally most accurately estimated using cat-LS, especially with fewer than five categories; however, factor correlations and model fitness were assessed equally well with ML. Cat-LS was found to be more sensitive to sample size and to violations of the assumption of normality of the underlying continuous variables. Normal theory ML was found to be more sensitive to asymmetric category thresholds and was especially biased when estimating large factor loadings. Rhemtulla et al. recommended cat-LS for datasets containing variables with fewer than five categories and ML when there are five or more categories, sample size is small, and category thresholds are approximately symmetric. With six to seven categories, results were similar across methods for many conditions; in these cases, either method is acceptable.

Beauducel and Herzberg (2006) through simulation study compared maximum likelihood (ML) estimation with weighted least squares means and variance adjusted (WLSMV) estimation based on confirmatory factor analyses. The simulation study was performed for four different samples sizes (250, 500, 750, 1000), with four different numbers of variables (5, 10, 20, and 40 with 1, 2, 4, and 8 latent factors, respectively) and five numbers of categories in the variables (2, 3, 4, 5, and 6). The distributions of the variables were generated on the basis of a binomial distribution. It was found that WLSMV estimation performed as well as ML estimation across all sample sizes. For all sample sizes and for all number of categories, the mean size of the WLSMV factor loadings was closer to the continuous variables population loading (0.50 for the orthogonal case; 0.55 for the oblique case) than the mean size of the ML loadings. Generally, a clear superiority of WLSMV over ML estimation was found for categorical variables with two and three categories. Fitness indexes indicated superior model fitness when based on WLSMV and two and three categories. When based on five and six categories, there was no difference in ML and WLSMV, which means that the performance of the ML-based fitness assessment increased with five and six categories. There was, however, a clear tendency to underestimate the size of the factor loadings with ML estimation when the variables had only two or three categories. This tendency diminished with increasing number of categories, but even with six categories, there was a slight tendency to underestimate the magnitude of the loadings with ML estimation. The standard errors of the loadings were a bit smaller for WLSMV than for ML estimation across all number of categories. With four and five categories, the performance of WLSMV estimation was slightly superior to the performance of ML estimation, especially with respect to the bias of the loadings.

DiStefano (2002) investigated the impact of categorization on confirmatory factor analysis (CFA) parameter estimates, standard errors, and five ad hoc fitness indexes. Simulated datasets were generated under various conditions such as model size, sample sizes, and loading values. Two estimators, weighted least squares (WLS; with polychoric correlation input) and maximum likelihood (ML;

with Pearson product-moment input) were employed in the study. CFA results obtained from analysis of normally distributed, continuous data were compared to results obtained from five-category Likert-type data with normal distributions. Results indicated that, ML parameter estimates reported moderate levels of negative bias for all conditions, WLS standard errors showed high amounts of bias, especially with a small sample size and moderate loading values. With non-normally distributed, ordered categorical data, ML parameter estimates, standard errors, and factor inter-correlation showed high levels of bias.

van der Eijk and Rose (2015) undertook a systematic assessment of the extent to which factor analysis produces the correct number of latent dimensions (factors) when applied to ordered-categorical survey items (so-called Likert items). The authors simulated 2400 datasets of unidimensional Likert items that vary systematically over a range of conditions such as the underlying population distribution, the number of items, the level of random error, and characteristics of items and item-sets. Each of these datasets was factor analysed on the basis of Pearson and polychoric correlations. They found that, irrespective of the particular mode of analysis, factor analysis applied to ordered-categorical survey data very often leads to over-dimensionalisation. The magnitude of this risk depends on the specific way in which factor analysis is conducted, the number of items, the properties of the set of items, and the underlying population distribution.

Comparison of FA and IRT on Item Analysis

Forero and Maydeu-Olivares (2009) examined the performance of parameter estimates and standard errors in estimating graded response (GR) model across various conditions. The authors compared Full information maximum likelihood (FIML) with a 3-stage estimator for categorical item factor analysis (CIFA) when the unweighted least squares method was used in CIFA's third stage. They found that CIFA is much faster in estimating multidimensional models, particularly with correlated dimensions. Results further showed that, generally, CIFA yields slightly more accurate parameter estimates, and FIML yields slightly more accurate standard errors. FIML was found to be the best estimator in small sample sizes (200 observations). Again, CIFA was the best estimator in larger samples (on computational grounds). Forero and Maydeu-Olivares noted that both methods failed in a number of conditions, most of which involved 200 observations, few indicators per dimension, highly skewed items, or low factor loadings and these conditions are to be avoided in applications.

Maydeu-Olivares, Cai, and Hernández (2011) compared the fitness of an FA model and of an IRT model to the same dataset using test statistics based on residual covariances. The authors suggested that IRT and FA models yield similar fitnesses when applied to a binary dataset. On the contrary, for ordinal polytomous dataset, IRT models yielded a better fit because they involve a higher number of parameters. Maydeu-Olivares et al., however, noted that when fitness is assessed using the root mean square error of approximation (RMSEA), similar results are obtained again. They explained that these test statistics have little power to distinguish between FA and IRT models; they are unable to detect that linear FA is misspecified when applied to ordinal data generated under an IRT model.

