

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

0238

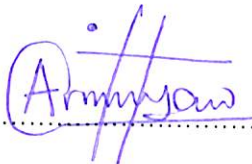
JULY 2018

**SAM JONAH LIBRARY**  
**UNIVERSITY OF CAPE COAST**  
**CAPE COAST**

DECLARATION

**Candidate's Declaration**

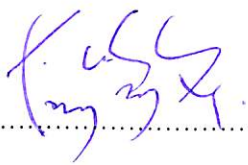
I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

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

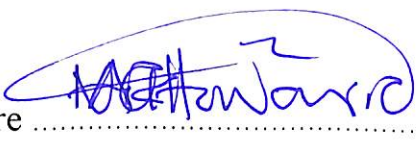
Name: Arimiyaw Zakaria

**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* ( $\alpha$ ,  $d$ ,  $N$ , *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

	Page
DECLARATION	ii
ABSTRACT	iii
KEY WORDS	iv
ACKNOWLEDGEMENTS	v
DEDICATION	vi
LIST OF TABLES	x
LIST OF FIGURES	xiv
CHAPTER ONE: INTRODUCTION	
Background to the Study	1
Statement of the Problem	5
Objectives of the Study	8
Description of Datasets Used in the Study	8
Organisation of the Thesis	10
Chapter Summary	11
CHAPTER TWO: LITERATURE REVIEW	
Introduction	12
IRT Analysis of Items	12
Factor Analysis of Items	18
Comparison of FA and IRT on Item Analysis	30
Chapter Summary	34
CHAPTER THREE: RESEARCH METHODS	
Introduction	37
Item Response Theory	37

Classification of IRT Models	38
IRT Graphical Techniques	53
Estimation of Parameters of IRT Models	72
Assessment of the Fitness of IRT Models	75
Multidimensional Item Response Theory	78
Multidimensional Person and Item Parameters	83
MIRT Graphical Representations	84
Factor Analysis	90
Estimation of Parameters of the Factor Model	94
Relationship between Factor Analysis and Item Response Theory	97
Measures of Correlation Coefficients	107
Chapter Summary	111
<b>CHAPTER FOUR: RESULTS AND DISCUSSION</b>	
Introduction	113
Data Simulation	113
Data Analysis	116
Assessment of Dichotomous Response Scale	117
Investigation of Three-Point Likert Scale	141
Examining Five-Point Likert Scale	160
Assessment of Seven-Point Likert Scale	179
Comparison of Results of Various Response Scales and Sample Sizes	199
Chapter Summary	213
<b>CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS</b>	
Overview	223



Summary	223
Conclusions	234
Recommendations	236
REFERENCES	237
APPENDIX: R CODES FOR DATA SIMULATION AND ANALYSIS	246

## LIST OF TABLES

Table	Page
1	9
2	108
3	110
4	115
5	118
6	121
7	125
8	127
9	130
10	133
11	134
12	137
13	139

14	<i>P</i> -values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Three-Point Scale	142
15	Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Three-Point Scale	144
16	Two-Factor Solutions of Unidimensional Datasets for Various Sample Sizes on Three-Point Scale	146
17	Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Three-Point Scale	149
18	<i>P</i> -values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Three-Point Scale	152
19	Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Three-Point Scale	153
20	<i>P</i> -values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Three-Point Scale	156
21	Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Three-Point Scale	158
22	<i>P</i> -values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Five-Point Scale	161
23	Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale	163
24	Two-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale	165
25	Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale	168

26	<i>P</i> -values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Five-Point Scale	171
27	Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Five-Point Scale	172
28	<i>P</i> -values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Five-Point Scale	175
29	Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Five-Point Scale	177
30	<i>P</i> -values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Seven-Point Scale	180
31	Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale	182
32	Two-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale	184
33	Three-Factor Solutions of Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale	187
34	<i>P</i> -values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Seven-point Scale	190
35	Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Seven-Point Scale	192
36	<i>P</i> -values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Seven-Point Scale	195
37	Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Seven-Point Scale	197

38	Summary Statistics for IRT Results Across Different Scales on Varied Dimensions	200
39	IRT Model Summary Statistics on Unidimensional Datasets for Various Response Scales and Sample Sizes	202
40	IRT Model Summary Statistics on Two-Dimensional Datasets for Various Response Scales and Sample Sizes	203
41	IRT Model Summary Statistics on Three-Dimensional Datasets for Various Response Scales and Sample Sizes	204
42	Summary Statistics for One-Factor Solutions on Unidimensional Datasets	205
43	Summary Statistics for Two-Factor Solutions on Unidimensional Datasets	206
44	Summary Statistics for Three-Factor Solutions on Unidimensional Datasets	208
45	Summary Statistics for Two-Factor Solutions on Two-Dimensional Datasets	210
46	Summary Statistics for Three-Factor Solutions on Three-Dimensional Datasets	212

## LIST OF FIGURES

Figure		Page
1	Diagrammatic representation of the relationship between the PC model's response categories and the category boundaries for two items	45
2	Representation of a set of RS model threshold parameters for two items	48
3	A hypothetical item characteristic curve	54
4	Item characteristic curves for ten items	55
5	Item characteristic curves for items of varying discriminations	56
6	Item characteristic curve for 3PL model	57
7	Graphs showing the effect of the guess parameter on probability of response for (a) item 1, (b) item 2, (c) item 3, (d) item 4, (e) item 5, (f) item 6, (g) item 7, (h) item 8, (i) item 9, and (j) item 10 on the brooding data	60
8	PC model's category characteristic curves for a five-category item	63
9	PC model's expected response for a five-category item	64
10	Total characteristic curve for ten items	65
11	Item information curve based on the Rasch model	67
12	Item information curves based on 2PL model for the brooding items	68
13	Total information curve for the brooding scale items	71
14	M2PL model's (a) ICS and (b) contour plot for an item	85

15	M2PL model's (a) total characteristic surface and (b) contour plot for ten items	87
16	Item information surface	89
17	Total information surface for ten items	90
18	Distribution of dichotomous responses	100
19	Distribution of polytomous response categories	104
20	Item characteristic curves of unidimensional 2PL model for sample sizes: (a) 30; (b) 100; (c) 150; (d) 200; (e) 500; (f) 800; and (g) 1000	123

# 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 ( $\alpha$ ), location ( $\delta$ ), 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  $\alpha$  value indicates the relevance of the item to the ability being measured by the questionnaire. An item with a positive  $\alpha$  value is, at least somewhat, consistent with the underlying ability (trait) being measured, and a relatively large  $\alpha$  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  $\alpha$  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,  $\delta$  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)  $\delta$  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)  $\delta$  value is considered to be "difficult" and persons with high ability levels tend to respond favourably to it. Respond-

ing 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 ( $\theta$ ) 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 in-

dicators 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 univari-

ate 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

Item	Description	Parameters	
		$\hat{\alpha}$	$\hat{\delta}$
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 personality 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 in-

creased, 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 1PPC 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  $\alpha$  parameters, standard errors were larger for the LI estimates when the bias was positive but smaller when the bias was negative. For the  $\delta$  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  $\chi^2$  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

sis 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 us-

ing 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 ( $y^*$ ) distribu-

tion, 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 for  $y^*$ : 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  $\lambda$ : 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 analy-

sis (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 ( $\rho = 0$ ), low ( $\rho = 0.3$ ), and high ( $\rho = 0.6$ ). 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  $\lambda$  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 1PPC 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,  $\theta$ . 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, while 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 ( $\delta$ ). 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

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

where  $X_{ij}$  is the response of the  $j$ th person to the  $i$ th 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  $p(X_{ij} = 1|\theta, \delta)$ , 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

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

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 ( $\theta$ ) for a group are transformed to a mean of zero and standard deviation of one,  $\delta_i$  vary from about  $-2.0$  to  $2.0$ . Values of  $\delta$  near  $-2.0$  correspond to items that are very easy. Values of  $\delta_i$  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

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

Under this model, items have different discrimination powers,  $\alpha_i$ . The  $\alpha_i$  are defined, theoretically, on the scale  $(-\infty, +\infty)$ . However, negatively discriminating items are discarded from ability tests. It is unusual to obtain  $\alpha_i$  values larger

than two. Hence,  $\alpha_i \in (0, 2)$  (Hambleton & Swaminathan, 1985). High values of  $\alpha_i$  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

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

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  $\theta = \delta_i$ ,

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

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  $\delta_i$ . 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,  $\delta_i$  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  $\theta$ . That is,

$$\begin{aligned}
 p'_i(\theta) &= \frac{\partial}{\partial \theta} p(x_{ij} = 1|\theta, \alpha, \delta, c) \\
 &= \frac{\partial}{\partial \theta} \left\{ c_i + (1 - c_i) \frac{1}{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]} \right\} \\
 &= \frac{\partial}{\partial \theta} \left\{ \frac{(1 - c_i)}{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]} \right\} \\
 &= (1 - c_i) \frac{\partial}{\partial \theta} \{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\}^{-1} \\
 &= -\frac{(1 - c_i)}{\{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\}^2} \times \frac{\partial}{\partial \theta} \{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\} \\
 &= -\frac{(1 - c_i)}{\{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta - \delta_i)]\} \times \\
 &\hspace{15em} \frac{\partial}{\partial \theta} [-1.702\alpha_i(\theta - \delta_i)] \\
 &= \frac{(1 - c_i)}{\{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta - \delta_i)]\} \times 1.702\alpha_i \\
 &= \frac{1.702\alpha_i(1 - c_i)}{\{1 + \exp[-1.702\alpha_i(\theta - \delta_i)]\}^2} \times \{\exp[-1.702\alpha_i(\theta - \delta_i)]\}. \tag{3.6}
 \end{aligned}$$

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  $X_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, n$ , has  $r_i + 1$  response categories represented by category scores  $k = \{0, 1, 2, \dots, g, \dots, r_i\}$  and  $r_i$  boundaries denoted by  $h = \{1, 2, \dots, g, \dots, k\}$ . 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 ( $\delta$ ) 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,  $\delta_{ih}$  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 ( $\delta_{ih}$ ), for a four-category item, may be represented diagrammatically as shown in Figure 1.

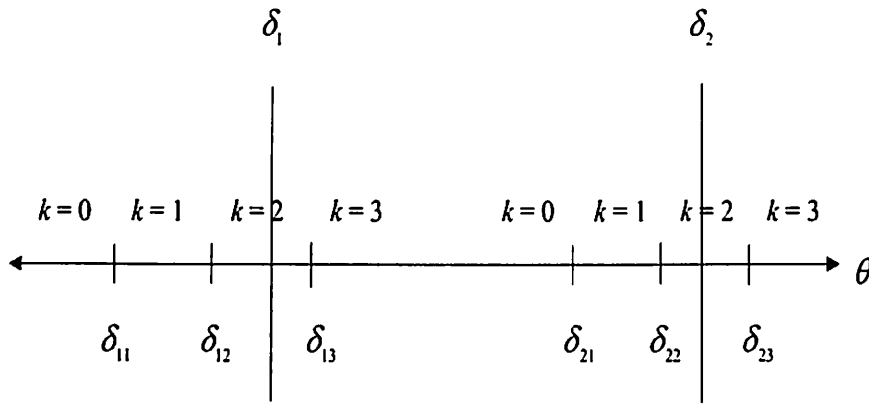


Figure 1: Diagrammatic representation of the relationship between the PC model's response categories and the category boundaries for two items

In Figure 1,  $\delta_1$  shows the location of Item 1, whereas  $\delta_2$  indicates the location of Item 2. The values  $\delta_{11}$ ,  $\delta_{12}$ , and  $\delta_{13}$  represent the locations of the category boundaries for Item 1. For Item 2,  $\delta_{21}$ ,  $\delta_{22}$ , and  $\delta_{23}$  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

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

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  $\delta_{ih}$  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 ( $\delta_{ih}$ ) 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

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

where,

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

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  $\psi$  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)

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

where  $\alpha_h$  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  $r_i = r, i = 1, 2, \dots, p$ , 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 ( $\delta_{ih}$ ) is partitioned into two components: (a) the item location parameter ( $\delta_i$ ) and (b) the threshold parameter ( $\tau_h$ ) which defines the boundary between the categories of the rating scale, relative to each item's location. The  $\tau_h$  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 ( $\delta_i$ ) and the threshold (offset) value, ( $\tau_h$ ) that determines the category boundary's location,  $\delta_{ih}$  on the continuum (de Ayala, 2009). Mathematically,

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

Figure 2 schematically represents the locations of two items,  $\delta_1$  and  $\delta_2$ , and how the thresholds for a four-point Likert scale relate to these two items.

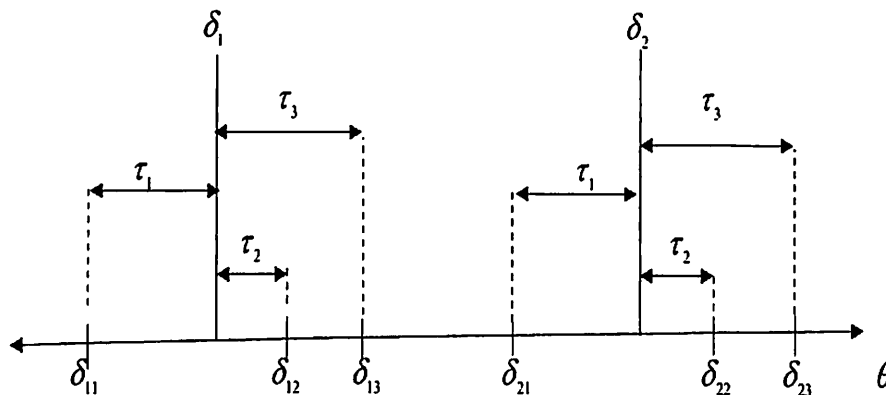


Figure 2: Representation of a set of RS model threshold parameters for two items

In Figure 2,  $\delta_1$  and  $\delta_2$  show the locations of Item 1 and Item 2, respectively. The values  $\delta_{11}$ ,  $\delta_{12}$ , and  $\delta_{13}$  indicate the category boundaries for Item 1. Similarly,  $\delta_{21}$ ,  $\delta_{22}$ , and  $\delta_{23}$  represent the category boundary locations for Item 2. For Item 1,  $\tau_1$  shows how far category boundary 1 ( $\delta_{11}$ ) is from the item's location ( $\delta_1$ ). Thus, the sum of ( $\delta_1$ ) and ( $\tau_1$ ) determines the location of category boundary 1 (i.e.,  $\delta_{11} = \delta_1 + \tau_1$ ). The probability of an individual with ability  $\theta$

responding in category  $g$  on item  $j$  with thresholds,  $\tau_h$  is given by

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

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

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

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 ( $\alpha_h$ ) among the item's category boundaries. To this end, the probability of a person responding in category  $g$  on item  $i$  is obtained as

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

where  $\alpha_h$  measures the extent to which categorical responses vary among items as  $\theta$  changes (Muraki, 1992). From Equation (3.13),

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

Let  $\beta_g = \sum_{h=0}^g \alpha_h$  and  $c_g = - \sum_{h=0}^g \alpha_h \tau_h$ . 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

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

where  $c_g$ , a function of  $\alpha_h$  and  $\tau_h$ , 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

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

where  $\delta_{ig}$  is the category boundary location for category score  $g$  and  $\alpha_i$  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,

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

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

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

In the GR model, the probit link function is used instead of the logit link function. That is, the function used is the cumulative density function of the normal distribution.

### *The nominal response model*

The nominal response (NR) model handles responses to items with two or more nominal categories, such as a multiple-choice item. In the NR model, the probability of responding in any given category is modelled directly by implementing the multinomial logistic function of the latent ability,  $\theta$ . Conceptually, each of the item's response categories has an associated probability. The sum of these response probabilities is one. Suppose item  $i$  has four response categories,  $X_i = 1$ ,  $X_i = 2$ ,  $X_i = 3$ , and  $X_i = 4$ , with respective probabilities  $p_1$ ,  $p_2$ ,  $p_3$ , and  $p_4$ . For this set of probabilities, the odds of one response against another can be determined. For instance,

$$\text{odds} \left( \frac{X_i = 2}{X_i = 1} \right) = \frac{p_2}{p_1}.$$

For convenience, the odds can be transformed into a logarithmic scale through the logit function. The logit transformation makes room for expressing the log odds of one response category versus another in terms of a respondent's ability  $\theta$ . That is,

$$\log \left( \frac{p_2}{p_1} \right) = \gamma_2 + \alpha_2 \theta, \quad (3.17)$$

where  $\alpha_2$  and  $\gamma_2$  characterise the  $X_i = 2$  response category. The notation  $\gamma_2$  is the intercept and reflects the propensity to respond in Category 2 over Category 1 regardless of the ability level. The value,  $\alpha_2$  is the slope and interpreted as the change in the log odds as the ability level  $\theta$  changes by one unit. In a similar fashion, the log odds of a response in Category 3 versus Category 1 may be

obtained as

$$\log \left( \frac{p_3}{p_1} \right) = \gamma_3 + \alpha_3 \theta. \quad (3.18)$$

In the case of Category 4 as opposed to Category 1,

$$\log \left( \frac{p_4}{p_1} \right) = \gamma_4 + \alpha_4 \theta. \quad (3.19)$$

The three logit equations uses response Category 1 as baseline response category (i.e a criterion variable).

The probability of a response in a given category can be directly expressed using Equations (3.17), (3.18), and (3.19). Thus,

$$p_2 = p_1 \exp [\gamma_2 + \alpha_2 \theta] \quad (3.20)$$

$$p_3 = p_1 \exp [\gamma_3 + \alpha_3 \theta] \quad (3.21)$$

$$p_4 = p_1 \exp [\gamma_4 + \alpha_4 \theta]. \quad (3.22)$$

The sum of response probabilities across the categories of an item is 1. That is,

$$p_1 + p_2 + p_3 + p_4 = 1.$$

This implies that

$$p_1 + p_1 \exp [\gamma_2 + \alpha_2 \theta] + p_1 \exp [\gamma_3 + \alpha_3 \theta] + p_1 \exp [\gamma_4 + \alpha_4 \theta] = 1.$$

Making  $p_1$  the subject gives

$$p_1 = \frac{1}{1 + \exp [\gamma_2 + \alpha_2 \theta] + \exp [\gamma_3 + \alpha_3 \theta] + \exp [\gamma_4 + \alpha_4 \theta]}$$

Thus, Equations (3.20), (3.21), and (3.22) become

$$p_2 = \frac{\exp [\gamma_2 + \alpha_2 \theta]}{1 + \exp [\gamma_2 + \alpha_2 \theta] + \exp [\gamma_3 + \alpha_3 \theta] + \exp [\gamma_4 + \alpha_4 \theta]}$$

$$p_3 = \frac{\exp [\gamma_3 + \alpha_3 \theta]}{1 + \exp [\gamma_2 + \alpha_2 \theta] + \exp [\gamma_3 + \alpha_3 \theta] + \exp [\gamma_4 + \alpha_4 \theta]}$$

$$p_4 = \frac{\exp[\gamma_4 + \alpha_4\theta]}{1 + \exp[\gamma_2 + \alpha_2\theta] + \exp[\gamma_3 + \alpha_3\theta] + \exp[\gamma_4 + \alpha_4\theta]}.$$

In general, if item  $i$  has  $r_i$  response categories, the probability that person  $j$  with ability  $\theta$  will respond in category  $g$  is

$$p(X_{ij} = g|\theta) = \frac{\exp[\gamma_{ig} + \alpha_{ig}\theta]}{1 + \sum_{k=2}^{r_i} \exp[\gamma_{ik} + \alpha_{ik}\theta]}. \quad (3.23)$$

For simplicity in representation, let  $\gamma_1 + \alpha_1\theta = 0$ , i.e.,  $\gamma_1 = \alpha_1 = 0$ , so that

$$\sum_{k=1}^1 \exp[\gamma_k + \alpha_k\theta] = 1.$$

Equation (3.23) becomes

$$p(X_{ij} = g|\theta) = \frac{\exp[\gamma_{ig} + \alpha_{ig}\theta]}{\sum_{k=1}^{r_i} \exp[\gamma_{ik} + \alpha_{ik}\theta]}. \quad (3.24)$$

Equation (3.24) is the nominal response model (Bock, 1972). Two constraints are imposed on the parameters  $\alpha$  and  $\gamma$ : (1)  $\sum_{k=1}^{r_i} \alpha_{ik} = 0$ , and (2)  $\sum_{k=1}^{r_i} \gamma_{ik} = 0$ . Alternatively, the parameters  $\alpha$  and  $\gamma$  may be set to zero for the baseline category. Consequently, the number of estimated category slopes and intercepts for an item is  $2(r_i - 1)$  (de Ayala, 2009).

## IRT Graphical Techniques

In this section, we present some graphical methods that are used to describe various characteristics of items. These graphical methods include item characteristic curve, total characteristic curve, item information curve and total information curve.

### Item characteristic curve

Item response theory postulates that the relationship between persons' item response and the set of abilities underlying item response can be expressed

by a monotonically increasing function, called the item characteristic curve (ICC) or item response function. This function specifies that as the level of the ability increases, the probability of a favourable response to an item increases (Hambleton et al., 1991). In IRT analysis, ICCs are often used to present and evaluate characteristics of items in a questionnaire. An ICC can be drawn based on an IRT model. It is obtained by plotting response probabilities,  $p_i(\theta)$  against persons' abilities,  $\theta$ . A hypothetical ICC is presented in Figure 3.

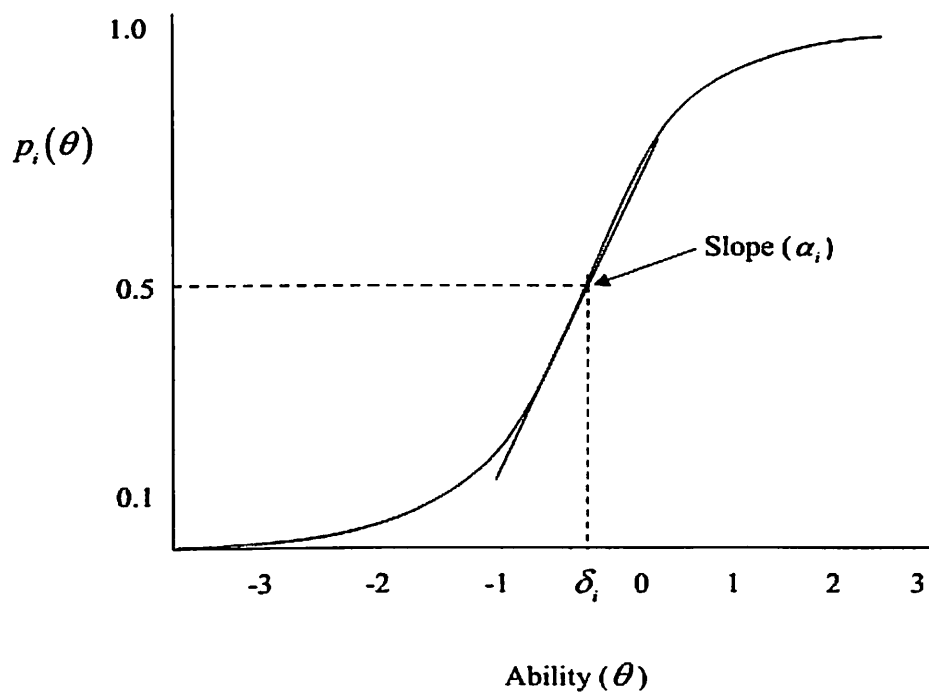


Figure 3: A hypothetical item characteristic curve

In Figure 3, the probability of a positive response is almost zero at the lowest levels of ability. This probability increases until at the highest levels of ability, where the probability of a positive response approaches 1. Thus, the ICC has two asymptotes: (a) lower asymptote, the probability of a positive response is zero, and (b) upper asymptote, where the probability is 1. There are two features of an ICC that are used to describe its general form: (1) the difficulty ( $\delta_i$ ), and (2) the discrimination ( $\alpha_i$ ) of the item. The difficulty of an item describes

where the item functions along the ability scale. For instance, an easy item functions among low-ability persons and a hard item functions among high-ability persons. This makes the difficulty of an item a location index. The ICC may contain as many curves as there are items in the questionnaire measuring the ability,  $\theta$ . As an illustration, the brooding scale dataset with ten items for a group of 2,569 females (Reeve, 2002) is considered. An ICC for a 2PL model based on the brooding scale dataset is presented in Figure 4.

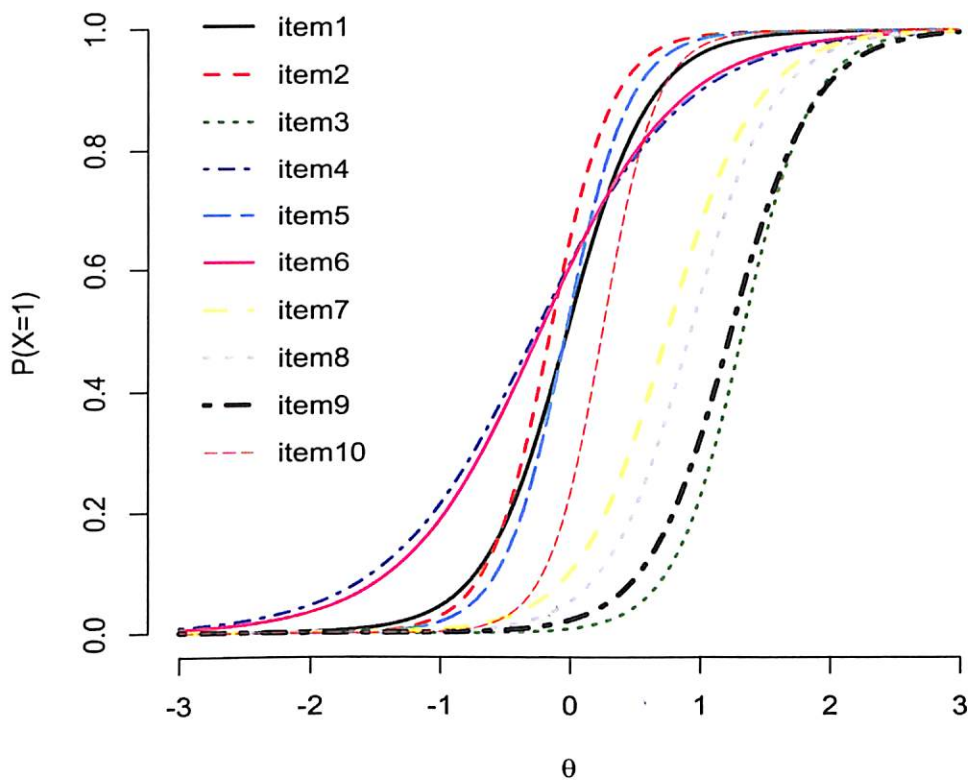


Figure 4: Item characteristic curves for ten items

In Figure 4, each curve corresponds to an item. All the items have different difficulty levels. The order of the curves, from left to right on the ability-axis, reflect their difficulty levels. The curve at the extreme left is considered to be easy, since the probability of responding favourably to the item is high for low-ability respondents. The curves at the centre is viewed to be averagely difficult.

The curve at the extreme right represents a hard item because the probability of a positive response is low for most parts of the ability axis, and only increases at higher ability levels.

The discrimination of an item describes how well an item can distinguish between persons having abilities below the item location and those having abilities above the item location. The discrimination reflects the steepness of the ICC in its middle part. The steeper the curve, the better the item can discriminate. The flatter the curve, the less the item is able to discriminate. In this case, the probability of a positive response at low-ability level is approximately the same as it is at high-ability levels. Figure 5 shows the probability of item response as a function of person ability for items of low and high discriminating values.

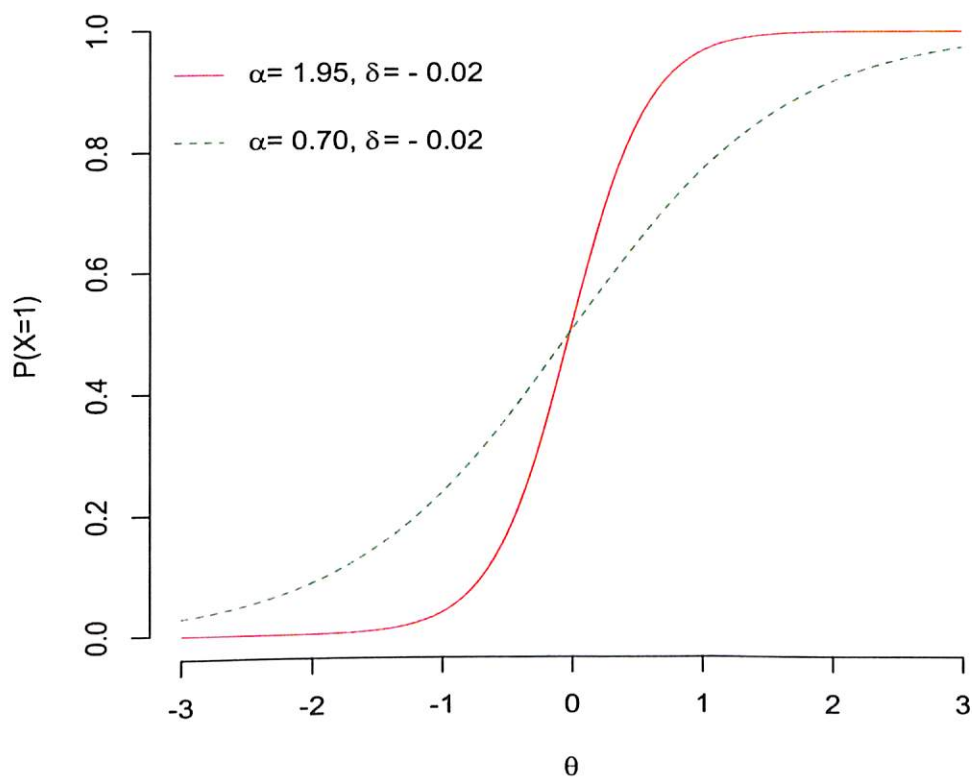


Figure 5: Item characteristic curves for items of varying discriminations

It can be observed in Figure 5 that the ICCs have different discrimination

values but the same difficulty parameters. The  $\alpha$  value affects the slope of the ICC at the point of inflexion (at the item's difficulty). For large  $\alpha$  values the slope of the ICC at the point of inflexion increases, thereby making the curve steeper. This means that as  $\alpha$  increases, virtually smaller differences in person ability,  $\theta$  yield larger differences in the probability of response. The effect of  $\alpha$  on the probabilities of response is maximum at the item's difficulty,  $\delta$ , and minimises as  $\theta$  and  $\delta$  become clearly separate. As  $\alpha$  gets closer to zero, an item's discriminating power decreases.

For a 3PL model, an ICC would have a non-zero lower asymptote that reflects the amount of guess. A typical item characteristic curve for a 3PL model is shown in Figure 6.

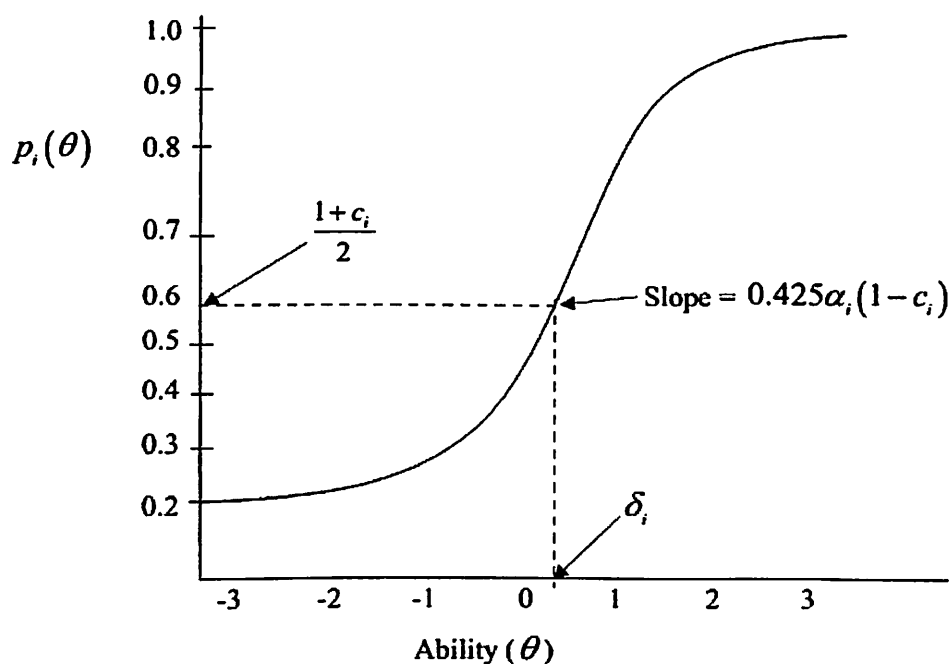
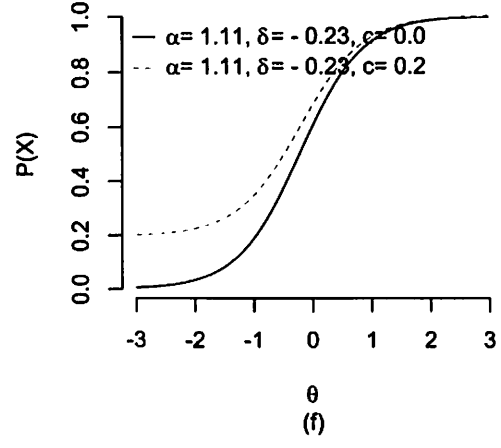
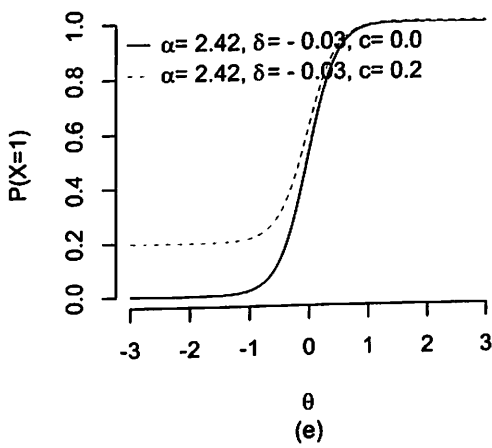
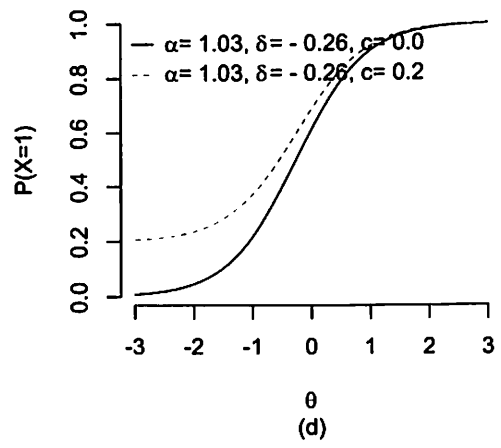
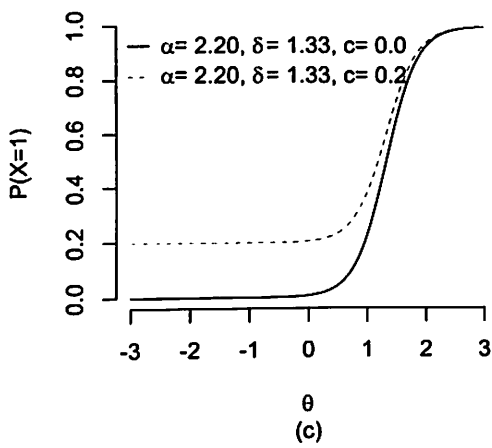
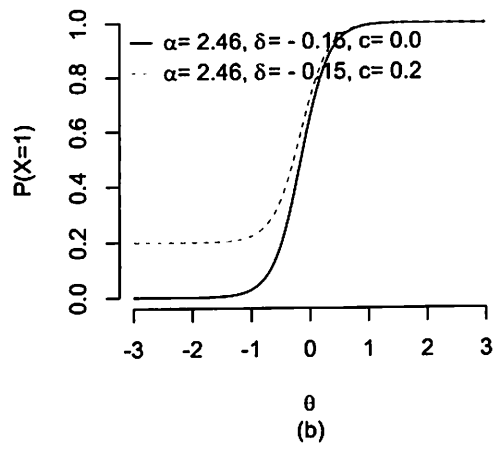
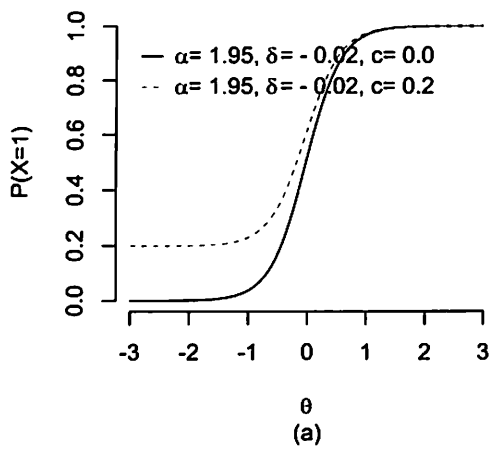


Figure 6: Item characteristic curve for 3PL model

From Figure 6, the guess value  $c_i = 0.2$  indicates the probability of a positive response among low-ability individuals. As the value of  $c_i$  increases, the probability of a positive response increases, and vice-versa. At  $\theta = \delta_i$ , the probability of a positive response is a function of  $c_i$ , i.e.  $(1 + c_i)/2$ . This probability



is high when the amount of guess is high. When  $c_i = 0$ , the probability is 0.5, as in the case of the 2PL model. The slope of the curve at  $\delta_i$ ,  $0.425\alpha_i(1 - c_i)$ , is a maximum. It can be observed that an item's discrimination power is affected by the value of  $c_i$ . In this case, as  $c_i$  increases, an item's discrimination power decreases. As a way of assessing the effect of the guess parameter value on probability of positive response, the relationship between the 2PL and the 3PL models is examined. Figure 7 illustrates graphs showing the effect of the guess parameter by setting  $c_i = 0.2$  for all ten items under the brooding scale data.



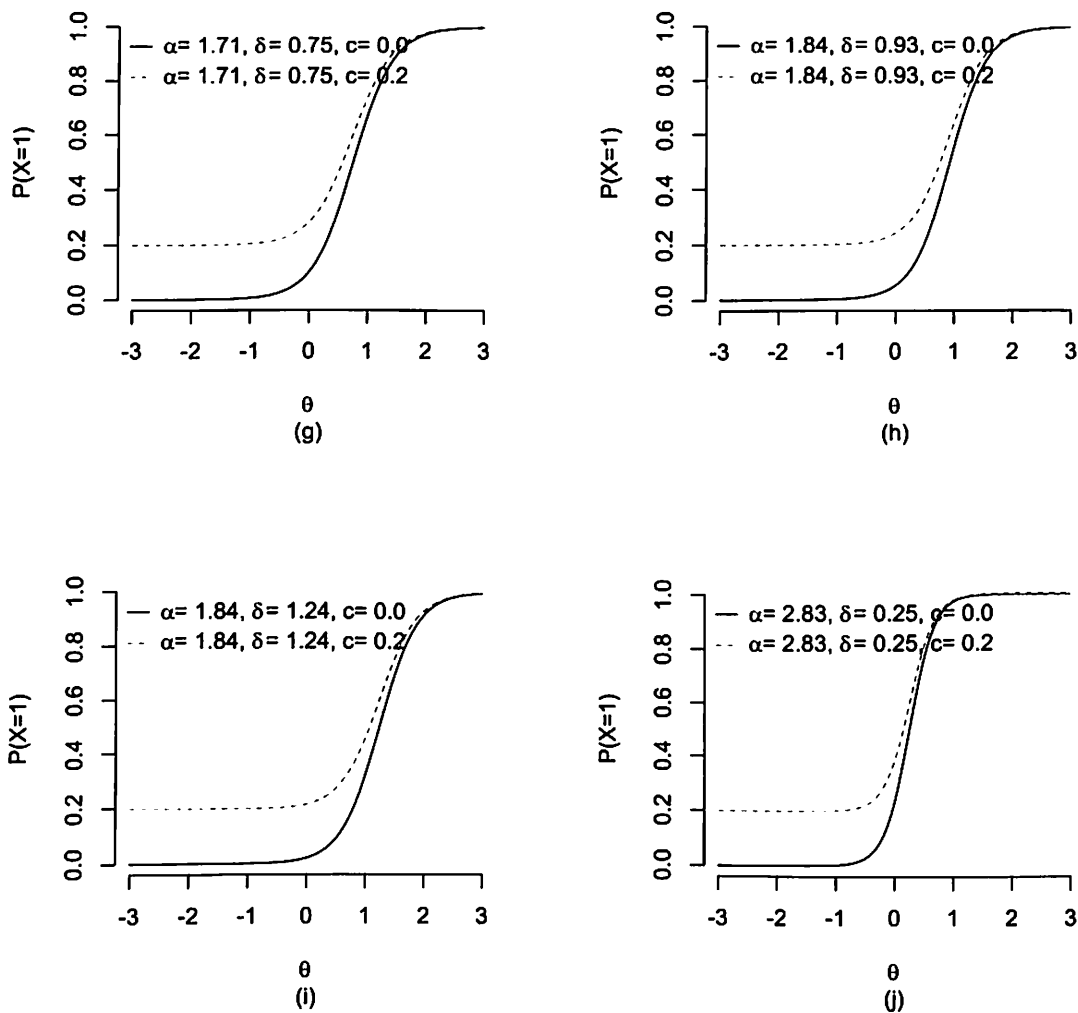


Figure 7: Graphs showing the effect of the guess parameter on probability of response for (a) item 1, (b) item 2, (c) item 3, (d) item 4, (e) item 5, (f) item 6, (g) item 7, (h) item 8, (i) item 9, and (j) item 10 on the brooding data

From Figure 7, it can be observed that for the set of Items 1, 2 and 5, there is a quite visible effect of guess work on the probability of positive response for extremely low ability (i.e.,  $\theta < 0$ ). The two curves are approximately the same for  $\theta$  close to zero, and are exactly the same for  $\theta$  greater than one. This indicates that difficulty parameter ( $\delta$ ) must be low for such items, since the probability of positive response does not so much depend on guess work. There is also a

clear indication of discrimination between respondents at 0 ability level. Thus, though the item difficulty is low, the amount of discrimination is high. It is noteworthy that for the set of Items 3 and 9, there is a clear effect of guess work on probability of positive response for extremely low ability (i.e.,  $\theta < 0$ ). The effect of guess work is constant and equal to 0.2 for the specified range of ability values. However, the two curves are approximately the same for ability levels close to 2, and are exactly the same for extremely high values of ability (i.e.,  $\theta \rightarrow 3$ ). This indicates that difficulty parameter must be quite high for such items, since probability of positive response appears to depend quite largely on guess work. There is a remarkable difference between respondents at 0 and 1 ability levels. Thus, both item difficulty and discrimination are quite high. Figure 7 shows that, for Items 4 and 6, the effect of guess is clear and only visible for extremely low ability level (i.e.,  $\theta < -2$ ). The two curves are quite different for a wide range of ability (from  $-1$  to  $2$ ), and are the same for extremely high values of ability (i.e.,  $\theta > 2$ ). This indicates that item difficulty parameter must be quite high with a low discrimination. Thus, there does not appear to be a visible discrimination among respondents on Items 4 and 6. Further, for the set of Items 7, 8 and 10, there is a clear effect of guess work on probability of positive response for extremely low ability levels (i.e.,  $\theta < 0$ ). The effect of guess work is constant and equal to 0.2 for the specified range of ability values. However, the two curves are approximately the same for ability levels close to 2, and are exactly the same for extremely high values of ability (i.e.,  $\theta > 2$ ). This indicates that difficulty parameter must be a bit high for these items, since probability of positive response seems to depend largely on guess parameter value. There is a remarkable difference between respondents at ability level of 1. Thus, both item difficulty and discrimination are quite high.

Thus, in the brooding scale data four main categories of items could be identified. There is a group (1, 2, 5) of items that show low difficulty level but

have high discrimination. For these items, there is a quite visible effect of guess work on the probability of positive response for abilities less than zero. Another group (3, 9) of items indicate that both difficulty and discrimination are quite high. Here, the effect of guess work on probability of positive response is the same as in the first group of items. Again, the group (4, 6) of items indicate that difficulty parameter is quite high with a low discrimination. In this case, the effect of guess is clear and only visible for extremely low ability level less than  $-2$ . Further, group (7, 8, 10) of items possess quite high difficulty and discrimination values, but the difficulty level is a bit lower than that of the second group (3, 9). The effect of guess work on the probability of positive response is quite clear for low abilities less than zero.

For Likert-type items with an ordered responses, an ICC can be plotted for each category. The resulting graph is the category characteristic curves (CCC), category response curves (CRC), or category response function (CRF) (DeMars, 2010). They represent the probability of a person responding in a particular category given the ability level. The item parameters in the chosen polytomous model dictate the shape and location of the CCCs. In general, the higher the slope parameters ( $\alpha_{ih}$ ), the steeper the CCCs. In addition, the narrower and peaked the category response curves, the more the response categories differentiate among ability levels. The CCCs peak in the middle of two adjacent category boundary location parameters ( $\delta_{ih}$ ). For polytomous items, the category response curves are not exclusively monotonic functions. In the case of items with ordered categories, only the curves for the extreme negative and extreme positive categories are, respectively, monotonically decreasing and increasing (Ostini & Nering, 2006). Figure 8 displays PC model category characteristic curves for a five-category item with category boundary parameter values  $\delta_{i1} = -2$ ,  $\delta_{i2} = -1$ ,  $\delta_{i3} = 0$ , and  $\delta_{i4} = 2$ .

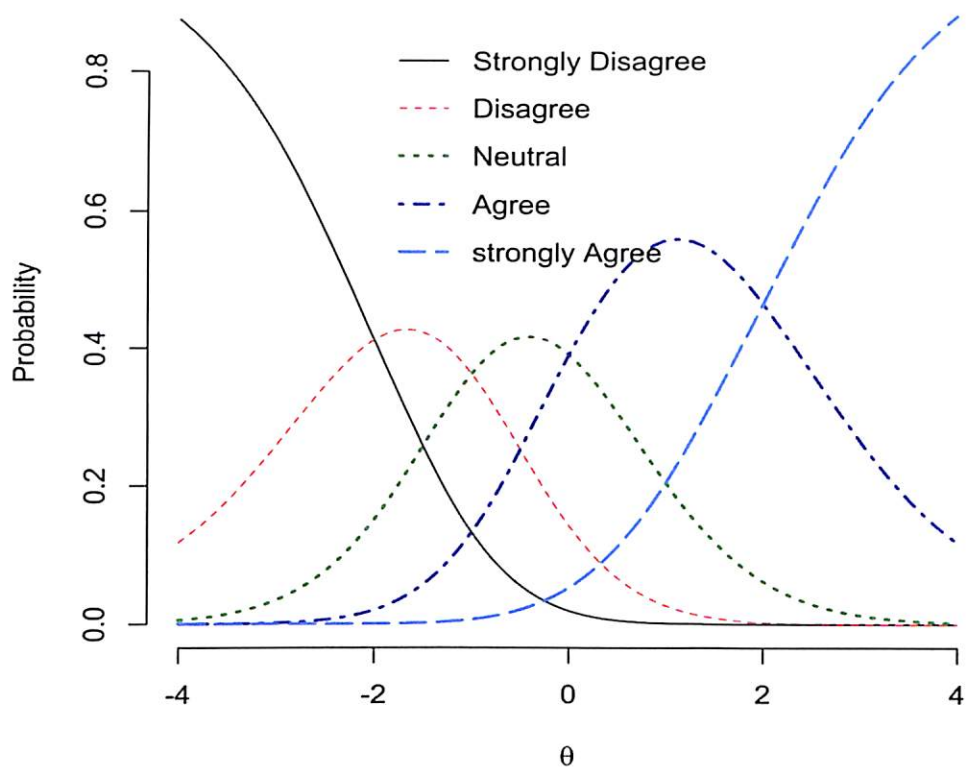


Figure 8: PC model's category characteristic curves for a five-category item

The curves in Figure 8 show the probability of each response category for persons at a given ability level. For instance, if a person has ability level of  $-3$ , the probability of a response in strongly disagree category is 0.70, a response in disagree category is 0.26, a response in neutral category is 0.04, a response in agree and strongly agree categories become much smaller and smaller. At this ability level, a response in strongly disagree category is the most likely, but the other categories are also possible. The threshold parameter values indicate where adjacent response category curves intersect. At the point of intersection, adjacent response categories have equal probabilities of response. Strongly disagree and disagree response categories have equal probabilities of response at  $-2$  ability level. Also, disagree and neutral response categories have equal probabilities of response at  $-1$  ability level. Figure 8 shows that the curve for the second

category (disagree) rises as the probability of responding in the first category (strongly disagree) decreases, but only up to a point, at which time it decreases as the probability of responding in the third category (neutral) increases. The curve for the last category (strongly agree) is a monotonically increasing function.

For the PC model, another way to describe the relationship between person ability level and item responses is to graph the expected response on the item as a function of ability level,  $\theta$  given by (Reckase, 2009)

$$E(X_{ij} = g|\theta) = \sum_{k=0}^{r_i} kP(X_{ij} = g|\theta, \delta_{ih}).$$

The expected response,  $E(X_{ij} = g|\theta)$  ranges from 0 to  $r_i$  as a function of  $\theta$ . Figure 9 displays the expected response for an item with the same parameters as used in Figure 8.

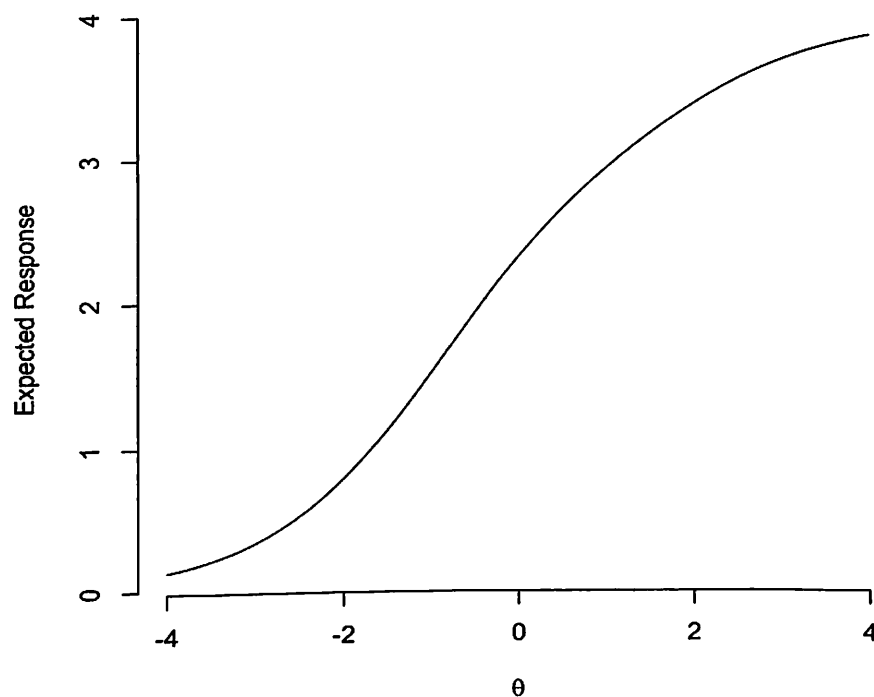


Figure 9: PC model's expected response for a five-category item

The curve in Figure 9 represents the expected item response category for persons at a particular ability level. This curve can be used to predict a person's

response to the specific item given the ability level. For instance, an individual with ability level of about 1.8 is expected to respond to the item in Category 3 (agree) whereas a person at low ability level, say  $-3$ , would respond in Category 0 (strongly disagree).

### Total characteristic curve

The aggregation of the item characteristic curves yields the total characteristic curve (TCC). The TCC is the sum of the probabilities of responding positively to the items in the questionnaire given the person's ability level. It ranges from the sum of the lower asymptotes (of the ICCs) to the number of items. Thus, the TCC indicates the expected number of items endorsed as a function of the individual's ability level (DeMars, 2010; Reeve, 2002). The corresponding TCC for the ICCs in Figure 4 is presented in Figure 10.

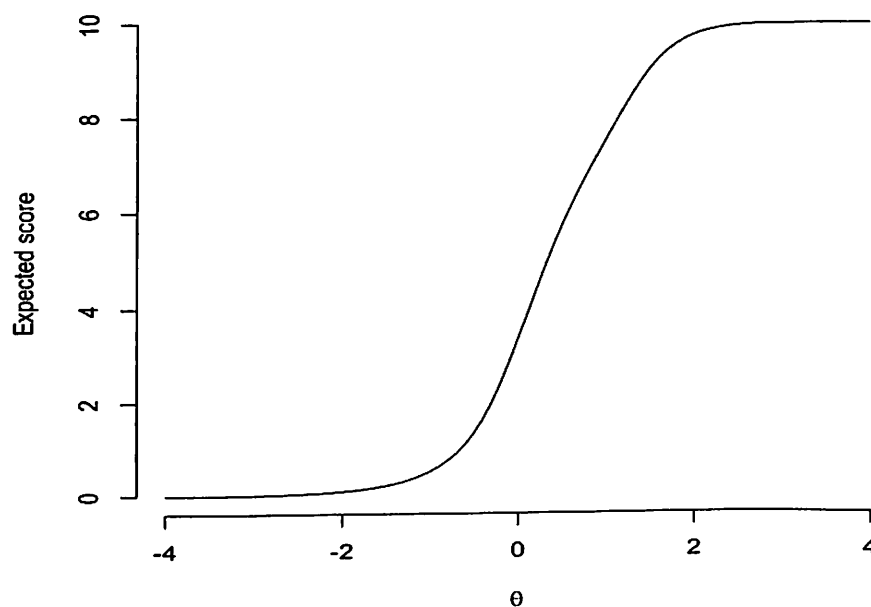


Figure 10: Total characteristic curve for ten items

Figure 10 presents a TCC for ten items. On the average, an individual with extremely high ability level ( $\theta > 3$ ) is expected to endorse all ten items.



Meanwhile, a person with extremely low ability is expected to endorse only a few or none of the items. Generally, at high-ability levels, nearly all items get endorsed and vice-versa.

### **Item information curve**

The term “information” as used in IRT is an indicator of the quality or certainty of the estimate of a parameter, most often the ability of a person,  $\theta$ . Information is usually represented as a function of the parameter being estimated rather than just a single value (Reckase, 2009). The response to any item in a questionnaire provides some information about the ability of a person. The amount of this information depends on how closely the difficulty of the item matches the ability of the person. For the Rasch model, this is the only parameter influencing item information, whereas in other models it is combined with other parameters (Partchev, 2004).

The item information of the Rasch model can be computed as

$$I_i(\theta) = p_i(\theta)[1 - p_i(\theta)]. \quad (3.25)$$

From Equation (3.25), the maximum item information for the Rasch model is 0.25. It occurs at the point where the probability of a positive and of a negative response are both equal 0.5. A graph of the item information function in Equation (3.25) is the item information curve (IIC). A hypothetical IIC for the Rasch model is presented in Figure 11.

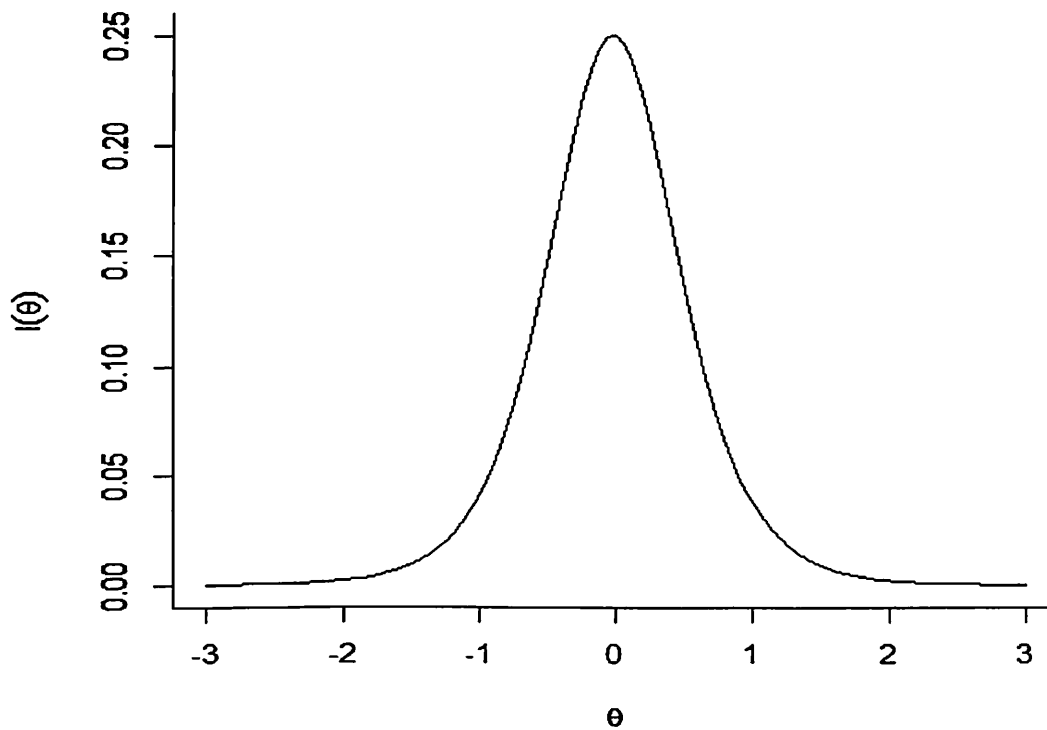


Figure 11: Item information curve based on the Rasch model.

Figure 11 shows that any item in the Rasch model is most informative for persons whose abilities is equal to the difficulty of the item. The item information curve is symmetric about the value of the item's difficulty parameter. As ability becomes either less or greater than the item difficulty, the item information decreases.

In the case of the 1PL model, item information can be calculated as

$$I_i(\theta) = \alpha^2 p_i(\theta) [1 - p_i(\theta)]. \quad (3.26)$$

The item information function in Equation (3.26) reflects the discrimination of the item, which is the same across all items in the questionnaire. The item information value of a 1PL model is influenced by the discrimination value. This so because, a discrimination value blow 1 decreases the item information considerably, and a value above 1 increases the item information remarkably. The maximum item information for the 1PL model is  $0.25\alpha^2$ . The shape of the IIC for Equation (3.26) is very similar to that for the preceding Rasch model.

For the 2PL model, the item information function is given by

$$I_i(\theta) = \alpha_i^2 p_i(\theta) [1 - p_i(\theta)]. \quad (3.27)$$

The discrimination parameter values for the 2PL model differs across the different items, as evident in Equation (3.27). The discrimination value greatly affects the item information as it appears as a square. In this case, items with high discriminating values ( $\alpha_i > 1$ ) will provide more information for estimating ability than those with low discriminating values ( $\alpha_i < 1$ ). The maximum item information of a 2PL model is  $0.25\alpha_i^2$ . As an illustration, IICs for a 2PL model based on the brooding sub-scale is shown in Figure 12.

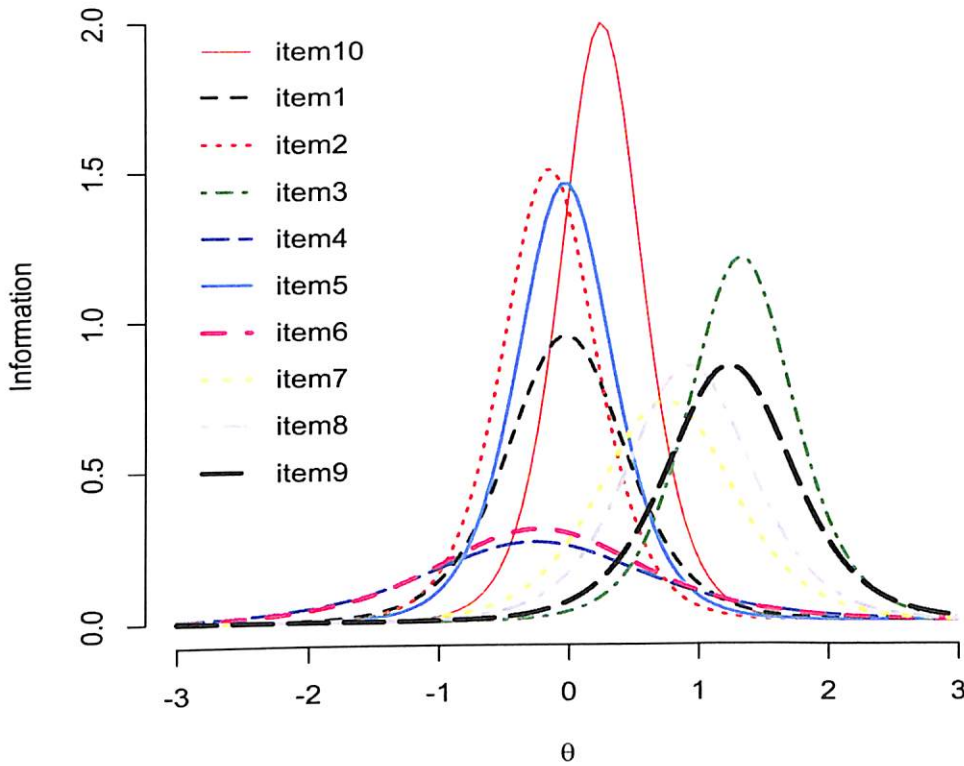


Figure 12: Item information curves based on 2PL model for the ten brooding items.

Under 3PL model, the amount of item information is determined by

$$I_i(\theta) = \alpha_i^2 \left[ \frac{1 - p_i(\theta)}{p_i(\theta)} \right] \frac{[p_i(\theta) - c_i]^2}{(1 - c_i)^2}. \quad (3.28)$$

The shape of the item information function in Equation (3.28) is quite similar to that of 2PL model. However, due to the involvement of the terms  $[p_i(\theta) - c_i]$  and  $(1 - c_i)$  in Equation (3.28), the amount of item information under 3PL model will be less than that of 2PL model having the same difficulty ( $\delta_i$ ) and discrimination ( $\alpha_i$ ) values (Baker, 2001). When both 3PL and 2PL models have common values of  $\delta_i$  and  $\alpha_i$ , item information will be the same when  $c_i = 0$ . In the case where  $c_i > 0$ , 3PL model will always yield less item information.

Under polytomous IRT models, the amount of information for estimating a person's ability provided by a given item can be determined. For PC model, item information is

$$I_i(\theta) = \sum_{k=1}^{r_i} k^2 p_{ik} - \left( \sum_{k=1}^{r_i} k p_{ik} \right)^2, \quad (3.29)$$

where  $p_{ik}$  is the probability of responding in category  $k$  of item  $i$ . Equation (3.29) shows that items with the same number of response categories provide the same amount of information. However, items possessing more response categories yield more information across  $\theta$  than do items with fewer categories (Dodd & Koch, 1987). In PC model, the maximum item information occurs within the range of transition locations (de Ayala, 2009).

For RS model, the item information is given by

$$I_i(\theta) = \left( \sum_{k=0}^r k p_{ik} \right)^2 - \sum_{k=0}^r k^2 p_{ik}. \quad (3.30)$$

The location of the maximum of the item information function is influenced by the symmetry of the thresholds about the item location, the number of thresholds, and the range of the thresholds. Generally, items with six thresholds produce more total information across the ability continuum than items with five thresholds.

The item information function for GR model is defined as

$$I_i(\theta) = \sum_{k=1}^{r_i} \frac{(p'_k)^2}{p_k}. \quad (3.31)$$

Equation (3.31) indicates that there is an increase in item information if a response category is added between two adjacent categories. In general, the amount of information available by treating an item in a polytomous graded form is at least equal to, and more likely greater than, the amount of information available when the item is scored in a dichotomous form (de Ayala, 2009).

In the case of NR model, the amount of information provided by a given response category is

$$I_{ik}(\theta) = \alpha \mathbf{W} \alpha' p_{ik}, \quad (3.32)$$

where

$$\mathbf{W} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{r_i} \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{r_i} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{r_i}p_1 & -p_{r_i}p_2 & \cdots & p_{r_i}(1-p_{r_i}) \end{pmatrix}$$

The item information function is given by

$$\begin{aligned} I_i(\theta) &= \sum_{k=1}^{r_i} \alpha \mathbf{W} \alpha' p_{ik} \\ &= \alpha \mathbf{W} \alpha'. \end{aligned} \quad (3.33)$$

In the matrix  $\mathbf{W}$ ,  $p_{r_i}$  denotes the probability of response in the highest category ( $r_i$ ) of the item.

### Total information curve

The item information values can be combined to form the total information or test information values. Specifically, item information values at a particular ability level can be added together to obtain a total information value,  $I(\theta)$  at that ability level. That is,

$$I(\theta) = \sum_{i=1}^p I_i(\theta). \quad (3.34)$$

An instrument's total information reflects that each of the items potentially contributes some amount of information to improve the certainty about a person's ability independent of the other items on the instrument. A graph of the total information function in Equation (3.34) is the total information curve (TIC). A total information curve is useful for illustrating the extent to which an instrument (questionnaire) provides different amount of information at different ability levels. For 2PL model, the total information is given by

$$I(\theta) = \sum_{i=1}^p \alpha_i^2 p_i(\theta) [1 - p_i(\theta)] \quad (3.35)$$

The values of  $I(\theta)$  ranges between zero and the maximum number of items in the dataset. As an illustration, the total information curve of 2PL model for the TCCs in Figure 10 on the brooding scale data is displayed in Figure 13.

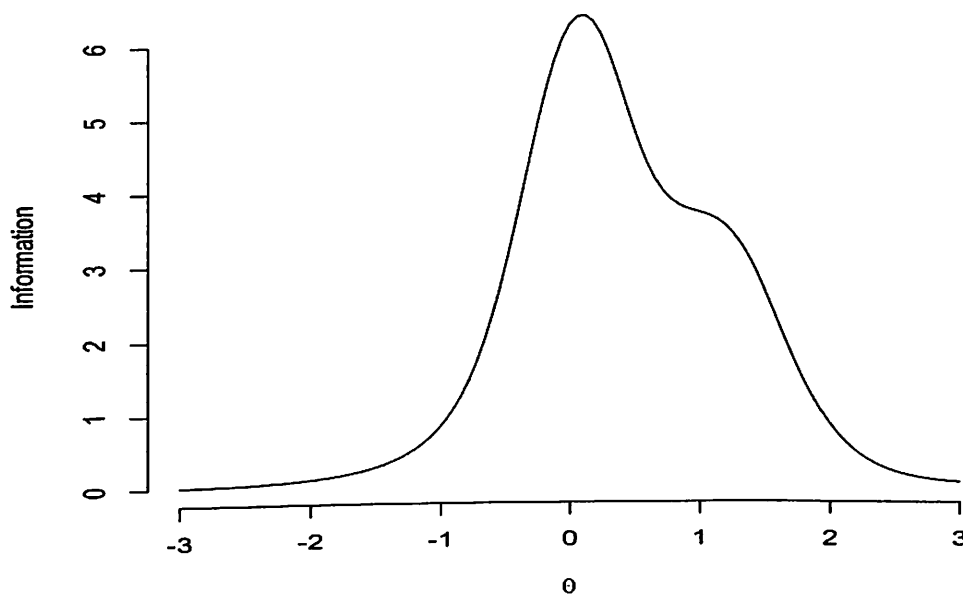


Figure 13: Total information curve for the brooding scale items

In Figure 13, the instrument provides greater information at an average ability level, and it provides less information at more extreme ability levels. That is, the instrument does well at differentiating between people who have ability

levels within 1 or 2 standard deviations of the mean. In contrast, it is relatively poor at differentiating among people who have ability levels that are more than 1 standard deviation below the mean, and it is relatively poor at differentiating among people who have ability levels that are 2 standard deviations above the mean.

### **Estimation of Parameters of IRT Models**

Undoubtedly, when it comes to the utilisation of a specified IRT model, it is particularly relevant to know the values of the parameters in the model. Most often, the values of these parameters are unknown, and which must be estimated through the use of sampled data. The parameters in the IRT models can be broadly classified into two: the person's ability parameter, and the item's parameter(s).

Several estimation procedures are available for estimating the IRT model parameters. In this study, the maximum likelihood estimation technique is considered. With this technique, the estimation of the ability parameter ( $\theta$ ) when the item parameters are considered known is referred to as conditional estimation of  $\theta$ . The case of estimating only item parameters is called the marginal maximum likelihood estimation. This procedure is achieved by integrating the maximum likelihood function with respect to the ability parameter which is assumed to be continuous (Hambleton & Swaminathan, 1985). After obtaining the estimates of item parameters, one proceeds to determine the estimates of the ability via other methods of estimation. In this section, the simultaneous estimation of both ability and item parameters, known as the joint maximum likelihood estimation, is presented. For the sake of generality, the 3PL model is considered. The presentation of the joint maximum likelihood estimation is done as in the following paragraphs.

Suppose an instrument that contains  $p$  items is administered to a group of  $n$  individuals of varied abilities. Suppose, further, that a response to a dichotomous item is scored as 1 if response is favourable, and 0 if response is not favourable. The probability of the  $j$ th person's response to the  $i$ th item is given by

$$p(x_{ij}|\theta, \delta_i, \alpha_i, c_i) = p(x_{ij} = 1)^{x_{ij}} p(x_{ij} = 0)^{1-x_{ij}}. \quad (3.36)$$

Without loss of generality, let  $p(x_{ij} = 1) = p_{ij}$ . The probability of the responses across an instrument's items is obtained by

$$p(\mathbf{x}_j|\theta, \delta, \alpha, \mathbf{c}) = \prod_{i=1}^p p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}}. \quad (3.37)$$

In Equation (3.37), the term  $p(\mathbf{x}_j|\theta, \delta, \alpha, \mathbf{c})$  is the probability of the response vector for person  $j$ ,  $\mathbf{x}_j$ , given the person's ability,  $\theta$ , a vector of item difficulties,  $\delta$ , a vector of item discriminations,  $\alpha$ , and a vector of guess values,  $\mathbf{c}$ . For item  $i$ , the probability  $p_i$  is calculated based on the specified IRT model.

The joint likelihood function,  $L$  across both persons and items, is given by

$$L = \prod_{j=1}^n \prod_{i=1}^p p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}}. \quad (3.38)$$

Applying the natural logarithmic transformation to Equation (3.38), the joint log likelihood function ( $l$ ) is obtained

$$l = \sum_{j=1}^n \sum_{i=1}^p [x_{ij} \ln p_{ij} + (1 - x_{ij}) \ln (1 - p_{ij})]. \quad (3.39)$$

The values of  $\theta$ ,  $\delta$ ,  $\alpha$  and  $\mathbf{c}$  that maximise Equation (3.39) are taken as the ability and item parameter estimates. These estimates are obtained by solving the likelihood equations

$$\frac{\partial l}{\partial v_s} = 0, \quad (s = 1, 2, \dots, n + 3p - 2), \quad (3.40)$$

where  $v_s$  is an element of the parameter vector  $\mathbf{v}$ , defined as

$$\mathbf{v}' = [\theta', \delta', \alpha', \mathbf{c}'],$$



and  $n + 3p - 2$  is the total number of parameters that have to be estimated in the 3PL model.

Equation (3.40) is a system of non-linear equations. To approximate the solution of the non-linear system, the multivariate Newton-Raphson iteration methods is applied. Suppose that  $\mathbf{v}$  is a  $p$ -dimensional vector that maximises  $f(\mathbf{v})$ . If  $\mathbf{v}^{(t)}$  is the  $t$ th approximation to the value of  $\mathbf{v}$ , then, according to the Newton-Raphson procedure, a better approximation is given by

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} - \mathbf{J}^{-1}[\mathbf{v}^{(t)}]f(\mathbf{v}^{(t)}), \quad (3.41)$$

where  $f(\mathbf{v}^{(t)})$  is the  $(p \times 1)$  vector of first partial derivatives of  $l$  (as given in Equation (3.40)) evaluated at  $\mathbf{v}^{(t)}$ . That is,  $f(\mathbf{v}^{(t)}) = \left( \frac{\partial l}{\partial \theta} \quad \frac{\partial l}{\partial \delta} \quad \frac{\partial l}{\partial \alpha} \quad \frac{\partial l}{\partial c} \right)'$ . The matrix  $\mathbf{J}[\mathbf{v}^{(t)}]$  is the Jacobian matrix of  $f$ , and it is the  $(p \times p)$  matrix of second partial derivatives of  $l$  evaluated at  $\mathbf{v}^{(t)}$ . That is,

$$\mathbf{J}[\mathbf{v}^{(t)}] = \begin{pmatrix} \frac{\partial^2 l}{\partial \theta^2} & \frac{\partial^2 l}{\partial \theta \partial \delta} & \frac{\partial^2 l}{\partial \theta \partial \alpha} & \frac{\partial^2 l}{\partial \theta \partial c} \\ \frac{\partial^2 l}{\partial \delta \partial \theta} & \frac{\partial^2 l}{\partial \delta^2} & \frac{\partial^2 l}{\partial \delta \partial \alpha} & \frac{\partial^2 l}{\partial \delta \partial c} \\ \frac{\partial^2 l}{\partial \alpha \partial \theta} & \frac{\partial^2 l}{\partial \alpha \partial \delta} & \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial c} \\ \frac{\partial^2 l}{\partial c \partial \theta} & \frac{\partial^2 l}{\partial c \partial \delta} & \frac{\partial^2 l}{\partial c \partial \alpha} & \frac{\partial^2 l}{\partial c^2} \end{pmatrix}$$

The iteration procedure in Equation (3.41) is terminated when  $\mathbf{v}^{(t)}$  does not change considerably.

To determine the maximum likelihood estimates of the parameters, the iterative procedure is carried out in two stages: (1) starting with initial values for  $\delta$ ,  $\alpha$ ,  $c$  and treating the item parameters as known,  $\theta$  is estimated, and (2) treating the ability parameter,  $\theta$  as known, the item parameters are estimated. These two stages are continued until the ability and item values converge, with the final values taken as the maximum likelihood estimates (Hambleton & Swaminathan, 1985).

In respect of the current circumstance, suppose that for item  $i$ ,

$$\mathbf{v}_i = [\delta_i \quad \alpha_i \quad c_i]'$$

If  $\mathbf{v}_i^{(t)}$  is the  $t$ th approximation, then

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} - \mathbf{J}^{-1}[\mathbf{v}_i^{(t)}]f(\mathbf{v}_i^{(t)}), \quad (3.42)$$

where  $f(\mathbf{v}_i^{(t)}) = \left[ \frac{\partial L}{\partial \delta_i}[\mathbf{v}_i^{(t)}] \quad \frac{\partial L}{\partial \alpha_i}[\mathbf{v}_i^{(t)}] \quad \frac{\partial L}{\partial c_i}[\mathbf{v}_i^{(t)}] \right]'$  and

$$\mathbf{J}[\mathbf{v}_i^{(t)}] = \begin{pmatrix} \frac{\partial^2 L}{\partial \delta_i^2} & \frac{\partial^2 L}{\partial \delta_i \partial \alpha_i} & \frac{\partial^2 L}{\partial \delta_i \partial c_i} \\ \frac{\partial^2 L}{\partial \alpha_i \partial \delta_i} & \frac{\partial^2 L}{\partial \alpha_i^2} & \frac{\partial^2 L}{\partial \alpha_i \partial c_i} \\ \frac{\partial^2 L}{\partial c_i \partial \delta_i} & \frac{\partial^2 L}{\partial c_i \partial \alpha_i} & \frac{\partial^2 L}{\partial c_i^2} \end{pmatrix}$$

The process in Equation (3.42) is terminated when the difference between the  $(t + 1)$ th and the  $t$ th approximations is sufficiently small. The iteration method is carried out for  $n$  items. When convergence takes place, the item parameters are treated as known, and the ability parameters are estimated.

### Assessment of the Fitness of IRT Models

Assessing the fitness of models to data are routine in statistical procedures and involve determining whether the model could have generated the observed data. In IRT, the task of assessing model fitness requires: (1) checking underlying assumptions such as dimensionality, local independence, and monotonicity; and (2) assessing the agreement between observations and model predictions. Numerous procedures are available for checking model assumptions (e.g., see Embretson & Reise, 2000; Hambleton et al., 1991). In this current discussion, we focus on the goodness-of-fit of IRT models. Assessment of the fitness of the model to data is multi-faceted and must be carried out at the overall model level as well as the item level.

#### Overall model fitness

The assessment of the overall fitness of IRT model can be accomplished using different approaches (Hambleton et al., 1991; Reise & Revicki, 2015). A

typical method is the standard Chi-square test which, in IRT context, relies on all the observed and expected response patterns of items being modelled. The test statistic can be used to test the null hypothesis that the IRT model for the pattern of responses is correctly specified, and given by

$$Q = \sum_{r=1}^R \frac{(O_r - E_r)^2}{E_r},$$

where  $O_r$  and  $E_r$  are observed and expected frequencies of the  $r$ th response pattern, and  $R$  is the total number of response patterns. When the null hypothesis holds, the  $Q$  statistic has an approximate Chi-square distribution with  $(R - s - 1)$  degrees of freedom, where  $s$  is the number of item parameters estimated in the IRT model. The null hypothesis is rejected (i.e., the IRT model does not fit the data) when the value of  $Q$  exceeds a critical value obtained from the Chi-square distribution with the specified degrees of freedom.

#### Assessment of item fitness

The methods for assessing the fitness of items to item response models are based on discrepancy measures between observed and expected probability of a favourable response (i.e., yes response on a two-point scale) at various points on the ability continuum or at response category scores (Hambleton et al., 1991). When standard Chi-square test is applied, examinees are ordered based on each ability level, and sorted into specific number of groups (treating each subgroup as though all examinees were at the same ability level). Within each ability subgroup, the exact observed and expected probabilities of favourable responses are compared based on the IRT model. In the case of ordinal polytomous responses, differences can be computed within each response category (Dodeen, 2004). The test statistic,  $Q_i$  for assessing item fitness is given by

$$Q_i = \sum_{j=1}^J \sum_{g=0}^k n_{jg} \frac{(O_{jg} - E_{jg})^2}{E_{jg}}, \quad (3.43)$$

where  $j$  denotes the ability interval and  $g$  denotes the response category. In terms of dichotomous item responses, Equation (3.43) reduces to

$$Q_i = \sum_{j=1}^J \frac{n_j (O_j - E_j)^2}{E_j(1 - E_j)},$$

where  $n_j$  denotes number of persons falling into the  $j$ th ability interval. Different forms of  $Q_i$  can be obtained by varying the numbers and methods of constructing ability intervals. The test statistic  $Q_i$ , under the null hypothesis, has an approximate Chi-square distribution with  $J(k - 1) - s$  degrees of freedom.

A number of problems arise in using Chi-square statistic as tests of model fitness to data in IRT context. Firstly, the ability intervals,  $J$  used in the computation of the test statistic is arbitrary. This indicates that, different choices of  $J$  can be made, which can lead to different values of the test statistic and ostensibly different conclusions about the fitness of items. A related issue is the of the minimum interval size needed for a Chi-square approximation to be valid.

Another problem with Chi-square statistics has to do with sample size. When a large sample size is used, the power of the test increases, and many items tend to misfit the model (Dodeen, 2004). The number of response patterns increases considerably when the number of items or number of response categories increase. For example, ten items with seven response categories each allow for 10,000,000 response patterns. Practically, the number of response patterns might be markedly larger than the sample size, which in this case, leads to sparseness in the data. Sparseness is present when the ratio of sample size to the total number of response patterns is small and there are many response patterns with expected frequencies less than one (Agresti & Yang, 1987). When data are sparse, the Chi-square approximation does not hold for the distribution of the fitness statistics. In addition, the goodness-of-fit statistic is inflated and rates of rejection of the null hypothesis are too high. In order to improve model fitness to sparse data, the following are recommended: (a) adding small constants to cells,

(b) merging cells, (c) considering only cells with observed or expected frequencies that exceed a certain value, and (d) deriving the small sample distribution of fitness statistic by bootstrapping (Kraus, 2012).

### **Multidimensional Item Response Theory**

With respect to the IRT models presented in the previous sections, a person's response to a set of items is accounted for by only one latent ability. However, a questionnaire usually comprises some groups of items measuring different but related abilities. In this case, the IRT unidimensionality assumption of the person ability may be excessively obstructive.

Multidimensional item response theory (MIRT) is a generalisation of unidimensional item response theory (UIRT). When the unidimensionality assumption required by IRT models is violated, MIRT can be used to model the relationship between two or more abilities and the probability of responses to items in a questionnaire. In the realm of MIRT, each person's response to an item is influenced by a combination of two or more abilities. In MIRT, several parameters are used to measure person abilities and a vector of parameters are used to characterise the items (Duong, Subedi, & Lee, 2008).

### **Assumptions of MIRT models**

The assumptions underlying the MIRT models are:

1. **Functional form:** The probability of response increases monotonically when there is an increase in any one or any combination of a person's abilities, and that for infinitely low abilities the probability of response approaches zero (de Ayala, 2009).
2. **Conditional independence:** For any group of individuals that are characterised by the same abilities, the conditional distributions of the item

responses are all independent of one another.

3. Dimensionality: The probabilities of responses are functions of a set of continuous person latent abilities.

### **Types of MIRT models**

The types of MIRT models are defined by the way in which information from person abilities is combined with item parameters to compute the probability of responses to the item. In this regard, a MIRT model can be classified broadly as either compensatory or non-compensatory. For a compensatory MIRT model, a linear combination of abilities is required to obtain the probability of a response. That is, a high ability on one dimension can compensate for a low ability on another dimension when calculating the probability of a response. Under the compensatory MIRT models, persons with different sets of abilities can have the same probability of response to an item. These models are said to be additive.

Non-compensatory MIRT model separates different dimensions in ability into parts and uses a unidimensional model for each part. The probability of response for the item is the product of the probabilities for each part. This product of probabilities results in a non-linear form of the non-compensatory models. These models are considered to be multiplicative. The non-compensatory MIRT models are also appropriately referred to as partially compensatory models because an increase in one of the ability values can improve the overall probability, but only up to the limit set by the lowest term in the product. Thus, the compensation effect is not totally removed (Reckase, 2009).

With these two major types of MIRT models, there are a variety of models. In this work, only compensatory MIRT models are discussed due to its wider applications. Whether compensatory or non-compensatory, MIRT model can

either be dichotomous or polytomous just as the unidimensional case.

### *Dichotomous MIRT models*

The simplest MIRT model is the Rasch multidimensional item response theory model. The Rasch MIRT model is an extension of the unidimensional Rasch model (see Equation (3.1)). According to this model, the probability of a positive response is given by

$$p(X_{ij} = 1 | \boldsymbol{\theta}_j, d_i) = \frac{1}{1 + \exp[-(\mathbf{1}'\boldsymbol{\theta}_j + d_i)]}, \quad (3.44)$$

where  $\boldsymbol{\theta}_j$  is an  $m \times 1$  vector of abilities for person  $j$  with  $m$  dimensions in the ability space,  $d_i = -m\delta_i$  is a measure of difficulty for item  $i$  and  $\mathbf{1}$  is an  $m \times 1$  vector of 1s. Equation (3.44) shows that the Rasch MIRT model differs from the unidimensional Rasch model with respect to the manner in which the ability parameter,  $\theta$  is measured. For the Rasch MIRT model, the ability parameter is a value that is obtained by summing the different ability dimensions rather than just a single construct.

As indicated in Equation (3.2), the 1PL model considers the discrimination parameter to be fixed for all items. This model is extended by the multidimensional 1PL (MIPL) model. In the MIPL model the probability of a favourable response is obtained as

$$p(X_{ij} = 1 | \boldsymbol{\theta}_j, d_i) = \frac{1}{1 + \exp[-(\boldsymbol{\alpha}_i'\boldsymbol{\theta}_j + d_i)]}, \quad (3.45)$$

where  $d_i = -\boldsymbol{\alpha}_i'\boldsymbol{\theta}_j$  is a measure of difficulty for item  $i$ .