Finch (2010) examined the ability of two confirmatory factor analysis models, specifically for dichotomous data, to properly estimate item parameters using common formulae for converting factor loadings and thresholds to discrimination and difficulty indices. The author considered unweighted least squares (ULS) and robust weighted least squares (RWLS) (MIRT estimation methods), and the unidimensional estimation approach which are implemented in software packages NOHARM , Mplus, and BILOGM G, respectively. Finch assessed these techniques in terms of the overall accuracy, bias, and standard error of item parameter estimates under a variety of sample sizes, test lengths, inter-trait correlations, pseudo-guess, and latent trait distribution conditions. The results indicated that performance of MPlus estimation was compromised, when guess (*c*) was present in the data, for both item discrimination and difficulty parameters, but such effect on bias was not seen with NORHAM. The author explained that, NOHARM provides *c* parameter estimates as it estimates item difficulty and discrimination, whereas such is not the case for MPlus. Again, the study found that estimates provided by both methods were influenced by the distribution of the latent traits, with larger standard errors in the skewed case for NOHARM and MPlus estimates of item difficulty and discrimination. For the unidimensional results produced by BILOGMG,, item difficulty bias is near 0 for the 60-item case, but has the largest such bias of the three approaches studied for 15 and 30 items. It was revealed that, there was greater precision in the discrimination estimates for larger sample sizes for both ULS and RWLS.

Knol and Berger (1991) used a simulation study to compare the ability of NOHARM, TESTFACT, standard principal factor analysis (based on tetrachoric correlations), and an MIRT parameter estimation approach to recover item parameter values. A total of 10 replications of each set of studied conditions were conducted, where the manipulated factors included sample size (250, 500, 1,000), number of items (15, 30) and number of dimensions (1, 2, 3). They reported that NOHARM and the standard factor-analytic approaches using the tetrachoric correlation performed as well as TESTFACT , and actually better than the MIRT estimation. De Bruin (2004) examined problems encountered in the factor analysis of items and demonstrated two methods that may be used to address these problems, namely the Rasch rating scale model, and the factor analysis of item parcels. The results showed that the Rasch rating scale model and the factoring of parcels produce superior results to the factor analysis of items.

Gosz and Walker (2002) conducted a Monte Carlo simulation in which they compared the ability of TESTFACT and NOHARM to estimate the probabilities of correct responses to a set of items for a group of simulated examinees. The authors assessed the performance of the methods by calculating root mean square deviation between the estimated and actual probabilities of correct responses for 2,500 examinees. Six different 40-item exams were simulated and replicated 100 times each. The exams differed in terms of the number of two-dimensional and unidimensional items that were generated. The correlation between the two latent traits was varied at 0.5, 0.75, and 0.9. Gosz and Walker found that when a test contained a large number of items associated with two factors, full information estimation using TESTFACT was better able to re-create examinees' response probabilities that matched those in the population than was the partial information approach carried out in NOHARM. In contrast, when fewer items exhibited this non-simple structure, NOHARM more accurately re-created item response probabilities across the examinees.

Asún, Rdz-Navarro, and Alvarado (2016) compared the performance of two approaches in analysing four-point Likert rating scales with a factorial model: the classical factor analysis (FA) and the item factor analysis (IFA). For FA, maximum likelihood (ML) and weighted least squares (WLS) estimations using Pearson correlation matrices among items were considered. For IFA, diagonally weighted least squares (DWLS) and unweighted least squares (ULS) estimations using items polychoric correlation matrices were considered. Data were generated for one, two, and three dimensional structures. For multidimensional conditions, three degrees of correlation among factors were considered, namely, zero (ρ = 0), low (ρ =0.3), and high (ρ =0.6). Six items were created for each dimension; thus, 6, 12, and 18 items were created for unidimensional, two-dimensional, and three-dimensional conditions, respectively. Factor loadings were adjusted to represent low ($\lambda = 0.3$) and medium ($\lambda = 0.6$) quality items. Continuous items were recoded into four categories forming three distributions with different degrees of asymmetry: Type I items represented symmetric distributions, Type II items represented mild asymmetry, and Type III items represented high asymmetry of responses. Finally, sample sizes were adjusted to represent variation from small to large sample sizes namely, 100, 200, 500, 1,000, and 2,000 subjects. Results indicated that although all estimation procedures showed similar capacity for producing valid solutions and stable λ and correlation parameter estimates, ULS and DWLS yielded remarkably lower bias in both parameter estimates and were robust in extreme conditions: asymmetric item distributions, low item quality ($\lambda = 0.3$), and small sample sizes. The study confirmed that classical estimation procedures in ordinal data with four-point scales is inappropriate. Asún et al. maintained that if one expects the quality of the items in the scale to be low ($\lambda = 0.3$), a sample of 500 subjects might be selected in order to ensure a large probability of achieving admissible results (i.e., a convergent solution) and relatively unbiased and stable estimation of key parameters in the model. And, if the items are suspected to reflect the latent construct in a better fashion ($\lambda = 0.6$), accurate estimations can be reached for small samples (200 or even 100 subjects) if item distributions are symmetric or mildly asymmetric.