The multidimensional two-parameter logistic (M2PL) model is an extension of the 2PL model. Reckase expressed the M2PL model as

$$p(X_{ij} = 1 | \boldsymbol{\theta}_j, \boldsymbol{\alpha}_i, d_i) = \frac{1}{1 + \exp[-1.702(\boldsymbol{\alpha}_i'\boldsymbol{\theta}_j + d_i)]}, \quad (3.46)$$

where  $\boldsymbol{\alpha}_i$  is a vector of discrimination parameters for item  $i$  and  $d_i = -\mathbf{1}'\boldsymbol{\alpha}_i\delta_i$  is a scalar parameter that is related to the item's difficulty. The exponent in

Equation (3.46) can be expanded to give

$$\begin{aligned}\alpha'_i \theta_j + d_i &= \alpha_{i1} \theta_{j1} + \alpha_{i2} \theta_{j2} + \cdots + \alpha_{il} \theta_{jl} + \cdots + \alpha_{im} \theta_{jm} + d_i \\ &= \sum_{l=1}^m \alpha_{il} \theta_{jl} + d_i,\end{aligned}\tag{3.47}$$

which indicates how the elements of the  $\alpha$  and  $\theta$  vectors interact. Equation (3.47) shows that the exponent in the M2PL model is a linear combination of the elements of  $\theta$ . This feature reflects the compensatory nature of the M2PL model. Equation (3.47) defines a line in an  $m$ -dimensional space. If the exponent is set to some constant,  $k$ , that is

$$\alpha'_i \theta_j + d_i = k,\tag{3.48}$$

then all  $\theta$ -vectors satisfying Equation (3.48) will fall along the same straight line with the same probability of a favourable response for the model.

The M2PL model provides the possibility for extension of the unidimensional 3PL model to a multidimensional sense. The multidimensional three-parameter logistic (M3PL) model is given by (Reckase, 2009)

$$p(X_{ij} = 1 | \theta_j, \alpha_i, c_i, d_i) = c_i + (1 - c_i) \frac{1}{1 + \exp[-1.702(\alpha'_i \theta_j + d_i)]}\tag{3.49}$$

The M3PL model accounts for the probability of a favourable response to an item among persons having low abilities, with a guess,  $c_i$ .

### *Polytomous MIRT models*

Polytomous IRT models have been presented under the unidimensional IRT models. In this section, polytomous IRT models have been developed to encompass a multidimensional person ability.

The multidimensional partial credit (MPC) model is an extension of the partial credit (PC) model. The PC model has the properties of the Rasch model where the only item characteristic is the difficulty parameter. In MPC model,



the ability score,  $\theta$ , of the PC model is considered to be multidimensional. The probability of a response to item  $i$  in category  $g$  is given by (Kelderman & Rijkes, 1994)

$$P(X_{ij} = g | \theta, \delta_{ih}) = \frac{\exp \left[ \sum_{l=1}^m (\theta_{jl} - \delta_{ilg}) \right]}{\sum_{h=0}^{r_i} \exp \left[ \sum_{l=1}^m (\theta_{jl} - \delta_{ilh}) \right]}, \quad (3.50)$$

where  $\delta_{ilg}$  is the difficulty for item  $i$  on dimension  $l$  ( $l = 1, 2, \dots, m$ ) for category  $g$  ( $g = 0, 1, 2, \dots, k$ ). In Equation (3.50), each response score  $g$  may be seen as the result of performing a series of steps. To obtain a response score  $g$ ,  $g$  steps must be performed. Each step in the MPC model depends on a different ability dimension,  $\theta_{jl}$ . The probability of a response in a given category utilises a series of dichotomous Rasch model for each dimension with a difficulty for that dimension.

In the estimation of the category difficulty parameters,  $\delta_{ilg}$ , Kelderman and Rijkes (1994) note that indeterminacy exists between  $\delta_{ilg}$  of different response scores  $g$  within the same ability dimension  $l$  and item  $i$ . However, this indeterminacy can be eliminated by setting the difficulty parameters equal across the different response categories.

The multidimensional generalised partial credit (MGPC) model which is an extension of the generalised partial credit (GPC) model or the 2PPL model is meant to describe the interaction of persons with items that have been scored polytomously. The MGPC model allows the person characteristics to be represented multidimensionally. The score assigned to person  $j$  on item  $i$  is denoted by  $k = 0, 1, 2, \dots, g, \dots, r_i$ . The MGPC model expresses the probability of a response in category  $g$  as

$$P(X_{ij} = g | \theta_j, \alpha_i, \delta_{ih}) = \frac{\exp \left( g \alpha_i' \theta_j - \sum_{h=0}^g \delta_{ih} \right)}{\sum_{k=0}^{r_i} \exp \left( k \alpha_i' \theta_j - \sum_{h=0}^k \delta_{ih} \right)}. \quad (3.51)$$

Another technique to multidimensional modelling of polytomous item responses is the multidimensional graded response (MGR) model. The MGR model is an extension of the unidimensional graded response (GR) model. The MGR model assumes that answering in a given response category of an item requires a number of steps and reaching step  $g$  requires success on step  $g - 1$ . To construct the MGR model the response scale is dichotomised at  $g$ , scoring response category  $g$  and above as 1 and below  $g$  as 0, and fitting a dichotomous model to the result. The probability of responding in category  $g$  and above is modelled by a M2PL model, and expressed as

$$P(X_{ij} \geq g | \boldsymbol{\theta}_j) = \frac{1}{1 + \exp[-(\boldsymbol{\alpha}'_i \boldsymbol{\theta}_j + d_{ig})]}. \quad (3.52)$$

Equation (3.52) shows that the probability of responding in category  $g$  and above increases with an increase in any of the elements of the person ability vector,  $\boldsymbol{\theta}_j$ . The probability of a person responding in a specific category  $g$  is obtained as

$$\begin{aligned} p(X_{ij} = g | \boldsymbol{\theta}_j) &= P(X_{ij} \geq g | \boldsymbol{\theta}_j) - P(X_{ij} \geq g + 1 | \boldsymbol{\theta}_j) \\ &= \frac{1}{1 + \exp[-(\boldsymbol{\alpha}'_i \boldsymbol{\theta}_j + d_{ig})]} - \frac{1}{1 + \exp[-(\boldsymbol{\alpha}'_i \boldsymbol{\theta}_j + d_{i,g+1})]}, \end{aligned} \quad (3.53)$$

where  $d_{ig}$  measures the ease with which a person will respond in category  $g$  of item  $i$ . The  $d_{ig}$  parameter will have high positive values when it is relatively easy to respond in a specific category and large negative values when it is difficult to answer in a given category.

### **Multidimensional Person and Item Parameters**

The nature of items in a questionnaire differ from survey to survey. Some items are simple and require just a single person ability to respond to them. Other items are complex and may require multiple abilities to respond to them. In this case, a model based on multiple dimensions of person ability is needed.

An ideal model of the relationship between the probability of response and the ability of a person uses a vector to represent the ability of the person in a multidimensional space. The ability of person  $j$  is denoted by the vector  $\theta_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jl}, \dots, \theta_{jm})'$ , with  $m$  dimensions. The elements of  $\theta_j$  provide coordinates for the location of a person in a multidimensional space.

In MIRT, the characteristics of an item is described by parameters of the model. For instance, in M3PL model the characteristics of the item can be summarised by  $\alpha$ -vector, and the scalar  $c$  and  $d$  parameters. However, unlike unidimensional IRT models, parameters in the MIRT models do not have intuitive meaning. To aid the interpretation of MIRT models, some measures have been derived from the parameters. For the M2PL model, Reckase and McKinley (1991) showed a measure of item discrimination, the multidimensional discrimination (MDISC), given as

$$MDISC = \sqrt{\alpha'_i \alpha_i}, \quad (3.54)$$

where  $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{il}, \dots, \alpha_{im})'$ . Equation (3.54) shows that when  $\alpha_{il} > 0$  and  $\alpha_{ik} = 0$  for all  $k \neq l$ ,  $MDISC = \alpha_{il}$ . In this case, MDISC is equal and interpreted as the unidimensional discrimination parameter ( $\alpha_i$ ).

The multidimensional difficulty (MDIFF) of an item is a composite measure of the difficulty level of an item and defined by (Reckase, 1985)

$$MDIFF = -\frac{d_i}{\sqrt{\alpha'_i \alpha_i}}. \quad (3.55)$$

The MDIFF can be interpreted just as the unidimensional item difficulty,  $\delta_i$ .

### MIRT Graphical Representations

Characteristics of items in MIRT models can be represented in a Cartesian coordinate system known as item characteristic surface (ICS) or item response surface (IRS). Item characteristic surface is an analogue to the unidimensional

item characteristic curve (ICC). For M3PL model, three characteristics of an item can be represented using item characteristic surface: difficulty ( $d_i$ ), discrimination ( $\alpha_i$ ) and guess ( $c_i$ ). The probability of positive response for a specified ability vector,  $\theta_j$  can be depicted graphically using ICS. Unlike unidimensional ICC, a single item can be represented in the ICs at a time. For a given ICS, a corresponding contour line representing probabilities of positive response can be constructed. Ackerman (1996) underpinned three features of an item that are more revealing in equiprobable contour plots than ICS These are as follows:

1. The ability vector,  $\theta_j$  the item is best measuring;
2. The region in the ability space the item is most discriminating. The more discriminating the item, the closer together the equiprobability contours;  
and
3. The difficulty of the item.

Illustratively, ICS and contour plot for two-dimensional M2PL model with  $\alpha_i = (1.5, 0.4)'$  and  $d_i = 1.2$  is shown in Figure 14.

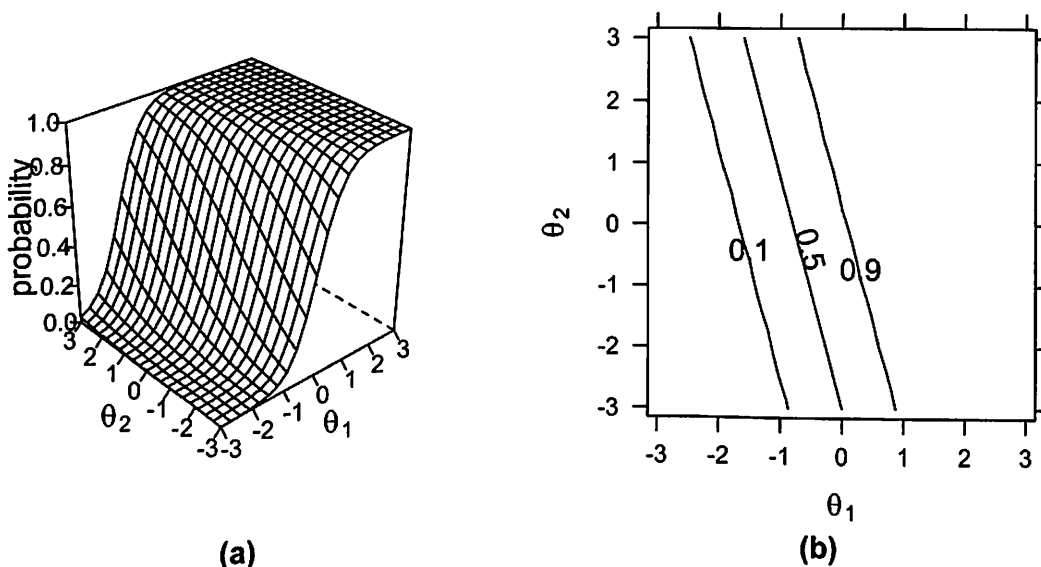


Figure 14: M2PL model's (a) ICS and (b) contour plot for an item

Figure 14 (a) shows the item characteristic surface, which represents the probability of positive response to the item in the  $\theta$  space, and (b) shows the probabilities of positive response as contours of the surface. The plots indicate that probability of positive response increases with one or both dimensions of  $\theta$ . The probability of positive response increases more rapidly along  $\theta_1$  dimension than along  $\theta_2$  dimension owing to the differences in the elements of  $\alpha$ . This means that the form of ICS and corresponding equiprobable contours are influenced by the  $\alpha$  parameter for the item. Contours of equal probabilities form straight lines as a result of the linear form of the exponent in the M2PL model (see Equation (3.47)). That is,

$$1.5\theta_1 + 0.4\theta_2 + 1.2 = k. \quad (3.56)$$

In Equation (3.56) and from Equation (3.46), when  $k$  equals 0, the probability of positive response is 0.5. Any combination of the elements of  $\theta$  that yields  $k$  equals 0 defines the coordinates of the equiprobable contour of 0.5. Several equiprobable contours are drawn by changing the value of  $k$ . The difficulty of an item is the distance the contour lines are from the origin which is dependent on  $d$  and  $\alpha$  parameters (Reckase, 2009). It can be deduced from Equation (3.55) that along  $\theta_1$  axis (where  $\theta_2$  is zero), relative difficulty of the item would be 0.80, but along  $\theta_2$  axis (where  $\theta_1$  is zero), it would be 3.0. This means that along  $\theta_2$  axis, a person would require an ability of 3.0 to obtain a 0.5 probability of positive response. An overall estimate of item difficulty is 0.77.

The total characteristic surface (TCS) generalises the unidimensional total characteristic curve (TCC). Total characteristic surface is the aggregation of item characteristic surfaces. It is found by summing the probabilities of positive response for each of  $p$  items in a questionnaire. The total characteristic surface is defined by

$$T(\theta) = \sum_{i=1}^p P_i(\theta), \quad (3.57)$$

where  $T(\boldsymbol{\theta})$  is interpreted as the expected number of positive responses specified by  $\boldsymbol{\theta}$ , and  $P_i(\boldsymbol{\theta})$  is the probability of positive response to item  $i$  defined by the underlying MIRT model. Figure 15 shows total characteristic surface and its associated equiprobable contours for ten items based on M2PL model.

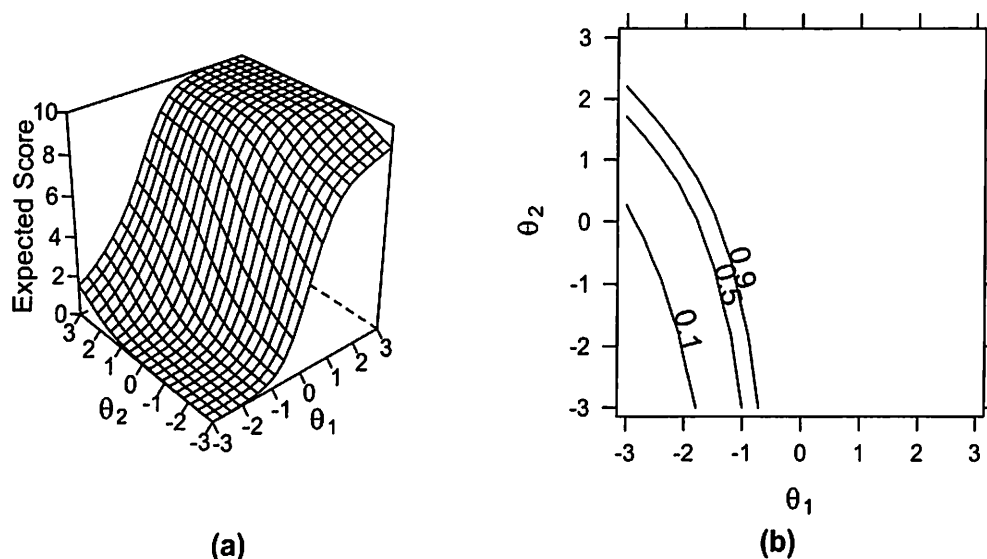


Figure 15: M2PL model's (a) total characteristic surface and (b) contour plot for ten items

The way in which an item functions as its capability to differentiate between two locations in the  $\boldsymbol{\theta}$  space can be represented graphically, known as item information surface (IIS). Item information surface is a multidimensional counterpart of unidimensional item information curve (IIC). Item information surface for item  $i$  is given by (Reckase, 2009)

$$I_i(\boldsymbol{\theta}) = \frac{[\nabla P_i(\boldsymbol{\theta})]^2}{P_i(\boldsymbol{\theta})[1 - P_i(\boldsymbol{\theta})]}, \quad (3.58)$$

where  $\nabla P_i(\boldsymbol{\theta})$  is the gradient of the item characteristic surface and measures the rate of change in the probability of positive response, given  $\boldsymbol{\theta}$ . The information provided by the item about  $\boldsymbol{\theta}$  depends on the desired direction since the gradient

of the item response surface depends on the direction. As a result, the numerator of Equation (3.58) is replaced with the directional derivative of the response surface. In this case, the item information surface in the direction of  $\alpha$  is given by

$$I_{i\alpha}(\theta) = \frac{[D_{\alpha}P_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \quad (3.59)$$

where  $D_{\alpha}P_i(\theta)$  is the directional derivative of the item characteristic surface in the direction of  $\alpha$ , and defined as

$$D_{\alpha}P_i(\theta) = \nabla P_i(\theta) \cdot \hat{\alpha}, \quad (3.60)$$

and  $\hat{\alpha}$  is the normalised vector of  $\alpha$ . That is,

$$\begin{aligned} D_{\alpha}P_i(\theta) &= \left[ \frac{\partial}{\partial \theta_1} P_i(\theta), \frac{\partial}{\partial \theta_2} P_i(\theta), \dots, \frac{\partial}{\partial \theta_l} P_i(\theta), \dots, \frac{\partial}{\partial \theta_m} P_i(\theta) \right]' \cdot \frac{\alpha}{|\alpha|} \\ &= \left[ \frac{\partial}{\partial \theta_1} P_i(\theta), \frac{\partial}{\partial \theta_2} P_i(\theta), \dots, \frac{\partial}{\partial \theta_l} P_i(\theta), \dots, \frac{\partial}{\partial \theta_m} P_i(\theta) \right]' \cdot |\hat{\alpha}| \cos \beta \\ &= \left[ \frac{\partial}{\partial \theta_1} P_i(\theta), \frac{\partial}{\partial \theta_2} P_i(\theta), \dots, \frac{\partial}{\partial \theta_l} P_i(\theta), \dots, \frac{\partial}{\partial \theta_m} P_i(\theta) \right]' \cdot \cos \beta, \end{aligned} \quad (3.61)$$

where  $\beta = (\beta_1, \beta_2, \dots, \beta_l, \dots, \beta_m)'$  is the vector of angles between the gradient of the surface and the normalised vector. In other words,  $\beta_l$  is the angle between  $\theta_l$  coordinate axis and the line from the origin to the person's location in the  $\theta$  space. Equation (3.61) becomes

$$\begin{aligned} D_{\alpha}P_i(\theta) &= \frac{\partial}{\partial \theta_1} P_i(\theta) \cos \beta_1 + \frac{\partial}{\partial \theta_2} P_i(\theta) \cos \beta_2 + \dots + \frac{\partial}{\partial \theta_l} P_i(\theta) \cos \beta_l + \dots \\ &\quad + \frac{\partial}{\partial \theta_m} P_i(\theta) \cos \beta_m. \end{aligned} \quad (3.62)$$

For M2PL model,

$$\begin{aligned} D_{\alpha}P_i(\theta) &= \alpha_1 P_i(\theta) [1 - P_i(\theta)] \cos \beta_1 + \alpha_2 P_i(\theta) [1 - P_i(\theta)] \cos \beta_2 + \\ &\quad \dots + \alpha_l P_i(\theta) [1 - P_i(\theta)] \cos \beta_l + \dots + \alpha_m P_i(\theta) [1 - P_i(\theta)] \cos \beta_m \\ &= P_i(\theta) [1 - P_i(\theta)] \sum_{l=1}^m \alpha_l \cos \beta_l. \end{aligned} \quad (3.63)$$

Substituting Equation (3.63) into Equation (3.59) gives

$$\begin{aligned}
 I_{i\alpha}(\boldsymbol{\theta}) &= \frac{\{P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})] \sum_{l=1}^m \alpha_l \cos \beta_l\}^2}{P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})]} \\
 &= P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})] \left( \sum_{l=1}^m \alpha_l \cos \beta_l \right)^2. \quad (3.64)
 \end{aligned}$$

In Equation (3.63), the maximum value of  $D_{\alpha}P_i(\boldsymbol{\theta})$  occurs if  $\cos \beta_l = 1$ ; that is  $\beta_l = 0$ , in which case the normalised vector  $\hat{\alpha}$  has the same direction as the gradient of the response surface,  $\nabla P_i(\boldsymbol{\theta})$ . In this direction,

$$D_{\alpha}P_i(\boldsymbol{\theta}) = P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})] \sum_{l=1}^m \alpha_l \quad (3.65)$$

In this case, the maximum item information value for M2PL model is given by

$$\begin{aligned}
 I_{i\alpha\max}(\boldsymbol{\theta}) &= P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})] \sum_{l=1}^m \alpha_l^2 \\
 &= \boldsymbol{\alpha}'\boldsymbol{\alpha}P_i(\boldsymbol{\theta}) [1 - P_i(\boldsymbol{\theta})]. \quad (3.66)
 \end{aligned}$$

The form of item information in Equation (3.66) has the same orientation as that for the unidimensional 2PL model. Figure 16 shows information surface for an item with the same parameters as used in Figure 14.

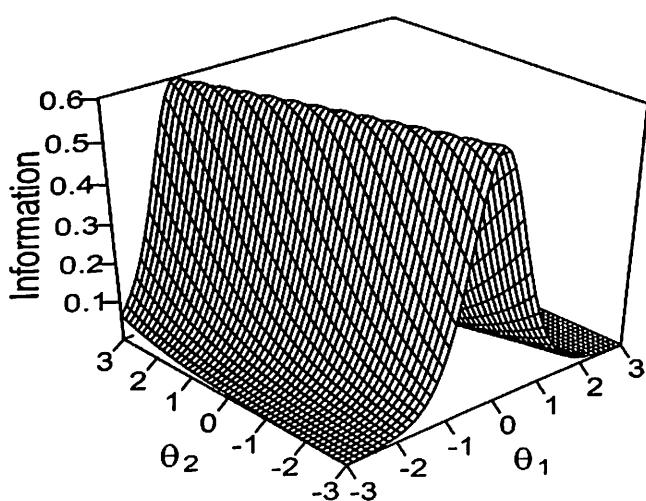


Figure 16: Item information surface



For a set of items in a questionnaire, total information surface about  $\theta$  can be computed by summing the item information values in a particular direction. Figure 17 shows total information for a set of ten items with same parameters as used in Figure 15.

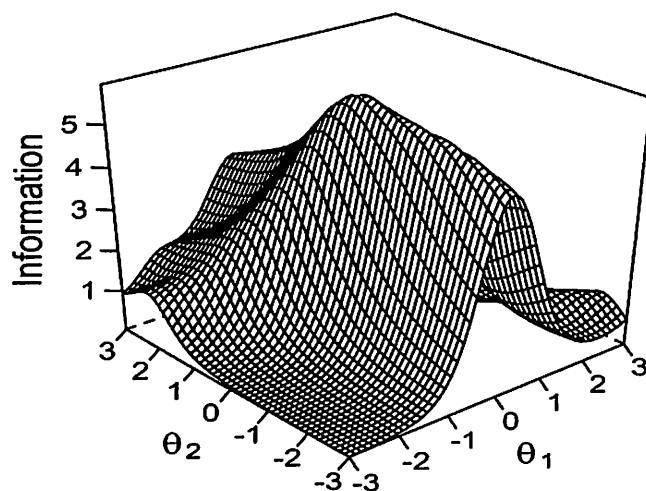


Figure 17: Total information surface for ten items

### Factor Analysis

Factor analysis is a multivariate statistical technique that is employed to discover which variables (indicators) in a set form meaningful subsets that are relatively independent of one another. Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors (abilities in IRT). These factors are thought to reflect underlying processes that have created the correlations among variables (Tabachnick & Fidell, 2013).

### Factor model definition

Suppose that a random sample  $y_1, y_2, \dots, y_n$  from a homogeneous population with mean vector,  $\boldsymbol{\mu}$  and covariance matrix,  $\boldsymbol{\Sigma}$ . For  $y_1, y_2, \dots, y_p$  in any observation vector,  $\mathbf{y}$ , the model is given by

$$\begin{aligned}
 y_1 &= \lambda_{11}\theta_1 + \lambda_{12}\theta_2 + \dots + \lambda_{1l}\theta_l + \dots + \lambda_{1m}\theta_m + \varepsilon_1 \\
 y_2 &= \lambda_{21}\theta_1 + \lambda_{22}\theta_2 + \dots + \lambda_{2l}\theta_l + \dots + \lambda_{2m}\theta_m + \varepsilon_2 \\
 &\vdots \\
 y_i &= \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \dots + \lambda_{il}\theta_l + \dots + \lambda_{im}\theta_m + \varepsilon_i \\
 &\vdots \\
 y_p &= \lambda_{p1}\theta_1 + \lambda_{p2}\theta_2 + \dots + \lambda_{pl}\theta_l + \dots + \lambda_{pm}\theta_m + \varepsilon_p
 \end{aligned} \tag{3.67}$$

where  $\theta_1, \theta_2, \dots, \theta_m$  are the common factors (latent abilities in IRT); the coefficients  $\lambda_{il}$  are the loadings; and the error terms  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$  are unique factors. The system of Equations 3.67 expresses each  $y_i$  as a linear combination of the factors  $\theta_l$  with an accompanying error term  $\varepsilon_i$  to account for the part of the variable  $y_i$  that is unique. The loadings  $\lambda_{il}$  can be used in the interpretation of the factors. For instance,  $\theta_m$  may be interpreted by examining its loadings  $\lambda_{1m}, \lambda_{2m}, \dots, \lambda_{pm}$  and noting the  $y$ 's that have large loadings on  $\theta_m$ . This subset of  $y$ 's gives an identification to  $\theta_m$ .

In matrix notation, Equation (3.67) can be written as

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{3.68}$$

where  $\mathbf{y} = (y_1, y_2, \dots, y_p)'$ ,  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)'$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)'$ , and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \dots & \lambda_{pm} \end{pmatrix}. \tag{3.69}$$

## Assumptions of the factor model

Without loss of generality, it is assumed that for  $i = 1, 2, \dots, p$  and  $l = 1, 2, \dots, m$ :

1.  $E(\theta_l) = 0$ ,  $E(\epsilon_i) = 0$ , and  $E(y_i) = 0$ .
2.  $\text{var}(\theta_l) = 1$  and  $\text{var}(\epsilon_i) = \psi_i$ .
3.  $\text{cov}(\theta_l, \theta_k) = 0$ , for  $l \neq k$ ;  $\text{cov}(\epsilon_i, \epsilon_k) = 0$ , for  $i \neq k$ ; and  $\text{cov}(\epsilon_i, \theta_l) = 0$ , for all  $i$  and  $l$ .

Assumption 1 shows that the means of the common factors, unique factors, and indicator variables are zero. The assumptions for  $\epsilon_i$  are similar to those of  $\theta_l$  except that each  $\epsilon_i$  is allowed to have different variance  $\psi_i$ . Assumption 3 indicates that the unique factors are uncorrelated among themselves or with the common factor. The assumption  $\text{cov}(\epsilon_i, \epsilon_k) = 0$  implies that the factors account for all the correlations among the  $y$ 's. Thus, the emphasis in factor analysis is on modelling the covariances or correlations among the  $y$ 's.

Assumptions 1, 2 and 3 can be expressed concisely using vector and matrix notation

$$E(\boldsymbol{\theta}) = \mathbf{0}_{(m \times 1)} \quad (3.70)$$

$$\text{var}(\boldsymbol{\theta}) = E(\boldsymbol{\theta}\boldsymbol{\theta}') = \mathbf{I}_{(m \times m)} \quad (3.71)$$

$$E(\boldsymbol{\epsilon}) = \mathbf{0}_{(p \times 1)} \quad (3.72)$$

$$\text{cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Psi}_{(p \times p)} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix} \quad (3.73)$$

$$\text{cov}(\boldsymbol{\varepsilon}, \boldsymbol{\theta}) = \mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\theta}') = \mathbf{0}_{(p \times m)} \quad (3.74)$$

$$\mathbf{E}(\mathbf{y}) = \mathbf{0}_{(p \times 1)} \quad (3.75)$$

### Covariance structure of the factor model

The variance-covariance of the observed variables,  $\boldsymbol{\Sigma}$  can be expressed in terms of the factor loadings and the unique factors. From Equation (3.68),

$$\begin{aligned} \boldsymbol{\Sigma} = \text{cov}(\mathbf{y}) &= \text{cov}(\boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \\ &= \text{cov}(\boldsymbol{\Lambda}\boldsymbol{\theta}) + \text{cov}(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\Lambda}\text{cov}(\boldsymbol{\theta})\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \end{aligned} \quad (3.76)$$

Thus, Equation (3.76) represents a simplified structure for  $\boldsymbol{\Sigma}$ , in which the covariances are modelled by the  $\lambda_{il}$ s alone since  $\boldsymbol{\Psi}$  is diagonal. From Equation (3.76),

$$\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2 + \psi_i \quad (3.77)$$

Thus, the variance of  $y_i$  is partitioned into a part that is due to the common factors, called the communality (denoted  $h_i^2$ ), and a part that is unique to  $y_i$ , known as the specific variance ( $\psi_i$ ). That is,

$$h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2. \quad (3.78)$$

Again, considering Equation (3.76),

$$\text{cov}(y_i, y_k) = \lambda_{i1}\lambda_{k1} + \lambda_{i2}\lambda_{k2} + \cdots + \lambda_{im}\lambda_{km}. \quad (3.79)$$

Thus, the covariance between any two variables is the sum of cross-product of corresponding factor loadings.

The covariances of the  $y$ 's with the  $\theta$ 's can also be found in terms of the  $\lambda$ 's. That is,

$$\begin{aligned}
 \text{cov}(\mathbf{y}, \boldsymbol{\theta}) &= E(\mathbf{y}\boldsymbol{\theta}') \\
 &= E[(\boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon})\boldsymbol{\theta}'] \\
 &= E(\boldsymbol{\Lambda}\boldsymbol{\theta}\boldsymbol{\theta}' + \boldsymbol{\varepsilon}\boldsymbol{\theta}') \\
 &= E(\boldsymbol{\Lambda}\boldsymbol{\theta}\boldsymbol{\theta}') + E(\boldsymbol{\varepsilon}\boldsymbol{\theta}') \\
 &= \boldsymbol{\Lambda}E(\boldsymbol{\theta}\boldsymbol{\theta}') + E(\boldsymbol{\varepsilon}\boldsymbol{\theta}') \\
 &= \boldsymbol{\Lambda}.
 \end{aligned} \tag{3.80}$$

Since  $\lambda_{il}$  is the  $(i - l)$ th element of  $\boldsymbol{\Lambda}$ , Equation (3.80) can be written as

$$\text{cov}(y_i, \theta_l) = \lambda_{il}. \tag{3.81}$$

Thus, in Equation (3.81), the loadings represent correlations of the variables with the factors.

If standardised variables are used, Equation (3.76) is replaced by a model for the correlation matrix,  $\mathbf{R}$ , as

$$\mathbf{R} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi}. \tag{3.82}$$

### **Estimation of Parameters of the Factor Model**

The covariance,  $\boldsymbol{\Sigma}$  and correlation,  $\mathbf{R}$  matrices of the observed variables are functions of the parameter matrices  $\boldsymbol{\Lambda}$  and  $\boldsymbol{\psi}$ . An initial problem in factor analysis is estimating the factor loadings  $\lambda_{il}$  and specific variances  $\psi_i$ . In this section, we present two of the several techniques of parameter estimation, Principal Component and Principal Axis methods. The principal axis method (which is a modification of the principal component method) will be employed in Chapter 4 to extract the factor structure.

## Principal component method

For the random sample  $y_1, y_2, \dots, y_n$ , we obtain the sample covariance matrix  $\mathbf{S}$ . Then, we find an estimator  $\hat{\mathbf{\Lambda}}$  that will approximate Equation (3.76) with  $\mathbf{S}$  in place of  $\mathbf{\Sigma}$ . That is,

$$\mathbf{S} \simeq \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\boldsymbol{\psi}}. \quad (3.83)$$

In the principal component method, we exclude  $\hat{\boldsymbol{\psi}}$  and factor  $\mathbf{S}$  into

$$\mathbf{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}',$$

using spectral decomposition. Let  $\mathbf{S}$  have eigenvalue-eigenvector pairs  $(a_i, \mathbf{e}_i)$  with  $a_1 \geq a_2 \geq \dots \geq a_m \geq 0$  such that  $\mathbf{e}_i' \mathbf{e}_i = 1$ . Then,

$$\begin{aligned} \mathbf{S} &= a_1 \mathbf{e}_1 \mathbf{e}_1' + a_2 \mathbf{e}_2 \mathbf{e}_2' + \dots + a_m \mathbf{e}_m \mathbf{e}_m' \\ &= (\sqrt{a_1} \mathbf{e}_1 \quad \sqrt{a_2} \mathbf{e}_2 \quad \dots \quad \sqrt{a_m} \mathbf{e}_m) \begin{pmatrix} \sqrt{a_1} \mathbf{e}_1' \\ \sqrt{a_2} \mathbf{e}_2' \\ \vdots \\ \sqrt{a_m} \mathbf{e}_m' \end{pmatrix}, \end{aligned} \quad (3.84)$$

where  $a_1, a_2, \dots, a_m$  are the first eigenvalues of  $\mathbf{S}$  and  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$  are the corresponding normalised eigenvectors. Thus,  $\hat{\mathbf{\Lambda}}$  has the form

$$\hat{\mathbf{\Lambda}} = \begin{pmatrix} \sqrt{a_1} e_{11} & \sqrt{a_2} e_{12} & \dots & \sqrt{a_m} e_{1m} \\ \sqrt{a_1} e_{21} & \sqrt{a_2} e_{22} & \dots & \sqrt{a_m} e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{a_1} e_{m1} & \sqrt{a_2} e_{m2} & \dots & \sqrt{a_m} e_{mm} \end{pmatrix}. \quad (3.85)$$

This indicates that, the estimate of the factor loading is given as

$$\hat{\lambda}_{il} = \sqrt{a_l} e_{il}.$$

The  $i$ th diagonal element of  $\hat{\Lambda}\hat{\Lambda}'$  is the sum of squares of the  $i$ th row of  $\hat{\Lambda}$ . This sum of squares is an estimator of the  $i$ th communality,  $\hat{h}_i^2$ . That is,

$$\begin{aligned}\hat{h}_i^2 &= \hat{\lambda}_i' \hat{\lambda}_i \\ &= \sum_{l=1}^m a_l e_{il}^2 \\ &= \sum_{l=1}^m \hat{\lambda}_{il}^2\end{aligned}$$

### Principal axis method

The principal axis (PA) is an improvement of the principal component (PC) method. In the PC method, the term  $\hat{\psi}$  was excluded from Equation (3.83). The PA method begins by finding an initial estimate of  $\hat{\psi}$  and decomposes  $\mathbf{S} - \hat{\psi}$  or  $\mathbf{R} - \hat{\psi}$  into  $\hat{\Lambda}\hat{\Lambda}'$ . That is,

$$\begin{aligned}\mathbf{S} - \hat{\psi} &\cong \hat{\Lambda}\hat{\Lambda}' \\ \mathbf{R} - \hat{\psi} &\cong \hat{\Lambda}\hat{\Lambda}'\end{aligned}$$

Thus, the PA method starts with eigenvalue-eigenvector pairs of  $\mathbf{S} - \hat{\psi}$  or  $\mathbf{R} - \hat{\psi}$ .

The diagonal elements of  $\mathbf{S} - \hat{\psi}$  are the communalities

$$\hat{h}_i^2 = s_{ii} - \hat{\psi}_i, \quad i = 1, 2, \dots, p$$

where  $s_{ii}$  is the  $i$ th diagonal element of  $\mathbf{S}^{-1}$ . Similarly, the diagonal elements of  $\mathbf{R} - \hat{\psi}$  are

$$\hat{h}_i^2 = 1 - \hat{\psi}_i, \quad i = 1, 2, \dots, p.$$

A reasonable estimator for a communality in  $\mathbf{R} - \hat{\psi}$  is

$$\hat{h}_i^2 = R_i^2 = 1 - \frac{1}{r^{ii}}, \quad i = 1, 2, \dots, p,$$

where  $r^{ii}$  is the  $i$ th diagonal element of  $\mathbf{R}^{-1}$ , and  $R_i^2$  is the squared multiple correlation coefficient between  $y_i$  and all other variables (Rencher, 2002; R. A. Johnson & Wichern, 2007).

## Relationship between Factor Analysis and Item Response Theory

Item response models and factor analysis techniques have widely been applied in analysing questionnaire survey data, which are mainly item responses. In what follows, we present the relationship between the parameters of factor analysis and item response models (Takane & de Leeuw, 1987) under various conditions such as item response format (dichotomous and polytomous) and dimensionality of the underlying factor/ability (unidimensional and multidimensional).

### Dichotomous items

When performing factor analysis, it is assumed that both the underlying latent factor  $\theta$  and the response variables  $Y_i$ ,  $i = 1, 2, \dots, p$ , are continuous. Suppose that  $\theta$  and  $Y_i$  possess a joint normal probability distribution, then their density function,  $f(y, \theta)$ , is defined by

$$f(y, \theta) = \frac{1}{2\pi\sigma_y\sigma_\theta\sqrt{1-\rho_{y_i,\theta}^2}} \exp \left[ -\frac{1}{2(1-\rho_{y_i,\theta}^2)} \left\{ \left( \frac{y-\mu_y}{\sigma_y} \right)^2 - 2\rho_{y_i,\theta} \left( \frac{y-\mu_y}{\sigma_y} \right) \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right) + \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right\} \right], \quad (3.86)$$

where  $\mu_y$  and  $\sigma_y$  are, respectively, the mean and standard deviation of  $Y_i$ ,  $\mu_\theta$  and  $\sigma_\theta$  are, respectively, the mean and standard deviation of  $\theta$ , and  $\rho_{y_i,\theta}^2$  measures the correlation between  $Y_i$  and  $\theta$ . The distribution of  $\theta$  is assumed to be normal and defined as

$$g(\theta) = \frac{1}{\sigma_\theta\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right] \quad (3.87)$$

The conditional distribution of  $y_i$  given  $\theta$ ,  $f(y|\theta)$ , is given by



$$\begin{aligned}
f(y|\theta) &= \frac{f(y, \theta)}{g(\theta)} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2(1-\rho_{y_i,\theta}^2)} \left\{ \left( \frac{y-\mu_y}{\sigma_y} \right)^2 - \right. \right. \\
&\quad \left. \left. 2\rho_{y_i,\theta} \left( \frac{y-\mu_y}{\sigma_y} \right) \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right) + \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right\} + \frac{1}{2} \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right] \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2(1-\rho_{y_i,\theta}^2)} \left\{ \left( \frac{y-\mu_y}{\sigma_y} \right)^2 - \right. \right. \\
&\quad \left. \left. 2\rho_{y_i,\theta} \left( \frac{y-\mu_y}{\sigma_y} \right) \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right) + \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 + (1-\rho_{y_i,\theta}^2) \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right\} \right] \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2(1-\rho_{y_i,\theta}^2)} \left\{ \left( \frac{y-\mu_y}{\sigma_y} \right)^2 - \right. \right. \\
&\quad \left. \left. 2\rho_{y_i,\theta} \left( \frac{y-\mu_y}{\sigma_y} \right) \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right) + \rho_{y_i,\theta}^2 \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right)^2 \right\} \right] \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2(1-\rho_{y_i,\theta}^2)} \left\{ \left( \frac{y-\mu_y}{\sigma_y} \right) - \right. \right. \\
&\quad \left. \left. \rho_{y_i,\theta} \left( \frac{\theta-\mu_\theta}{\sigma_\theta} \right) \right\}^2 \right] \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2\sigma_y^2(1-\rho_{y_i,\theta}^2)} \times \right. \\
&\quad \left. \left\{ y - \mu_y - \rho_{y_i,\theta} \frac{\sigma_y}{\sigma_\theta} (\theta - \mu_\theta) \right\}^2 \right] \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_y^2(1-\rho_{y_i,\theta}^2)}} \exp \left[ -\frac{1}{2\sigma_y^2(1-\rho_{y_i,\theta}^2)} \times \right. \\
&\quad \left. \left\{ y - \left( \mu_y + \rho_{y_i,\theta} \frac{\sigma_y}{\sigma_\theta} (\theta - \mu_\theta) \right) \right\}^2 \right].
\end{aligned} \tag{3.88}$$

Equation (3.88) is a density function of a normal random variable with mean

$$\mu_y + \rho_{y_i,\theta} \frac{\sigma_y}{\sigma_\theta} (\theta - \mu_\theta), \tag{3.89}$$

and variance

$$\sigma_y^2 (1 - \rho_{y_i, \theta}^2). \quad (3.90)$$

Therefore, the conditional distribution of  $Y_i$  given  $\theta$  is normally distributed. That is,

$$Y_i | \theta \sim N \left[ \mu_y + \rho_{y_i, \theta} \frac{\sigma_y}{\sigma_\theta} (\theta - \mu_\theta), \sigma_y^2 (1 - \rho_{y_i, \theta}^2) \right]. \quad (3.91)$$

An objective of factor analysis is to model the relationship between the underlying latent ability  $\theta$  and the observed response variable  $Y_i$ ,  $i = 1, 2, \dots, p$ .

A one-factor model is given by

$$y_i = \lambda_i \theta + \varepsilon_i, \quad i = 1, 2, \dots, p, \quad (3.92)$$

where  $\lambda_i$  is the loading of  $y_i$  on  $\theta$ . By the assumptions,

$$\theta \sim N(0, 1) \quad \text{and} \quad \varepsilon_i \sim N(0, \psi_i). \quad (3.93)$$

For variable  $i$ ,

$$Y_i \sim N(0, \lambda_i^2 + \psi_i). \quad (3.94)$$

Making use of Equation (3.81) and Equation (3.93), Equation (3.91) becomes

$$Y_i | \theta \sim N(\lambda_i \theta, 1 - \lambda_i^2). \quad (3.95)$$

In factor analysis the response variable,  $Y$  is assumed to be continuous and normally distributed. However, responses to close-ended items in questionnaires are categorical and, for that matter, result in categorical data. Many researchers have described the relationship between item responses to be non-linear and declared the standard factor models in Equations (3.92) and (3.67) as inappropriate (Bernstein & Teng, 1989; Ferrando & Lorenzo-Seva, 2013). In order to apply factor analysis model to item response data, it is assumed that the continuous response variable,  $Y$  is discretised to yield the categorical response variable,  $X$ .