Chapter Summary

The review of related literature shows that overwhelming number of studies on IRT and factor analyses of item responses are based on simulation studies using one or combinations of various conditions. An issue that has engaged the attention of researchers has to do with investigating the relationship between number of response categories employed and internal-consistency reliability of

Likert-type questionnaires. It was found that in situations where low total score variability is achieved with a small number of categories, reliability can be increased through increasing the number of categories employed. In situations where opinion is widely divided toward the content being measured, reliability appeared to be independent of the number of response categories.

A great concern in the literature is about the effect of item parameters on item-fitness statistics. Results showed that item discrimination and guess but not difficulty level parameters affected item-fitness. That is, as the level of item discrimination or guess parameter increased, item-fitness values increased.

One of the problems in IRT that has been studied has to do with the comparison of the performances of one-parameter and two- parameter partial credit (1PPC and 2PPC) models. Results showed that the 2PPC model alone or in combination with the 3PL model provided uniformly better fitness than did the 1PPC model used alone or in combination with the 1PL model. It was noted that the poorer fit performance by the lPPC model alone or in combination with the 1PL model is likely produced by the considerable variability in item discrimination, as well as guessing on the multiple-choice items. Further, the percentages of items with good fitness tended to be larger when the 3PL-2PPC model combination was used. Also, this model combination tended to produce better item fitness across datasets with dissimilar properties.

The literature also assessed how violations of the normality assumption impact the item discrimination and difficulty parameter estimates. It was revealed that when the latent variable was negatively skewed, for the most discriminating easy or difficult items, estimates of both parameters were considerably biased coupled with large standard errors.

The review of literature indicated that an issue to consider when conducting factor analysis is the characteristics of the sample from which the measurements of the indicator variables are taken. Obviously, an aspect of the sample that is worth considering is how large the sample should be in order to perform factor analysis. It has been found that correlations – which are used as input data in factor analysis – are less reliable when estimated from small samples. Studies showed that samples of size 50 give very inadequate reliability of correlation coefficients, while samples of size 1000 are more than adequate for factor analysis. With regards to evaluating the adequacy of the sample size, the literature provided some guidelines: 50 is very poor, 100 is poor, 200 is fair, 300 is good, 500 is very good, and 1000 or greater is excellent.

The comparison of the performance of two approaches in analysing four-point Likert rating scales – the classical factor analysis (FA) and the item factor analysis (IFA) – has been advanced in the literature. The FA employs Pearson correlation matrices among items, whereas IFA considers polychoric correlation matrices. The literature confirms that classical estimation procedures in ordinal data with four-point scales is inappropriate. For factor analysis of ordered polytomous data, it is recommended to use polychoric correlations.

CHAPTER THREE

RESEARCH METHODS

Introduction

This chapter focuses on key concepts and methods used in item response theory (IRT) and factor analyses. It presents various IRT models and their graphical representations. The chapter also presents theoretical connection between the parameters of factor analysis and item response models under item response format and dimensionality of the underlying ability. Two measures of correlation coefficients – tetrachoric and polychoric – are presented. In what follows, we present the assumptions and class of IRT models.

Item Response Theory

Item response theory provides a framework for modelling and analysing item response data. IRT is based on statistical assumptions, and only when these assumptions are met that the IRT model can reasonably be implemented. In what follows, we present the assumptions of IRT models.

Assumptions of IRT models

The assumptions underlying IRT models are:

1. Unidimensionality: The set of items are measuring a single continuous latent ability, θ. A requirement for this assumption to be met adequately by a set of response data is the presence of a "dominant" factor that influences responses to items (Hambleton et al., 1991). This dominant factor is the ability measured by the instrument.

2. Local (Conditional) independence: The response to an item is independent of the responses to other items conditional on the ability level. For

this assumption to hold, a person's response to one item must not affect his or her responses to any other items in the questionnaire. For instance, the content of an item must not provide clues to the responses of other items. When local independence exists, the probability of any pattern of item scores occurring for an individual is simply the product of the probability of occurrence of the scores on each item (Hambleton & Swaminathan, 1985). This assumption is needed to guarantee the uniqueness of the maximum likelihood estimation of parameters in a given IRT model. When the assumption of unidimensionality holds, local independence is achieved. However, local independence can be achieved even when the dataset is not unidimensional.

3. Monotonicity: The probability of a positive response is a non-decreasing function of an individual's ability. This assumption can be interpreted to mean that respondents with high ability levels are more likely to endorse items than those with low ability level (M. S. Johnson, Sinharay, & Bradlow, 2007).

Classification of IRT Models

The item response theory models may be classified broadly in three essential ways. Firstly, in terms of the item characteristics or parameters that are included in the models. In this regard, some models are designed to account for one parameter, whiles other more complex models account for two or more parameters. Secondly, IRT models can also differ in terms of the response option format. Along these lines, some models are designed to be used for dichotomous items, whereas others are designed for items with more than two response options (i.e., polytomous items), such as Likert scale items. Thirdly, IRT models are classified in terms of the number of dimensions that define the person ability parameter. In this case, an IRT model is either unidimensional or multidimensional. In what follows, a discussion of unidimensional item response theory (UIRT) models, in terms of response option format, is presented.

Dichotomous IRT models

Dichotomous items have only two response categories, namely, true-false, yes-no, agree-disagree, or right-wrong.

The Rasch model

According to this model, a person's response to a dichotomous item is determined by the individual's ability level and only a single item parameter - the item difficulty (δ). One way of stating the model is in terms of the probability that a person with a given ability level will endorse an item that has a particular difficulty (Embretson & Reise, 2000). The model is given by

$$p\left(X_{ij}=1|\theta,\delta\right)=\frac{1}{1+\exp\left[-(\theta-\delta_i)\right]}, \tag{3.1}$$

where X_{ij} is the response of the jth person to the ith item. This model assumes that all items have the same discrimination power. In other words, all items are assumed to be equally good measures of the ability. For purposes of simplicity in notation, $p_i(\theta)$ is used to represent $p\left(X_{ij}=1|\theta,\delta\right)$, the probability of responding positively to the item. At $\theta=\delta_i$, $p_i(\theta)=0.5$, which means that when the ability level of an individual matches the difficulty of an item, there is 50% chance that the person will respond positively to the item. This gives the meaning of item difficulty under the Rasch model. That is, the item difficulty is the point on the ability scale at which an individual has a 0.5 probability of item endorsement. When $\theta>\delta_i$, $p_i(\theta)>0.5$, which shows that when the ability of the person exceeds the item location (difficulty), there will be more than 0.5 probability of endorsing the item. At this point, the item is considered to be "easy" for that particular individual. On the other hand, when $\theta < \delta_i$, $p_i(\theta) < 0.5$, which suggests that when the item location (difficulty) exceeds the person's ability, there will be less than 50% chance of responding favourably to the item. At this instance, the item is said to be "difficult" for the individual.

The one-parameter logistic model

In the one-parameter logistic (1PL) model, the probability of a respondent providing a positive response to item *i* is given by

$$p\left(X_{ij} = 1|\theta, \delta\right) = \frac{1}{1 + \exp\left[-\alpha(\theta - \delta_i)\right]}. \tag{3.2}$$

The 1PL model requires that all items related to the ability being measured have common discrimination, but not fixed at one. The item difficulty parameter has the same interpretation as in the Equation (3.1). When the ability scores (θ) for a group are transformed to a mean of zero and standard deviation of one, δ_i vary from about -2.0 to 2.0. Values of δ near -2.0 correspond to items that are very easy. Values of δ_i near 2.0 correspond to items that are very difficult for the group of examinees (Hambleton & Swaminathan, 1985).

The two-parameter logistic model

In the two-parameter logistic (2PL) model, the probability of a positive response to an item incorporates how well the item differentiates between low-ability and high-ability respondents. The model is defined as

$$p\left(X_{ij} = 1|\theta, \alpha, \delta\right) = \frac{1}{1 + \exp\left[-1.702\alpha_i(\theta - \delta_i)\right]}. \tag{3.3}$$

Under this model, items have different discrimination powers, α_i. The α_i are defined, theoretically, on the scale $(-\infty, +\infty)$. However, negatively discriminating items are discarded from ability tests. It is unusual to obtain α_i values larger than two. Hence, $\alpha_i \in (0,2)$ (Hambleton & Swaminathan, 1985). High values of α_i result in steeper item characteristic curves. In Equation (3.3), 1.702 is a scaling factor that ensures the value of the item discriminating parameter in logistic models comparable to a normal-ogive model. This scaling is important for linking IRT parameters with factor analysis results (Reise & Revicki, 2015).