This means that the continuous response variable,  $Y$  underlies each categorical response variable,  $X_i$ . Specifically, for binary items, each response score  $X$  (0 and 1) is considered to arise from an arbitrary dichotomisation of the continuous underlying response variable  $Y$ . Figure 18 illustrates (Mehta, Neale, & Flay, 2004; Ferrando & Lorenzo-Seva, 2013) the relationship between observed item response  $X$  and underlying response variable  $Y$ .

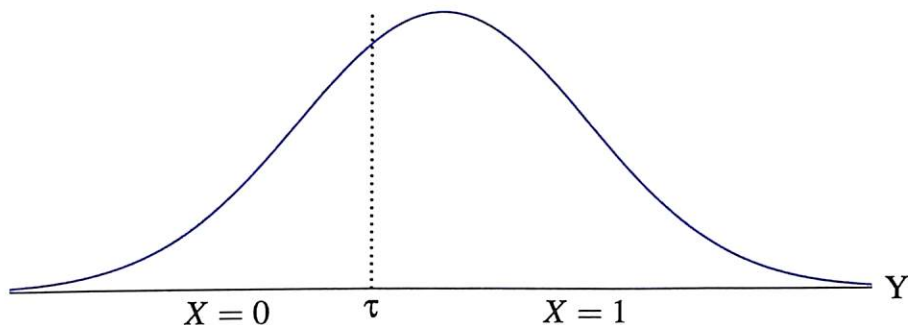


Figure 18: Distribution of dichotomous responses

Figure 18 indicates that the relationship between continuous variable,  $Y$  and the dichotomous response variable,  $X$  is defined by

$$X_i = \begin{cases} 1, & \text{if } Y \geq \tau \\ 0, & \text{if } Y < \tau \end{cases} \quad (3.96)$$

where  $\tau$  denotes threshold between the two response categories. The definition implies that, to obtain a positive response (i.e. Yes, represented by  $X_i = 1$ ), then

$$\begin{aligned} p(X_i = 1|\theta) &= p(Y \geq \tau) \\ &= p \left[ \frac{Y - \lambda_i \theta}{\sqrt{1 - \lambda_i^2}} \geq \frac{\tau - \lambda_i \theta}{\sqrt{1 - \lambda_i^2}} \right] \\ &= p \left[ Z \geq \frac{\tau - \lambda_i \theta}{\sqrt{1 - \lambda_i^2}} \right] \\ &= 1 - \Phi \left( \frac{\tau - \lambda_i \theta}{\sqrt{1 - \lambda_i^2}} \right) \end{aligned}$$

$$\begin{aligned}
p(X_i = 1|\theta) &= \Phi\left(\frac{\lambda_i\theta - \tau}{\sqrt{1 - \lambda_i^2}}\right) \\
&= \Phi\left[\frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}\left(\theta - \frac{\tau}{\lambda_i}\right)\right], \tag{3.97}
\end{aligned}$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution.

Now consider a random variable  $X$  which is logistically distributed with parameters  $\mu$  and  $\sigma(> 0)$ , with a density function defined by (Balakrishnan, 1991)

$$f(x; \mu, \sigma) = \frac{\pi}{\sigma\sqrt{3}} \cdot \frac{\exp\left\{-\frac{\pi}{\sigma\sqrt{3}}(x - \mu)\right\}}{\left[1 + \exp\left\{-\frac{\pi}{\sigma\sqrt{3}}(x - \mu)\right\}\right]^2}. \tag{3.98}$$

By letting  $Z = \frac{1}{\sigma}(x - \mu)$ , which has a standard Normal distribution with mean 0 and variance 1, Equation (3.98) becomes

$$f(z; 0, 1) = 1.702 \cdot \frac{\exp(-1.702z)}{[1 + \exp(-1.702z)]^2}. \tag{3.99}$$

This equation is the standard logistic distribution function. The cumulative distribution function of  $Z$ ,  $F(z; 0, 1)$ , is given by

$$F(z; 0, 1) = \int_{-\infty}^z f(t; 0, 1) dt. \tag{3.100}$$

That is,

$$F(z; 0, 1) = 1.702 \int_{-\infty}^z \frac{\exp(-1.702t)}{[1 + \exp(-1.702t)]^2} dt. \tag{3.101}$$

Let  $u = 1 + \exp(-1.702t)$ . Differentiating  $u$  with respect to  $t$  gives

$$du = -1.702 \exp(-1.702t) dt,$$

or

$$dt = \frac{-1}{1.702(u - 1)} du. \tag{3.102}$$

Substituting  $dt$  into Equation (3.101) yields

$$\begin{aligned}
 F(z; 0, 1) &= 1.702 \int_{\infty}^{1+\exp(-1.702z)} \frac{(u-1)}{u^2} \cdot \frac{-1}{1.702(u-1)} du \\
 &= \int_{\infty}^{1+\exp(-1.702z)} -\frac{1}{u^2} du \\
 &= \left[ \frac{1}{u} \right]_{\infty}^{1+\exp(-1.702z)} \\
 &= \frac{1}{1+\exp(-1.702z)}, \quad z \in \mathfrak{R},
 \end{aligned} \tag{3.103}$$

which is the cumulative distribution function of the logistic distribution.

The unidimensional 2PL model

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

has the form of the logistic cumulative distribution function in Equation (3.103) evaluated at  $\alpha_i(\theta - \delta_i)$ . Thus, for the  $j$ th individual

$$p(X_{ij} = 1 | \theta) = \Phi[\alpha_i(\theta - \delta_i)]. \tag{3.104}$$

Therefore, appropriately equating the probabilities in Equation (3.97) and Equation (3.104) yields

$$\alpha_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}, \quad |\lambda_i| < 1, \tag{3.105}$$

and

$$\delta_i = \frac{\tau}{\lambda_i}. \tag{3.106}$$

Equation (3.105) indicates that  $\alpha_i$  is directly a function of  $\lambda_i$ . This means that, an item that greatly discriminates between individuals at lower and higher ability levels will be highly influential in the formation of the corresponding factor. However, if the item has poor discriminatory power then, it will not contribute significantly to the formation of the factor. This relationship will be examined empirically in Chapter 4.

Equation (3.106) shows that an item's difficulty  $\delta_i$  is a function of its category threshold value ( $\tau$ ) and  $\lambda_i$ . In this case, there is no clear relationship between the difficulty parameter and that of the factor model.

Sometimes the responses to a set of items in a questionnaire is not characterised by only one ability, but a combination of several abilities of the respondent. To this end, the relationship between the multidimensional item response theory and factor analysis can be determined. Considering the  $m$ -factor model (see Equation (3.68)), each  $X_i$  can be written as

$$X_i = \boldsymbol{\lambda}'_i \boldsymbol{\theta} + \varepsilon_i, \quad (3.107)$$

where,  $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{im})'$ . The distribution of  $X_i$  in Equation (3.94) becomes

$$X_i \sim N \left( 0, \sum_{j=1}^m \lambda_{ij}^2 + \psi_i \right). \quad (3.108)$$

Also, the conditional distribution of  $X_i$  given  $\boldsymbol{\theta}$  (see Equation (3.95)) is given by

$$X_i | \boldsymbol{\theta} \sim N \left( \boldsymbol{\lambda}'_i \boldsymbol{\theta}, 1 - \sum_{j=1}^m \lambda_{ij}^2 \right). \quad (3.109)$$

Following similar algebraic steps in Equation (3.97), it is determined that

$$\begin{aligned} p(X_i = 1 | \boldsymbol{\theta}) &= \Phi \left( \frac{\boldsymbol{\lambda}'_i \boldsymbol{\theta} - \tau}{\sqrt{1 - \sum_{j=1}^m \lambda_{ij}^2}} \right) \\ &= \Phi \left( \frac{\boldsymbol{\lambda}_i}{\sqrt{1 - \sum_{j=1}^m \lambda_{ij}^2}} \boldsymbol{\theta} - \frac{\tau}{\sqrt{1 - \sum_{j=1}^m \lambda_{ij}^2}} \right). \end{aligned} \quad (3.110)$$

Using M2PL model and following Equation (3.104), it shows that

$$p(X_{ij} = 1 | \boldsymbol{\theta}) = \Phi \left( \boldsymbol{\alpha}'_i \boldsymbol{\theta} + d_i \right). \quad (3.111)$$

Comparing Equation (3.110) and Equation (3.111) gives

$$\alpha_i = \frac{\lambda_i}{\sqrt{1 - \sum_{j=1}^m \lambda_{ij}^2}}, \quad (3.112)$$

and

$$d_i = -\frac{\tau}{\sqrt{1 - \sum_{j=1}^m \lambda_{ij}^2}}. \quad (3.113)$$

### Polytomous items

In the case of polytomous data, the categorical response variable,  $X$  is a realisation of the continuous response variable,  $Y$  by means of a series of thresholds,  $\tau_h, h = 1, 2, \dots, g, g + 1, \dots, k$ . Schematically, the distribution of polytomous response categories may be represented as shown in Figure 19.

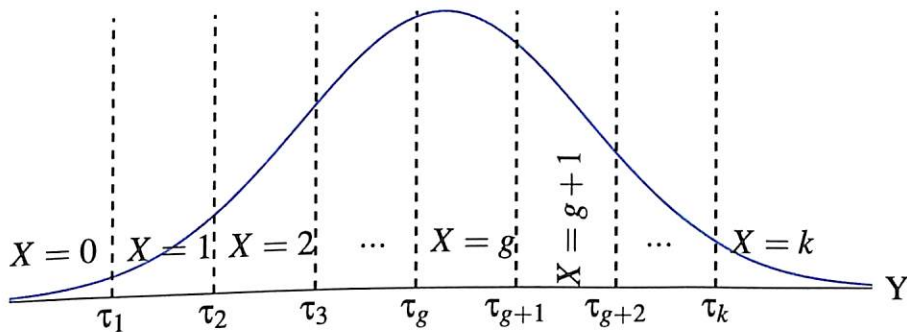


Figure 19: Distribution of polytomous response categories

Figure 19 implies that, the relationship between continuous variable,  $Y$  and the polytomous response variable,  $X$  is defined by

$$X_i = \begin{cases} g, & \text{if } \tau_g \leq Y_i < \tau_{g+1} \\ 0, & \text{otherwise} \end{cases} \quad (3.114)$$

To obtain a response in category  $g$  of the item, then

$$\begin{aligned}
 p(Y_i = g|\theta) &= p(\tau_g \leq Y_i < \tau_{g+1}) \\
 &= \Phi \left( \frac{\tau_{g+1} - \lambda_i \theta_j}{\sqrt{1 - \lambda_i^2}} \right) - \Phi \left( \frac{\tau_g - \lambda_i \theta_j}{\sqrt{1 - \lambda_i^2}} \right) \\
 &= \Phi \left[ \frac{-\lambda_i}{\sqrt{1 - \lambda_i^2}} \left( \theta_j - \frac{\tau_{g+1}}{\lambda_i} \right) \right] - \Phi \left[ \frac{-\lambda_i}{\sqrt{1 - \lambda_i^2}} \left( \theta_j - \frac{\tau_g}{\lambda_i} \right) \right] \\
 &= \Phi \left[ \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}} \left( \theta_j - \frac{\tau_g}{\lambda_i} \right) \right] - \Phi \left[ \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}} \left( \theta_j - \frac{\tau_{g+1}}{\lambda_i} \right) \right].
 \end{aligned} \tag{3.115}$$

In order to link factor analysis and item response theory models for polytomous data, the graded response (GR) model is considered. The form of the model makes it tractable. The GR model is stated as

$$p(X_i = g|\theta) = P(X_i \geq g|\theta) - P(X_i \geq g + 1|\theta).$$

Thus,

$$\begin{aligned}
 p(X_i = g | \theta) &= \frac{1}{1 + \exp[-\alpha_i(\theta - \delta_{ig})]} - \frac{1}{1 + \exp[-\alpha_i(\theta - \delta_{i,g+1})]} \\
 &= \Phi[\alpha_i(\theta - \delta_{ig})] - \Phi[\alpha_i(\theta - \delta_{i,g+1})].
 \end{aligned} \tag{3.116}$$

By comparing Equations (3.115) and (3.116), it can be observed that

$$\alpha_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}, \tag{3.117}$$

$$\delta_{ig} = \frac{\tau_{ig}}{\lambda_i}, \tag{3.118}$$

and

$$\delta_{i,g+1} = \frac{\tau_{i,g+1}}{\lambda_i}. \tag{3.119}$$

These establish the relationship among the parameters of factor analysis and IRT models for polytomous items (de Leeuw, 1983).



In many practical situations, responses to polytomous items are characterised by multidimensional person ability. Thus, an equivalent of Equation (3.115) for  $m$ -factor model is given by

$$p(X_i = g | \boldsymbol{\theta}) = \Phi \left( \frac{\lambda'_i}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}} \boldsymbol{\theta} - \frac{\tau_g}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}} \right) - \Phi \left( \frac{\lambda'_i}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}} \boldsymbol{\theta} - \frac{\tau_{g+1}}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}} \right). \quad (3.120)$$

This can be likened to a multidimensional IRT model, say MGR model, given by

$$p(X_i = g | \boldsymbol{\theta}_j) = \frac{1}{1 + \exp[-(\boldsymbol{\alpha}'_i \boldsymbol{\theta} + d_{ig})]} - \frac{1}{1 + \exp[-(\boldsymbol{\alpha}'_i \boldsymbol{\theta} + d_{i,g+1})]}.$$

That is,

$$p(X_i = g | \boldsymbol{\theta}) = \Phi[\boldsymbol{\alpha}'_i \boldsymbol{\theta} + d_{ig}] - \Phi[\boldsymbol{\alpha}'_i \boldsymbol{\theta} + d_{i,g+1}]. \quad (3.121)$$

Equations (3.120) and (3.121) show that

$$\boldsymbol{\alpha}_i = \frac{\lambda'_i}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}}, \quad (3.122)$$

$$d_{ig} = -\frac{\tau_g}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}}, \quad (3.123)$$

and

$$d_{i,g+1} = -\frac{\tau_{g+1}}{\sqrt{1 - \sum_{l=1}^m \lambda_{il}^2}}. \quad (3.124)$$

## Measures of Correlation Coefficients

A measure of correlation provides a platform for expressing the degree of linear relationship between two variables. Expression of the degree of relationship requires determining the coefficient of correlation. Several techniques have been devised for calculating correlation coefficient, notably, the Pearson correlation ( $r$ ) and Spearman's rank correlation coefficients. The Pearson  $r$  is often used to calculate correlation coefficient when the two variables involved are measured on at least the interval scale and are jointly normally distributed. Suppose that for the joint distribution of two random variables, say,  $X$  and  $Y$ , a random sample of  $n$  paired data  $(x_1, y_1), (x_2, y_2), \dots$ , and  $(x_n, y_n)$  is drawn, the Pearson  $r$  is given by

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] \left[ n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \right]}}. \quad (3.125)$$

Theoretically,  $r$  can assume a maximum value of  $+1$  and a minimum value of  $-1$ . There is a perfect degree of relationship between  $X$  and  $Y$  when  $r$  assumes a value of  $1$  in absolute terms. However, it is practically impossible for  $r$  to attain the value of  $+1$  or  $-1$  unless the marginal distributions of  $X$  and  $Y$  are identical in shape (Glass & Hopkins, 1996).

In this study, two methods of calculating correlation coefficient – tetrachoric and polychoric correlations – have been applied due to the nature of items involved.

### Tetrachoric correlation

The Pearson  $r$  serves as the basis for most other methods of calculating correlation coefficients. For instance, when  $X$  and  $Y$  are both dichotomies, one of the classes is scored zero and the other as one, as illustrated in Table 2.

Table 2: Two Dichotomous Variables

		Variable X		
		0	1	Total
Variable Y	0	$a$	$b$	$a + b$
	1	$c$	$d$	$c + d$
Total		$a + c$	$b + d$	$n$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the observed frequencies. The quantities in Equation (3.125) have the following equivalents in the table.

$$\sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 = b + d,$$

$$\sum_{i=1}^n y_i = \sum_{i=1}^n y_i^2 = c + d,$$

$$\sum_{i=1}^n x_i y_i = d,$$

and

$$n = a + b + c + d.$$

Substituting these equivalents into Equation (3.125) gives the special case known as the Phi coefficient ( $\phi$ ). That is,

$$\begin{aligned} \phi &= \frac{(a + b + c + d)d - (b + d)(c + d)}{\sqrt{[(a + b + c + d)(b + d) - (b + d)^2][(a + b + c + d)(c + d) - (c + d)^2]}} \\ &= \frac{ad + bd + cd + d^2 - bc - bd - cd - d^2}{\sqrt{[(a + b + c + d - b - d)(b + d)][(a + b + c + d - c - d)(c + d)]}} \\ &= \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}. \end{aligned} \tag{3.126}$$

By letting  $p_x$  be the proportion of observing score 1 on X;  $p_y$  the proportion of observing score 1 on Y; and  $p_{xy}$  the proportion of observing score 1 on both X and Y, the phi coefficient becomes

$$\phi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}}, \tag{3.127}$$

where  $q_x = 1 - p_x$  denotes the proportion of observing score 0 on  $X$  and  $q_y = 1 - p_y$  the proportion of observing score 0 on  $Y$ . The phi coefficient has the same interpretation as Pearson  $r$ , and in a like manner, attain a maximum theoretical absolute value of one only if  $p_x = p_y$ . If  $p_x$  differs substantially from  $p_y$ , the maximum value of  $\phi$  can be much lower than one (Glass & Hopkins, 1996).

If the dichotomous nature of  $X$  and  $Y$  are practically compelled, then it becomes inappropriate to use phi coefficient as a measure of correlation. In this case, a measure of correlation between the artificially dichotomised variables is the tetrachoric correlation coefficient ( $r_{tet}$ ). Tetrachoric correlation coefficient provides an estimate of Pearson  $r$  when both  $X$  and  $Y$  are artificial dichotomies with underlying normal distributions. It can be calculated by (Chen & Popovich, 2002; Glass & Hopkins, 1996)

$$r_{tet} = \frac{ad - bc}{\tau_x \tau_y n^2}, \quad (3.128)$$

where  $\tau_x$  and  $\tau_y$  are the ordinates of the standard normal distributions at  $p_x$  and  $p_y$ , respectively. The value  $r_{tet}$  can be interpreted to mean an estimate of what the observed correlation would be if both variables were measured continuously. The sample size  $n$  must be above 400 (Glass & Hopkins, 1996) for the estimate of  $r$  provided by  $r_{tet}$  to be substantially accurate.

### **Polychoric correlation**

An ordinal variable can be thought of as a crude representation of an unobserved continuous variable. The polychoric correlation coefficient is an estimate of a measure of association between ordinal variables which rests upon an assumption of an underlying continuous bivariate normal distribution (Basto & Pereira, 2012). As noted in Equation (3.114), we suppose that the ordinal response variable,  $X$  (with category scores  $0, 1, 2, \dots, g, \dots, k$ ) is a representation of an underlying continuous response variable,  $Y$  by means of series of thresh-



Taking partial derivative of  $l$  with respect to  $\rho$ , and setting the result to zero gives

$$\frac{n\rho}{1-\rho^2} - \frac{\rho}{(1-\rho^2)^2} \left( \sum_{i=1}^{k^1} X_{1i}^2 - 2\rho \sum_{i=1}^{k^1} \sum_{j=1}^{k^2} X_{1i}X_{2j} + \sum_{j=1}^{k^2} X_{2j}^2 \right) + \frac{1}{1-\rho^2} \sum_{i=1}^{k^1} \sum_{j=1}^{k^2} X_{1i}X_{2j} = 0.$$

Further simplification yields

$$\rho(1-\rho^2) + \frac{1}{n}(1+\rho^2) \sum_{i=1}^{k^1} \sum_{j=1}^{k^2} X_{1i}X_{2j} - \frac{\rho}{n} \left( \sum_{i=1}^{k^1} X_{1i}^2 + \sum_{j=1}^{k^2} X_{2j}^2 \right) = 0. \quad (3.130)$$

Equation (3.130) has three roots for  $\rho$ , the value of  $\rho$  which maximizes the likelihood function,  $L$  (see Equation (3.129)) is taken as the polychoric correlation between  $X_1$  and  $X_2$  (Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010; Kendall & Stuart, 1961).

## Chapter Summary

The chapter discussed key concepts and methods used in IRT. It presented various IRT models and their graphical representations. The graphical properties were explored using the brooding scale dataset. By the graphs, four groups of brooding items were identified: (1) items that show low difficulty level but have high discrimination; (2) items indicate that both the difficulty and discrimination are quite high; (3) items indicate that the difficulty parameter is quite high with a low discrimination; and (4) items that possess quite high difficulty and discrimination values, but the difficulty level a bit lower than that of the second group. The chapter also presented the procedures used in assessing the fitness of IRT models. It was found that this task is multi-faceted and must be carried out at the overall model level as well as the item level. The assessment of the fitness of IRT models can be accomplished using different approaches. In this study, we employed the widely used standard Chi-square test, which relies on all the observed and expected response patterns of items being modelled. In this

chapter, we also discussed the concept of multidimensional IRT models, which can be used to model the relationship between two or more person abilities and the probability of responses to items.

The methods of factor analysis that are employed in this study have been discussed in the chapter. The factor model and its underlying assumptions were reviewed. Further, two methods – the Principal Component and Principal Axis – for estimating the parameters of the factor model have been discussed with the intention of implementing them in the study to extract the factor structure.

In this chapter, it has been shown that the parameters of IRT and factor analysis models are closely linked. It is noted that, item discrimination parameter is directly a function of factor loading, such that items that greatly discriminate between individuals at lower and higher ability levels are highly influential in the formation of factors.

Two measures of correlation – tetrachoric and polychoric coefficients – were presented. It is found that tetrachoric and polychoric correlation coefficients are maximum likelihood estimates of the correlation between two ordinal variables having bivariate normal distribution.

## CHAPTER FOUR

### RESULTS AND DISCUSSION

#### Introduction

The purpose of this study is to examine the effects of measurement scales on results of item response theory models and multivariate techniques. The study is based on simulated datasets under various conditions such as item response format, number of ability dimensions underlying response scales, and sample size using R package *mirt* command: *simdata (a, d, N, itemtype)*. Two main statistical techniques – Item Response Theory (IRT) models and Factor Analysis – are employed in analysing the simulated datasets using standard R 3.4.3 codes. We will first present a description of the generation of the datasets. Next, we present the analyses of the datasets.

#### Data Simulation

In order to examine the effect of response scale on results of factor analysis, data is simulated using *mirt* package in R software. Several datasets are generated under different conditions for a total of twenty items. Datasets are generated for varied response scales, namely dichotomous, three-point, five-point, and seven-point scales. For each response scale, different sample sizes are considered. These sample sizes include 30, 100, 150, 200, 500, 800, and 1000. The rationale is to investigate the effect of sample size on factor analysis results. In addition, for each response scale, different dimensions of underlying person-ability are considered, particularly unidimensional, two-dimensional and three-dimensional. The datasets are generated using the command: *simdata (a, d, N, itemtype)* (Chalmers, 2012), where argument *a* denotes a vector/matrix of discrimination parameter values, *d* vector/matrix of difficulty parameter values,



$N$  sample size and “itemtype” the underlying IRT model. These arguments are specified to generate the desired dataset.

Firstly, unidimensional dichotomous response dataset is simulated using 2PL model. The 2PL model is considered, because it assumes that items have different discrimination powers. In questionnaires, items differ in terms of content, and so are their discriminations. In this system, a  $20 \times 1$  vector of discrimination values are specified for  $a$ , and another  $20 \times 1$  vector of difficulty values for  $d$  for all sample sizes. Table 4 shows the discrimination and difficulty parameter values used in simulating unidimensional item response data for twenty items.

In the table, we have  $0.4 < \alpha < 3.0$  and  $-2.5 < \delta < 2.5$ . High values of  $\alpha$  means that the item is discriminating largely between low-ability and high-ability persons. High positive value of  $\delta$  means the item is “difficult” and only high-ability persons can respond to it in higher response categories. Conversely, an item with high negative  $\delta$  value is considered to be “easy” and persons with high ability levels tend to respond favourably to it. A  $\delta$  value of zero or close to zero means that the item is averagely difficult and persons of average ability could respond to it in higher response categories.

Table 4: Discrimination and Difficulty Levels of Each Item

Item	Discrimination ( $\alpha$ )	Difficulty ( $\delta$ )
1	0.5	0.00
2	0.7	0.12
3	0.8	-2.30
4	0.6	0.10
5	0.4	2.00
6	2.2	-2.50
7	1.5	-2.00
8	2.7	-1.50
9	1.8	-2.20
10	1.6	2.50
11	2.0	2.30
12	2.9	1.50
13	3.0	2.20
14	2.1	0.30
15	2.8	0.50
16	1.4	0.25
17	1.9	0.40
18	1.2	0.42
19	1.3	0.56
20	2.9	0.20

To generate a two-dimensional dichotomous dataset,  $a$  argument is modified to a  $20 \times 2$  matrix of discrimination values. Here, same vector of discrimination values on first dimension is repeated on the second dimension. Thus, the two dimensions have the same discrimination values. The intent is to deter-

mine how the information will manifest in factor model. In this case, one of two possibilities is expected in the factor model. On one hand, a factor solution is expected to be dominated by two repeating factors since the same information is contained on the two dimensions that underlie the data. On the other hand, a single dominant factor is expected with the other influenced by few items. To simulate a three-dimensional dichotomous dataset,  $20 \times 3$  matrix of discrimination values is specified for  $a$ . The three dimensions have same discrimination values, with similar rationale as for the two-dimensional case.

Next, a unidimensional three-point scale data is generated using a polytomous IRT model, specifically GPC model. Polytomous models incorporates category boundaries to cater for the multi-category nature of items. For instance, a three-point scale requires two category boundaries, five-point scale requires four category boundaries, etc. For a given scale, number of response categories is specified using argument  $d$ . For three-point scale,  $d$  consists of a  $20 \times 2$  matrix of difficulty values with  $20 \times 1$  discrimination values depicting unidimensional fashion. The GPC model considers items of varying discriminations just as 2PL model.

Higher response scale datasets, five and seven-point, are also simulated in the same manner as three-point scale.

## Data Analysis

Two statistical analytical techniques – item response theory (IRT) models and factor analysis – are employed in analysing simulated datasets using standard R 3.4.3 codes (R Core Team, 2017). The IRT analysis is conducted using R package *mirt* (Chalmers, 2012). The analyses of unidimensional dichotomous item response datasets are based on two-parameter logistic (2PL) IRT model, whereas multidimensional two-parameter logistic (M2PL) IRT model is employed in the analyses of multidimensional dichotomous datasets. For unidi-

mensional polytomous item response datasets, we employ the generalised partial credit (GPC) model in the analyses. In terms of multidimensional polytomous datasets, the multidimensional generalised partial credit (MGPC) model is used to conduct the analyses.

Factor analysis is also performed on each simulated item response dataset using R package *psych* (Revelle, 2017). Factor analyses of dichotomous item response datasets are based on tetrachoric correlation matrices. On the other hand, polychoric correlation matrix is used as input in factor analysis of polytomous item response datasets. The R codes used in all the analyses are provided in Appendix.

### **Assessment of Dichotomous Response Scale**

This section presents the results of IRT and factor analysis on dichotomous response scale. In regard of this scale, three different dimensions of the underlying ability are considered. These are unidimensional, two-dimensional and three-dimensional. For unidimensional dichotomous scale, three types of factor analyses are carried out. That is, one-factor, two-factor and three-factor solutions. For all three dimensions considered, factor solutions are obtained at various sample sizes. We begin with an exploration of characteristics of items in the unidimensional dichotomous item response dataset.

#### **Unidimensional item response datasets**

Table 5 shows the estimates of discrimination and difficulty parameters of unidimensional 2PL model.

Table 5: Discrimination and Difficulty Estimates of Unidimensional 2PL Model  
for Various Sample Sizes

Item	Sample Size							
	30		100		150		200	
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$
1	0.684	-0.153	0.567	0.719	0.744	0.331	0.490	0.148
2	1.287	0.147	1.539	-0.001	0.845	0.184	0.828	0.045
3	0.679	-1.514	0.267	-3.273	0.450	-1.944	0.344	-1.857
4	0.321	0.274	1.008	-0.005	0.696	0.059	0.654	0.131
5	-0.377	1.655	0.076	1.588	-0.044	2.537	0.492	2.237
6	0.501	-1.458	1.718	-1.715	1.737	-2.281	2.146	-2.204
7	0.432	-2.269	0.367	-1.491	2.250	-2.283	0.927	-1.723
8	1.866	-0.880	2.266	-0.743	3.201	-1.353	2.346	-0.927
9	1.298	-1.325	1.852	-2.412	2.175	-2.173	2.296	-2.504
10	0.967	2.187	9.008	12.668	1.522	2.352	1.118	1.942
11	4.018	3.418	2.143	2.214	2.092	2.015	2.084	2.271
12	3.056	2.370	2.537	1.760	3.153	1.735	3.964	1.750
13	8.339	6.370	5.092	3.717	3.521	2.889	2.694	1.687
14	1.469	0.522	2.730	0.522	2.221	0.157	1.963	0.506
15	4.389	1.254	3.263	1.062	3.648	0.374	3.289	0.411
16	0.863	0.622	0.732	-0.002	1.339	0.100	1.524	0.420
17	1.697	0.346	1.258	0.467	1.807	0.437	2.523	0.659
18	1.316	0.147	0.778	0.363	1.287	-0.041	1.091	0.346
19	1.874	-0.056	1.606	0.759	1.353	0.536	0.921	0.525
20	4.551	-0.557	3.292	0.402	2.440	0.406	4.249	0.041

Table 5, continued

Item	Sample Size					
	500		800		1000	
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$
1	0.460	0.034	0.504	0.090	0.453	-0.029
2	0.710	0.187	0.081	-0.001	0.784	0.100
3	0.814	-2.227	0.895	-2.210	0.700	-2.396
4	0.690	-0.160	0.638	0.132	0.656	0.159
5	0.558	1.976	0.274	1.985	0.443	2.056
6	2.276	-2.385	2.196	-2.353	2.654	-2.520
7	1.641	-2.120	1.723	-1.864	1.886	-2.248
8	2.679	-1.435	2.835	-1.370	2.720	-1.362
9	2.275	-2.554	2.486	-2.866	1.739	-2.000
10	1.786	2.529	1.599	2.551	1.663	2.468
11	1.594	2.158	2.050	2.437	2.271	2.595
12	3.385	1.874	2.898	1.678	2.971	1.805
13	3.268	2.344	3.496	2.880	3.108	2.420
14	2.069	0.458	2.272	0.410	2.120	0.378
15	3.723	0.340	2.438	0.649	2.865	0.771
16	1.331	0.365	1.480	0.430	1.456	0.352
17	1.905	0.347	1.930	0.453	1.765	0.410
18	1.069	0.297	1.053	0.421	1.371	0.724
19	1.333	0.720	1.165	0.641	1.127	0.720
20	2.889	0.061	2.920	0.198	2.900	0.377

The results show that the estimated values of discrimination ( $\hat{\alpha}$ ) and difficulty ( $\hat{\delta}$ ) parameters generally fluctuates with increasing sample size. That is, the

estimates of discrimination and difficulty parameters change depending on the sample. There is a marked difference between the specified and estimated item parameter values at lower samples ( $n = 30$  and  $100$ ). However, the differences tend to reduce at sample sizes of  $150$  and beyond. In addition, the differences become negligible at larger samples ( $n = 500, 800, 1000$ ). For example, from Table 4 the specified discrimination value ( $\alpha$ ) of Item 10 is  $1.6$ , which is quite close to the estimated values ( $\hat{\alpha}$ ) at samples of sizes  $150$  through  $1000$ .

Table 6 shows the significances of the fit of items for unidimensional 2PL model. The Table also shows the fitness of 2PL model for unidimensional dichotomous response dataset for various sample sizes.

The hypothesis based on which the Table 6 is generated is that

$H_0$  : Items fit the model; against

$H_1$  : Items do not fit the model.

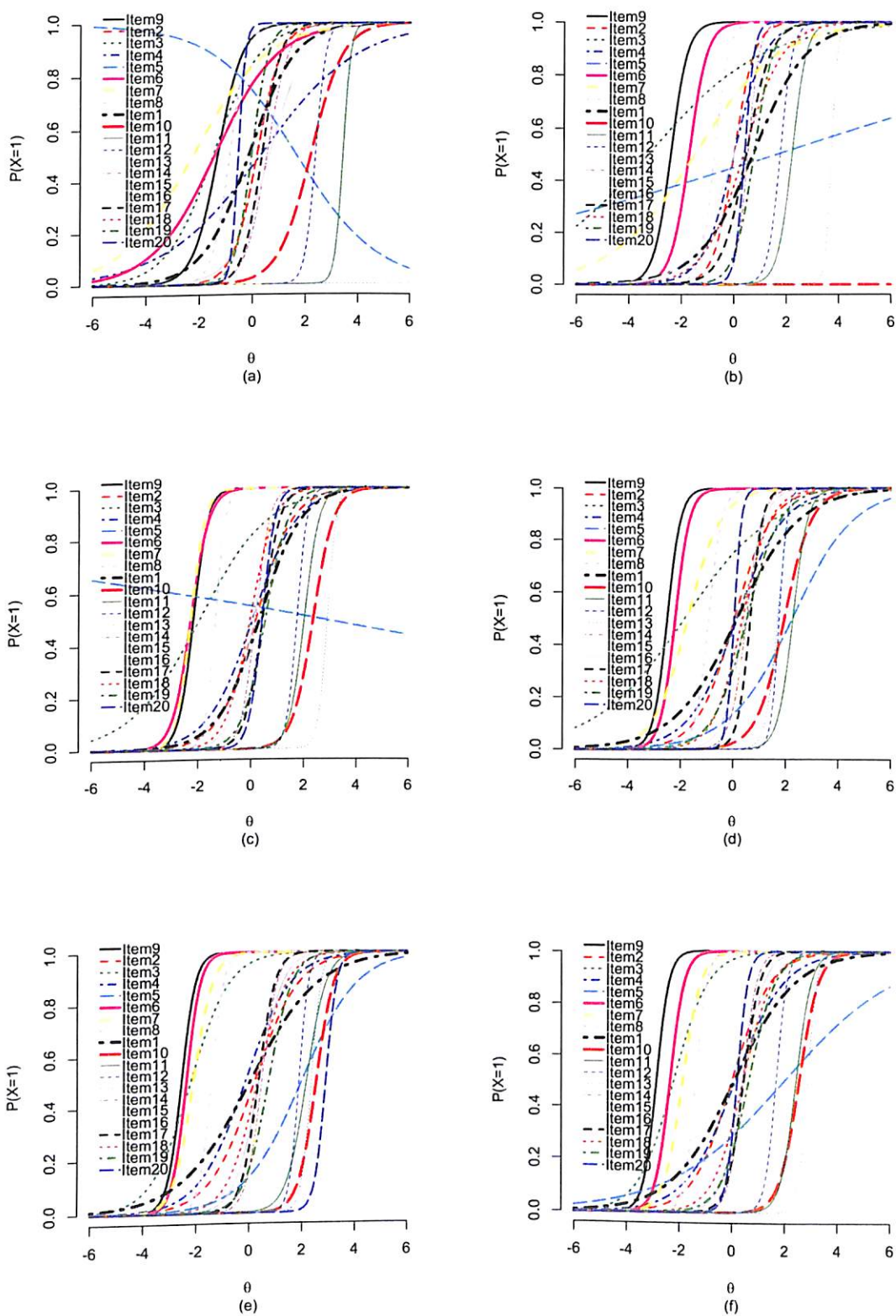
Items fit the unidimensional 2PL model since the  $p$ -values are generally much higher than  $0.05$ . Only at  $n = 150, 200$  and  $1000$  it is detected that three items ( $6, 12,$  and  $7,$  respectively) do not fit the model. The 2PL model significantly fits the unidimensional dichotomous item response data for all sample sizes.

Table 6: *P*-values for Item Fitness for Unidimensional 2PL Model for Various Sample Sizes

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.377	0.251	0.666	0.446	0.202	0.507	0.178
2	0.360	0.694	0.459	0.314	0.988	0.564	0.368
3	0.344	0.155	0.113	0.274	0.141	0.530	0.768
4	0.792	0.553	0.260	0.055	0.890	0.665	0.609
5	0.530	0.851	0.979	0.668	0.681	0.559	0.901
6	0.325	0.304	0.001	0.465	0.905	0.849	0.818
7	NaN	0.317	0.608	0.377	0.476	0.165	0.019
8	0.767	0.931	0.595	0.939	0.483	0.177	0.997
9	0.592	0.535	0.382	0.648	0.708	0.467	0.353
10	NaN	NaN	0.433	0.296	0.269	0.893	0.188
11	NA	0.877	0.682	0.486	0.490	0.318	0.079
12	NaN	0.234	0.008	0.033	0.783	0.429	0.182
13	NA	0.188	0.856	0.530	0.605	0.275	0.758
14	0.310	0.294	0.973	0.437	0.890	0.657	0.414
15	0.457	0.750	0.812	0.210	0.317	0.906	0.639
16	0.131	0.136	0.729	0.240	0.282	0.363	0.457
17	0.303	0.387	0.256	0.961	0.856	0.560	0.559
18	0.228	0.253	0.071	0.989	0.677	0.558	0.428
19	0.550	0.433	0.675	0.217	0.947	0.768	0.424
20	0.669	0.562	0.177	0.907	0.314	0.225	0.421
Model Fit	0.114	0.966	0.514	0.381	0.363	0.387	0.938



Figure 20 illustrates graphical representations of items in unidimensional dichotomous response dataset for various sample sizes.



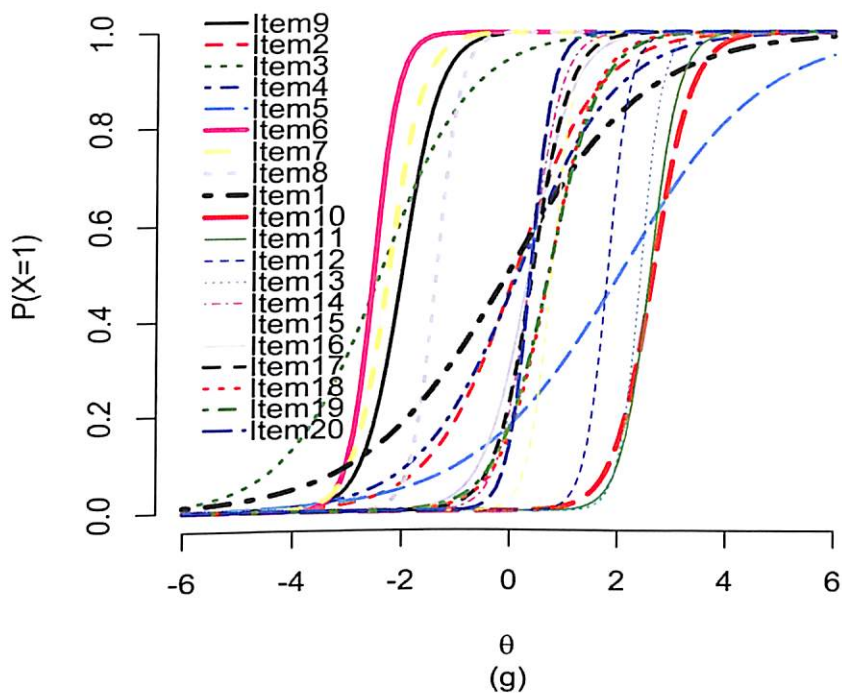


Figure 20: Item characteristic curves for unidimensional 2PL model for sample sizes: (a) 30; (b) 100; (c) 150; (d) 200; (e) 500; (f) 800; and (g) 1000

In Figure 20, the curve for Item 9 is extremely steep at a low ability level (at -2). For individuals with extremely low ability, Item 9 sharply discriminates. The only item with similar nature is Item 6. However, Item 6 discriminates at a little higher ability than Item 9. Items 2, 14, 16 through 20 are discriminating maximally at average ability levels. Meanwhile, Item 11 is discriminating among individuals at high ability levels for small samples but, tend to discriminate at average levels for large sample sizes. Item 13 discriminates the most at extremely high ability levels for  $n = 30$ , but conforms to discriminating at just high ability levels like other items for  $n \geq 150$ . For  $n = 30$ , the curves of all items have the desired shape (monotonic increasing), except for Item 5. It can be seen that this is as a result of negative estimate of discrimination parameter. In item response modelling, this implies that those with high ability rather have

a low probability of response. As sample size increases, the curve for Item 5 conforms to the expected form. Thus, the expected discrimination of the items is better achieved in larger sample sizes.

Table 7 displays the loadings of one-factor solutions for unidimensional item response datasets for various sample sizes on dichotomous scale.

Table 7: Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Dichotomous Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.400	0.346	0.399	0.283	0.269	0.307	0.277
2	0.597	0.405	0.464	0.457	0.398	0.454	0.434
3	0.249	0.507	0.211	0.183	0.402	0.424	0.359
4	0.149	0.544	0.405	0.386	0.380	0.373	0.372
5	-0.220	0.000	0.000	0.251	0.275	0.143	0.229
6	0.228	0.639	0.658	0.774	0.769	0.745	0.810
7	0.156	0.211	0.773	0.434	0.660	0.697	0.701
8	0.608	0.755	0.831	0.787	0.828	0.853	0.831
9	0.422	0.640	0.734	0.753	0.784	0.789	0.676
10	0.474	0.796	0.626	0.520	0.700	0.637	0.678
11	0.889	0.774	0.775	0.760	0.661	0.749	0.781
12	0.744	0.865	0.857	0.888	0.874	0.864	0.855
13	0.931	0.894	0.902	0.826	0.878	0.891	0.870
14	0.529	0.839	0.773	0.738	0.765	0.805	0.774
15	0.865	0.865	0.855	0.858	0.894	0.814	0.849
16	0.402	0.429	0.618	0.663	0.626	0.665	0.654
17	0.677	0.627	0.721	0.823	0.736	0.750	0.725
18	0.603	0.430	0.575	0.555	0.540	0.539	0.636
19	0.640	0.685	0.627	0.473	0.615	0.573	0.565
20	0.809	0.861	0.806	0.907	0.848	0.863	0.852
Prop Var	0.339	0.424	0.450	0.429	0.455	0.461	0.4458
Fit	0.706	0.870	0.8964	0.881	0.916	0.922	0.924

By comparing Table 5 and Table 7 we see that there is direct relationship between parameters of IRT and those of factor models, particularly the discrimination parameter and the factor loadings. We note that items with high discrimination values load highly on factors. The discrimination values of these items are greater than one. Thus, for an item to be influential in the formation of a factor, it should possess a discriminatory power with absolute value greater than one. The result also shows that specific indicators have no influence or do influence in a different direction, for example Variable 5. However, this observation is associated with low sample size ( $n = 30$ ). The negative loading changes to positive for higher sample sizes. The item interpretation changes accordingly for Variable 5. Also, we observe from Table 7 that the number of influential indicators appear to converge (at 15) for higher sample size starting at  $n = 150$ . The indicators are the same at point of convergence. The proportion of variance accounted for by the single factor increases from 33.9% (for  $n = 30$ ) to a highest of 46.1% (for  $n = 800$ ).