The three-parameter logistic model

The three-parameter logistic model is an extension of the 2PL model. Under three-parameter logistic (3PL) model, a provision is made to account for low-ability persons that will respond positively to the item. The probability of a positive response to an item is given by

$$p\left(X_{ij} = 1|\theta, \alpha, \delta, c\right) = c_i + (1 - c_i)\frac{1}{1 + \exp\left[-1.702\alpha_i(\theta - \delta_i)\right]}, \qquad (3.4)$$

where c_i denotes the guess parameter value for the *i*th item. The values of c_i lies between zero and one, both inclusive (i.e., $0 \leq c_i \leq 1$). Typically, c_i assume values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item (Hambleton & Swaminathan, 1985). The interpretation of the item difficulty parameter in the 3PL model differs from the 1PL and 2PL models. From Equation 3.4, when $(\theta - \delta_i)$ approaches $+\infty$, $p_i(\theta)$ approaches one, indicating that when the ability level of a person far exceeds the difficulty of the item, it is almost certain that such an individual will respond positively (without guess) to the item. Also, when $(\theta - \delta_i)$ approaches $-\infty$, $p_i(\theta)$ approaches c_i, showing that when the difficulty of an item far exceeds the ability of an individual, he or she will only respond favourably by guessing at the item. In other words, if $(\theta - \delta_i)$ is negative or low, the guess parameter is expected to be high. This means that guess is expected to be high among individuals with low ability levels. At $\theta = \delta_i$,

$$\begin{aligned} p\left(X_{ij} = 1|\theta, \alpha, \delta, c\right) &= c_i + (1-c_i)\frac{1}{1+\exp[0]} \\ &= c_i + (1-c_i)\frac{1}{2} \\ &= \frac{1+c_i}{2}. \end{aligned} \tag{3.5}$$

Thus, $c_i = f(\theta - \delta_i)$, a function of the difference, $(\theta - \delta_i)$. Equation (3.5) gives the probability of an individual responding favourably to the item at the value of δ_i. When $c_i = 0$, $p_i(\theta) = 0.5$, as in the 1PL and 2PL models. Also when $c_i > 0$, $p_i(\theta) > 0.5$. This means that when a respondent whose ability matches the item's difficulty guesses at the item, he or she would have more than 50% chance of responding positively.

For the 3PL model, δ_i is located at a point on the ability scale where the slope of the item characteristic curve is a maximum. The slope of the 3PL model is obtained by finding the first partial derivative of the probability function with respect to θ. That is,

$$\begin{aligned} p_i'(\theta) &= \frac{\partial}{\partial\theta} p(x_{ij} = 1|\theta, \alpha, \delta, c) \\ &= \frac{\partial}{\partial\theta}\left\{c_i + (1-c_i)\frac{1}{1+\exp[-1.702\alpha_i(\theta-\delta_i)]}\right\} \\ &= \frac{\partial}{\partial\theta}\left\{\frac{(1-c_i)}{1+\exp[-1.702\alpha_i(\theta-\delta_i)]}\right\} \\ &= (1-c_i)\frac{\partial}{\partial\theta}\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\}^{-1} \\ &= -\frac{(1-c_i)}{\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\}^2} \times \frac{\partial}{\partial\theta}\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\} \\ &= -\frac{(1-c_i)}{\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta-\delta_i)]\} \times \\ &\qquad \frac{\partial}{\partial\theta}[-1.702\alpha_i(\theta-\delta_i)] \\ &= \frac{(1-c_i)}{\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta-\delta_i)]\} \times 1.702\alpha_i \\ &= \frac{1.702\alpha_i(1-c_i)}{\{1+\exp[-1.702\alpha_i(\theta-\delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta-\delta_i)]\}. \end{aligned} \tag{3.6}$$

Equation (3.6) measures the rate of change in item endorsement with respect to different ability levels. When $(\theta - \delta_i)$ approaches $+\infty$, $p_i'(\theta)$ approaches zero. This means that an individual whose ability is far above the item's difficulty level, would almost surely endorse (without guess) the item. Since the probability of endorsing the item is almost certain, the rate of change in responding positively is expected to be zero. Also, when $(\theta - \delta_i)$ approaches $-\infty$, $p_i''(\theta)$ approaches zero. That is, an individual whose ability is far lower than the item's difficulty level, would endorse the item by guessing. The amount of guess, among low ability persons, is constant, and therefore, the rate of change in endorsing the item would be zero. It is noteworthy from Equation (3.6) that, when $\theta = \delta_i$,

$$\begin{aligned} p_i'(\theta) &= \frac{1.702\alpha_i(1-c_i)}{\{1+\exp[0]\}^2} \times \{\exp[0]\} \\ &= \frac{1.702\alpha_i(1-c_i)}{4} \\ &= 0.4255\alpha_i(1-c_i). \end{aligned} \tag{3.7}$$

At $c_i = 1$, $p_i'(\theta) = 0$. This means that, if c_i is at its maximum, the rate of endorsement for respondents whose ability matches exactly with the item's difficulty would be zero. Thus, guess work is not helpful (or undesirable) for respondents whose ability matches with the difficulty level of items. Suppose that $c_i = 0$, $p_i'(\theta)$ is a maximum. This indicates that when there is no guess work, among persons whose ability level matches with the item's difficulty, the tendency to endorse the item would be very high.

Polytomous IRT models

Polytomous items are categorical items with more than two possible response categories. Categorical data can be described effectively in terms of the number of categories into which data can be placed. For ordered polytomous items, the response categories have an explicit rank ordering with respect to the ability. Ordered categories are defined by boundaries that separate the categories. Intuitively, there is always one less boundary than there are categories. For instance, a five-point Likert-type item requires four boundaries to separate the five possible response categories (Ostini & Nering, 2006). In general, each response variable X_{ij}, $i = 1,2,...,p$; $j = 1,2,...,n$, has $r_i + 1$ response categories represented by category scores $k = \{0,1,2,...,g,...,r_i\}$ and r_i boundaries denoted by $h = \{1,2,...,g,...,k\}$. Polytomous models results in a general expression for the probability of a person responding in a given item category. Mathematically, the various polytomous models for ordered response categories differ in terms of the expressions that are used to represent the location parameter (δ) of the category boundaries.

The partial credit model

To construct the partial credit (PC) model for ordered polytomous data, one may decompose the responses into a series of ordered pairs of adjacent categories, and then successively apply a dichotomous model to each pair. The PC model assumes that there is a point, δ_{ih} on the latent ability continuum below which an individual provides a particular response and above which the person provides the next higher response. This point indicates the transition from one category to the next category. In the PC model, there is a separate location parameter for each category boundary fo each item (Ostini & Nering, 2006; Reeve, 2002). The relationship between response categories and category boundaries (δ_{ih}), for a four-category item, may be represented diagrammatically as shown in Figure 1.

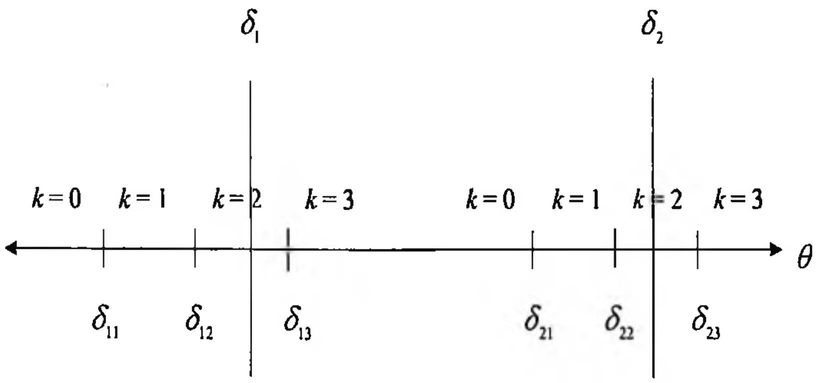

Figure 1: Diagrammatic representation of the relationship between the PC model's response categories and the category boundaries for two items

In Figure 1, δ_1 shows the location of Item 1, whereas δ_2 indicates the location of Item 2. The values δ_{11}, δ_{12}, and δ_{13} represent the locations of the category boundaries for Item 1. For Item 2, δ_{21}, δ_{22}, and δ_{23} indicate the category boundary locations. Thus, each of the two items has four response categories.

For a given pair of adjacent response categories, the probability of observing a response in category g over category $g-1$ for item j is given by

$$P\left(X_{ij}=g|\theta,\delta_{ih}\right)=\frac{\exp\left[\sum_{h=0}^{g}(\theta-\delta_{ih})\right]}{\sum_{k=0}^{r_i}\exp\left[\sum_{h=0}^{k}(\theta-\delta_{ih})\right]}. \qquad (3.8)$$

For notational convenience,

$$\sum_{h=0}^{0}(\theta-\delta_{ih})=0.$$

So that

$$\sum_{h=0}^{k}(\theta-\delta_{ih})\equiv\sum_{h=1}^{k}(\theta-\delta_{ih}).$$

The value δ_{ih} is the category boundary location parameter, and governs the probability of an individual scoring in category g relative to category $g-1$ for item i. In Equation (3.8), g is the count of the boundary locations up to the category under consideration. The numerator contains only the locations of the boundaries prior to the specific category, g, being modelled. The denominator is the sum of all $r_i + 1$ possible numerators (Ostini & Nering, 2006). The expression $\sum(\theta - \delta_{ih})$ indicates the sum of the differences between a given ability level and the location of each category boundary up to the category (g) being modelled. Equation (3.8) utilises only one parameter, category boundary (δ_{ih}) to characterise the item, and referred to as the Rasch partial credit model.

For a higher probability, the difference $(\theta - \delta_{ih})$ should be large. The difference $(\theta - \delta_{ih})$ measures the extent of ease with which an individual can respond favourably to the particular item. For a higher probability in Equation (3.8), we expect the difference $(\theta - \delta_{ih})$ to be positive and large. On the other hand, if $(\theta - \delta_{ih})$ approaches zero, it indicates that a respondent could barely respond favourably. In this case, probability of endorsing the item is expected to be low.