Table 8 contains loadings of two-factor solutions for unidimensional item response datasets for various sample sizes on dichotomous scale.

Table 8: Loadings of Two-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Dichotomous Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.371	0.196	0.232	0.268	0.389	0.000	0.277	0.104
2	0.136	0.858	0.000	0.582	0.443	0.142	0.326	0.324
3	0.519	-0.255	0.608	0.000	0.126	0.381	0.271	0.000
4	0.338	-0.198	0.196	0.677	0.454	-0.138	0.217	0.350
5	0.152	-0.576	0.188	-0.170	0.141	-0.660	0.000	0.453
6	0.000	0.322	0.392	0.552	0.634	0.174	0.530	0.577
7	-0.132	0.437	0.000	0.339	0.757	0.155	0.000	0.673
8	0.495	0.345	0.761	0.242	0.821	0.139	0.603	0.504
9	0.515	0.000	0.405	0.536	0.734	0.000	0.371	0.768
10	0.215	0.506	0.594	0.533	0.683	-0.135	0.625	0.000
11	0.651	0.605	0.560	0.545	0.763	0.139	0.715	0.322
12	0.406	0.705	0.647	0.579	0.829	0.214	0.758	0.470
13	0.722	0.576	0.811	0.404	0.927	0.000	0.729	0.407
14	0.565	0.136	0.909	0.195	0.721	0.307	0.708	0.295
15	0.565	0.687	0.803	0.332	0.825	0.220	0.703	0.490
16	0.488	0.000	0.218	0.432	0.522	0.496	0.537	0.388
17	0.715	0.188	0.400	0.517	0.656	0.364	0.757	0.370
18	0.618	0.185	0.420	0.152	0.525	0.280	0.301	0.520
19	0.855	0.000	0.560	0.392	0.629	0.000	0.505	0.124
20	0.614	0.522	0.799	0.365	0.774	0.226	0.650	0.641
Prop Var	0.256	0.195	0.300	0.183	0.427	0.072	0.286	0.194
Cum Var		0.451		0.483		0.499		0.480
Fit		0.801		0.896		0.917		0.904

Table 8, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.185	0.204	0.304	0.000	0.211	0.185
2	0.268	0.309	0.285	0.482	0.394	0.188
3	0.289	0.286	0.271	0.433	0.223	0.321
4	0.355	0.156	0.262	0.327	0.344	0.151
5	0.000	0.471	0.000	0.487	0.000	0.383
6	0.793	0.222	0.688	0.285	0.791	0.271
7	0.624	0.263	0.683	0.183	0.713	0.190
8	0.891	0.195	0.812	0.276	0.801	0.293
9	0.677	0.399	0.686	0.397	0.679	0.195
10	0.440	0.596	0.686	0.000	0.507	0.468
11	0.521	0.405	0.683	0.307	0.688	0.372
12	0.575	0.709	0.811	0.304	0.695	0.498
13	0.700	0.529	0.830	0.326	0.677	0.558
14	0.573	0.513	0.691	0.425	0.640	0.434
15	0.686	0.576	0.703	0.420	0.636	0.585
16	0.422	0.487	0.615	0.253	0.553	0.348
17	0.645	0.361	0.674	0.327	0.532	0.517
18	0.523	0.196	0.584	0.000	0.489	0.416
19	0.521	0.327	0.534	0.209	0.522	0.231
20	0.682	0.501	0.796	0.334	0.683	0.513
Prop Var	0.314	0.172	0.387	0.105	0.333	0.145
Cum Var		0.486		0.492		0.478
Fit		0.925		0.934		0.930

In Table 8, with the exception of sample sizes  $n = 150$  and  $n = 800$ , there is generally the incidence of repetition of high loadings on the same indicator variable of the two factors which can distract interpretation. However, for  $n = 150$ , the first factor loads highly on as many as 15 indicators and explains 42.7% of variation. The second factor loads highly on only one indicator (Variable 5) and is a contrast to its representation in IRT. In addition, amount of variance explained by the second factor appears to be negligible. These observations are consistent with the correlation matrix as Variable 5 has negative correlation with most of the other variables. The sample size of  $n = 150$  thus gives a more plausible factor solution than all other samples. The  $n = 150$  also explains the highest cumulative variation. For  $n = 800$  the second factor is rather considered as redundant.

Table 9 are loadings of three-factor solutions for unidimensional item response datasets for various sample sizes on dichotomous scale. From Table 9, the result becomes less meaningful and even unrealistic for sample sizes beyond 30. There is generally the incidence of repeating indicators on multiple factors. There is also the incidence of unrealistic loadings that are greater than one in higher factor numbers, particularly for Factor 3. Specifically, the loadings of Item 5 on Factor 3 is greater than one for  $n = 150$  and 200 (1.019 and 1.200, respectively). This incidence is as a result of an extraction of higher factor structure from a lesser dimensional dataset.



Table 9: Loadings of Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Dichotomous Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.785	0.000	-0.106	0.340	0.000	0.536
2	0.302	0.000	0.818	0.000	0.581	0.141
3	0.152	0.515	-0.232	0.635	0.000	0.000
4	0.127	0.312	-0.202	0.231	0.531	0.475
5	0.000	0.205	-0.544	0.125	0.000	-0.379
6	0.000	0.000	0.360	0.407	0.480	0.251
7	-0.119	0.000	0.509	0.000	0.270	0.199
8	0.495	0.265	0.245	0.773	0.233	0.000
9	0.000	0.724	0.155	0.437	0.429	0.343
10	0.610	0.176	0.334	0.519	0.662	0.000
11	0.742	0.289	0.453	0.558	0.509	0.202
12	0.377	0.280	0.672	0.621	0.594	0.114
13	0.593	0.494	0.499	0.730	0.607	-0.285
14	0.000	0.737	0.253	0.920	0.202	0.000
15	0.374	0.484	0.708	0.805	0.383	0.000
16	0.000	0.554	0.000	0.145	0.541	0.000
17	0.753	0.346	0.000	0.339	0.605	0.000
18	0.721	0.246	0.000	0.433	0.127	0.000
19	0.498	0.675	0.000	0.547	0.401	0.000
20	0.453	0.465	0.486	0.773	0.411	0.000
Prop Var	0.206	0.170	0.169	0.288	0.189	0.053
Cum Var			0.544			0.530
Fit			0.864			0.916

Table 9, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.435	0.000	0.000	0.206	0.189	-0.189
2	0.301	0.379	0.000	0.265	0.389	0.000
3	-0.120	0.617	-0.165	0.184	0.000	-0.168
4	0.398	0.000	0.212	0.215	0.341	0.000
5	0.000	-0.121	1.019	0.188	0.264	1.200
6	0.622	0.249	0.000	0.406	0.718	-0.125
7	0.572	0.513	0.158	0.000	0.709	0.124
8	0.809	0.285	0.000	0.564	0.546	0.000
9	0.741	0.204	0.000	0.385	0.717	0.000
10	0.401	0.473	0.429	0.519	0.194	-0.198
11	0.753	0.268	0.000	0.665	0.392	0.000
12	0.644	0.568	0.000	0.756	0.481	0.000
13	0.696	0.539	0.301	0.709	0.440	0.000
14	0.595	0.502	0.000	0.727	0.292	0.000
15	0.911	0.205	0.000	0.751	0.448	0.101
16	0.472	0.452	-0.172	0.499	0.431	0.000
17	0.421	0.689	0.000	0.682	0.465	-0.174
18	0.336	0.539	0.000	0.236	0.579	-0.107
19	0.486	0.385	0.114	0.636	0.000	0.130
20	0.790	0.279	0.000	0.633	0.652	0.000
Prop Var	0.328	0.170	0.074	0.266	0.214	0.084
Cum Var			0.572			0.563
Fit			0.937			0.918

## Two-dimensional item response datasets

We now consider an assessment of two-factor model on two-dimensional dichotomous datasets. In this system, datasets are generated by specifying the same vector of item discrimination parameter values on both dimensions of the underlying ability. Here, we expect that a good factor solution should have two repeating factors since the same information is contained on the two underlying dimensions of the dataset. Alternatively, we could expect a single dominant first factor in the two-factor solution with similar reasoning as in the former instance. We begin with an assessment of the fitness of items for datasets as shown in Table 10.

It can be observed from Table 10 that almost all items significantly fit the two-dimensional model. The only exception is when  $n = 30$  where the fitness of majority of items have not been possible to evaluate due to sparseness in the data. Correspondingly, the overall model fitness could not be determined due to low degrees of freedom. The item response model significantly fits the two-dimensional dichotomous response data for all other sample sizes.

Table 10: *P*-values for Item Fitness for Two-Dimensional 2PL Model for Various Sample Sizes

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.284	0.094	0.097	0.716	0.199	0.319	0.319
2	0.259	0.094	0.144	0.270	0.672	0.383	0.761
3	NaN	0.322	0.654	0.258	0.684	0.178	0.813
4	0.297	0.063	0.467	0.221	0.715	0.620	0.523
5	NaN	0.906	0.547	0.268	0.166	0.313	0.446
6	NA	0.744	0.832	0.205	0.694	0.443	0.329
7	NaN	0.199	0.209	0.338	0.224	0.509	0.752
8	NaN	0.112	0.093	0.466	0.148	0.159	0.841
9	NA	0.023	0.121	0.186	0.438	0.692	0.019
10	NaN	0.604	0.590	0.601	0.677	0.063	0.965
11	NaN	0.255	0.158	0.683	0.928	0.322	0.017
12	NA	0.651	0.079	0.197	0.254	0.690	0.232
13	NaN	0.940	0.460	0.354	0.344	0.237	0.340
14	NA	0.992	0.317	0.319	0.587	0.543	0.333
15	NaN	0.302	0.430	0.800	0.518	0.291	0.747
16	NaN	0.290	0.036	0.020	0.586	0.860	0.284
17	NaN	0.945	0.403	0.617	0.308	0.872	0.768
18	0.712	0.730	0.932	0.674	0.259	0.336	0.067
19	NA	0.182	0.239	0.609	0.924	0.153	0.402
20	NaN	0.026	0.347	0.686	0.486	0.488	0.877
Model Fit	-	0.920	0.406	0.249	0.953	0.974	0.123

Table 11 contains two-factor solutions for two-dimensional item response

datasets for various sample sizes on dichotomous scale.

Table 11: Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Dichotomous Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.162	0.470	0.567	0.170	0.255	0.225	0.478	-0.172
2	0.219	-0.313	0.555	0.136	0.375	0.428	0.440	0.219
3	0.383	0.660	0.441	0.000	0.245	0.395	0.351	0.654
4	0.326	0.315	0.571	0.000	-0.127	0.828	0.611	0.271
5	0.153	-0.556	0.000	0.939	0.826	0.000	0.000	0.543
6	0.540	0.175	0.847	0.000	0.540	0.684	0.852	0.234
7	0.271	0.659	0.848	0.000	0.483	0.545	0.694	0.297
8	0.693	0.287	0.902	-0.158	0.696	0.603	0.877	0.342
9	0.187	0.335	0.861	0.000	0.563	0.597	0.538	0.731
10	0.531	0.249	0.739	0.297	0.653	0.450	0.622	0.555
11	0.851	0.129	0.898	0.000	0.540	0.601	0.751	0.304
12	0.778	0.531	0.953	0.110	0.607	0.714	0.798	0.488
13	0.868	0.271	0.888	0.116	0.649	0.612	0.761	0.555
14	0.699	0.274	0.859	0.000	0.583	0.696	0.670	0.563
15	0.855	-0.136	0.927	0.000	0.545	0.664	0.811	0.447
16	0.831	-0.273	0.835	0.152	0.377	0.630	0.756	0.182
17	0.851	0.201	0.898	0.000	0.709	0.558	0.763	0.367
18	0.240	0.383	0.793	0.000	0.6377	0.468	0.547	0.543
19	0.787	0.379	0.773	0.000	0.676	0.248	0.676	0.417
20	0.814	0.342	0.930	0.191	0.705	0.641	0.755	0.552
Prop Var	0.378	0.144	0.619	0.058	0.322	0.317	0.447	0.204
Cum Var		0.522		0.677		0.639		0.651
Fit		0.873		0.973		0.965		0.971

Table 11, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.362	0.185	0.324	0.161	0.348	0.217
2	0.561	0.234	0.460	0.223	0.525	0.261
3	0.467	0.169	0.666	0.000	0.135	0.924
4	0.502	0.000	0.303	0.405	0.456	0.203
5	0.000	0.885	0.000	0.343	0.354	0.000
6	0.838	0.187	0.681	0.624	0.762	0.402
7	0.797	0.000	0.645	0.523	0.710	0.282
8	0.900	0.178	0.665	0.574	0.817	0.414
9	0.821	0.117	0.604	0.523	0.694	0.372
10	0.808	0.141	0.604	0.457	0.731	0.315
11	0.881	0.119	0.673	0.519	0.809	0.277
12	0.883	0.348	0.758	0.482	0.848	0.340
13	0.901	0.265	0.813	0.411	0.879	0.351
14	0.834	0.221	0.748	0.457	0.789	0.363
15	0.884	0.165	0.795	0.462	0.846	0.390
16	0.767	0.195	0.494	0.577	0.672	0.354
17	0.843	0.134	0.818	0.300	0.786	0.324
18	0.651	0.000	0.586	0.460	0.732	0.000
19	0.742	0.240	0.687	0.290	0.659	0.383
20	0.909	0.128	0.793	0.455	0.813	0.439
Prop Var	0.314	0.172	0.387	0.105	0.333	0.145
Cum Var		0.486		0.492		0.478
Fit		0.925		0.934		0.930

The  $n = 150$  generates a repeating factor consistent with two repeating dimensions underlying the dataset. The cumulative variation is also highest for this sample size.

### **Three-dimensional item response datasets**

In this system, we assess the performance of a three-factor model based on three-dimensional dichotomous item response dataset for various sample sizes. Table 12 shows the significance of the fitness of items for three-dimensional dichotomous datasets.

Table 12 indicates that items significantly fit the three-dimensional 2PL model. However, for smaller samples the fitness of items is appalling. The fitness of items to the model get better as sample size increases. We observe that the three-dimensional 2PL model significantly fits the data for all sample sizes, except at  $n = 30$  where the fitness of the model could not be determined due low degrees of freedom.

Table 12: *P*-values for Item Fitness for Three-Dimensional 2PL Model for Various Sample Sizes

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.316	0.470	0.061	0.089	0.959	0.3279
2	NaN	0.243	0.251	0.074	0.773	0.140	0.024
3	NA	0.745	0.152	0.054	0.518	0.244	0.173
4	0.280	0.540	0.917	0.036	0.662	158	670
5	NA	0.180	0.146	0.481	0.101	0.055	0.078
6	NaN	0.105	0.061	0.138	0.590	0.698	0.495
7	NaN	0.440	0.488	0.121	0.127	0.447	0.004
8	NA	0.580	0.124	0.059	0.184	0.633	0.300
9	NA	0.117	0.023	0.022	0.495	0.500	0.716
10	NaN	NaN	0.163	0.907	0.569	0.872	0.848
11	NaN	NaN	0.215	0.615	0.907	0.368	0.455
12	NA	0.571	NaN	0.438	0.045	0.244	0.071
13	NA	NaN	0.025	0.469	0.233	0.091	0.473
14	NA	0.082	0.558	0.441	0.096	0.723	0.255
15	NA	0.037	0.266	0.395	0.419	0.331	0.003
16	NaN	0.120	0.444	0.834	0.469	0.109	0.698
17	NA	0.747	0.551	0.220	0.271	0.686	0.566
18	NaN	0.091	0.465	0.204	0.182	0.662	0.679
19	NaN	0.859	0.518	0.720	0.946	0.353	0.861
20	NA	0.039	0.104	0.176	0.381	0.031	0.061
Model Fit	-	0.882	0.462	0.885	0.556	0.783	0.371



Table 13 are loadings of three-Factor solutions for three-dimensional item response datasets for various sample sizes on dichotomous scale. Lower sample sizes show two dominant factors in the result. This pattern is inconsistent with the expected dimension of the scale. However, for  $n = 150$  and  $200$ , as expected only the first factor is highly influenced by the indicator variables. The factor solution for  $n = 150$  is more conceivable as it accounts for as high as 78.9% cumulative variation. The fitness of the model is almost perfect for all sample sizes.

Table 13: Loadings of Three-Factor Solutions for Three-Dimensional Datasets  
for Various Sample Sizes on Dichotomous Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.000	0.853	0.239	0.618	0.000	0.000
2	0.495	0.363	0.000	0.482	0.389	0.257
3	0.000	0.000	-0.710	0.000	0.707	0.000
4	0.000	0.172	0.728	0.493	0.540	0.000
5	0.408	0.217	-0.253	0.000	0.000	0.919
6	0.353	0.700	0.293	0.593	0.566	0.418
7	0.375	0.578	-0.401	0.363	0.692	0.514
8	0.620	0.649	-0.200	0.502	0.611	0.476
9	0.737	0.522	-0.177	0.641	0.560	0.333
10	0.858	0.000	0.000	0.658	0.596	0.142
11	0.776	0.231	0.246	0.555	0.712	0.133
12	0.757	0.491	0.171	0.591	0.598	0.357
13	0.859	0.340	0.000	0.826	0.476	0.214
14	0.759	0.600	0.149	0.747	0.419	0.315
15	0.861	0.456	0.125	0.600	0.699	0.206
16	0.933	0.135	0.000	0.781	0.326	0.220
17	0.677	0.665	0.169	0.689	0.437	0.387
18	0.381	0.661	0.309	0.386	0.580	0.366
19	0.687	0.587	-0.135	0.429	0.689	0.000
20	0.705	0.671	0.000	0.560	0.585	0.516
Prop Var	0.401	0.257	0.088	0.321	0.300	0.132
Cum Var			0.746			0.753
Fit			0.973			0.986

Table 13, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.200	0.945	0.000	0.485	0.183	0.106
2	0.644	0.204	0.191	0.507	0.552	0.000
3	0.514	0.165	0.117	0.262	0.695	0.112
4	0.425	0.551	0.000	0.239	0.160	0.723
5	0.267	0.100	0.969	0.114	0.613	0.183
6	0.849	0.436	0.000	0.779	0.276	0.405
7	0.791	0.183	0.216	0.679	0.321	0.280
8	0.891	0.304	0.000	0.815	0.421	0.231
9	0.761	0.393	0.111	0.636	0.590	0.182
10	0.764	0.191	0.285	0.841	0.153	0.000
11	0.776	0.426	0.292	0.788	0.383	0.253
12	0.868	0.351	0.272	0.729	0.464	0.246
13	0.849	0.253	0.358	0.765	0.346	0.300
14	0.814	0.294	0.172	0.601	0.469	0.488
15	0.843	0.335	0.284	0.736	0.415	0.342
16	0.783	0.356	0.285	0.630	0.277	0.423
17	0.813	0.384	0.223	0.748	0.438	0.213
18	0.823	0.000	0.174	0.743	0.125	0.284
19	0.806	0.283	0.327	0.661	0.338	0.338
20	0.841	0.407	0.234	0.760	0.435	0.279
Prop Var	0.552	0.144	0.094	0.432	0.171	0.099
Cum Var			0.789			0.702
Fit			0.990			0.983

## Investigation of Three-Point Likert Scale

Three-point Likert scale consists of three response categories. This scale can also take on different dimensions. Similar to dichotomous scale, three different dimensions of the underlying ability are considered. These are one, two and three dimensions. From now on, we dwell on the performance of various factor solutions on three-point response scale at different sample sizes. As a start, we consider unidimensional three-point Likert scale.

### Unidimensional three-point Likert scale

Table 14 shows the  $p$ -values of fitness of items for unidimensional three-point scale datasets based on generalised partial credit (GPC) model.

We observe from Table 14 that in migrating from dichotomous to three-point scale, the significance of fitness of items for the model generally fluctuates for all sample sizes. While the  $p$ -values of some items improve for given sample size, those of other items deteriorate. For instance, at  $n = 30$  the fitness of Item 1 for three-point scale improves over that of dichotomous scale. On the contrary, at  $n = 30$  the fitness of Item 2 for three-point scale decreases as compared to dichotomous scale. The overall fitness of GPC model for unidimensional three-point response data is significant at sample sizes of 100 and over.

Table 14: *P*-values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Three-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.662	0.533	0.343	0.685	0.598	0.629	0.823
2	0.176	0.632	0.706	0.013	0.818	0.460	0.690
3	0.566	0.467	0.872	0.773	0.741	0.277	0.055
4	0.731	0.925	0.199	0.286	0.409	0.454	0.288
5	0.218	0.365	0.308	0.604	0.246	0.866	0.359
6	0.190	0.502	0.773	0.598	0.153	0.453	0.558
7	0.065	0.391	0.791	0.320	0.536	0.887	0.514
8	0.267	0.455	0.927	0.426	0.340	0.770	0.826
9	0.459	0.419	0.930	0.715	0.894	0.228	0.006
10	0.332	0.619	0.934	0.711	0.663	0.505	0.851
11	NaN	0.278	0.743	0.742	0.272	0.391	0.245
12	NaN	0.152	0.018	0.101	0.258	0.799	0.105
13	0.017	0.495	0.462	0.295	0.263	0.016	0.923
14	0.114	0.345	0.364	0.078	0.568	0.038	0.482
15	NaN	0.464	0.645	0.894	0.073	0.635	0.895
16	0.233	0.690	0.323	0.176	0.198	0.617	0.930
17	0.332	0.398	0.394	0.516	0.218	0.076	0.557
18	0.221	0.852	0.171	0.123	0.356	0.586	0.967
19	0.116	0.732	0.191	0.674	0.401	0.554	0.277
20	NaN	0.088	0.919	0.144	0.489	0.564	0.691
Model Fit	0.002	0.477	0.198	0.589	0.581	0.603	0.911

Table 15 gives the loadings of one-factor solutions for various sample sizes of unidimensional datasets on three-point scale. It also gives the proportion of variation (Prop Var) explained by the factor as well as the fitness of the factor model to the data.

Table 15 shows that there is increased number of indicators that influence the factors from a highest of 15 (in two-point scale) to 18 (under three-point scale). Even though there does not appear to be converging number of the influential indicators, the dominant number is 17. Incidentally, for  $n = 150$ , the number of influential indicators is also 17. The proportion of variation accounted for by the factors increased from 50.7% (for  $n = 30$ ) to a high of 59.6% (for  $n = 200$ ) then fluctuates afterwards. There is a high level of fit of the model for all sample sizes. It appears, therefore, that the model is as good for smaller sample size as for higher samples.

Table 15: Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Three-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.209	0.473	0.457	0.464	0.386	0.417	0.409
2	0.753	0.564	0.514	0.533	0.510	0.577	0.501
3	0.343	0.569	0.561	0.583	0.565	0.570	0.602
4	0.148	0.559	0.453	0.454	0.477	0.450	0.508
5	0.522	0.298	0.416	0.408	0.359	0.332	0.409
6	0.768	0.869	0.841	0.861	0.863	0.855	0.890
7	0.475	0.686	0.803	0.747	0.791	0.803	0.781
8	0.824	0.886	0.915	0.879	0.897	0.914	0.899
9	0.769	0.747	0.821	0.853	0.824	0.847	0.816
10	0.789	0.802	0.759	0.818	0.813	0.791	0.806
11	0.922	0.884	0.808	0.884	0.865	0.846	0.865
12	0.962	0.911	0.929	0.933	0.907	0.918	0.931
13	0.858	0.888	0.910	0.904	0.928	0.913	0.929
14	0.748	0.868	0.843	0.845	0.861	0.879	0.865
15	0.963	0.855	0.931	0.913	0.922	0.896	0.910
16	0.547	0.617	0.753	0.717	0.759	0.765	0.769
17	0.713	0.715	0.810	0.883	0.840	0.844	0.827
18	0.786	0.682	0.694	0.726	0.694	0.682	0.724
19	0.561	0.802	0.713	0.732	0.757	0.707	0.712
20	0.821	0.891	0.883	0.929	0.897	0.927	0.916
Prop Var	0.507	0.558	0.577	0.596	0.588	0.589	0.597
Fit	0.900	0.953	0.962	0.968	0.968	0.969	0.971

The loadings, proportion of variation explained and fitness of two-factor model for various sample sizes of unidimensional datasets on three-point scale are presented in Table 16. There is higher number of indicator variables on the factors, particularly for the first factor than its counterpart under the two-point scale. The proportion of variation explained also increases up to  $n = 150$  and decreases thereafter. There is a general repetition of indicators on factors at all samples, with exception of  $n = 150$ . Unlike all other sample sizes, the results for  $n = 150$  is more plausible as the first factor accounts for almost all cumulative variation explained by the solution. The fitness of the model is almost perfect for all sample sizes.



Table 16: Two-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Three-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.000	0.534	0.410	0.243	0.463	0.114	0.300	0.359
2	0.746	0.201	0.150	0.734	0.517	0.000	0.338	0.419
3	0.274	0.214	0.547	0.228	0.539	0.155	0.199	0.651
4	0.000	0.186	0.291	0.535	0.413	0.192	0.430	0.205
5	0.223	0.726	0.238	0.179	0.148	0.939	0.373	0.198
6	0.806	0.126	0.701	0.512	0.783	0.308	0.711	0.501
7	0.719	-0.311	0.530	0.433	0.764	0.245	0.629	0.421
8	0.772	0.298	0.788	0.434	0.841	0.369	0.620	0.623
9	0.508	0.707	0.521	0.543	0.803	0.188	0.627	0.577
10	0.792	0.193	0.712	0.393	0.699	0.303	0.650	0.501
11	0.902	0.269	0.656	0.592	0.816	0.113	0.586	0.669
12	0.896	0.360	0.585	0.731	0.901	0.234	0.699	0.617
13	0.763	0.387	0.767	0.462	0.857	0.303	0.642	0.637
14	0.690	0.289	0.845	0.341	0.811	0.233	0.616	0.578
15	0.859	0.429	0.709	0.481	0.882	0.296	0.706	0.580
16	0.316	0.588	0.318	0.595	0.730	0.191	0.706	0.299
17	0.631	0.327	0.411	0.636	0.834	0.000	0.542	0.715
18	0.670	0.409	0.725	0.193	0.658	0.218	0.388	0.652
19	0.255	0.754	0.538	0.612	0.623	0.387	0.421	0.625
20	0.765	0.304	0.815	0.410	0.858	0.216	0.841	0.464
Prop Var	0.417	0.179	0.357	0.242	0.522	0.098	0.331	0.287
Cum Var		0.596		0.599		0.620		0.618
Fit		0.937		0.962		0.968		0.971

Table 16, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.505	0.000	0.373	0.194	0.367	0.210
2	0.376	0.345	0.442	0.374	0.373	0.335
3	0.340	0.459	0.358	0.487	0.582	0.268
4	0.407	0.267	0.241	0.448	0.333	0.387
5	0.000	0.464	0.140	0.385	0.185	0.396
6	0.692	0.529	0.777	0.380	0.704	0.553
7	0.449	0.672	0.711	0.383	0.562	0.543
8	0.682	0.586	0.738	0.540	0.626	0.645
9	0.611	0.554	0.732	0.430	0.556	0.598
10	0.506	0.644	0.699	0.380	0.583	0.557
11	0.618	0.605	0.751	0.401	0.536	0.691
12	0.639	0.644	0.774	0.493	0.658	0.658
13	0.642	0.669	0.740	0.533	0.661	0.652
14	0.630	0.588	0.682	0.560	0.653	0.569
15	0.668	0.635	0.709	0.549	0.633	0.653
16	0.406	0.672	0.589	0.495	0.572	0.514
17	0.697	0.492	0.704	0.464	0.700	0.470
18	0.457	0.525	0.669	0.234	0.376	0.656
19	0.527	0.544	0.524	0.488	0.547	0.459
20	0.670	0.597	0.788	0.489	0.610	0.686
Prop Var	0.304	0.301	0.404	0.199	0.311	0.295
Cum Var		0.605		0.603		0.606
Fit		0.971		0.972		0.973

In Table 17, three-factor solutions for unidimensional response datasets for various sample sizes on three-point scale are displayed.

Table 17 shows that there is increased number of indicators on factors particularly for the first factor (over that of two-point scale). There is, however, the incidence of repeating indicators of multiple factors for all sample sizes. The sample size of  $n = 150$  appears to produce a more reasonable result as the last factor (Factor 3) accounts for a negligible proportion of the cumulative variation. The fitness of the three-factor model increases as sample size increase.

Table 17: Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Three-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.000	0.000	0.848	0.459	0.231	0.000
2	0.699	0.411	0.000	0.205	0.821	0.000
3	0.256	0.217	0.102	0.565	0.181	0.000
4	0.000	0.261	0.000	0.249	0.454	0.434
5	0.115	0.886	0.239	0.109	0.000	0.710
6	0.777	0.268	0.000	0.704	0.438	0.255
7	0.745	-0.264	-0.120	0.538	0.376	0.194
8	0.766	0.223	0.244	0.779	0.343	0.290
9	0.462	0.681	0.317	0.575	0.511	0.000
10	0.742	0.529	-0.227	0.711	0.321	0.225
11	0.883	0.278	0.153	0.674	0.504	0.266
12	0.875	0.336	0.223	0.625	0.652	0.232
13	0.739	0.372	0.209	0.798	0.378	0.180
14	0.679	0.275	0.152	0.828	0.243	0.293
15	0.824	0.453	0.195	0.770	0.443	0.000
16	0.296	0.425	0.415	0.318	0.513	0.326
17	0.675	0.000	0.491	0.421	0.552	0.307
18	0.652	0.343	0.268	0.710	0.125	0.194
19	0.247	0.478	0.566	0.552	0.527	0.282
20	0.820	0.000	0.468	0.808	0.315	0.283
Prop Var	0.401	0.157	0.111	0.369	0.191	0.080
Cum Var			0.669			0.639
Fit			0.956			0.971

Table 17, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.448	0.149	0.124	0.244	0.323	0.241
2	0.255	0.605	0.000	0.315	0.362	0.242
3	0.264	0.624	0.131	0.217	0.636	0.174
4	0.235	0.404	0.200	0.134	0.200	0.508
5	0.176	0.112	0.780	0.309	0.145	0.245
6	0.690	0.375	0.306	0.602	0.394	0.485
7	0.564	0.546	0.238	0.760	0.224	0.262
8	0.799	0.310	0.379	0.495	0.551	0.482
9	0.768	0.298	0.166	0.671	0.438	0.333
10	0.617	0.334	0.302	0.499	0.424	0.498
11	0.681	0.448	0.100	0.664	0.535	0.298
12	0.809	0.415	0.214	0.646	0.494	0.457
13	0.741	0.439	0.289	0.705	0.496	0.329
14	0.672	0.470	0.203	0.535	0.481	0.441
15	0.707	0.538	0.287	0.542	0.493	0.554
16	0.649	0.354	0.160	0.399	0.239	0.647
17	0.770	0.365	0.000	0.579	0.596	0.334
18	0.601	0.294	0.194	0.245	0.671	0.382
19	0.477	0.406	0.414	0.476	0.535	0.238
20	0.844	0.280	0.203	0.619	0.367	0.633
Prop Var	0.388	0.168	0.082	0.265	0.206	0.170
Cum Var			0.637			0.641
Fit			0.973			0.975

## Two-dimensional three-point Likert scale

In two-dimensional three-point scale, two person-abilities underlie item responses. In this system, we expect two repeating factors as the two dimensions that underlie the generation of the data are just the same. Alternatively, we could expect a single dominant factor in the first factor with the other influenced by about one or at most two indicators and explains negligible amount of the cumulative variation. Table 18 shows the  $p$ -values of fitness of items for two-dimensional GPC model on three-point scale data for various sample sizes.

From Table 18, it can be observed that, generally, items significantly fit the two-dimensional GPC model on three-point scale. However, on the same three-point scale, the magnitude of  $p$ -values differ from unidimensional to two-dimensional case. There appears to be two groups of items: (1) items whose  $p$ -value decreases with an additional ability dimension, and (2) items whose  $p$ -value increases with additional ability dimension. Specifically, for  $n = 150$ , the first group contains the set of Items 2, 6, 7, 8, 9, 10, 11, 13, 14, 19 and 20, while the rest of the items constitute the second group. Items of the first group will require just one person-ability to get a response in higher categories, whereas those of the second group need multiple person-abilities to get a similar response. The overall fitness of the model to the two-dimensional three-point response data is significant at all sample sizes, except for  $n = 30$ .

Table 18: *P*-values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Three-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.405	0.049	0.493	0.737	0.777	0.462	0.142
2	0.142	0.472	0.102	0.102	0.592	0.472	0.081
3	NaN	0.337	0.968	0.439	0.104	0.436	0.889
4	0.023	0.438	0.243	0.630	0.512	0.154	0.155
5	0.260	0.304	0.892	0.304	0.985	0.163	0.682
6	NaN	0.239	0.704	0.201	0.618	0.065	0.709
7	NaN	0.286	0.171	0.092	0.006	0.202	0.810
8	NaN	0.112	0.737	0.848	0.408	0.504	0.645
9	NaN	0.001	0.133	0.798	0.523	0.606	0.477
10	NaN	0.412	0.101	0.641	0.867	0.204	0.712
11	0.206	0.356	0.493	0.404	0.830	0.187	0.955
12	NaN	0.243	0.228	0.054	0.159	0.328	0.496
13	NaN	0.193	0.010	0.184	0.028	0.108	0.179
14	NaN	0.460	0.267	0.034	0.908	0.212	0.191
15	NaN	0.343	0.878	0.017	0.716	0.621	0.122
16	0.180	0.169	0.437	0.363	0.967	0.067	0.431
17	0.051	0.145	0.782	0.615	0.100	0.469	0.122
18	0.005	0.038	0.294	0.262	0.146	0.389	0.545
19	NaN	0.192	0.158	0.590	0.685	0.191	0.520
20	NaN	0.782	0.774	0.985	0.807	0.167	0.003
Model Fit	0.006	0.469	0.969	0.482	0.872	0.924	0.779

The loadings of two-factor solutions for two-dimensional item response

dataset for various sample sizes on three-point scale are illustrated in Table 19.

Table 19: Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Three-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.237	0.318	0.673	0.196	0.444	0.199	0.140	0.749
2	-0.238	0.613	0.510	0.578	0.322	0.692	0.591	0.213
3	0.588	0.660	0.593	0.277	0.495	0.553	0.659	0.231
4	0.658	0.146	0.533	0.429	0.258	0.613	0.712	0.163
5	-0.140	0.000	0.000	0.632	0.491	0.430	0.469	0.000
6	0.623	0.658	0.874	0.346	0.741	0.578	0.809	0.428
7	0.809	0.442	0.837	0.295	0.640	0.610	0.855	0.249
8	0.540	0.684	0.900	0.344	0.738	0.560	0.860	0.407
9	0.899	0.334	0.860	0.381	0.791	0.482	0.803	0.409
10	0.780	0.522	0.682	0.527	0.671	0.588	0.801	0.429
11	0.638	0.558	0.888	0.272	0.728	0.525	0.795	0.418
12	0.623	0.703	0.925	0.308	0.736	0.617	0.833	0.505
13	0.828	0.523	0.908	0.295	0.773	0.593	0.907	0.356
14	0.722	0.508	0.817	0.358	0.748	0.614	0.886	0.332
15	0.781	0.466	0.828	0.480	0.691	0.635	0.866	0.445
16	0.412	0.829	0.819	0.409	0.682	0.482	0.722	0.467
17	0.522	0.659	0.900	0.281	0.684	0.661	0.748	0.494
18	0.857	0.000	0.870	0.205	0.824	0.335	0.772	0.297
19	0.504	0.784	0.814	0.301	0.776	0.367	0.810	0.342
20	0.646	0.674	0.863	0.429	0.741	0.632	0.829	0.475
Prop Var	0.406	0.308	0.615	0.148	0.445	0.304	0.582	0.160
Cum Var		0.714		0.763		0.749		0.742
Fit		0.974		0.990		0.989		0.989



Table 19, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.301	0.446	0.425	0.276	0.570	0.000
2	0.554	0.377	0.642	0.250	0.630	0.288
3	0.487	0.577	0.660	0.323	0.652	0.327
4	0.392	0.466	0.372	0.452	0.578	0.202
5	0.235	0.471	0.165	0.462	0.230	0.739
6	0.752	0.541	0.701	0.603	0.858	0.382
7	0.669	0.567	0.638	0.607	0.837	0.267
8	0.647	0.700	0.648	0.671	0.879	0.391
9	0.664	0.609	0.636	0.582	0.831	0.395
10	0.669	0.603	0.661	0.570	0.852	0.326
11	0.696	0.579	0.639	0.650	0.870	0.310
12	0.657	0.693	0.744	0.609	0.877	0.379
13	0.773	0.579	0.708	0.651	0.897	0.373
14	0.614	0.693	0.721	0.581	0.856	0.378
15	0.696	0.644	0.728	0.620	0.861	0.417
16	0.644	0.565	0.556	0.600	0.815	0.302
17	0.723	0.571	0.733	0.535	0.802	0.448
18	0.740	0.356	0.599	0.539	0.727	0.335
19	0.641	0.579	0.647	0.479	0.792	0.292
20	0.785	0.551	0.711	0.637	0.884	0.354
Prop Var	0.403	0.320	0.400	0.301	0.611	0.136
Cum Var		0.723		0.700		0.747
Fit		0.988		0.986		0.991

In the table, we observe that there is increased number of indicators on factors, particularly for Factor 1 over that of two-point scale. The amounts of variation explained are almost the same for the two factors for sample sizes 30, 150, 500, and 800. The amount of cumulative variation explained by the two-factor model generally fluctuates with increasing sample size, but highest at  $n = 100$ . At this point, the fitness of the model is also highest.

### **Three-dimensional three-point Likert scale**

Table 20 shows the  $p$ -values of fitness of items for three-dimensional GPC model on three-point scale for various sample sizes. On the three-dimensional response datasets, most items fit the model, particularly for larger sample sizes ( $n \geq 150$ ). For smaller sample sizes ( $n \leq 100$ ), the  $p$ -values of items worsened on high dimensions, especially for items that may require only one person-ability to get a response in higher categories. Table 20 indicates that items whose fitness is quite high or almost perfect may require up to three person-abilities to get a response in higher categories. At  $n = 30$ , even though items misfit the model, the overall fitness of the model is significant. The plausibility is that the IRT model yields better results on high response scales with large number of dimensions.

Table 20: *P*-values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Three-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.238	0.527	0.893	0.022	0.507	0.257
2	NaN	0.275	0.084	0.065	0.890	0.223	0.228
3	0.063	0.012	0.565	0.153	0.310	0.285	0.078
4	0.007	0.046	0.095	0.468	0.910	0.671	0.111
5	0.041	0.048	0.194	0.866	0.033	0.531	0.049
6	NaN	NaN	0.214	0.570	0.510	0.128	0.718
7	NaN	0.292	0.902	0.796	0.741	0.107	0.076
8	NaN	0.065	0.054	0.184	0.921	0.011	0.126
9	NaN	0.167	0.351	0.843	0.315	0.138	0.722
10	NaN	0.067	0.043	0.068	0.123	0.631	0.733
11	NA	0.327	0.384	0.345	0.440	0.663	0.144
12	NaN	0.177	0.012	0.077	0.524	0.579	0.878
13	NaN	0.293	0.031	0.292	0.143	0.046	0.754
14	NaN	0.102	0.191	0.164	0.552	0.739	0.180
15	NA	0.013	0.024	0.347	0.526	0.325	0.180
16	NaN	0.101	0.060	0.062	0.039	0.174	0.660
17	NaN	0.085	0.041	0.210	0.387	0.379	0.909
18	NaN	0.359	0.182	0.769	0.678	0.490	0.642
19	NaN	0.409	0.097	0.556	0.210	0.675	0.304
20	NaN	0.019	0.101	0.576	0.011	0.146	0.013
Model Fit	0.830	0.623	0.717	1.000	0.948	0.998	0.766

The loadings of corresponding three-factor solutions for three dimensional three-point scale dataset for various sample sizes are shown in Table 21.

Table 21 demonstrates that there is increased number of indicators that influence the formation of factors, Factor 1 in particular, over that of two-point scale. In a like manner, the amount of cumulative variation is quite high in favour of three-point scale. There is a higher number of indicator variables on the factors, the first factor in particular. The solution for  $n = 150$  is consistent with our expectation as the first is dominant with 18 influential indicators and the others are influenced by a single indicator each. The amount of cumulative variation largely oscillates with increasing sample size, but peaks at  $n = 150$  with highest model fitness.