Consider a four-category item, the probability of an individual responding in Category 3 (i.e. $g = 2$) is computed as

$$P\left(X_{ij} = 2 | \theta, \delta_{ih}\right) = \frac{\exp\left[0 + (\theta - \delta_{i1}) + (\theta - \delta_{i2})\right]}{\psi}, \tag{3.9}$$

where,

$$\begin{aligned}\psi = \exp\left[0\right] + \exp\left[0 + (\theta - \delta_{i1})\right] + \exp\left[0 + (\theta - \delta_{i1}) + (\theta - \delta_{i2})\right] \\ + \exp\left[0 + (\theta - \delta_{i1}) + (\theta - \delta_{i2}) + (\theta - \delta_{i3})\right].\end{aligned}$$

In Equation (3.9), the numerator shows the odds of a person at a given ability level responding in the higher category of each dichotomisation up to the category in question. The denominator is the sum of the numerator values for every category in the item. In other words, it is the sum of the odds at every category in the item. The denominator ψ ensures that the probability of responding in any given category does not exceed one, and that the cumulative probabilities of responding in a category, across all the categories for an item sum to one.

The PC model can be written to include two item parameters – difficulty and discrimination parameters. In this case, the probability of observing a response in category g over category $g-1$ for item i is given by (Muraki, 1992)

$$P\left(X_{ij}=g|\theta,\alpha_i,\delta_{ih}\right)=\frac{\exp\left[\sum\limits_{h=0}^{g}\alpha_h(\theta-\delta_{ih})\right]}{\sum\limits_{k=0}^{r_i}\exp\left[\sum\limits_{h=0}^{k}\alpha_h(\theta-\delta_{ih})\right]}, \tag{3.10}$$

where α_h denotes the discrimination associated with response category h on item i. Equation (3.10) is the generalised partial credit (GPC) model or the two-parameter partial credit (2PPC) model, since it uses two parameters to describe the item.

The rating scale model

Although the rating scale (RS) model was proposed before the PC model, the former can be derived from the latter. The RS model is distinctively appropriate for a Likert scale where respondents are asked to respond to an item using a pre-defined set of responses and where the same set of response categories is applied to all the items in the questionnaire. The RS model assumes that all items in the questionnaire have the same kind of response categories (i.e. the same number of categories $r_i = r, i = 1,2,\ldots,p$, having the same meaning) (Bartolucci, Bacci, & Gnaldi, 2016). However, if items in a questionnaire use two or more rating scales with different number of response categories, or if the categories have different labels, then by definition, they are different scales, and the RS model would apply to each scale separately (Ostini & Nering, 2006). For the RS model, the distance between category boundaries is assumed to be equal across all items. This is what distinguishes the RS model from the PC model. In the RS model, the PC model's category boundary parameter (δ_{ih}) is partitioned into two components: (a) the item location parameter (δ_i) and (b) the threshold parameter (τ_h) which defines the boundary between the categories of the rating scale, relative to each item's location. The τ_h indicates how far each category boundary is from the location parameter. In other words, the threshold values may be viewed as offsets from an item's location. Hence, it is the combination of the item's location (δ_i) and the threshold (offset) value, (τ_h) that determines the category boundary's location, δ_{ih} on the continuum (de Ayala, 2009). Mathematically,

$$\delta_{ih} = \delta_i + \tau_h.$$

Figure 2 schematically represents the locations of two items, δ_1 and δ_2, and how the thresholds for a four-point Likert scale relate to these two items.

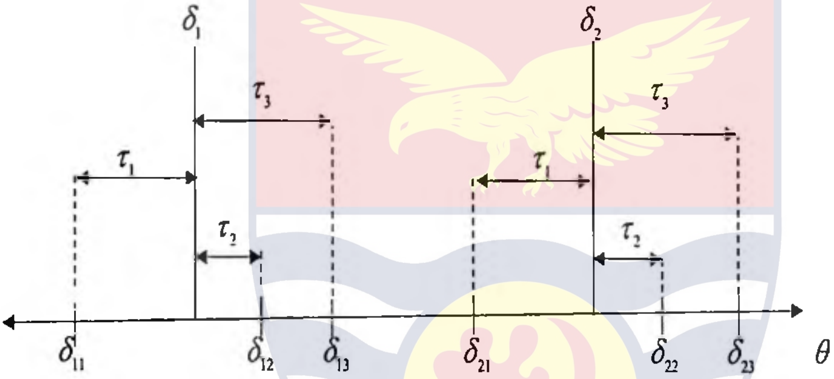

Figure 2: Representation of a set of RS model threshold parameters for two items

In Figure 2, δ_1 and δ_2 show the locations of Item 1 and Item 2, respectively. The values δ_{11}, δ_{12}, and δ_{13} indicate the category boundaries for Item 1. Similarly, δ_{21}, δ_{22}, and δ_{23} represent the category boundary locations for Item 2. For Item 1, τ_1 shows how far category boundary 1 (δ_{11}) is from the item's location (δ_1). Thus, the sum of (δ_1) and (τ_1) determines the location of category boundary 1 (i.e., $\delta_{11} = \delta_1 + \tau_1$). The probability of an individual with ability θ

responding in category g on item j with thresholds, τ_h is given by

$$P(X_{ij} = g|\theta) = \frac{\exp\left[\sum\limits_{h=0}^{g} \{\theta - (\delta_i + \tau_h)\}\right]}{\sum\limits_{k=0}^{r} \exp\left[\sum\limits_{h=0}^{k} \{\theta - (\delta_i + \tau_h)\}\right]}. \tag{3.11}$$

From Equation (3.11),

$$\sum_{h=0}^{g} \{\theta - (\delta_i + \tau_h)\} = -\sum_{h=0}^{g} \tau_h + g(\theta - \delta_i).$$

So that

$$P(X_{ij} = g|\theta) = \frac{\exp\left[-\sum\limits_{h=0}^{g} \tau_h + g(\theta - \delta_i)\right]}{\sum\limits_{k=0}^{r} \exp\left[-\sum\limits_{h=0}^{k} \tau_h + k(\theta - \delta_i)\right]}. \tag{3.12}$$

Equation (3.12) supposes that all the categories of an item are discriminating equally among the responses. However, that RS model can be re-stated to reflect unequal discrimination values (α_h) among the item's category boundaries. To this end, the probability of a person responding in category g on item i is obtained as

$$P(X_{ij} = g|\theta) = \frac{\exp\left[\sum\limits_{h=0}^{g} \alpha_h \{\theta - (\delta_i + \tau_h)\}\right]}{\sum\limits_{k=0}^{r} \exp\left[\sum\limits_{h=0}^{k} \alpha_h \{\theta - (\delta_i + \tau_h)\}\right]}, \tag{3.13}$$

where α_h measures the extent to which categorical responses vary among items as θ changes (Muraki, 1992). From Equation (3.13),

$$\begin{aligned}\sum_{h=0}^{g} \alpha_h \{\theta - (\delta_i + \tau_h)\} &= \sum_{h=0}^{g} \theta\alpha_h - \sum_{h=0}^{g} \alpha_h\delta_h - \sum_{h=0}^{g} \alpha_h\tau_h \\ &= \sum_{h=0}^{g} \alpha_h (\theta - \delta_i) - \sum_{h=0}^{g} \alpha_h\tau_h.\end{aligned}$$

Let $\beta_g = \sum\limits_{h=0}^{g} \alpha_h$ and $c_g = -\sum\limits_{h=0}^{g} \alpha_h\tau_h$. This implies that

$$\sum_{h=0}^{g} \alpha_h \{\theta - (\delta_i + \tau_h)\} = c_g + \beta_g (\theta - \delta_i).$$

Equation (3.13) becomes

$$P(X_{ij} = g|\theta) = \frac{\exp\left[c_g + \beta_g\left(\theta - \delta_i\right)\right]}{\sum_{k=0}^{r} \exp\left[c_k + \beta_k\left(\theta - \delta_i\right)\right]}, \tag{3.14}$$

where c_g, a function of α_h and τ_h, is a category coefficient. By definition, $c_g = \beta_g = 0$ when $g = 0$.

The graded response model

In the graded response (GR) model, the approach to modelling the probability of response categories is such that the ordered polytomous scores are turned into a series of cumulative comparisons (i.e., below a given category as opposed to at and above this category). The GR model specifies the probability of an individual responding in category g or higher versus responding in category lower than k. According to the GR model, the probability of responding in category g or higher is

$$P(X_{ij} \geq g \mid \theta) = \frac{1}{1 + \exp\left[-\alpha_i(\theta - \delta_{ig})\right]}, \tag{3.15}$$

where δ_{ig} is the category boundary location for category score g and α_i is the discrimination parameter which is constant across an item's response categories. In essence, Equation (3.15) is a 2PL model applied to the categories of item i. This model measures the cumulative probability of a person obtaining category g or higher on item i. To calculate the probability of a person responding in a given category g, the difference between the cumulative probabilities for adjacent categories must be determined. That is,

$$p(X_{ij} = g \mid \theta) = P(X_{ij} \geq g \mid \theta) - P(X_{ij} \geq g + 1 \mid \theta),$$

where $P(X_{ij} \geq g + 1 \mid \theta)$ is the probability of responding in category $g + 1$ or higher. Generally,

$$p(X_{ij} = g \mid \theta) = \frac{1}{1 + \exp\left[-\alpha_i(\theta - \delta_{ig})\right]} - \frac{1}{1 + \exp\left[-\alpha_i(\theta - \delta_{i,g+1})\right]}. \tag{3.16}$$