Table 21: Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Three-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.779	0.403	-0.147	0.205	0.258	0.820
2	0.835	0.000	0.109	0.613	0.400	0.261
3	0.102	0.000	1.034	0.805	0.268	0.000
4	0.231	0.296	0.536	0.678	0.383	0.229
5	0.198	0.945	0.000	0.189	0.558	0.167
6	0.924	0.245	0.238	0.727	0.548	0.366
7	0.844	0.392	0.172	0.592	0.694	0.290
8	0.575	0.513	0.526	0.735	0.579	0.228
9	0.708	0.493	0.369	0.645	0.597	0.387
10	0.735	0.610	0.223	0.611	0.634	0.353
11	0.745	0.530	0.343	0.766	0.487	0.284
12	0.762	0.489	0.346	0.593	0.706	0.253
13	0.664	0.581	0.422	0.674	0.599	0.336
14	0.804	0.446	0.344	0.663	0.619	0.303
15	0.721	0.592	0.323	0.747	0.553	0.304
16	0.570	0.655	0.366	0.568	0.666	0.269
17	0.802	0.365	0.383	0.523	0.744	0.297
18	0.756	0.217	0.321	0.686	.494	0.252
19	0.752	0.302	0.301	0.759	0.361	0.281
20	0.865	0.396	0.222	0.660	0.735	0.156
Prop Var	0.497	0.225	0.155	0.412	0.316	0.107
Cum Var			0.878			0.835
Fit			0.995			0.995

Table 21, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.327	0.163	0.877	0.556	0.232	0.189
2	0.655	0.260	0.318	0.553	0.522	0.191
3	0.757	0.282	0.153	0.692	0.359	0.000
4	0.615	0.232	0.432	0.288	0.361	0.562
5	0.323	0.921	0.184	0.200	0.760	0.239
6	0.871	0.236	0.289	0.767	0.300	0.468
7	0.853	0.254	0.274	0.830	0.280	0.278
8	0.889	0.208	0.341	0.804	0.334	0.423
9	0.869	0.261	0.247	0.734	0.375	0.367
10	0.853	0.261	0.296	0.784	0.271	0.363
11	0.894	0.219	0.287	0.779	0.309	0.440
12	0.886	0.274	0.337	0.751	0.372	0.469
13	0.888	0.208	0.321	0.760	0.380	0.467
14	0.845	0.195	0.354	0.741	0.400	0.407
15	0.881	0.255	0.363	0.730	0.364	0.503
16	0.832	0.239	0.318	0.705	0.220	0.536
17	0.869	0.276	0.228	0.767	0.366	0.370
18	0.820	0.215	0.265	0.787	0.205	0.279
19	0.806	0.330	0.298	0.785	0.193	0.388
20	0.881	0.264	0.341	0.821	0.291	0.318
Prop Var	0.638	0.126	0.100	0.506	0.152	0.134
Cum Var			0.864			0.792
Fit			0.997			0.994

## Examining Five-Point Likert Scale

In this section, we present and discuss the results from five-point Likert scale datasets. On this scale, 16 datasets have been generated under different conditions such as number of dimensions and sample size. For these datasets, we assess the performance of GPC item response model and that of factor analysis. We begin with five-point response scale with underlying unidimensional person-ability.

### Unidimensional five-point Likert scale

Table 22 illustrates the  $p$ -values of item fitness for the unidimensional generalised partial credit (GPC) item response model at various sample sizes. The Table also contains the  $p$ -values of the fitness of the overall GPC model to the five-point response data at various sample sizes.

We note from Table 22 that overwhelming majority of items significantly fit the unidimensional GPC model at all sample sizes, except  $n = 30$  where the  $p$ -values of some items could not be determined. Not unexpectedly, the overall GPC model significantly fits the five-point response dataset, with the exception of  $n = 30$ . As sample size increase from 100 up to 1000, the fitness of the item response model fluctuates, but highest at  $n = 1000$ .

Table 22: *P*-values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Five-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.097	0.146	0.519	0.007	0.747	0.247	0.808
2	0.234	0.063	0.481	0.348	0.583	0.143	0.838
3	0.003	0.773	0.932	0.425	0.199	0.537	0.982
4	0.384	0.581	0.135	0.158	0.293	0.268	0.979
5	0.183	0.189	0.090	0.499	0.734	0.227	0.600
6	NaN	0.526	0.746	0.328	0.705	0.540	0.918
7	0.013	0.753	0.378	0.132	0.858	0.036	0.380
8	NaN	0.330	0.201	0.452	0.590	0.480	0.306
9	NaN	0.047	0.914	0.813	0.709	0.693	0.006
10	NaN	0.097	0.185	0.045	0.421	0.768	0.434
11	NaN	0.096	0.277	0.221	0.329	0.494	0.265
12	NaN	0.918	0.122	0.598	0.945	0.111	0.117
13	NaN	0.070	0.174	0.154	0.193	0.233	0.596
14	0.025	0.125	0.128	0.296	0.232	0.429	0.641
15	NaN	0.029	0.013	0.247	0.907	0.833	0.311
16	0.381	0.038	0.060	0.154	0.925	0.454	0.463
17	0.229	0.190	0.178	0.035	0.592	0.700	0.964
18	0.001	0.125	0.007	0.459	0.232	0.899	0.601
19	0.031	0.304	0.053	0.876	0.172	0.821	0.486
20	NaN	0.223	0.118	0.037	0.660	0.067	0.635
Model Fit	0.007	0.253	0.368	0.809	0.435	0.150	0.990



The loadings of corresponding one-factor model at various sample sizes are represented in Table 23. In Table 23, there is increased number of indicators that influence the factor from a highest of 18 (on three-point scale) to 20 (under five-point scale). Even though there does not appear to be converging number of influential indicators, the dominant number is 20. Incidentally, the dominant number of influential indicators starts with  $n = 150$ . The proportion of variation accounted increases from 58% (for  $n = 30$ ) to highest of 72.9% (for  $n = 1000$ ). It is relevant to note that proportion of variation explained is almost the same for  $n = 200$  as for  $n = 1000$ . There is a high level of fitness of the model for all sample sizes. The model is as good for smaller samples as for larger ones. There is also improved fitness for  $n = 30$  over smaller scales.

The loadings of corresponding one-factor model at various sample sizes are represented in Table 23. In Table 23, there is increased number of indicators that influence the factor from a highest of 18 (on three-point scale) to 20 (under five-point scale). Even though there does not appear to be converging number of influential indicators, the dominant number is 20. Incidentally, the dominant number of influential indicators starts with  $n = 150$ . The proportion of variation accounted increases from 58% (for  $n = 30$ ) to highest of 72.9% (for  $n = 1000$ ). It is relevant to note that proportion of variation explained is almost the same for  $n = 200$  as for  $n = 1000$ . There is a high level of fitness of the model for all sample sizes. The model is as good for smaller samples as for larger ones. There is also improved fitness for  $n = 30$  over smaller scales.

Table 23: Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.399	0.600	0.591	0.613	0.560	0.588	0.573
2	0.664	0.661	0.699	0.615	0.678	0.706	0.667
3	0.409	0.726	0.643	0.704	0.723	0.719	0.738
4	0.295	0.710	0.582	0.596	0.637	0.617	0.656
5	0.628	0.393	0.512	0.542	0.488	0.476	0.557
6	0.865	0.914	0.907	0.922	0.927	0.914	0.940
7	0.559	0.775	0.875	0.861	0.872	0.875	0.881
8	0.901	0.934	0.942	0.938	0.944	0.952	0.946
9	0.847	0.844	0.898	0.908	0.910	0.917	0.895
10	0.827	0.863	0.860	0.885	0.878	0.880	0.890
11	0.927	0.920	0.887	0.935	0.920	0.915	0.924
12	0.942	0.950	0.945	0.967	0.951	0.957	0.962
13	0.875	0.935	0.951	0.941	0.958	0.956	0.963
14	0.780	0.906	0.908	0.915	0.927	0.931	0.923
15	0.948	0.922	0.958	0.962	0.962	0.948	0.955
16	0.688	0.803	0.871	0.845	0.868	0.884	0.866
17	0.768	0.827	0.885	0.929	0.915	0.918	0.908
18	0.816	0.764	0.798	0.832	0.805	0.816	0.845
19	0.690	0.856	0.833	0.816	0.842	0.829	0.839
20	0.926	0.935	0.933	0.953	0.947	0.955	0.955
Prop Var	0.580	0.678	0.697	0.714	0.717	0.720	0.729
Fit	0.949	0.982	0.985	0.987	0.988	0.988	0.990

Table 24 shows the loadings of two-factor solutions for unidimensional five-point response data at various sample sizes. The Table also shows the proportion and cumulative variation explained as well as fitness of the two-factor model.

There is a higher number of indicator variables on the two factors for the five-point scale than those of the three-point scale. The cumulative variations accounted for by the two-factor model are consistently higher for a five-point scale than its three-point scale counterpart. Thus, for unidimensional datasets, a five-point scale gives optimum results over three-point scale. Generally, cumulative variation explained increases remarkably from 64.5% (for  $n = 30$ ) up to 73.2% (for  $n = 150$ ), slight increment to 73.9% (for  $n = 200$  and 500), then fluctuates thereafter. The indicator variables greatly influence Factor 1 than Factor 2 for  $n = 150, 500,$  and 1000 which is in line with the underlying dimension of the scale. The two-factor model nearly perfectly fits the five-point scale dataset for all sample sizes.

Table 24: Two-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.126	0.519	0.462	0.382	0.544	0.248	0.508	0.354
2	0.573	0.339	0.222	0.766	0.529	0.468	0.323	0.559
3	0.157	0.490	0.575	0.444	0.620	0.230	0.290	0.733
4	0.228	0.187	0.419	0.603	0.218	0.747	0.397	0.448
5	0.228	0.787	0.288	0.268	0.419	0.295	0.552	0.202
6	0.766	0.413	0.687	0.602	0.731	0.539	0.710	0.589
7	0.886	-0.242	0.520	0.583	0.685	0.551	0.720	0.419
8	0.813	0.411	0.774	0.533	0.832	0.449	0.683	0.642
9	0.571	0.662	0.591	0.607	0.764	0.471	0.652	0.632
10	0.723	0.409	0.758	0.445	0.701	0.499	0.743	0.501
11	0.798	0.474	0.678	0.620	0.740	0.487	0.685	0.636
12	0.866	0.410	0.622	0.734	0.763	0.559	0.703	0.663
13	0.709	0.510	0.778	0.531	0.819	0.485	0.672	0.659
14	0.597	0.504	0.844	0.415	0.821	0.405	0.670	0.623
15	0.805	0.500	0.726	0.569	0.829	0.483	0.766	0.588
16	0.481	0.510	0.554	0.587	0.755	0.436	0.678	0.511
17	0.671	0.381	0.567	0.609	0.722	0.511	0.633	0.684
18	0.708	0.412	0.793	0.261	0.803	0.238	0.486	0.702
19	0.370	0.682	0.586	0.631	0.615	0.582	0.559	0.597
20	0.778	0.500	0.750	0.562	0.797	0.458	0.811	0.528
Prop Var	0.409	0.237	0.398	0.306	0.493	0.225	0.395	0.337
Cum Var		0.645		0.704		0.732		0.739
Fit		0.966		0.985		0.987		0.989

Table 24, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.554	0.124	0.511	0.298	0.293	0.652
2	0.625	0.262	0.546	0.448	0.552	0.377
3	0.687	0.231	0.509	0.522	0.623	0.397
4	0.586	0.249	0.346	0.574	0.599	0.276
5	0.236	0.757	0.242	0.477	0.519	0.214
6	0.880	0.296	0.794	0.465	0.802	0.491
7	0.795	0.360	0.735	0.477	0.804	0.370
8	0.902	0.286	0.744	0.595	0.834	0.448
9	0.858	0.307	0.781	0.486	0.789	0.423
10	0.804	0.351	0.730	0.492	0.802	0.392
11	0.878	0.281	0.772	0.495	0.818	0.431
12	0.856	0.421	0.783	0.550	0.843	0.465
13	0.893	0.346	0.749	0.593	0.835	0.478
14	0.882	0.291	0.740	0.565	0.812	0.441
15	0.900	0.340	0.747	0.584	0.836	0.461
16	0.792	0.358	0.735	0.493	0.741	0.449
17	0.859	0.315	0.749	0.530	0.771	0.482
18	0.751	0.290	0.761	0.345	0.760	0.374
19	0.801	0.265	0.605	0.577	0.723	0.425
20	0.891	0.322	0.781	0.549	0.811	0.505
Prop Var	0.621	0.118	0.470	0.262	0.550	0.190
Cum Var		0.739		0.731		0.740
Fit		0.990		0.990		0.991

Table 25 indicates the loadings of three-factor solutions for unidimensional five-point response data at various sample sizes. The Table also presents the proportion and cumulative variation explained as well as fitness of the three-factor model.

There is increased number of influential indicator variables, especially on Factor 1, over that of three-point scale. Also, cumulative variation explained by the three-factor model are higher for five-point than a three-point scale. Table 25 shows that there is the incidence of several repeating indicators on multiple factors for all sample sizes, except for  $n = 150$  which has few. This is a clear violation of the factor principle. The sample size of 150 seems to produce most realistic result as it contains highest number of influential indicators on Factor 1, which accounts for largest (46.2%) proportion of variation. In addition, the remaining two factors are influenced by just one or two indicators. The three-factor model largely fits the unidimensional five-point scale datasets.

Table 25: Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Five-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.135	0.532	0.000	0.492	0.366	0.000
2	0.582	0.350	0.000	0.260	0.758	0.115
3	0.164	0.501	0.000	0.590	0.399	0.155
4	0.131	0.000	0.892	0.367	0.524	0.446
5	0.220	0.772	0.160	0.156	0.125	0.711
6	0.785	0.432	0.000	0.692	0.531	0.271
7	0.889	-0.245	0.000	0.520	0.518	0.270
8	0.789	0.371	0.299	0.763	0.451	0.312
9	0.556	0.638	0.221	0.624	0.568	0.142
10	0.704	0.380	0.236	0.767	0.387	0.192
11	0.797	0.470	0.106	0.681	0.552	0.274
12	0.858	0.400	0.142	0.639	0.664	0.273
13	0.705	0.503	0.122	0.789	0.470	0.216
14	0.578	0.477	0.233	0.843	0.347	0.232
15	0.800	0.491	0.138	0.753	0.517	0.172
16	0.492	0.527	0.000	0.546	0.509	0.312
17	0.652	0.352	0.228	0.576	0.542	0.253
18	0.679	0.364	0.335	0.766	0.156	0.333
19	0.344	0.648	0.286	0.596	0.570	0.246
20	0.785	0.508	0.000	0.761	0.495	0.234
Prop Var	0.399	0.225	0.069	0.403	0.244	0.086
Cum Var			0.694			0.732
Fit			0.975			0.989



Table 25, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.513	0.201	0.233	0.283	0.279	0.594
2	0.539	0.431	0.166	0.338	0.518	0.184
3	0.588	0.171	0.254	0.218	0.739	0.289
4	0.234	0.682	0.217	0.396	0.396	0.212
5	0.236	0.227	0.647	0.484	0.146	0.295
6	0.665	0.468	0.424	0.711	0.501	0.330
7	0.671	0.491	0.283	0.672	0.413	0.372
8	0.777	0.384	0.378	0.675	0.557	0.346
9	0.764	0.426	0.222	0.724	0.554	0.214
10	0.670	0.450	0.298	0.638	0.418	0.471
11	0.759	0.444	0.178	0.673	0.555	0.345
12	0.767	0.512	0.232	0.626	0.582	0.451
13	0.804	0.429	0.277	0.652	0.583	0.353
14	0.786	0.333	0.331	0.573	0.547	0.463
15	0.811	0.420	0.296	0.637	0.496	0.541
16	0.734	0.375	0.282	0.574	0.440	0.448
17	0.740	0.456	0.199	0.595	0.606	0.383
18	0.736	0.130	0.428	0.459	0.630	0.341
19	0.580	0.533	0.319	0.543	0.525	0.314
20	0.800	0.439	0.223	0.698	0.444	0.500
Prop Var	0.462	0.177	0.098	0.333	0.262	0.150
Cum Var			0.737			0.745
Fit			0.989			0.990

## **Two-dimensional five-point Likert scale**

In this structure, we expect a good two-factor solution to possess either two repeating factors or a single dominant first factor since the same information is contained on both dimensions that underlie item responses. Table 26 shows the  $p$ -values of item fitness to the two-dimensional generalised partial credit (GPC) item response model at various sample sizes. The Table also contains the  $p$ -values of the fitness of the overall GPC model to the five-point response data at various sample sizes.

Table 26 indicates that most items significantly fit the two-dimensional GPC model at sample size of 100 and beyond. The fitnesses of almost all items could not be evaluated at  $n = 30$  due to low degrees of freedom. Surprisingly, the IRT model significantly fits the data at  $n = 30$ . The overall fitness of the IRT model generally fluctuates with increasing sample size. Although the fitnesses of items to the IRT model are significant, the amount of significance on the two-dimensional five-point scale has deteriorated as compared to that of the three-point scale.

Table 26: *P*-values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Five-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.070	0.229	0.449	0.239	0.427	0.894	0.046
2	0.011	0.085	0.107	0.141	0.034	0.581	0.823
3	NaN	0.148	0.192	0.256	0.195	0.375	0.416
4	0.064	0.323	0.424	0.215	0.467	0.621	0.236
5	0.139	0.372	0.230	0.476	0.884	0.267	0.532
6	NaN	0.151	0.218	0.371	0.091	0.246	0.213
7	NaN	0.107	0.143	0.843	0.202	0.492	0.080
8	NaN	0.017	0.022	0.168	0.083	0.883	0.082
9	NaN	0.146	0.240	0.009	0.300	0.361	0.172
10	NaN	0.276	0.538	0.386	0.226	0.112	0.618
11	NaN	0.001	0.243	0.490	0.426	0.268	0.569
12	NaN	0.192	0.018	0.033	0.297	0.367	0.475
13	NaN	0.130	0.313	0.831	0.118	0.309	0.068
14	NaN	0.078	0.209	0.072	0.448	0.797	0.142
15	NaN	0.079	0.625	0.118	0.454	0.899	0.836
16	NaN	0.136	0.067	0.430	0.473	0.068	0.815
17	NaN	0.335	0.425	0.045	0.776	0.009	0.795
18	NaN	0.007	0.187	0.422	0.324	0.060	0.330
19	NaN	0.440	0.514	0.337	0.145	0.355	0.641
20	NaN	0.129	0.052	0.346	0.841	0.045	0.001
Model Fit	0.550	0.896	0.956	0.579	0.326	0.969	0.864

Table 27 shows the corresponding two-factor solutions for the two dimen-

sional five-point response data at various sample sizes.

Table 27: Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Five-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.556	0.000	0.731	0.280	0.253	0.634	0.249	0.711
2	0.540	0.000	0.742	0.384	0.727	0.375	0.625	0.488
3	0.931	-0.148	0.753	0.205	0.580	0.588	0.635	0.543
4	0.724	0.000	0.696	0.313	0.726	0.309	0.688	0.386
5	0.000	0.967	0.282	0.978	0.600	0.454	0.597	0.205
6	0.915	0.111	0.930	0.244	0.680	0.690	0.699	0.654
7	0.952	0.000	0.886	0.261	0.693	0.633	0.799	0.491
8	0.924	0.000	0.931	0.299	0.659	0.692	0.781	0.590
9	0.905	0.000	0.922	0.273	0.644	0.711	0.733	0.602
10	0.923	0.251	0.837	0.402	0.736	0.587	0.708	0.632
11	0.880	0.124	0.913	0.284	0.673	0.681	0.732	0.606
12	0.932	0.182	0.946	0.263	0.715	0.672	0.703	0.684
13	0.976	0.000	0.929	0.310	0.712	0.680	0.805	0.578
14	0.908	0.149	0.901	0.281	0.698	0.684	0.785	0.571
15	0.907	0.198	0.924	0.322	0.777	0.592	0.699	0.695
16	0.871	0.102	0.887	0.353	0.588	0.697	0.645	0.667
17	0.883	0.000	0.930	0.273	0.738	0.636	0.660	0.673
18	0.780	0.134	0.868	0.294	0.501	0.805	0.766	0.486
19	0.903	0.113	0.870	0.284	0.612	0.678	0.723	0.595
20	0.953	0.187	0.901	0.379	0.763	0.626	0.740	0.636
Prop Var	0.719	0.062	0.726	0.136	0.440	0.399	0.488	0.344
Cum Var		0.781		0.862		0.839		0.832
Fit		0.990		0.997		0.996		0.996

Table 27, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.591	0.382	0.540	0.398	0.667	0.266
2	0.691	0.380	0.691	0.413	0.705	0.420
3	0.633	0.550	0.699	0.444	0.720	0.424
4	0.613	0.453	0.520	0.512	0.645	0.363
5	0.329	0.606	0.305	0.548	0.342	0.715
6	0.799	0.529	0.746	0.606	0.835	0.488
7	0.695	0.615	0.726	0.595	0.840	0.412
8	0.710	0.675	0.708	0.661	0.834	0.514
9	0.756	0.571	0.728	0.572	0.831	0.483
10	0.776	0.535	0.701	0.611	0.838	0.452
11	0.704	0.644	0.709	0.644	0.855	0.438
12	0.755	0.619	0.750	0.633	0.843	0.496
13	0.808	0.565	0.734	0.653	0.864	0.474
14	0.761	0.589	0.769	0.580	0.826	0.504
15	0.768	0.594	0.736	0.645	0.839	0.509
16	0.681	0.610	0.664	0.610	0.811	0.454
17	0.787	0.540	0.809	0.517	0.794	0.525
18	0.713	0.521	0.672	0.585	0.787	0.433
19	0.739	0.556	0.700	0.535	0.820	0.420
20	0.750	0.621	0.750	0.629	0.849	0.482
Prop Var	0.505	0.317	0.478	0.330	0.618	0.222
Cum Var		0.822		0.808		0.840
Fit		0.996		0.995		0.997

There is increased number of influential indicators on factors, notably Factor 1, over that of a three-point scale. In addition, cumulative variations explained are higher for a five-point scale than a three-point scale. Thus, for two-dimensional datasets, five-point scale yields more enhanced results than lower scales. Table 27 shows that the two factors are substantially influenced by comparable sets of indicator variables for  $n = 150, 200,$  and  $800$ . This occurrence is consistent with number of ability dimensions underlying the scale. For these samples,  $n = 150$  has highest 83.9% of cumulative variation, and as such gives most plausible result.

### **Three-dimensional five-point Likert scale**

In this system, we expect that a plausible three-factor solution to possess either three repeating factors or a single dominant first factor since the same information is contained on three dimensions that underlie item responses. The  $p$ -values of item fitness for three-dimensional GPC item response model for various sample sizes are illustrated in Table 28.

Table 28: *P*-values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Five-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.210	0.448	0.240	0.376	0.181	0.000
2	NaN	0.265	0.143	0.330	0.330	0.391	0.000
3	NaN	0.447	0.847	0.326	0.095	0.046	0.000
4	NaN	0.025	0.943	0.013	0.884	0.638	0.000
5	NaN	0.146	0.466	0.480	0.008	0.825	0.000
6	NaN	NaN	0.029	0.736	0.215	0.406	0.000
7	NaN	0.138	0.355	0.115	0.332	0.248	0.000
8	NaN	0.011	0.141	0.119	0.004	0.350	0.000
9	NaN	0.025	0.123	0.147	0.090	0.194	0.000
10	NaN	0.064	0.064	0.004	0.633	0.668	0.000
11	NaN	0.058	0.303	0.213	0.224	0.136	0.000
12	NaN	0.053	0.006	0.080	0.297	0.356	0.000
13	NaN	0.043	0.244	0.007	0.501	0.448	0.001
14	NaN	0.120	0.070	0.300	0.482	0.065	0.000
15	NA	NaN	0.009	0.509	0.393	0.002	0.001
16	NaN	0.120	0.182	0.062	0.186	0.000	0.000
17	NaN	0.041	0.602	0.299	0.214	0.037	0.000
18	NaN	0.012	0.086	0.398	0.161	0.745	0.000
19	NaN	0.321	0.264	0.487	0.023	0.424	0.000
20	NaN	0.007	0.002	0.011	0.017	0.002	0.016
Model Fit	0.728	0.977	0.997	1.000	0.887	0.997	1.000

Table 28 shows that the fitness of items cannot be determined for all items at  $n = 30$  due to sparseness in the data. Generally, the  $p$ -values of item fitnesses to the model are much reduced for all sample sizes in this scenario. Particularly, for  $n = 1000$ , all items do not fit the data. This is an indication that majority of the items do not fit the higher dimensional IRT model on higher response scales. Surprisingly, the three-dimensional GPC model almost perfectly fits the five-point response data at all sample sizes. We observe that the item response model yields better results on high response scales. Generally, the overall fitness of the IRT model fluctuates with increasing sample size.

Table 29 presents the loadings of three-factor solutions for three dimensional item response datasets for various sample sizes on five-point scale. There is comparable number of indicator variables that influence formation of factors on both three and five-point scales for all sample sizes. However, there is a moderate improvement in the amount of variation explained over a three-point scale. It is worthy of note that indicator variables greatly influence the formation of two factors for lower sample sizes. This incidence is not appealing in terms of expected dimension of the scale. On the contrary, for  $n = 150$  and  $200$ , only Factor 1 considerably dominates the other two. The result for  $n = 150$  is most desirable as it possesses highest 90% cumulative variation. The fitness of the three-factor model on the three-dimensional five-point scale is almost perfect for all sample sizes, and peaks at  $n = 150$ .



Table 29: Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Five-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.784	0.000	0.627	0.608	0.262	0.329
2	0.851	0.413	0.000	0.661	0.404	0.396
3	0.215	0.841	0.000	0.382	0.801	0.292
4	0.285	0.564	0.319	0.747	0.402	0.281
5	0.269	0.460	0.676	0.286	0.240	0.716
6	0.778	0.469	0.319	0.773	0.496	0.396
7	0.719	0.464	0.398	0.674	0.477	0.494
8	0.574	0.680	0.311	0.592	0.617	0.495
9	0.593	0.637	0.374	0.717	0.448	0.488
10	0.683	0.569	0.378	0.675	0.452	0.535
11	0.651	0.628	0.361	0.661	0.547	0.451
12	0.612	0.593	0.444	0.648	0.468	0.544
13	0.615	0.644	0.383	0.689	0.523	0.462
14	0.696	0.583	0.380	0.693	0.497	0.478
15	0.610	0.658	0.425	0.704	0.545	0.433
16	0.576	0.651	0.393	0.622	0.442	0.571
17	0.674	0.551	0.390	0.647	0.453	0.563
18	0.666	0.462	0.372	0.562	0.589	0.441
19	0.674	0.513	0.318	0.625	0.573	0.356
20	0.743	0.467	0.424	0.651	0.508	0.552
Prop Var	0.404	0.320	0.156	0.410	0.251	0.226
Cum Var			0.880			0.887
Fit			0.998			0.998

Table 29, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.382	0.748	0.242	0.690	0.262	0.266
2	0.673	0.387	0.369	0.703	0.257	0.422
3	0.769	0.291	0.342	0.725	0.217	0.372
4	0.615	0.512	0.339	0.403	0.799	0.313
5	0.388	0.261	0.757	0.337	0.271	0.802
6	0.822	0.392	0.348	0.839	0.371	0.300
7	0.801	0.362	0.395	0.836	0.315	0.324
8	0.828	0.431	0.327	0.856	0.362	0.332
9	0.809	0.370	0.368	0.806	0.329	0.370
10	0.796	0.424	0.337	0.810	0.365	0.298
11	0.814	0.422	0.344	0.825	0.390	0.322
12	0.806	0.458	0.356	0.824	0.391	0.341
13	0.786	0.421	0.409	0.817	0.407	0.371
14	0.778	0.467	0.334	0.779	0.413	0.384
15	0.800	0.439	0.382	0.799	0.432	0.362
16	0.757	0.459	0.378	0.780	0.453	0.272
17	0.814	0.371	0.372	0.851	0.341	0.302
18	0.766	0.420	0.356	0.841	0.234	0.298
19	0.755	0.402	0.402	0.833	0.332	0.261
20	0.797	0.432	0.401	0.856	0.360	0.297
Prop Var	0.561	0.188	0.152	0.597	0.147	0.135
Cum Var			0.900			0.880
Fit			0.999			0.998

## Assessment of Seven-Point Likert Scale

In this segment, we present and discuss the results from seven-point Likert scale datasets. On this scale too, 16 datasets have been generated under different conditions such as number of ability dimensions and sample size. For these datasets, we assess the performance of GPC item response model and that of factor analysis. We begin with seven-point response scale with underlying unidimensional person-ability.

### Unidimensional seven-point Likert scale

Table 30 shows the  $p$ -values of fitness of items for unidimensional seven-point scale dataset based on generalised partial credit (GPC) model.

We note from Table 30 that majority of items generally fit the unidimensional GPC model, except for  $n = 30$  where the fitnesses of items could not be determined due to sparseness in the data. The overall GPC model significantly fits the unidimensional seven-point response dataset for sample sizes from 150 to 1000. For small samples ( $n = 30$  and 100), the IRT model misfits the data.

Table 30: *P*-values for Item Fitness for Unidimensional GPC Model for Various Sample Sizes on Seven-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.268	0.277	0.775	0.622	0.363	0.315
2	NaN	0.454	0.536	0.457	0.140	0.863	1.000
3	NaN	0.738	0.446	0.420	0.135	0.475	0.078
4	0.062	0.329	0.183	0.181	0.107	0.008	0.463
5	0.029	0.306	0.161	0.270	0.394	0.254	0.675
6	NaN	0.106	0.514	0.163	0.345	0.220	0.391
7	NaN	0.616	0.243	0.148	0.331	0.742	0.426
8	NaN	0.035	0.372	0.408	0.171	0.436	0.664
9	NaN	0.201	0.728	0.047	0.131	0.536	0.083
10	NaN	0.454	0.246	0.591	0.636	0.186	0.228
11	NaN	0.193	0.265	0.667	0.712	0.066	0.306
12	NaN	0.366	0.520	0.291	0.859	0.737	0.198
13	NaN	0.149	0.425	0.107	0.177	0.052	0.566
14	NaN	0.124	0.026	0.026	0.194	0.702	0.373
15	NaN	0.053	0.340	0.937	0.219	0.195	0.501
16	NaN	0.044	0.085	0.169	0.668	0.910	0.075
17	NaN	0.291	0.101	0.189	0.053	0.797	0.105
18	NaN	0.086	0.003	0.925	0.510	0.105	0.529
19	NaN	0.322	0.023	0.241	0.267	0.089	0.529
20	NaN	0.230	0.040	0.611	0.717	0.407	0.035
Model Fit	0.000	0.038	0.703	0.144	0.533	0.476	0.963

The loadings of corresponding one-factor model at various sample sizes are represented in Table 31. There is increased number of indicators that influence the factor even for lower sample size. For all  $n \geq 150$ , all twenty indicator variables are influential. The proportion of variation accounted for increases from 61.6% (for  $n = 30$ ) to a highest of 79.4% (for  $n = 1000$ ). It is also relevant to note that the amount of information explained remains as high (77.6%) for  $n = 200$  as for  $n = 1000$ . Generally, there is almost a perfect fit of the model for all sample sizes.

Table 31: Loadings of One-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	0.409	0.710	0.675	0.707	0.663	0.683	0.663
2	0.668	0.731	0.778	0.689	0.761	0.776	0.758
3	0.517	0.774	0.737	0.786	0.798	0.792	0.818
4	0.426	0.755	0.662	0.702	0.722	0.724	0.753
5	0.599	0.483	0.616	0.639	0.605	0.577	0.649
6	0.874	0.931	0.931	0.933	0.948	0.938	0.961
7	0.647	0.839	0.911	0.900	0.908	0.915	0.915
8	0.918	0.951	0.957	0.957	0.960	0.966	0.962
9	0.871	0.875	0.926	0.936	0.933	0.939	0.922
10	0.885	0.893	0.895	0.912	0.916	0.914	0.922
11	0.935	0.939	0.908	0.955	0.946	0.942	0.949
12	0.957	0.957	0.959	0.976	0.966	0.970	0.974
13	0.885	0.949	0.960	0.950	0.970	0.972	0.973
14	0.813	0.921	0.934	0.938	0.948	0.955	0.947
15	0.948	0.940	0.969	0.965	0.978	0.966	0.970
16	0.690	0.845	0.893	0.889	0.906	0.915	0.903
17	0.817	0.868	0.909	0.951	0.939	0.945	0.937
18	0.814	0.820	0.843	0.875	0.862	0.871	0.883
19	0.713	0.892	0.878	0.861	0.886	0.886	0.886
20	0.943	0.939	0.951	0.965	0.965	0.967	0.969
Prop Var	0.616	0.736	0.759	0.776	0.784	0.787	0.794
Fit	0.968	0.990	0.992	0.993	0.994	0.994	0.995

Table 32 shows the loadings, proportion and cumulative variations explained, and fitness of two-factor solutions for unidimensional seven-point response data at various sample sizes. Although the number of influential indicators on Factor 1 have not increased over that of five-point scale, those of Factor 2 have shot up for all sample sizes, except  $n = 100$ . Generally, under seven-point scale, there are comparable sets of indicator variables that largely contribute to the formation of both factors. This situation is a contravention of the expected underlying unidimensionality of the scale. Thus, the change in scale (from five to seven-point) seems to disrupt the dimensionality of underlying ability. The only reasonable factor solution for the seven-point scale is the case where  $n = 100$ , with Factor 1 accounting for as high as 64.7% of cumulative variation. An observation of Table 32 shows that there is increased amount of cumulative variation explained over five-point scales for all sample sizes. The fitness of the two-factor model is almost perfect for all sample sizes.

Table 32: Two-Factor Solution for Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.180	0.416	0.689	0.186	0.428	0.531	0.548	0.448
2	0.518	0.421	0.709	0.193	0.668	0.424	0.365	0.629
3	0.290	0.453	0.731	0.255	0.431	0.622	0.468	0.658
4	0.262	0.348	0.652	0.413	0.633	0.292	0.424	0.579
5	0.000	0.833	0.222	0.819	0.329	0.554	0.551	0.343
6	0.675	0.554	0.848	0.388	0.685	0.630	0.706	0.609
7	0.898	0.000	0.786	0.291	0.687	0.599	0.708	0.558
8	0.756	0.530	0.884	0.352	0.656	0.701	0.707	0.645
9	0.547	0.699	0.858	0.209	0.731	0.574	0.681	0.643
10	0.661	0.587	0.846	0.288	0.701	0.560	0.771	0.507
11	0.743	0.570	0.895	0.289	0.701	0.579	0.697	0.654
12	0.796	0.543	0.906	0.310	0.785	0.565	0.725	0.653
13	0.654	0.594	0.906	0.286	0.727	0.627	0.728	0.610
14	0.610	0.536	0.880	0.277	0.619	0.707	0.749	0.569
15	0.758	0.572	0.922	0.221	0.682	0.689	0.768	0.588
16	0.477	0.503	0.780	0.325	0.618	0.647	0.697	0.555
17	0.672	0.472	0.818	0.291	0.727	0.553	0.660	0.689
18	0.603	0.545	0.752	0.328	0.422	0.792	0.548	0.702
19	0.395	0.634	0.829	0.329	0.718	0.516	0.665	0.548
20	0.784	0.535	0.883	0.318	0.727	0.613	0.820	0.533
Prop Var	0.370	0.292	0.647	0.118	0.417	0.357	0.436	0.350
Cum Var		0.662		0.764		0.774		0.787
Fit		0.976		0.993		0.993		0.994



Table 32, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.613	0.279	0.591	0.359	0.363	0.603
2	0.644	0.405	0.573	0.526	0.533	0.544
3	0.655	0.456	0.555	0.573	0.562	0.604
4	0.607	0.392	0.429	0.621	0.620	0.432
5	0.307	0.629	0.301	0.545	0.539	0.367
6	0.798	0.513	0.781	0.528	0.717	0.641
7	0.713	0.565	0.729	0.554	0.746	0.534
8	0.847	0.462	0.726	0.637	0.738	0.617
9	0.785	0.504	0.757	0.558	0.740	0.552
10	0.712	0.582	0.738	0.542	0.749	0.541
11	0.810	0.490	0.773	0.542	0.767	0.563
12	0.747	0.620	0.762	0.601	0.749	0.622
13	0.801	0.546	0.738	0.631	0.743	0.628
14	0.804	0.504	0.719	0.629	0.746	0.584
15	0.804	0.556	0.719	0.647	0.737	0.630
16	0.713	0.563	0.731	0.551	0.679	0.595
17	0.789	0.510	0.730	0.600	0.686	0.641
18	0.729	0.461	0.762	0.446	0.729	0.506
19	0.767	0.449	0.640	0.616	0.693	0.553
20	0.802	0.537	0.756	0.603	0.741	0.624
Prop Var	0.535	0.258	0.472	0.324	0.471	0.329
Cum Var		0.793		0.796		0.800
Fit		0.994		0.995		0.995

Table 33 indicates the loadings of three-factor solutions for unidimensional seven-point response data at various sample sizes. The Table also presents the proportion and cumulative variation explained as well as fitness of the three-factor model.

There is equivalent number of influential indicators on factors for both five and seven-point scales for all sample sizes. Also, there is no improvement in the proportion of variation explained by the first factor over that of five-point scale. There is, however, moderate improvement in the proportion of variation accounted for by the other two factors. Even if there were appreciable increase in the proportion of variation explained by Factors 2 and 3, it could only mean a change in dimensionality of the scale. Lower sample sizes show two dominant factors, which contrasts the underlying ability dimension. However, for  $n = 150$  and 200, only the first factor is substantially influenced by indicator variables. The result for  $n = 150$  is most reasonable as Factor 1 explains highest (46.3%) proportion of cumulative variation. Even though the fitness of the three-factor model is almost perfect, as sample size increases, there is marginal or no improvement in the amount of fitness.

Table 33: Three-Factor Solutions for Unidimensional Datasets for Various Sample Sizes on Seven-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.114	0.172	0.780	0.535	0.441	0.167
2	0.513	0.444	0.000	0.327	0.750	0.160
3	0.285	0.388	0.244	0.619	0.404	0.241
4	0.269	0.333	0.000	0.439	0.498	0.418
5	0.000	0.844	0.221	0.202	0.160	0.758
6	0.680	0.584	0.000	0.656	0.547	0.374
7	0.879	0.000	0.000	0.583	0.535	0.280
8	0.751	0.442	0.312	0.729	0.515	0.333
9	0.546	0.618	0.321	0.655	0.562	0.186
10	0.665	0.632	0.000	0.702	0.487	0.268
11	0.739	0.505	0.272	0.670	0.602	0.270
12	0.790	0.475	0.281	0.641	0.658	0.292
13	0.653	0.553	0.227	0.741	0.535	0.264
14	0.615	0.515	0.156	0.806	0.420	0.251
15	0.756	0.515	0.252	0.762	0.534	0.197
16	0.471	0.460	0.215	0.573	0.539	0.316
17	0.676	0.337	0.375	0.589	0.582	0.273
18	0.602	0.438	0.343	0.779	0.249	0.321
19	0.397	0.534	0.336	0.595	0.593	0.311
20	0.782	0.426	0.352	0.738	0.504	0.296
Prop Var	0.367	0.241	0.089	0.403	0.272	0.104
Cum Var			0.696			0.779
Fit			0.981			0.994

Table 33, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.571	0.206	0.316	0.562	0.340	0.267
2	0.603	0.446	0.243	0.306	0.643	0.316
3	0.577	0.237	0.417	0.582	0.539	0.178
4	0.302	0.716	0.251	0.465	0.475	0.260
5	0.269	0.270	0.672	0.310	0.246	0.685
6	0.667	0.477	0.449	0.724	0.465	0.361
7	0.662	0.475	0.415	0.700	0.425	0.375
8	0.761	0.365	0.460	0.702	0.504	0.410
9	0.798	0.403	0.283	0.674	0.510	0.404
10	0.701	0.436	0.348	0.756	0.348	0.404
11	0.750	0.408	0.320	0.663	0.532	0.442
12	0.771	0.487	0.318	0.730	0.510	0.399
13	0.790	0.417	0.358	0.744	0.466	0.368
14	0.733	0.356	0.466	0.740	0.427	0.393
15	0.772	0.399	0.430	0.734	0.444	0.447
16	0.726	0.345	0.394	0.660	0.430	0.415
17	0.740	0.458	0.290	0.714	0.547	0.325
18	0.673	0.195	0.541	0.652	0.554	0.244
19	0.617	0.517	0.374	0.639	0.423	0.395
20	0.811	0.393	0.326	0.825	0.365	0.389
Prop Var	0.463	0.174	0.157	0.433	0.219	0.150
Cum Var			0.794			0.802
Fit			0.994			0.995

## Two-dimensional seven-point Likert scale

In this scheme, we expect an ideal two-factor solution to possess either two repeating factors or a single dominant first factor since the same information is contained on both dimensions that underlie item responses. Table 34 shows the  $p$ -values of item fitness to the two-dimensional generalised partial credit (GPC) item response model at various sample sizes. The Table also contains the  $p$ -values of the fitness of the overall GPC model for the seven-point response data at various sample sizes.

On the seven-point response scale, there is reduced fitness of most items to the two-dimensional GPC model as compared to the unidimensional case. At lower sample sizes ( $n = 30$  and  $100$ ), the fitness of items has worsened. On the contrary, the fitness of the overall GPC model to the two-dimensional seven-point response data have become much significant. The item response model significantly fits the data at all sample sizes.

Table 34: *P*-values for Item Fitness for Two-Dimensional GPC Model for Various Sample Sizes on Seven-point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.365	0.198	0.550	0.823	0.015	0.508
2	NaN	0.021	0.896	0.298	0.135	0.753	0.311
3	NaN	0.094	0.020	0.422	0.122	0.936	0.436
4	NaN	0.204	0.266	0.168	0.679	0.435	0.530
5	NaN	0.090	0.667	0.119	0.292	0.004	0.575
6	NaN	0.130	0.116	0.060	0.694	0.039	0.010
7	NaN	0.009	0.198	0.268	0.346	0.587	0.554
8	NaN	0.023	0.038	0.476	0.421	0.412	0.150
9	NaN	0.016	0.060	0.629	0.254	0.392	0.072
10	NaN	0.098	0.017	0.356	0.130	0.427	0.051
11	NaN	0.003	0.030	0.182	0.054	0.881	0.813
12	NaN	0.009	0.025	0.624	0.169	0.786	0.306
13	NaN	0.002	0.131	0.413	0.153	0.185	0.232
14	NaN	0.210	0.068	0.051	0.138	0.706	0.536
15	NaN	NaN	0.083	0.101	0.271	0.576	0.479
16	NaN	0.344	0.258	0.209	0.018	0.301	0.547
17	NaN	0.042	0.082	0.009	0.771	0.081	0.252
18	NaN	0.169	0.248	0.174	0.055	0.222	0.245
19	NaN	0.085	0.238	0.541	0.560	0.285	0.496
20	NaN	NaN	0.036	0.295	0.413	0.075	0.073
Model Fit	0.842	0.854	0.979	0.930	0.252	0.994	0.667

Table 35 shows corresponding two-factor solutions for the two-dimensional seven-point response data at various sample sizes.

There is a similar number of indicator variables that highly influence the formation of factors, especially Factor 1, for both five and seven-point scales. In addition, there is no substantial increase in the proportion of variation explained by the first factor. In this case, seven-point scale has no substantial information over a five-point scale for two-dimensional datasets. Two factors are highly dominated by similar sets of influential indicators for  $n = 150, 200$  and  $500$ . For these samples,  $n = 150$  provides most desirable result as it accounts for largest cumulative variation (87.5%). The two-factor model is nearly perfectly significant, and remains the same at all sample sizes except for  $n = 30$ .

Table 35: Two-Factor Solutions for Two-Dimensional Datasets for Various Sample Sizes on Seven-Point Scale

Item	Sample Size 30		Sample Size 100		Sample Size 150		Sample Size 200	
	PA1	PA2	PA1	PA2	PA1	PA2	PA1	PA2
1	0.659	0.000	0.764	0.347	0.324	0.723	0.325	0.703
2	0.597	0.000	0.743	0.462	0.742	0.432	0.600	0.603
3	0.945	0.000	0.784	0.311	0.685	0.547	0.652	0.612
4	0.753	0.000	0.738	0.341	0.773	0.346	0.688	0.481
5	0.147	1.101	0.343	0.879	0.635	0.480	0.659	0.305
6	0.902	0.163	0.906	0.360	0.742	0.632	0.656	0.712
7	0.953	0.118	0.883	0.337	0.727	0.614	0.768	0.574
8	0.930	0.138	0.894	0.405	0.685	0.685	0.758	0.628
9	0.892	0.184	0.904	0.358	0.725	0.639	0.706	0.654
10	0.920	0.280	0.832	0.449	0.760	0.568	0.693	0.661
11	0.883	0.176	0.890	0.371	0.712	0.656	0.716	0.642
12	0.900	0.253	0.893	0.398	0.737	0.654	0.671	0.722
13	0.959	0.204	0.892	0.417	0.745	0.647	0.750	0.647
14	0.903	0.238	0.879	0.385	0.746	0.638	0.726	0.652
15	0.902	0.247	0.901	0.398	0.804	0.561	0.682	0.715
16	0.870	0.130	0.858	0.432	0.688	0.624	0.662	0.675
17	0.879	0.176	0.902	0.368	0.764	0.609	0.632	0.725
18	0.824	0.214	0.851	0.391	0.616	0.734	0.722	0.590
19	0.905	0.221	0.849	0.399	0.659	0.654	0.707	0.633
20	0.941	0.285	0.866	0.465	0.794	0.587	0.713	0.675
Prop Var	0.728	0.093	0.702	0.184	0.504	0.370	0.463	0.406
Cum Var		0.821		0.886		0.875		0.869
Fit		0.994		0.998		0.998		0.998



Table 35, continued

Item	Sample Size 500		Sample Size 800		Sample Size 1000	
	PA1	PA2	PA1	PA2	PA1	PA2
1	0.624	0.476	0.648	0.384	0.711	0.342
2	0.712	0.470	0.782	0.356	0.766	0.428
3	0.597	0.653	0.784	0.389	0.781	0.422
4	0.648	0.525	0.696	0.404	0.715	0.391
5	0.404	0.643	0.379	0.734	0.417	0.766
6	0.779	0.581	0.859	0.456	0.842	0.493
7	0.689	0.647	0.850	0.440	0.851	0.429
8	0.671	0.722	0.846	0.486	0.840	0.518
9	0.744	0.608	0.847	0.424	0.846	0.478
10	0.733	0.618	0.833	0.457	0.849	0.466
11	0.682	0.687	0.830	0.506	0.863	0.452
12	0.725	0.663	0.852	0.502	0.857	0.484
13	0.763	0.632	0.843	0.514	0.866	0.484
14	0.724	0.648	0.861	0.454	0.845	0.493
15	0.709	0.680	0.848	0.506	0.844	0.514
16	0.679	0.650	0.803	0.477	0.839	0.443
17	0.723	0.639	0.876	0.413	0.823	0.509
18	0.707	0.588	0.815	0.440	0.823	0.442
19	0.689	0.647	0.814	0.431	0.840	0.428
20	0.731	0.653	0.860	0.482	0.856	0.484
Prop Var	0.477	0.390	0.646	0.220	0.656	0.230
Cum Var		0.868		0.866		0.886
Fit		0.998		0.998		0.998

### **Three-dimensional seven-point Likert scale**

In this system, we expect that a plausible three-factor solution should possess either three repeating factors or a single dominant first factor since the same information is contained on three dimensions that underlie item responses. The  $p$ -values of item fitness for three-dimensional GPC item response model for various sample sizes are illustrated in Table 36.

There is further deterioration of the fitness of items to three-dimensional GPC model as compared to two-dimensional case on the seven-point scale. Meanwhile, the item response model significantly fits the three-dimensional seven-point response data for all sample sizes. The fitness of the IRT model largely fluctuates with increasing sample size.

Table 36: *P*-values for Item Fitness for Three-Dimensional GPC Model for Various Sample Sizes on Seven-Point Scale

Item	Sample Size						
	30	100	150	200	500	800	1000
1	NaN	0.300	0.094	0.018	0.083	0.570	0.000
2	NaN	0.031	0.644	0.024	0.100	0.605	0.000
3	NaN	0.068	0.501	0.000	0.175	0.075	0.000
4	NaN	0.113	0.105	0.054	0.323	0.782	0.000
5	NaN	0.032	0.749	0.736	0.030	0.703	0.000
6	NaN	NaN	0.001	0.060	0.612	0.025	0.000
7	NaN	0.000	0.030	0.198	0.384	0.026	0.000
8	NaN	NaN	0.029	0.040	0.713	0.084	0.000
9	NaN	NaN	0.157	0.081	0.068	0.646	0.000
10	NaN	NaN	0.067	0.013	0.046	0.785	0.000
11	NaN	0.000	0.183	0.079	0.037	0.843	0.000
12	NaN	NaN	NaN	0.080	0.950	0.103	0.000
13	NaN	NaN	0.042	0.021	0.184	0.074	0.000
14	NaN	NaN	0.001	0.183	0.087	0.050	0.000
15	NA	NaN	0.095	0.080	0.406	0.159	0.000
16	NaN	0.020	0.013	0.160	0.011	0.135	0.000
17	NaN	0.002	0.084	0.504	0.569	0.389	0.000
18	NaN	0.013	0.009	0.742	0.152	0.569	0.000
19	NaN	0.048	0.357	0.007	0.149	0.525	0.000
20	NaN	0.001	0.006	0.050	0.007	0.001	0.000
Model Fit	0.834	0.594	0.437	0.489	0.981	0.997	1.000

Table 37 illustrates the loadings of three-factor solutions for three dimensional response datasets for various sample sizes on seven-point scale. There are comparable sets of indicator variables that significantly contribute to the formation of Factor 1 for both five and seven-point scales. Even though there is increased number of indicators on other two factors, the corresponding cumulative variation explained have increased only marginally. To this end, seven-point scale has negligible amount of information over a five-point scale. There appears to be three dominant factors for all sample sizes except  $n = 200$ . Factors 2 and 3 explain equivalent proportions of variation for all sample sizes. In any case, amount of cumulative variation explained increases from 87.8% (for  $n = 30$ ) up to 91.6% (for  $n = 150$ ), and decreases afterwards. Thus,  $n = 150$  offers most credible result.

Table 37: Three-Factor Solutions for Three-Dimensional Datasets for Various Sample Sizes on Seven-Point Scale

Item	Sample Size 30			Sample Size 100		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.277	0.887	0.288	0.339	0.603	0.431
2	0.747	0.437	0.261	0.490	0.632	0.404
3	0.687	0.168	0.363	0.741	0.414	0.353
4	0.528	0.202	0.514	0.496	0.714	0.311
5	0.340	0.422	0.664	0.314	0.300	0.743
6	0.718	0.560	0.334	0.588	0.669	0.445
7	0.637	0.538	0.436	0.611	0.563	0.495
8	0.743	0.344	0.482	0.661	0.567	0.459
9	0.709	0.452	0.447	0.570	0.598	0.517
10	0.678	0.492	0.482	0.580	0.577	0.527
11	0.706	0.456	0.495	0.635	0.581	0.458
12	0.674	0.517	0.462	0.586	0.589	0.509
13	0.691	0.429	0.538	0.603	0.621	0.466
14	0.678	0.518	0.492	0.583	0.594	0.515
15	0.660	0.436	0.613	0.639	0.617	0.438
16	0.670	0.449	0.533	0.574	0.538	0.542
17	0.702	0.499	0.415	0.580	0.546	0.564
18	0.642	0.517	0.370	0.713	0.454	0.441
19	0.646	0.503	0.403	0.639	0.577	0.375
20	0.694	0.577	0.394	0.640	0.550	0.524
Prop Var	0.426	0.241	0.212	0.346	0.327	0.234
Cum Var			0.878			0.907
Fit			0.998			0.999

Table 37, continued

Item	Sample Size 150			Sample Size 200		
	PA1	PA2	PA3	PA1	PA2	PA3
1	0.392	0.706	0.346	0.709	0.314	0.304
2	0.586	0.479	0.463	0.710	0.346	0.395
3	0.794	0.362	0.350	0.756	0.226	0.404
4	0.572	0.570	0.405	0.462	0.760	0.355
5	0.411	0.380	0.630	0.388	0.301	0.796
6	0.714	0.508	0.435	0.810	0.416	0.347
7	0.695	0.464	0.492	0.818	0.374	0.336
8	0.725	0.516	0.436	0.824	0.379	0.395
9	0.671	0.491	0.503	0.787	0.393	0.387
10	0.665	0.529	0.472	0.794	0.386	0.361
11	0.716	0.518	0.432	0.812	0.422	0.338
12	0.683	0.544	0.477	0.806	0.413	0.377
13	0.681	0.501	0.508	0.784	0.430	0.420
14	0.671	0.550	0.436	0.767	0.445	0.393
15	0.688	0.524	0.482	0.785	0.446	0.387
16	0.662	0.518	0.485	0.752	0.468	0.342
17	0.710	0.451	0.487	0.823	0.381	0.345
18	0.691	0.491	0.440	0.819	0.320	0.333
19	0.659	0.469	0.526	0.803	0.401	0.324
20	0.667	0.532	0.513	0.820	0.408	0.349
Prop Var	0.435	0.260	0.221	0.578	0.171	0.158
Cum Var			0.916			0.907
Fit			0.999			0.999

## **Comparison of Results of Various Response Scales and Sample Sizes**

In this section, we carry out a comparison of IRT results across four different response scales and sample sizes based on the dimensionality of datasets. We also compare the results of factor solutions under different dimensions of the underlying latent ability. For each dimension, a comparison of factor solutions is done at various response scales and sample sizes.

### **IRT results across different scales and sample sizes**

We assess IRT results under various conditions such as the number of points on response scales, number of ability dimensions, and sample size. Table 38 illustrates summary statistics for IRT results across different response scales with different number of dimensions underlying datasets.

We note from Table 38 that on unidimensional dichotomous response scale overwhelming majority of items fit the IRT model for all samples. The corresponding overall fitness of the IRT model is significant at all sample sizes. This result is similar to the two-dimensional case, except for  $n = 30$  where the fitness of the model could not be determined due to low degrees of freedom. On three-dimensional dichotomous response data, the number of fit items decreases sharply for  $n = 30$ , but marginally at higher samples. We observe that on dichotomous scale, the fitness of the model generally fluctuates as the number of underlying dimensions increase.

Table 38: Summary Statistics for IRT Results Across Different Scales on Varied Dimensions

Scale	Measures	Sample Size							
		30	100	150	200	500	800	1000	
Two-Point	Unidim.	Fit Items	15	19	18	19	20	20	19
		Model Fit	0.114	0.966	0.514	0.381	0.363	0.387	0.938
	Two-dim.	Fit Items	4	18	19	19	20	20	18
		Model Fit	-	0.920	0.406	0.249	0.953	0.974	0.123
	Three-dim.	Fit Items	1	15	17	18	19	19	17
		Model Fit	-	0.882	0.462	0.885	0.556	0.783	0.371
Three-Point	Unidim.	Fit Items	15	20	19	19	20	18	19
		Model Fit	0.002	0.477	0.198	0.589	0.581	0.603	0.911
	Two-dim.	Fit Items	6	17	19	18	18	20	19
		Model Fit	0.006	0.469	0.969	0.482	0.872	0.924	0.779
	Three-dim.	Fit Items	1	14	15	20	16	18	18
		Model Fit	0.830	0.623	0.717	1.000	0.948	0.998	0.766
Five-Point	Unidim.	Fit Items	6	16	18	16	20	19	19
		Model Fit	0.007	0.253	0.368	0.809	0.435	0.150	0.990
	Two-dim.	Fit Items	3	17	18	17	19	18	18
		Model Fit	0.550	0.896	0.956	0.579	0.326	0.969	0.864
	Three-dim.	Fit Items	0	11	16	16	16	15	0
		Model Fit	0.728	0.977	0.997	1.000	0.887	0.997	1.000
Seven-Point	Unidim.	Fit Items	1	18	16	18	20	18	19
		Model Fit	0.000	0.038	0.703	0.144	0.533	0.476	0.963
	Two-dim.	Fit Items	0	10	14	19	19	17	19
		Model Fit	0.842	0.854	0.979	0.930	0.252	0.994	0.667
	Three-dim.	Fit Items	0	3	12	13	15	17	0
		Model Fit	0.834	0.594	0.437	0.489	0.981	0.997	1.000



On unidimensional three-point response scale, almost all items fit the item response model at all sample sizes, except  $n = 30$ . The model significantly fits the unidimensional three-point response data at various samples, except  $n = 30$ . Similar result holds for two-dimensional three-point response scale. With respect to three-dimensional three-point response scale, the number of fit items declines at all samples, except for  $n = 200$ . The overall fitness of the IRT model on three dimensions improves considerably over that of two dimensions. This may be attributed to the increase in the number of dimensions.

In respect of unidimensional five-point response scale, overwhelming majority of items fit the model at sample size of 100 and beyond. For these samples, the overall fitness of the item response model is significant. With additional dimension to the scale, the number of fit items remains fairly comparable, but with a gain in overall fitness of the model. In the case of three-dimensional five-point scale, the number of fit items decreases with further gain in overall fitness of the model.

There is greater number of items that significantly fit the item response model on the unidimensional seven-point response scale at all samples, except  $n = 30$ . In the two-dimensional case, the number of fit items is virtually the same as that of unidimensional, except for  $n = 30$  and 100. With three dimensions, the number of fit items decreases slightly, though the overall fitness of the model continues to improve.

We now dwell on the effects of response scales and sample size on IRT results for specific dimensionality of datasets. Table 39 shows IRT model summary statistics for various response scales and sample sizes on unidimensional datasets.

Table 39: IRT Model Summary Statistics on Unidimensional Datasets for Various Response Scales and Sample Sizes

Scale	Measures	Sample Size						
		30	100	150	200	500	800	1000
Two-Point	Fit Items	15	19	18	19	20	20	19
	Model Fit	0.114	0.966	0.514	0.381	0.363	0.387	0.938
Three-Point	Fit Items	15	20	19	19	20	18	19
	Model Fit	0.002	0.477	0.198	0.589	0.581	0.603	0.911
Five-Point	Fit Items	6	16	18	16	20	19	19
	Model Fit	0.007	0.253	0.368	0.809	0.435	0.150	0.990
Seven-Point	Fit Items	1	18	16	18	20	18	19
	Model Fit	0.000	0.038	0.703	0.144	0.533	0.476	0.963

Table 39 indicates that on unidimensionality across various scales, overall fitness of item response model deteriorates with increasing scale points for small samples, particularly at  $n = 30$  and  $100$ . This suggests that on unidimensional higher response scales, samples of sizes  $100$  and below would not produce reliable results. Meanwhile, too large a sample, particularly  $n = 1000$ , may produce almost the same IRT results since the performance of the model does not change for all response scales. In some instances on polytomous response scales with one dimension, overall fitness of the IRT model increases with increasing points on the scale.

Table 40 displays item response model summary statistics for various response scales and sample sizes on two-dimensional datasets.

Table 40: IRT Model Summary Statistics on Two-Dimensional Datasets for Various Response Scales and Sample Sizes

Scale	Measures	Sample Size						
		30	100	150	200	500	800	1000
Two-point	Fit Items	4	18	19	19	20	20	18
	Model Fit	-	0.920	0.406	0.249	0.953	0.974	0.123
Three-point	Fit Items	6	17	19	18	18	20	19
	Model Fit	0.006	0.469	0.969	0.482	0.872	0.924	0.779
Five-point	Fit Items	3	17	18	17	19	18	18
	Model Fit	0.550	0.896	0.956	0.579	0.326	0.969	0.864
Seven-point	Fit Items	0	10	14	19	19	17	19
	Model Fit	0.842	0.854	0.979	0.930	0.252	0.994	0.667

We observe from Table 40 that in respect of two-dimensional response scales, the number of fit items is almost the same at sample sizes of 150 and beyond. For some of these samples, the overall fitness of the model increases with increasing points on the scale.

Table 41 presents item response model summary statistics for various response scales and sample sizes on three-dimensional datasets.

Table 41: IRT Model Summary Statistics on Three-Dimensional Datasets for Various Response Scales and Sample Sizes

Scale	Measures	Sample Size						
		30	100	150	200	500	800	1000
Two-point	Fit Items	1	15	17	18	19	19	17
	Model Fit	-	0.882	0.462	0.885	0.556	0.783	0.371
Three-point	Fit Items	1	14	15	20	16	18	18
	Model Fit	0.830	0.623	0.717	1.000	0.948	0.998	0.766
Five-point	Fit Items	0	11	16	16	16	15	0
	Model Fit	0.728	0.977	0.997	1.000	0.887	0.997	1.000
Seven-point	Fit Items	0	3	12	13	15	17	0
	Model Fit	0.834	0.594	0.437	0.489	0.981	0.997	1.000

There is almost the same number of fit items at sample sizes of 150 and beyond across all response scales with three dimensions. For polytomous response scales with three dimensions, model fitness seems to be quite high at larger samples, particularly for three and five-point scales. However, there does not appear to be any relationship between the model fitness and the number of fit items.

In summary, the overall fitness of the IRT model increases, in some cases, with increasing number of points on the scale. For a given response scale, the fitness of the model generally fluctuates with increasing sample size.

### One-factor solutions on unidimensional datasets for various scales

Table 42 presents summary statistics for one-factor solutions on unidimensional datasets at various response scales and sample sizes.

Table 42: Summary Statistics for One-Factor Solutions on Unidimensional Datasets

Scale	Measures	Sample Size						
		30	100	150	200	500	800	1000
Two-Point	No. of Ind.	11	14	15	13	15	15	15
	Cum. Var	0.339	0.424	0.450	0.429	0.455	0.461	0.446
	Fit	0.706	0.870	0.900	0.880	0.916	0.922	0.924
Three-Point	No. of Ind.	16	18	17	17	17	17	18
	Cum. Var	0.507	0.558	0.577	0.596	0.588	0.589	0.597
	Fit	0.900	0.953	0.962	0.968	0.968	0.969	0.971
Five-Point	No. of Ind.	17	19	20	20	20	19	20
	Cum. Var	0.580	0.678	0.697	0.714	0.717	0.720	0.729
	Fit	0.949	0.982	0.985	0.987	0.988	0.988	0.990
Seven-Point	No. of Ind.	18	19	20	20	20	20	20
	Cum. Var	0.616	0.736	0.759	0.776	0.784	0.787	0.794
	Fit	0.706	0.968	0.990	0.992	0.993	0.994	0.995

It is evident in Table 42 that the number of influential indicators on the factor increases as points on the scale increase across all sample sizes. In addition, the dominant number of indicators starts from  $n = 150$  in each case. The cumulative variation (Cum. Var) accounted for by the factor peaks at  $n = 150$  and fluctuates thereafter for two-point scales. In the others, the amount of cumulative variation is almost the same for  $n \geq 150$  within rounding errors. Similarly, the significance of the fit of the model also increases with increasing scale points and sample size. This result is consistent with what is observed in the IRT analysis.

## Two-factor solutions on unidimensional datasets for various scales

The summary statistics for two-factor solutions at different scale-points on unidimensional datasets are displayed in Table 43. The Table also shows summary statistics for various sample sizes.

Table 43: Summary Statistics for Two-Factor Solutions on Unidimensional Datasets

Scale	Measures	Factors	Sample Size						
			30	100	150	200	500	800	1000
Two-Point	No. of Ind.	PA1	10	10	15	12	13	15	15
		PA2	8	8	1	6	6	0	4
	Cum. Var Fit		0.339	0.424	0.450	0.429	0.455	0.461	0.446
			0.706	0.870	0.900	0.880	0.916	0.922	0.924
Three-Point	No. of Ind.	PA1	14	14	17	13	13	15	15
		PA2	5	9	1	13	14	4	13
	Cum. Var Fit		0.596	0.599	0.620	0.618	0.605	0.603	0.606
			0.937	0.962	0.968	0.971	0.971	0.972	0.973
Five-Point	No. of Ind.	PA1	14	16	18	16	19	18	19
		PA2	9	14	6	16	1	10	2
	Cum. Var Fit		0.645	0.704	0.732	0.739	0.739	0.731	0.740
			0.966	0.985	0.987	0.989	0.990	0.990	0.991
Seven-Point	No. of Ind.	PA1	14	19	16	17	19	18	19
		PA2	14	1	18	18	12	18	18
	Cum. Var Fit		0.662	0.764	0.774	0.787	0.793	0.796	0.880
			0.976	0.993	0.993	0.994	0.994	0.995	0.995

We notice in Table 43 that the desired factor structure is observed at  $n = 150$  where there is the highest number of influential indicators and only one on

the second. However, at higher scale-points, the structure occurs at a small sample size of  $n = 100$ . It is notable that at the desire factor structure, the cumulative variation peaks and deteriorates thereafter. This is true either for the cumulative variation or for the value of fitness of the model.

### **Three-factor solutions on unidimensional datasets for various scales**

Table 44 presents the summary statistics for three-factor solutions for various scale-points on unidimensional datasets. We consider various statistics for sample sizes of 30, 100, 150 and 200. Higher sample sizes are ignored as their results do not show improvement over lower sample sizes.

Generally, we find results for  $n = 150$  to be consistent with underlying dimensionality of the data. It gives the first factor as the dominant one and the other two are just much fewer-indicator factor (or none) that contribute marginally to the cumulative proportion of variation explained. Again, at this sample size, cumulative variation (or the fitness) peaks and deteriorates thereafter.

Table 44: Summary Statistics for Three-Factor Solutions on  
Unidimensional Datasets

Scale	Measures	Factors	Sample Size			
			30	100	150	200
Two-Point	No. of Ind.	PA1	6	10	10	10
		PA2	5	8	7	6
		PA3	5	1	0	1
	Cum. Var		0.544	0.530	0.572	0.563
	Fit		0.864	0.916	0.937	0.918
Three-Point	No. of Ind.	PA1	13	14	14	10
		PA2	3	7	4	5
		PA3	2	1	1	4
	Cum. Var		0.669	0.639	0.637	0.641
	Fit		0.956	0.971	0.973	0.975
Five-Point	No. of Ind.	PA1	14	16	18	14
		PA2	8	12	2	12
		PA3	1	1	1	3
	Cum. Var		0.694	0.732	0.737	0.745
	Fit		0.975	0.989	0.989	0.990
Seven-Point	No. of Ind.	PA1	14	17	18	17
		PA2	9	13	2	8
		PA3	1	1	2	1
	Cum. Var		0.696	0.779	0.794	0.802
	Fit		0.981	0.994	0.994	0.995



## **Two-factor solutions on two-dimensional datasets for various scales**

Two-dimensional datasets are generated by specifying the same vector of item discrimination parameter values on both dimensions of the underlying ability. In this system, we expect that a good factor solution should have two repeating factors since the same information is contained on the two underlying dimensions of the dataset. Alternatively, we could expect a single dominant first factor in the two-factor solution with similar reasoning as in the former instance. Here, we compare the results of two-factor solutions on two-dimensional datasets at various sample sizes and scale-points. The results are summarised in Table 45.

Table 45: Summary Statistics for Two-Factor Solutions on Two-Dimensional Datasets

Scale	Measures	Factors	Sample Size						
			30	100	150	200	500	800	1000
Two-Point	No. of Ind.	PA1	12	18	14	16	17	15	16
		PA2	4	1	13	8	1	6	1
	Cum. Var		0.522	0.677	0.639	0.651	0.486	0.492	0.478
	Fit		0.873	0.973	0.965	0.971	0.925	0.934	0.930
Three-Point	No. of Ind.	PA1	16	19	15	18	16	17	19
		PA2	13	3	14	2	15	14	1
	Cum. Var		0.714	0.763	0.749	0.742	0.723	0.700	0.747
	Fit		0.974	0.990	0.989	0.989	0.988	0.986	0.991
Five-Point	No. of Ind.	PA1	19	19	19	19	19	19	19
		PA2	1	1	17	15	17	17	5
	Cum. Var		0.781	0.862	0.839	0.832	0.822	0.808	0.840
	Fit		0.990	0.997	0.996	0.996	0.996	0.995	0.997
Seven-Point	No. of Ind.	PA1	19	19	19	19	19	19	19
		PA2	1	1	17	18	18	5	4
	Cum. Var		0.821	0.886	0.875	0.869	0.868	0.866	0.886
	Fit		0.994	0.998	0.998	0.998	0.998	0.998	0.998

Table 45 shows that the cumulative variation and fitness of the model peak at  $n = 100$  for all scale points, and deteriorates or fluctuates thereafter. The desired factor structure is thus obtained at  $n = 100$ . It is also observed that the amount of cumulative variation explained by the model increases with increasing scale-point. The fitness of the model as well as the number of significant indicators are also generally high at higher scale-points.

### **Three-factor solutions on three-dimensional datasets for various scales**

Table 46 displays the summary statistics for three-factor solutions at various samples and scale-points on three-dimensional datasets. Since the results do not show improvement for higher sample sizes, the results for  $n = 500, 800$  and  $1000$  are excluded.

Generally, a sample size of 150 produces a more consistent factor solution based on the underlying dimensionality of the data. At this sample size, cumulative variation and/or model fitness peaks and deteriorates thereafter. The amount of cumulative variation explained increases with increasing scale-points. It follows that the number of influential indicators on factors increases with increasing scale-points, which is particularly true for the first factor. This means that factors are more well defined and could be more interpretable on larger scale-points. The results are, however, almost the same on higher scale-points of five and seven.

Table 46: Summary Statistics for Three-Factor Solutions on Three-Dimensional Datasets

Scale	Measures	Factors	Sample Size			
			30	100	150	200
Two-Point	No. of Ind.	PA1	12	13	17	16
		PA2	10	13	2	4
		PA3	2	3	1	1
	Cum. Var		0.746	0.753	0.789	0.702
	Fit		0.973	0.986	0.990	0.983
Three-Point	No. of Ind.	PA1	17	18	18	18
		PA2	8	12	1	2
		PA3	2	1	1	2
	Cum. Var		0.818	0.835	0.864	0.792
	Fit		0.995	0.995	0.997	0.994
Five-Point	No. of Ind.	PA1	17	18	18	18
		PA2	13	8	2	1
		PA3	1	6	1	0
	Cum. Var		0.880	0.887	0.900	0.880
	Fit		0.998	0.998	0.999	0.998
Seven-Point	No. of Ind.	PA1	18	16	18	18
		PA2	8	17	12	1
		PA3	5	8	5	1
	Cum. Var		0.878	0.907	0.916	0.907
	Fit		0.998	0.999	0.999	0.999

## Chapter Summary

The chapter investigated the effects of measurement scales on results of item response theory and factor analysis models. This was done through the analyses of simulated datasets under various conditions such as item response format, number of ability dimensions underlying response scales, and sample size.

The study reveals that estimated values of discrimination ( $\hat{\alpha}$ ) and difficulty ( $\hat{\delta}$ ) parameters generally fluctuates with increasing sample size. There is a marked difference between the specified and estimated item parameter values at lower samples ( $n = 30$  and  $100$ ), but the difference tends to reduce at sample sizes of  $150$  and beyond. In addition, the differences become negligible at larger samples ( $n = 500, 800, 1000$ ). This result is consistent with what Stone (1992) observed.

The study shows that there is a direct relationship between parameters of IRT and those of factor models, particularly item discrimination and factor loadings. This result is consistent with what has been found by de Leeuw (1983) and Takane and de Leeuw (1987).

The study shows that items fit the unidimensional 2PL model since the  $p$ -values are generally much higher than  $0.05$ . Only at  $n = 150, 200$  and  $1000$  it is detected that three items (6, 12, and 7, respectively) do not fit the model. The 2PL model significantly fits the unidimensional dichotomous item response data for all sample sizes.

We observe that almost all items significantly fit two-dimensional IRT model. The only exception is when  $n = 30$  where the fitness of majority of items have not been possible to evaluate due to sparseness in the data. Correspondingly, the overall model fitness could not be determined due to low degrees of freedom. The item response model significantly fits the two-dimensional di-

chotomous response data for all other sample sizes.

The study indicates that items significantly fit the three-dimensional 2PL model. However, for smaller samples the fitness of items is appalling. The fitness of items to the model get better as sample size increases. We observe that the three-dimensional 2PL model significantly fits the data for all sample sizes, except at  $n = 30$  where the fitness of the model could not be determined due low degrees of freedom.

We notice that in migrating from dichotomous to three-point scale, the significance of fitness of items to the model generally fluctuates for all sample sizes. While the  $p$ -values of fitness of some items improve for given sample size, those of other items deteriorate. The overall fitness of GPC model to unidimensional three-point response data is significant at sample sizes of 100 and over.

We observe that, generally, items significantly fit the two-dimensional GPC model on three-point scale. However, on this scale, the magnitude of  $p$ -values differ from unidimensional to two-dimensional case. There appears to be two groups of items: (1) items whose  $p$ -value decreases with an additional ability dimension, and (2) items whose  $p$ -value increases with additional ability dimension. Items of the first group will require just one person-ability to get a response in higher categories, whereas those of the second group need multiple person-abilities to get a similar response. The overall fitness of the model to the two-dimensional three-point response data is significant at all sample sizes, except for  $n = 30$ .

On three-dimensional response datasets, most items fit the model, particularly for larger sample sizes ( $n \geq 150$ ). For smaller sample sizes ( $n \leq 100$ ), the  $p$ -values of items worsened on high dimensions, especially for items that may require only one person-ability to get a response in higher categories. The results indicates that items whose fitness is quite high or almost perfect may require up to three person-abilities to get a response in higher categories. At  $n = 30$ , even

though items misfit the model, the overall fitness of the model is significant. The plausibility is that the IRT model yields better results on high response scales with large number of dimensions.

Considering unidimensional five-point scale, we detect that overwhelming majority of items significantly fit the unidimensional GPC model at all sample sizes, except  $n = 30$  where the  $p$ -values of some items could not be determined. Not unexpectedly, the overall GPC model significantly fits the five-point response dataset, with the exception of  $n = 30$ . As sample size increase from 100 up to 1000, the fitness of the item response model fluctuates, but highest at  $n = 1000$ .

On two-dimensional five-point scale, it indicates that most items significantly fit the two-dimensional GPC model at sample size of 100 and beyond. The fitnesses of almost all items could not be evaluated at  $n = 30$  due to low degrees of freedom. Surprisingly, the IRT model significantly fits the data at  $n = 30$ . The overall fitness of the IRT model generally fluctuates with increasing sample size.

In the case of three-dimensional five-point scale, it shows that the fitness of items cannot be determined for all items at  $n = 30$  due to sparseness in the data. Generally, the  $p$ -values of item fitnesses to the model are much reduced for all sample sizes in this scenario. Particularly, for  $n = 1000$ , all items do not fit the data. Surprisingly, the three-dimensional GPC model almost perfectly fits the five-point response data at all sample sizes. We observe that the item response model yields better results on high response scales. Generally, the overall fitness of the IRT model fluctuates with increasing sample size. This results are consistent with those of Beauducel and Herzberg (2006).

In terms of unidimensional seven-point scale, we note that majority of items generally fit the unidimensional GPC model, except for  $n = 30$  where the fitnesses of items could not be determined due sparseness in the data. The overall

GPC model significantly fits the unidimensional seven-point response dataset for sample sizes from 150 to 1000.

On the seven-point response scale, there is reduced fitness of most items to the two-dimensional GPC model as compared to the unidimensional case. However, the fitness of the overall GPC model to the two-dimensional seven-point response data have become much significant.

The IRT analysis of three-dimensional seven-point scale indicates that there is further deterioration of the fitness of items to the three-dimensional GPC model as compared to the two-dimensional case. Meanwhile, the IRT model significantly fits the three-dimensional seven-point response data for all sample sizes. The fitness of the IRT model largely fluctuates with increasing sample size.

We observe that, for one-factor solutions of unidimensional dichotomous scale, the number of influential indicators appear to converge (at 15) for higher sample size starting at  $n = 150$ . The indicators are the same at point of convergence. The proportion of variance accounted for by the single factor increases from 33.9% (for  $n = 30$ ) to a highest of 46.1% (for  $n = 800$ ).

With two-factor solutions of unidimensional dichotomous scale, there is generally the incidence of repetition of high loadings on the same indicator variable of the two factors, with exception of sample sizes  $n = 150$  and  $n = 800$ , which can distract interpretation. However, for  $n = 150$ , the first factor loads highly on as many as 15 indicators and explains 42.7% of variation. The second factor loads highly on only one indicator (Variable 5) and is a contrast to its representation in IRT. In addition, amount of variance explained by the second factor appears to be negligible. The sample size of  $n = 150$  thus gives a more plausible factor solution than all other samples. The  $n = 150$  also explains the highest cumulative variation.

In the case of three-factor solutions of unidimensional dichotomous scale,



the result becomes less meaningful and even unrealistic for sample sizes beyond 30. There is generally the incidence of repeating indicators on multiple factors. There is also the incidence of unrealistic loadings that are greater than one in higher factor numbers, particularly for Factor 3. This incidence is as a result of an extraction of higher factor structure from a lesser dimensional dataset.

With two-factor solutions of two-dimensional dichotomous scale, the  $n = 150$  generates a repeating factor consistent with two repeating dimensions underlying the dataset. The cumulative variation is also highest for this sample size.

Three-factor solutions of three-dimensional dichotomous scale indicates that lower sample sizes show two dominant factors in the result. This pattern is inconsistent with the expected dimension of the scale. However, for  $n = 150$  and 200, as expected only the first factor is highly influenced by the indicator variables. The factor solution for  $n = 150$  is more conceivable as it accounts for as high as 78.9% cumulative variation. The fitness of the three-factor model is almost perfect for all sample sizes.

The study shows that, for one-factor solutions of unidimensional three-point scale, there is increased number of indicators that influence the factors from a highest of 15 (in two-point scale) to 18 (under three-point scale). Even though there does not appear to be converging number of the influential indicators, the dominant number is 17. Incidentally, for  $n = 150$ , the number of influential indicators is also 17. The proportion of variation accounted for by the factors increased from 50.7% (for  $n = 30$ ) to a high of 59.6% (for  $n = 200$ ) then fluctuates afterwards. There is a high level of fit of the model for all sample sizes.

In regard of two-factor solutions of unidimensional three-point scale, there is higher number of indicator variables on the factors, particularly for the first factor than its counterpart under the two-point scale. The proportion of variation

explained also increases up to  $n = 150$ , and decreases thereafter. There is a general repetition of indicators on factors at all samples, with exception of  $n = 150$ . Unlike all other sample sizes, the results for  $n = 150$  is more plausible as the first factor accounts for almost all cumulative variation explained by the solution. The fitness of the two-factor model is almost perfect for all sample sizes.

Three-factor solutions of unidimensional three-point scale shows that there is increased number of indicators on factors, particularly for the first factor, over that of two-point scale. There is, however, the incidence of repeating indicators of multiple factors for all sample sizes. The sample size of  $n = 150$  appears to produce a more reasonable result as the last factor (i.e., Factor 3) accounts for a negligible proportion of the cumulative variation. The fitness of the three-factor model increases as sample size increase.

We note that for two-factor solutions of two-dimensional three-point scale, there is increased number of indicators on factors, particularly for Factor 1, over that of two-point scale. The amounts of variation explained are almost the same for the two factors for sample sizes 30, 150, 500, and 800. The amount of cumulative variation explained by the two-factor model generally fluctuates with increasing sample size, but highest at  $n = 100$ . At this point, the fitness of the model is also highest.

For two-factor solutions of two-dimensional three-point scale, there is increased number of indicators that influence the formation of factors, Factor 1 in particular, over that of two-point scale. In a like manner, the amount of cumulative variation is quite high in favour of three-point scale. The solution for  $n = 150$  is consistent with our expectation as the first is dominant with 18 influential indicators and the others are influenced by a single indicator each. The amount of cumulative variation largely oscillates with increasing sample size, but peaks at  $n = 150$  with highest model fitness.

Considering one-factor solutions of unidimensional five-point scale, there is increased number of indicators that influence the factor from a highest of 18 (on three-point scale) to 20 (under five-point scale). Even though there does not appear to be converging number of influential indicators, the dominant number is 20. Incidentally, the dominant number of influential indicators starts with  $n = 150$ . The proportion of variation accounted increases from 58% (for  $n = 30$ ) to highest of 72.9% (for  $n = 1000$ ). There is a high level of fitness of the one-factor model for all sample sizes.

With regards to two-factor solutions of unidimensional five-point scale, there is a higher number of indicator variables on the two factors for the five-point scale than those of the three-point scale. The cumulative variations accounted for by the two-factor model are consistently higher for a five-point scale than its three-point scale counterpart. Generally, cumulative variation explained increases remarkably from 64.5% (for  $n = 30$ ) up to 73.2% (for  $n = 150$ ), slight increment to 73.9% (for  $n = 200$  and 500), then fluctuates thereafter. The two-factor model nearly perfectly fits the five-point scale dataset for all sample sizes.

On the basis of three-factor solutions of unidimensional five-point scale, there is increased number of influential indicator variables, especially on Factor 1, over that of three-point scale. Also, cumulative variations explained by the three-factor model are higher for five-point than a three-point scale. There is the incidence of several repeating indicators on multiple factors for all sample sizes, except for  $n = 150$  which has few.

We notice that for two-factor solutions of two-dimensional five-point scale, there is increased number of influential indicators on factors, notably Factor 1, over that of a three-point scale. Further, cumulative variations explained are higher for a five-point scale than a three-point scale. The two factors are substantially influenced by comparable sets of indicator variables for  $n = 150, 200,$  and 800. This occurrence is consistent with number of ability dimensions un-

derlying the scale. For these samples,  $n = 150$  has highest 83.9% of cumulative variation, and as such gives most plausible result.

For three-dimensional datasets, there is comparable number of indicator variables that influence formation of factors on both three and five-point scales for all sample sizes. However, there is a moderate improvement in the amount of variation explained over a three-point scale. It is worthy of note that indicator variables greatly influence the formation of two factors for lower sample sizes. On the contrary, for  $n = 150$  and 200, only Factor 1 considerably dominates the other two. The result for  $n = 150$  is most desirable as it possesses highest 90% cumulative variation. The fitness of the three-factor model on the three-dimensional five-point scale is almost perfect for all sample sizes, and peaks at  $n = 150$ .

One-factor solutions of unidimensional seven-point scale show that there is increased number of indicators that influence the factor even for lower sample size. For all  $n \geq 150$ , all twenty indicator variables are influential. The proportion of variation accounted for increases from 61.6% (for  $n = 30$ ) to a highest of 79.4% (for  $n = 1000$ ). It is also relevant to note that the amount of information explained remains as high (77.6%) for  $n = 200$  as for  $n = 1000$ . Generally, there is almost a perfect fit of the one-factor model for all sample sizes.

An observation of two-factor solutions of unidimensional seven-point scale shows that although the number of influential indicators on Factor 1 have not increased over that of five-point scale, those of Factor 2 have shot up for all sample sizes, except  $n = 100$ . Generally, under seven-point scale, there are comparable sets of indicator variables that largely contribute to the formation of both factors, which is not in line with dimensionality of underlying ability. The only reasonable factor solution for the seven-point scale is the case where  $n = 100$ , with Factor 1 accounting for as high as 64.7% of cumulative variation. There is increased amount of cumulative variation explained over five-point scales for all

sample sizes. The fitness of the two-factor model is almost perfect for all sample sizes.

In respect of three-factor solutions of unidimensional seven-point scale, there is equivalent number of influential indicators on factors for both five and seven-point scales for all sample sizes. Also, there is no improvement in the proportion of variation explained by the first factor over that of five-point scale. Lower sample sizes show two dominant factors, which contrasts the underlying ability dimension. However, for  $n = 150$  and  $200$ , only the first factor is substantially influenced by indicator variables. The result for  $n = 150$  is most reasonable as Factor 1 explains highest (46.3%) proportion of cumulative variation. The fitness of the three-factor model is almost perfect, but as sample size increases, there is no appreciable increase in the amount of fitness.

With respect to two-factor solutions of two-dimensional seven-point scale, there is a similar number of indicator variables that highly influence the formation of factors, especially Factor 1, for both five and seven-point scales. Further, there is no substantial increase in the proportion of variation explained by the first factor. In this system, the two factors are highly dominated by similar sets of influential indicators for  $n = 150$ ,  $200$  and  $500$ . For these samples,  $n = 150$  provides most desirable result as it accounts for largest cumulative variation (87.5%). The two-factor model is nearly perfectly significant, and remains the same at all sample sizes except for  $n = 30$ .

Considering three-factor solutions of three-dimensional seven-point scale, there are comparable sets of indicator variables that significantly contribute to the formation of Factor 1 for both five and seven-point scales. Even though there is increased number of indicators on other two factors, the corresponding cumulative variation explained have increased only marginally. There appears to be three dominant factors for all sample sizes, except  $n = 200$ . The amount of cumulative variation explained increases from 87.8% (for  $n = 30$ ) up to 91.6%

(for  $n = 150$ ), and decreases afterwards. Thus,  $n = 150$  offers most credible result. The results of this part are in line with what Asún et al. (2016) observed.

## CHAPTER FIVE

### SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

#### Overview

This chapter presents a summary of the entire thesis. It highlights the main objectives of the study and research methods that have been taken to achieve them. The effects of number of points on response scales and sample size on item response theory and factor analysis results are highlighted in this chapter. From the summary, conclusions based on the findings of the study have been drawn and recommendations made.

#### Summary

The study investigates the effects of measurement scales on results of item response theory models and multivariate techniques. It is based on simulated datasets under various conditions such as item response format, number of dimensions underlying response scales, and sample size. Two main statistical techniques – Item Response Theory (IRT) models and Factor Analysis – are employed in analysing the simulated datasets.

The review of related literature shows that an 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. The literature showed 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. The literature establishes 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. It has been shown 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 1PPC 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.

The thesis has discussed key concepts and methods used in IRT and factor analyses. It has also presented various IRT models and their graphical representations. It was established that there is a 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 – tetrachoric and polychoric coefficients – were presented.

The study (Nkansah, Zakaria, & Howard, 2018) described the simulation and analyses of datasets under various conditions. It shows that estimated values of discrimination ( $\hat{\alpha}$ ) and difficulty ( $\hat{\delta}$ ) parameters generally fluctuates with increasing sample size. There is a marked difference between the specified and estimated item parameter values at lower samples ( $n = 30$  and  $100$ ), but the difference tends to reduce at sample sizes of  $150$  and beyond. In addition, the differences become negligible at larger samples ( $n = 500, 800, 1000$ ). Further, the study shows that there is a direct relationship between parameters of IRT and those of factor models, particularly item discrimination and factor loadings.

We performed item response theory analyses of unidimensional dichoto-

mous response datasets. The results indicate that items fit the unidimensional 2PL model, since the  $p$ -values are generally much higher than 0.05. Only at  $n = 150, 200$  and  $1000$  it is detected that three items (6, 12, and 7, respectively) do not fit the model. The 2PL model significantly fits the unidimensional dichotomous item response data for all sample sizes.

We observe that almost all items significantly fit two-dimensional IRT model. The only exception is when  $n = 30$  where the fitness of majority of items have not been possible to evaluate due to sparseness in the data. Correspondingly, the overall model fitness could not be determined due to low degrees of freedom. The item response model significantly fits the two-dimensional dichotomous response data for all other sample sizes.

The study indicates that items significantly fit the three-dimensional 2PL model. However, for smaller samples the fitness of items is appalling. The fitness of items to the model get better as sample size increases. We observe that the three-dimensional 2PL model significantly fits the data for all sample sizes, except at  $n = 30$  where the fitness of the model could not be determined due low degrees of freedom.

We notice that in migrating from dichotomous to three-point scale, the significance of fitness of items to the model generally fluctuates for all sample sizes. While the  $p$ -values of fitness of some items improve for given sample size, those of other items deteriorate. The overall fitness of GPC model to unidimensional three-point response data is significant at sample sizes of 100 and over.

We observe that, generally, items significantly fit the two-dimensional GPC model on three-point scale. However, on the same three-point scale, the magnitude of  $p$ -values differ from unidimensional to two-dimensional case. There appears to be two groups of items: (1) items whose  $p$ -value decreases with an additional ability dimension, and (2) items whose  $p$ -value increases with additional ability dimension. Items of the first group will require just one person-

ability to get a response in higher categories, whereas those of the second group need multiple person-abilities to get a similar response. The overall fitness of the model to the two-dimensional three-point response data is significant at all sample sizes, except for  $n = 30$ .

On three-dimensional response datasets, most items fit the model, particularly for larger sample sizes ( $n \geq 150$ ). For smaller sample sizes ( $n \leq 100$ ), the  $p$ -values of items worsened on high dimensions, especially for items that may require only one person-ability to get a response in higher categories. The results indicate that items whose fitness is quite high or almost perfect may require up to three person-abilities to get a response in higher categories. At  $n = 30$ , even though items misfit the model, the overall fitness of the model is significant. The plausibility is that the IRT model yields better results on high response scales with large number of dimensions.

Considering unidimensional five-point scale, we detect that overwhelming majority of items significantly fit the unidimensional GPC model at all sample sizes, except  $n = 30$  where the  $p$ -values of some items could not be determined. Not unexpectedly, the overall GPC model significantly fits the five-point response dataset, with the exception of  $n = 30$ . As sample size increase from 100 up to 1000, the fitness of the item response model fluctuates, but highest at  $n = 1000$ .

On two-dimensional five-point scale, it indicates that most items significantly fit the two-dimensional GPC model at sample size of 100 and beyond. The fitnesses of almost all items could not be evaluated at  $n = 30$  due to low degrees of freedom. Surprisingly, the IRT model significantly fits the data at  $n = 30$ . The overall fitness of the IRT model generally fluctuates with increasing sample size.

In the case of three-dimensional five-point scale, it shows that the fitness of items cannot be determined for all items at  $n = 30$  due to sparseness in the data.

Generally, the  $p$ -values of item fitnesses to the model are much reduced for all sample sizes in this scenario. Particularly, for  $n = 1000$ , all items do not fit the data. Surprisingly, the three-dimensional GPC model almost perfectly fits the five-point response data at all sample sizes. We observe that the item response model yields better results on high response scales. Generally, the overall fitness of the IRT model fluctuates with increasing sample size.

In terms of unidimensional seven-point scale, we note that majority of items generally fit the unidimensional GPC model, except for  $n = 30$  where the fitnesses of items could not be determined due to sparseness in the data. The overall GPC model significantly fits the unidimensional seven-point response dataset for sample sizes from 150 to 1000.

On the seven-point response scale, there is reduced fitness of most items to the two-dimensional GPC model as compared to the unidimensional case. However, the fitness of the overall GPC model to the two-dimensional seven-point response data has become much significant.

The IRT analysis of three-dimensional seven-point scale indicates that there is further deterioration of the fitness of items to the three-dimensional GPC model as compared to the two-dimensional case. Meanwhile, the IRT model significantly fits the three-dimensional seven-point response data for all sample sizes. The fitness of the IRT model largely fluctuates with increasing sample size.

The study indicates that the fitness of items to the IRT model largely varies as the dimensionality of the underlying ability changes. As the number of underlying dimensions increases, the overall fitness of the IRT model generally oscillates.

We conducted one factor analyses of unidimensional dichotomous response datasets. The results reveal that the number of influential indicators appears to converge (at 15) for higher sample size starting at  $n = 150$ . The indicators are

the same at point of convergence. The proportion of variance accounted for by the single factor increases from 33.9% (for  $n = 30$ ) to a highest of 46.1% (for  $n = 800$ ).

With two-factor solutions of unidimensional dichotomous scale, there is generally the incidence of repetition of high loadings on the same indicator variable of the two factors, with exception of sample sizes  $n = 150$  and  $n = 800$ , which can distract interpretation. However, for  $n = 150$ , the first factor loads highly on as many as 15 indicators and explains 42.7% of variation. The second factor loads highly on only one indicator (Variable 5) and is a contrast to its representation in IRT. In addition, amount of variance explained by the second factor appears to be negligible. The sample size of  $n = 150$  thus gives a more plausible factor solution than all other samples. The  $n = 150$  also explains the highest cumulative variation.

In the case of three-factor solutions of unidimensional dichotomous scale, the result becomes less meaningful and even unrealistic for sample sizes beyond 30. There is generally the incidence of repeating indicators on multiple factors. There is also the incidence of unrealistic loadings that are greater than one in higher factor numbers, particularly for Factor 3. This incidence is as a result of an extraction of higher factor structure from a lesser dimensional dataset.

With two-factor solutions of two-dimensional dichotomous scale, the  $n = 150$  generates a repeating factor consistent with two repeating dimensions underlying the dataset. The cumulative variation is also highest for this sample size.

Three-factor solutions of three-dimensional dichotomous scale indicates that lower sample sizes show two dominant factors in the result. This pattern is inconsistent with the expected dimension of the scale. However, for  $n = 150$  and 200, as expected only the first factor is highly influenced by the indicator variables. The factor solution for  $n = 150$  is more conceivable as it accounts for

as high as 78.9% cumulative variation. The fitness of the three-factor model is almost perfect for all sample sizes.

The study shows that, for one-factor solutions of unidimensional three-point scale, there is increased number of indicators that influence the factors from a highest of 15 (in two-point scale) to 18 (under three-point scale). Even though there does not appear to be converging number of the influential indicators, the dominant number is 17. Incidentally, for  $n = 150$ , the number of influential indicators is also 17. The proportion of variation accounted for by the factors increased from 50.7% (for  $n = 30$ ) to a high of 59.6% (for  $n = 200$ ) then fluctuates afterwards. There is a high level of fit of the model for all sample sizes.

In regard of two-factor solutions of unidimensional three-point scale, there is higher number of indicator variables on the factors, particularly for the first factor than its counterpart under the two-point scale. The proportion of variation explained also increases up to  $n = 150$ , and decreases thereafter. There is a general repetition of indicators on factors at all samples, with exception of  $n = 150$ . Unlike all other sample sizes, the results for  $n = 150$  is more plausible as the first factor accounts for almost all cumulative variation explained by the solution. The fitness of the two-factor model is almost perfect for all sample sizes.

Three-factor solutions of unidimensional three-point scale show that there is increased number of indicators on factors, particularly for the first factor, over that of two-point scale. There is, however, the incidence of repeating indicators of multiple factors for all sample sizes. The sample size of  $n = 150$  appears to produce a more reasonable result as the last factor (Factor 3) accounts for a negligible proportion of the cumulative variation. The fitness of the three-factor model increases as sample size increase.

Taking into account two-factor solutions of two-dimensional three-point

scale, there is increased number of indicators on factors, particularly for Factor 1 over that of two-point scale. The amounts of variation explained are almost the same for the two factors for sample sizes 30, 150, 500, and 800. The amount of cumulative variation explained by the two-factor model generally fluctuates with increasing sample size, but highest at  $n = 100$ . At this point, the fitness of the model is also highest.

For two-factor solutions of two-dimensional three-point scale, there is increased number of indicators that influence the formation of factors, Factor 1 in particular, over that of two-point scale. In a like manner, the amount of cumulative variation is quite high in favour of three-point scale. The solution for  $n = 150$  is consistent with our expectation as the first is dominant with 18 influential indicators and the others are influenced by a single indicator each. The amount of cumulative variation largely oscillates with increasing sample size, but peaks at  $n = 150$  with highest model fitness.

Considering one-factor solutions of unidimensional five-point scale, there is increased number of indicators that influence the factor from a highest of 18 (on three-point scale) to 20 (under five-point scale). Even though there does not appear to be converging number of influential indicators, the dominant number is 20. Incidentally, the dominant number of influential indicators starts with  $n = 150$ . The proportion of variation accounted increases from 58% (for  $n = 30$ ) to highest of 72.9% (for  $n = 1000$ ). There is a high level of fitness of the one-factor model for all sample sizes.

With regards to two-factor solutions of unidimensional five-point scale, there is a higher number of indicator variables on the two factors for the five-point scale than those of the three-point scale. The cumulative variations accounted for by the two-factor model are consistently higher for a five-point scale than its three-point scale counterpart. Generally, cumulative variation explained increases remarkably from 64.5% (for  $n = 30$ ) up to 73.2% (for  $n = 150$ ), slight

increment to 73.9% (for  $n = 200$  and  $500$ ), then fluctuates thereafter. The two-factor model nearly perfectly fits the five-point scale dataset for all sample sizes.

On the basis of three-factor solutions of unidimensional five-point scale, there is increased number of influential indicator variables, especially on Factor 1, over that of three-point scale. Also, cumulative variations explained by the three-factor model are higher for five-point than a three-point scale. There is the incidence of several repeating indicators on multiple factors for all sample sizes, except for  $n = 150$  which has few.

We notice that, for two-factor solutions of two-dimensional five-point scale, there is increased number of influential indicators on factors, notably Factor 1, over that of a three-point scale. Further, cumulative variations explained are higher for a five-point scale than a three-point scale. The two factors are substantially influenced by comparable sets of indicator variables for  $n = 150, 200,$  and  $800$ . This occurrence is consistent with number of ability dimensions underlying the scale. For these samples,  $n = 150$  has highest 83.9% of cumulative variation, and as such gives most plausible result.

For three-dimensional datasets, there is comparable number of indicator variables that influence formation of factors on both three and five-point scales for all sample sizes. However, there is a moderate improvement in the amount of variation explained over a three-point scale. It is worthy of note that indicator variables greatly influence the formation of two factors for lower sample sizes. On the contrary, for  $n = 150$  and  $200$ , only Factor 1 considerably dominates the other two. The result for  $n = 150$  is most desirable as it possesses highest 90% cumulative variation. The fitness of the three-factor model on the three-dimensional five-point scale is almost perfect for all sample sizes, and peaks at  $n = 150$ .

One-factor solutions of unidimensional seven-point scale show that there is increased number of indicators that influence the factor even for lower sample



size. For all  $n \geq 150$ , all twenty indicator variables are influential. The proportion of variation accounted for increases from 61.6% (for  $n = 30$ ) to a highest of 79.4% (for  $n = 1000$ ). It is also relevant to note that the amount of information explained remains as high (77.6%) for  $n = 200$  as for  $n = 1000$ . Generally, there is almost a perfect fit of the one-factor model for all sample sizes.

Two-factor solutions of unidimensional seven-point scale reveal that although the number of influential indicators on Factor 1 have not increased over that of five-point scale, those of Factor 2 have shot up for all sample sizes, except  $n = 100$ . Generally, under seven-point scale, there are comparable sets of indicator variables that largely contribute to the formation of both factors, which is not in line with dimensionality of underlying ability. The only reasonable factor solution for the seven-point scale is the case where  $n = 100$ , with Factor 1 accounting for as high as 64.7% of cumulative variation. There is increased amount of cumulative variation explained over five-point scales for all sample sizes. The fitness of the two-factor model is almost perfect for all sample sizes.

In respect of three-factor solutions of unidimensional seven-point scale, there is equivalent number of influential indicators on factors for both five and seven-point scales for all sample sizes. Also, there is no improvement in the proportion of variation explained by the first factor over that of five-point scale. Lower sample sizes show two dominant factors, which contrasts the underlying ability dimension. However, for  $n = 150$  and 200, only the first factor is substantially influenced by indicator variables. The result for  $n = 150$  is most reasonable as Factor 1 explains highest (46.3%) proportion of cumulative variation. The fitness of the three-factor model is almost perfect, but as sample size increases, there is no appreciable increase in the amount of fitness.

With respect to two-factor solutions of two-dimensional seven-point scale, there is a similar number of indicator variables that highly influence the formation of factors, especially Factor 1, for both five and seven-point scales. Fur-

ther, there is no substantial increase in the proportion of variation explained by the first factor. In this system, the two factors are highly dominated by similar sets of influential indicators for  $n = 150, 200$  and  $500$ . For these samples,  $n = 150$  provides most desirable result as it accounts for largest cumulative variation (87.5%). The two-factor model is nearly perfectly significant, and remains the same at all sample sizes except for  $n = 30$ .

Considering three-factor solutions of three-dimensional seven-point scale, there are comparable sets of indicator variables that significantly contribute to the formation of Factor 1 for both five and seven-point scales. Even though there is increased number of indicators on other two factors, the corresponding cumulative variation explained have increased only marginally. There appears to be three dominant factors for all sample sizes, except  $n = 200$ . The amount of cumulative variation explained increases from 87.8% (for  $n = 30$ ) up to 91.6% (for  $n = 150$ ), and decreases afterwards. Thus,  $n = 150$  offers most credible result.

## **Conclusions**

The study primarily investigates the effects of response scales of items on results of item response theory models and multivariate techniques. A number of datasets have been simulated under various conditions such as item response format, number of dimensions underlying response scales, and sample size to address the issues in the study. Two main statistical techniques – Item Response Theory (IRT) models and Factor Analysis – were employed to analyse datasets from various response scales of items.

The study made use of important methods in item response theory such as item and model fitness statistics in order to appreciate the effects of response scales, sample size, and dimensionality of the underlying person ability on re-

sults of IRT models. Further, the effects of the conditions highlighted on the results of factor analysis were examined by the plausibility of the factor solution. In order to understand the reasonableness of the factor solution, we made use of the correct representation of the dimensionality of the datasets, the amount of cumulative variation explained, and the fitness of the factor model.

A significant observation in the study is that there is a direct relationship between the parameters of IRT and those of factor models, particularly item discrimination and factor loadings. In this regard, items with high discrimination values load highly on factors. Such items possess a discriminatory power with absolute value greater than one.

As the number of points on the scale increases, the fitness of items to the IRT model largely fluctuates. In some cases, the fitness of items improves with increasing number of points on the scale, but in other cases the fits of items deteriorates. Further, in most cases the overall fitness of IRT models increases with increasing points on the scale.

In general, the fitness of items to IRT models may not be determinable at smaller sample sizes due to sparsity. As sample size increases, the fitness of items to IRT models generally fluctuates. This incidence is the same for the overall fitness of the IRT model. There does not appear to be any relationship between model fitness and number of fit items as sample size increases.

The results also show that the amount of cumulative variation explained increases with increasing scale-points. It follows that, the number of influential indicators on factors increases with increasing scale-points. This means that, factors are more well defined and could be most interpretable on larger scale-points. The results are, however, almost the same on higher scale-points of five and seven.

The study reveals that for smaller sample size, particularly below 100, items/indicators may not generate the desired dataset. That is, the data generated

may not follow the desired model. We may, therefore, not be able to obtain a reasonable factor solution. On the other hand, we could obtain unrealistic factor solution if we attempt to extract higher factor solution than the underlying dimensionality on few scale-points. This particularly shows that extracting more factors than necessary could run into difficulties, especially for low scale points. Generally, a sample size of 150 produces a more consistent factor solution based on the underlying dimensionality of the data. However, in some cases of the factor structure (particularly, high dimensional datasets), a sample size of 100 gives a more consistent result.

### **Recommendations**

In factor analysis, generally, results appear reasonable on higher scale points irrespective of sample size, even though a sample size of 150 stands out. However, on IRT model, results are particularly not good for small sample size and at higher scale points. It will therefore be important to examine the IRT model, along with factor models on Likert scale data. This has the potential to help obtain the right interpretations of factors.

Again, it can be suggested that the true dimensionality of data be determined in order to extract the appropriate factor structure. Further research in the area would investigate the effect of number of items on results IRT and multi-variate statistical techniques.

## REFERENCES

- Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20(4), 311–329.
- Agresti, A., & Yang, M. C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis*, 5, 9–21.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New Jersey, NJ: John Wiley and Sons.
- Asún, R. A., Rdz-Navarro, K., & Alvarado, J. M. (2016). Developing multidimensional likert scales using item factor analysis: The case of four-point items. *Sociological Methods & Research*, 45(1), 109–133.
- Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 24(2), 222–228.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.
- Balakrishnan, N. (1991). *Handbook of the logistic distribution*. New York, NY: Taylor and Francis.
- Bartolucci, F., Bacci, S., & Gnaldi, M. (2016). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Florida, FL: Taylor and Francis.
- Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of statistical software*, 46(4), 1–29.

- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Model*, 13(2), 186–203.
- Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, 105, 467–477.
- Blerkom, M. L. V. (2009). *Measurement and statistics for teachers*. New York, NY: Routledge.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. California, CA: Sage Publications.
- Cohen, B. H. (2001). *Explaining psychological statistics* (2nd ed.). New York, NY: John Wiley and Sons.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). New Jersey, NJ: Lawrence Erlbaum Associates.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- De Bruin, G. P. (2004). Problems with the factor analysis of items: Solutions based on item response theory and item parcelling. *SA Journal of Industrial Psychology*, 30(4), 16–26.
- de Leeuw, J. (1983). Models and methods for the analysis of correlation coefficients. *Journal of Econometrics*, 22, 113–137.
- DeMars, C. E. (2010). *Item response theory*. New York, NY: Oxford University Press.

- DeMars, C. E. (2012). A comparison of limited-information and full-information methods in Mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Model*, 19(4), 610–632.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9(3), 327–346.
- Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement*, 11, 371–384.
- Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, 41(3), 261–270.
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47(2), 309–326.
- Duong, M., Subedi, D., & Lee, J. (2008). *Item parameter estimation in multidimensional IRT models: A comparison of maximum likelihood and Bayesian approaches*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New York, NY.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum Associates.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York, NY: Oxford University Press.
- Ferrando, P. J., & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory* (Tech. Rep.). Department of Psychology, Universitat Rovira i Virgili, Tarragona.
- Finch, H. (2006). Comparison of the performance of Varimax and Promax rotations: Factor structure recovery for dichotomous items. *Journal of*

*Educational Development*, 43(1), 39–52.

- Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis—based models. *Applied Psychological Measurement*, 34(1), 10–26.
- Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, 33(3), 291–314.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16(4), 625–641.
- Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). London, England: Allyn Bacon.
- Gorsuch, R. (1974). *Factor analysis*. Philadelphia, PA: W. B. Saunders Company.
- Gosz, J. K., & Walker, C. M. (2002). An empirical comparison of multidimensional item response data using TESTFACT and NOHARM. In *annual meeting of the National Council on Measurement in Education, New Orleans, LA*.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles*



- and applications*. New York, NY: Springer Science and Business Media.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. T. (1991). *Fundamentals of item response theory*. California, CA: SAGE Publications.
- Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153.
- Jacoby, J., & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*, 495–500.
- Johnson, M. S., Sinharay, S., & Bradlow, E. T. (2007). Handbook of statistics on psychometrics. In C. R. Rao & S. Sinharay (Eds.), (pp. 587–606). Amsterdam, Netherlands: Elsevier.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). New Jersey, NJ: Pearson Education.
- Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, *59*(2), 149–176.
- Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics: Inference and relationship* (Vol. 2). New York, NY: Charles Griffin.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*(3), 457–477. doi: 10.1207/s15327906mbr2603\_5
- Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, *7*(1), 15–32.
- Kraus, K. (2012). *On the measurement of model fit for sparse categorical data* (Unpublished doctoral dissertation). Acta Universitatis Upsaliensis.
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological modeling*, 1(2), 130–149.
- Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research*, 10(3), 316–318.
- Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, 2(1), 49–53.
- Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(3), 333–356.
- Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among Parametric Item Response Models for polychotomous ordered data. *Applied Psychological Measurement*, 18(3), 245–256.
- McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9(1), 49–57.
- Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods*, 9(3), 301–333.
- Morata-Ramirez, M. D. L. A., & Holgado-Tello, F. P. (2013). Construct validity of Likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, 1, 54.
- Mount, R. E., & Schumacker, R. E. (1998). Identifying measurement disturbance effects using Rasch item fit statistics and the Logit Residual Index. *Journal of Outcomes Measurement*, 2(4), 338–350.
- Muraki, E. (1992). A generalised partial credit model: An application of an EM

- algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables ikert variables. *British Journal of Mathematical and Statistical Psychology*, 38(2), 171–189.
- Nkansah, B. K., Zakaria, A., & Howard, N. K. (2018). Effect of measurement scales on results of item response theory models and multivariate statistical techniques. *Manuscript submitted for publication*.
- Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, 1(2), 66–82.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. California, CA: Sage Publications.
- Parry, C. D. H., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, 15(1), 35–46.
- Partchev, I. (2004). *A visual guide to item response theory*. Retrieved from <http://www.metheval.uni-jena.de/irt/Visual/IRT.pdf>
- Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of mathematical and statistical psychology*, 46(2), 273–286.
- R Core Team. (2017). *R: A Language and Environment for Statistical Computing [Computer software manual]*. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401–412.

- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361–373.
- Reeve, B. B. (2002). An introduction to modern measurement theory. *National Cancer Institute, 1–67*.
- Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. New York, NY: Taylor and Francis.
- Reise, S. P., & Yu, Y. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133–144.
- Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York, NY: John Wiley and Sons.
- Revelle, W. (2017). psych: Procedures for Psychological, Psychometric, and Personality Research [Computer software manual]. Evanston, Illinois. Retrieved from <https://CRAN.R-project.org/package=psych> (R package version 1.7.8)
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373.
- Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*(1), 47–57.
- Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and psychological measurement, 48*(3), 657–667.

- Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. , *16*, 1–16.
- Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, *37*(3), 221–244.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). New Jersey, NJ: Pearson Education.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159–203.
- Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer.
- van der Eijk, C., & Rose, J. (2015). Risky business: Factor analysis of survey data - Assessing the probability of dimensionalisation. *PLoS ONE* *10*(3).
- Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392–423.

## APPENDIX

### R CODES FOR DATA SIMULATION AND ANALYSIS

```

library(mirt)
library(psych)

#Item parameter Values
#=====
a=c(0.5,0.7,0.8,0.6,0.4,2.2,1.5,2.7,1.8,1.6,2,2.9,3,
    2.1,2.8,1.4,1.9,1.2,1.3,2.9)
a2=matrix(c(0.5,0.7,0.8,0.6,0.4,2.2,1.5,2.7,1.8,1.6,
    2,2.9,3,2.1,2.8,1.4,1.9,1.2,1.3,2.9,0.5,
    0.7,0.8,0.6,0.4,2.2,1.5,2.7,1.8,1.6,2,
    2.9,3,2.1,2.8,1.4,1.9,1.2,1.3,2.9),20,2)
a3<=matrix(c(0.5,0.7,0.8,0.6,0.4,2.2,1.5,2.7,1.8,
    1.6,2,2.9,3,2.1,2.8,1.4,1.9,1.2,1.3,
    2.9,0.5,0.7,0.8,0.6,0.4,2.2,1.5,2.7,
    1.8,1.6,2,2.9,3,2.1,2.8,1.4,1.9,1.2,
    1.3,2.9,0.5,0.7,0.8,0.6,0.4,2.2,1.5,
    2.7,1.8,1.6,2,2.9,3,2.1,2.8,1.4,1.9,
    1.2,1.3,2.9),20,3)

d=c(0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,2.5,2.3,
    1.5,2.2,0.3,0.5,0.25,0.4,0.42,0.56,0.2)

```

```
d3=matrix(c(0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,  
2.5,2.3,1.5,2.2,0.3,0.5,0.25,0.4,0.42,  
0.56,0.2,0,0.12,-2.3,0.1,2.0,-2.5,-2,  
-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,0.5,0.25,  
0.4,0.42,0.56,0.2,0,0.12,-2.3,0.1,2.0,  
-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,  
0.5,0.25,0.4,0.42,0.56,0.2),20,3)
```

```
d5=matrix(c(0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,  
2.5,2.3,1.5,2.2,0.3,0.5,0.25,0.4,0.42,  
0.56,0.2,0,0.12,-2.3,0.1,2.0,-2.5,-2,  
-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,0.5,0.25,  
0.4,0.42,0.56,0.2,0,0.12,-2.3,0.1,2.0,  
-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,  
0.5,0.25,0.4,0.42,0.56,0.2,0,0.12,-2.3,  
0.1,2.0,-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,  
2.2,0.3,0.5,0.25,0.4,0.42,0.56,0.2,  
0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,  
2.5,2.3,1.5,2.2,0.3,0.5,0.25,0.4,0.42,  
0.56,0.2),20,5)
```

```
d7=matrix(c(0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,  
2.5,2.3,1.5,2.2,0.3,0.5,0.25,0.4,0.42,  
0.56,0.2,0,0.12,-2.3,0.1,2.0,-2.5,-2,  
-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,0.5,0.25,  
0.4,0.42,0.56,0.2,0,0.12,-2.3,0.1,2.0,  
-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,  
0.5,0.25,0.4,0.42,0.56,0.2,0,0.12,-2.3,
```

```

0.1,2.0,-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,
2.2,0.3,0.5,0.25,0.4,0.42,0.56,0.2,
0,0.12,-2.3,0.1,2.0,-2.5,-2,-1.5,-2.2,
2.5,2.3,1.5,2.2,0.3,0.5,0.25,0.4,0.42,
0.56,0.2,0,0.12,-2.3,0.1,2.0,-2.5,-2,
-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,0.5,0.25,
0.4,0.42,0.56,0.2,0,0.12,-2.3,0.1,2.0,
-2.5,-2,-1.5,-2.2,2.5,2.3,1.5,2.2,0.3,
0.5,0.25,0.4,0.42,0.56,0.2),20,7)

```

*#Simulating Dichotomous Data*

*#Unidimensional;*

```

set.seed(201);Dichot30=simdata(a=a,d=d,N=30,
                                itemtype ="2PL")
set.seed(201);Dichot100=simdata(a=a,d=d,N=100,
                                   itemtype ="2PL")
set.seed(201);Dichot150=simdata(a=a,d=d,N=150,
                                   itemtype ="2PL")
set.seed(201);Dichot200=simdata(a=a,d=d,N=200,
                                   itemtype ="2PL")
set.seed(201);Dichot500=simdata(a=a,d=d,N=500,
                                   itemtype ="2PL")
set.seed(201);Dichot800=simdata(a=a,d=d,N=800,
                                   itemtype ="2PL")
set.seed(201);Dichot1000=simdata(a=a,d=d,N=1000,
                                    itemtype ="2PL")

```



```

#Two-dimensional;
set.seed(201); Dichot30=simdata(a=a2,d=d,N=30,
                                itemtype = "2PL")
set.seed(201); Dichot100=simdata(a=a2,d=d,N=100,
                                   itemtype ="2PL")
set.seed(201); Dichot150=simdata(a=a2,d=d,N=150,
                                   itemtype ="2PL")
set.seed(201); Dichot200=simdata(a=a2,d=d,N=200,
                                   itemtype ="2PL")
set.seed(201); Dichot500=simdata(a=a2,d=d,N=500,
                                   itemtype ="2PL")
set.seed(201); Dichot800=simdata(a=a2,d=d,N=800,
                                   itemtype ="2PL")
set.seed(201); Dichot1000=simdata(a=a2,d=d,N=1000,
                                   itemtype ="2PL")

```

*#IRT Analyses of Dichotomous Data*

*#=====*

*#Unidimensional;*

```

Modell1=mirt(data=Dichot30,model = 1,
             IRT.param = TRUE,itemtype ="2PL")

```

```

print(coef(Modell1),digits = 3)

```

```

itemfit(Modell1) #Item fitness

```

```

M2(Modell1) #Overall model fitness

```

```

plot(Modell1,type ='trace') #trace lines

```

```

Modell2=mirt(data=Dichot100,model = 1,
             IRT.param = TRUE,itemtype ="2PL")

```

```

print (coef (Model2), digits = 3)
itemfit (Model2) #Item fitness
M2 (Model2)      #Overall model fitness
plot (Model2, type = 'trace') #trace lines
#Unidimensional IRT models for other samples
#are evaluated by changing the corresponding
#dataset.

#Two-dimensional;
Model1=mirt (data=Dichot30, model = 2,
             IRT.param = TRUE, itemtype ="2PL")
print (coef (Model1), digits = 3)
itemfit (Model1) #Item fitness
M2 (Model1)      #Overall model fitness
plot (Model1, type = 'trace') #trace lines

Model2=mirt (data=Dichot100, model = 2,
             IRT.param = TRUE, itemtype ="2PL")
print (coef (Model2), digits = 3)
itemfit (Model2) #Item fitness
M2 (Model2)      #Overall model fitness
plot (Model2, type = 'trace') #trace lines
#Two-dimensional IRT models for other samples
#are evaluated by changing the corresponding
#dataset.

```

## #Factor Analyses of Dichotomous Data

#=====

```
FA1<-fa(r=Dichot30,nfactors = 1,n.obs=30,  
        rotate = "varimax",scores = "regression",  
        fm="pa",cor = "tet")
```

```
print(FA1$r,digits = 3)
```

```
print(FA1$Structure,digits = 3)
```

```
print(FA1$fit)
```

```
FA2<-fa(r=Dichot100,nfactors = 1,n.obs=100,  
        rotate = "varimax",scores = "regression",  
        fm="pa",cor = "tet")
```

```
print(FA2$r,digits = 3)
```

```
print(FA2$Structure,digits = 3)
```

```
print(FA2$fit)
```

#One-factor solutions for other samples are  
#obtained by substituting the corresponding

#dataset and number of observations.

#Two-factor solutions are obtained by

#changing 'nfactors' to 2.

## #Simulating three-point scale Data

#=====

#Unidimensional;

```
set.seed(201);poly3_30=simdata(a=a,d=d3,  
                               N=30,itemtype = "gpcm")
```

```
set.seed(201);poly3_100=simdata(a=a,d=d3,N=100,  
                                itemtype = "gpcm")
```

```
#Three-point response datasets for other samples  
#can be generated by changing the value of 'N'.
```

```
#Two-dimensional;
```

```
set.seed(201); poly3_30=simdata(a=a2,d=d3,N=30,  
                               itemtype = "gpcm")
```

```
set.seed(201); poly3_100=simdata(a=a2,d=d3,N=100,  
                                 itemtype = "gpcm")
```

```
#Two-dimensional Three-point response datasets  
#for other samples can be generated by changing  
#the value of 'N'.
```

```
#IRT Analyses of Three-point Scale Data
```

```
#=====
```

```
#Unidimensional;
```

```
Model1<-mirt(data=poly3_30,model=1,
```

```
             IRT.param = TRUE,
```

```
             itemtype="gpcm")
```

```
print(coef(Model1),digits = 3)
```

```
itemfit(Model1) #Item fitness
```

```
M2(Model1)      #Overall model fitness
```

```
plot(Model1,type='trace') #trace lines
```

```
Model2=mirt(data=poly3_100,model = 1,
```

```
           IRT.param = TRUE,
```

```
           itemtype = "gpcm")
```

```
print(coef(Model2),digits = 3)
```

```
itemfit(Model2) #Item fitness
```

```

M2 (Model2)      #Overall model fitness
plot (Model2, type = 'trace') #trace lines
#Unidimensional "gpcm" models for other
#samples are evaluated by changing the
#corresponding dataset.

#Two-dimensional;
Model1<-mirt (data=poly3_30,model=2,
              IRT.param = TRUE,
              itemtype="gpcm")
print (coef (Model1), digits = 3)
itemfit (Model1) #Item fitness
M2 (Model1)      #Overall model fitness
plot (Model1, type = 'trace') #trace lines

Model2=mirt (data=poly3_100,model = 2,
             IRT.param = TRUE, itemtype = "gpcm")
print (coef (Model2), digits = 3)
itemfit (Model2) #Item fitness
M2 (Model2)      #Overall model fitness
plot (Model2, type = 'trace') #trace lines
#Two-dimensional "gpcm" models for other
#samples are evaluated by changing the
#corresponding dataset.

```

```

#Factor Analyses of Three-Point Scale Data
#=====
FA1<-fa(r=poly3_30,nfactors = 1,n.obs=30,
        rotate = "varimax",scores = "regression",
        fm="pa",cor = "poly")
print(FA1$r,digits = 3)
print(FA1$Structure,digits = 3)
print(FA1$fit)

```

```

FA2<-fa(r=poly3_100,nfactors = 1,n.obs=100,
        rotate = "varimax",scores = "regression",
        fm="pa",cor = "poly")
print(FA2$r,digits = 3)
print(FA2$Structure,digits = 3)
print(FA2$fit)

```

```

#One-factor solutions for other samples are
#obtained by substituting the corresponding
#dataset and number of observations.
#Two-factor solutions are obtained by
#changing 'nfactors' to 2.

```

```

#Simulating Five-point scale Data
#=====
#Unidimensional;
set.seed(201);poly5_30=simdata(a=a,d=d5,
                             N=30,itemtype = "gpcm")
set.seed(201);poly5_100=simdata(a=a,d=d5,
                                N=100,itemtype = "gpcm")

```

```
#Five-point response datasets for other
#samples can be generated by changing the
#value of 'N'.
```

```
#Two-dimensional;
```

```
set.seed(201);poly5_30=simdata(a=a2,d=d5,
                               N=30,itemtype = "gpcm")
```

```
set.seed(201);poly5_100=simdata(a=a2,d=d5,
                                  N=100,itemtype = "gpcm")
```

```
#Two-dimensional Five-point response datasets
#for other samples can be generated by changing
#the value of 'N'.
```

```
#IRT Analyses of Five-point Scale Data
```

```
#=====
```

```
#Unidimensional;
```

```
Model1<-mirt(data=poly5_30,model=1,
```

```
              IRT.param = TRUE,
```

```
              itemtype="gpcm")
```

```
print(coef(Model1),digits = 3)
```

```
itemfit(Model1) #Item fitness
```

```
M2(Model1)      #Overall model fitness
```

```
plot(Model1,type = 'trace') #trace lines
```

```
Model2=mirt(data=poly5_100,model = 1,
```

```
            IRT.param = TRUE,
```

```
            itemtype = "gpcm")
```

```
print(coef(Model2),digits = 3)
```

```

itemfit(Model2) #Item fitness
M2(Model2)      #Overall model fitness
plot(Model2,type ='trace') #trace lines
#Unidimensional "gpcm" models for other
#samples are evaluated by changing the
#corresponding dataset.

#Two-dimensional;
Model1<-mirt(data=poly3_30,model=2,
            IRT.param = TRUE,
            itemtype="gpcm")
print(coef(Model1),digits = 3)
itemfit(Model1) #Item fitness
M2(Model1)      #Overall model fitness
plot(Model1,type ='trace') #trace lines

Model2=mirt(data=poly3_100,model = 2,
            IRT.param = TRUE,itemtype ="gpcm")
print(coef(Model2),digits = 3)
itemfit(Model2) #Item fitness
M2(Model2)      #Overall model fitness
plot(Model2,type ='trace') #trace lines
#Two-dimensional "gpcm" models for other
#samples are evaluated by changing the
#corresponding dataset.

```



```

#Factor Analyses of Five-Point Scale Data
#=====
FA1<-fa(r=poly5_30,nfactors = 1,n.obs=30,
        rotate = "varimax",
        scores = "regression",fm="pa",cor = "poly")
print(FA1$r,digits = 3)
print(FA1$Structure,digits = 3)
print(FA1$fit)

FA2<-fa(r=poly5_100,nfactors = 1,n.obs=100,
        rotate = "varimax",
        scores = "regression",fm="pa",cor = "poly")
print(FA2$r,digits = 3)
print(FA2$Structure,digits = 3)
print(FA2$fit)

#One-factor solutions for other samples are
#obtained by substituting the corresponding
#dataset and number of observations.
#Two-factor solutions are obtained by changing
#'nfactors' to 2.

#Simulating Seven-point Scale Data
#=====
#Unidimensional;
set.seed(201);poly7_30=simdata(a=a,d=d7,
                              N=30,itemtype = "gpcm")
set.seed(201);poly7_100=simdata(a=a,d=d7,
                                N=100,itemtype = "gpcm")

```

```
#Three-point response datasets for other samples  
#can be generated by changing the value of 'N'.
```

```
#Two-dimensional;
```

```
set.seed(201); poly7_30=simdata(a=a2,d=d7,  
                                N=30,itemtype = "gpcm")
```

```
set.seed(201); poly7_100=simdata(a=a2,d=d7,  
                                  N=100,itemtype = "gpcm")
```

```
#Two-dimensional Three-point response datasets  
#for other samples can be generated by changing  
#the value of 'N'.
```

```
#IRT Analyses of Seven-point Scale Data
```

```
#=====
```

```
#Unidimensional;
```

```
Model1<-mirt(data=poly7_30,model=1,
```

```
              IRT.param = TRUE,
```

```
              itemtype="gpcm")
```

```
print(coef(Model1),digits = 3)
```

```
itemfit(Model1) #Item fitness
```

```
M2(Model1)      #Overall model fitness
```

```
plot(Model1,type = 'trace') #trace lines
```

```
Model2=mirt(data=poly7_100,model = 1,
```

```
            IRT.param = TRUE,
```

```
            itemtype = "gpcm")
```

```
print(coef(Model2),digits = 3)
```

```
itemfit(Model2) #Item fitness
```

```
M2 (Model2)      #Overall model fitness
plot (Model2, type = 'trace') #trace lines
#Unidimensional "gpcm" models for other
#samples are evaluated by changing the
#corresponding dataset.
```

```
#Two-dimensional;
```

```
Model1<-mirt (data=poly7_30,model=2,
              IRT.param = TRUE,
              itemtype="gpcm")
```

```
print (coef (Model1), digits = 3)
```

```
itemfit (Model1) #Item fitness
```

```
M2 (Model1)      #Overall model fitness
```

```
plot (Model1, type = 'trace') #trace lines
```

```
Model2=mirt (data=poly7_100,model = 2,
             IRT.param = TRUE,itemtype ="gpcm")
```

```
print (coef (Model2), digits = 3)
```

```
itemfit (Model2) #Item fitness
```

```
M2 (Model2)      #Overall model fitness
```

```
plot (Model2, type = 'trace') #trace lines
```

```
#Two-dimensional "gpcm" models for other
```

```
#samples are evaluated by changing the
#corresponding dataset.
```

*#Factor Analyses of Seven-Point Scale Data*

*#=====*

```
FA1<-fa(r=poly7_30,nfactors = 1,n.obs=30,  
        rotate = "varimax",  
        scores = "regression",fm="pa",cor = "poly")  
print(FA1$r,digits = 3)  
print(FA1$Structure,digits = 3)  
print(FA1$fit)
```

```
FA2<-fa(r=poly7_100,nfactors = 1,n.obs=100,  
        rotate = "varimax",  
        scores = "regression",fm="pa",cor = "poly")
```

```
print(FA2$r,digits = 3)  
print(FA2$Structure,digits = 3)  
print(FA2$fit)
```

*#One-factor solutions for other samples are  
#obtained by substituting the corresponding  
#dataset and number of observations.  
#Two-factor solutions are obtained by  
#changing 'nfactors' to 2.*