A HIERARCHICAL NON-PARAMETRIC BAYESIAN TESTLET MODEL FOR DUAL LOCAL DEPENDENCE

A DISSERTATION SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS
FOR THE DOCTOR OF PHILOSOPHY IN SCIENCE (STATISTICS)
IN THE FACULTY OF SCIENCE

By
VONAI CHARAMBA

SUPERVISOR:Main Supervisor: Prof. Lawrence Kazembe (University of Namibia)
Co-Supervisor: Dr. Ndeyapo Martha Nickanor (University of Namibia)

UNIVERSITY OF NAMIBIA
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS
MAY 2021

To Tawanda, Tadiwa and Tarisai ...

may you learn to love learning

Table of Contents

List of Tables

List of Figures

ABSTRACT

The use of psychometric tests to measure the level of individuals on unobservable traits is common in many fields and item response theory (IRT) models are usually used for proficiency measurement. Standard IRT models assume local person and item independence and normality assumption for the true ability distribution. However, respondents are often clustered and test items are often grouped according to sub-content measuring the same stimuli or sub-component of the trait. This study presents a non-parametric polytomous multilevel testlet model for simultaneously modelling person and item clustering effects. The grouping variable is assumed unknown and determined from the data using the Dirichlet Process. The model was compared with a parametric dual model with groups assumed to be known, the testlet model, the Generalised Partial Credit Model and the multilevel model accounting for person dependence effects only in terms of systematic, random and total errors in person and item parameter estimation and test information and reliability, for simulated and real life data. The effects of ignoring dual dependence effects were evaluated for variant group, sample, testlet size, number of response options and mixed items tests. Groups of size 5, 20 and 40 were compared for 400, 1000 and 2000 respondents. The effects of ignoring dependence effects were compared for 6 testlets of 3, 6 and 10 items each for 3, 4 and 5 response categories were compared. Consequences of mis-specifying the slope and proficiency distributional parameters were evaluated where the competing models estimated item and person parameters for skewed, bimodal, normal and uniformly distributed traits and constant and stochastic slopes. Three dependency conditions (0,

none; 0.5, medium; 1, large), were considered for both local item dependency (LID) and local person dependency (LPD). For each simulation study, a fully-crossed factorial design was employed and the general linear model was employed for comparing estimation errors. Significant different means were detected by use of Cohen's effect size, f. In general, ignoring LPD effects resulted in increased bias and total errors in the estimation of ability parameters while ignorance of LID negatively impacted on item parameter estimation increasing with sample size, testlet size and number of options. Failure to account for LID resulted in underestimation of proficiency standard errors, thus resulting in overestimation of test information and reliability. When dual dependence effects were ignored, both item and ability parameter estimation accuracy was reduced. The non-parameter model detected the number of groups and group membership well especially for smaller groups, increasingly with sample size, testlet size and number of categories. However, the non-parametric models requires high computational performance Considering the consequences of ignoring random effects and the computation efficiency required by the non-parametric model, it is recommended that the model be used to detect dependence effects and groups, and standard IRT models be applied for independent persons.

ACKNOWLEDGEMENTS

My entire gratitude for the timely and continuous help from my supervisors Professor Lawrence Kazembe and Professor Ndeyapo Nickanor, or their guidance despite their very tight work schedules.

I acknowledge Professor Jonathan Crush and the Hungry Cities Partnership for allowing me to use their data for my studies

I acknowledge the Queen Elizabeth Scholarship and the Hungry Cities Partnership for the scholarship award to travel to Canada where I gained knowledge on the issue of global food insecurity and Statistical Machine Learning, a valuable asset to the completion of my studies

Lastly I appreciate my young sister Beatrice Charamba for always being there for the late night and early morning calls to discuss my project.

DECLARATIONS

I, Vonai Charamba, hereby declare that this is a true reflection of my own research and that this work or part thereof has not been submitted for a degree in any other institution of higher learning. No part of this dissertation may be reproduced, stored in any retrieval system, or transmitted in any form, or by any means without the prior permission of the author, or the University of Namibia.

I, Vonai Charamba grant the University of Namibia the right to reproduce this dissertation in whole or in part, in any manner or format, which the University of Namibia may deem fit, for any person or institution requiring it for study and research; providing that the University of Namibia shall waive this right if the whole dissertation has been or is being published in a manner satisfactory to the University.

..... Date: .....

Vonai Charamba

List of Abbreviations

AFSUN	African Food Security Urban Network
AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
BGR	Brooks-Gelman-Rubin
BIC	Bayesian Information Criteria
CCC	Category Characteristic Curves
DIC	Deviance Information Criteria
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
FANTA	Food and Nutrition Technical Assistance I
GPCM	Generalised Partial Credit Model
HCP	Hungry Cities Partnership
HDDS	Hosehold Dietary Diversity Score
HFIAS	Household Food Insecurity Access Score
ICC	Item Characteristic Curve
IRT	Item Response Theory
LID	Local Item Independence
LLTM	Linear Logistic Test Model
LPD	Local Person Dependence
MAHFP	Months of Adequate Household Food Provisioning
MC	Markov Chain
MCMC	Markov Chain Monet Carlo
MGM	Multiple Group IRT Model
MRM	Mixture Rasch Model
PCM	Partial Credit Model
PSU	Primary Sampling Unit
RMSE	Root Mean Square Error
SE	Standard Error
SPSS	Statistical Package for Social Scientists
TIF	Test Information Function

Chapter 1

Introduction

1.1 Orientation of study

The use of psychometric tests to measure the level of an individual for certain unobservable latent attributes is a common practice in different areas of educational (Bradlow, Wainer & Wang, 1999;) social and behaviour (Raudenbush,Johnson & Sampson, 2003; Rijmen, 2008), and biomedical clinical sciences (Prenovost et al., 2018). Most of these psychometric achievements involve items that can be clustered into subgroups that measure a sub-content or single common "stimulus" (Bradlow et al., 1999) of the general latent attribute measured by the whole test. In addition, respondents are usually drawn from populations comprising of clusters with different characteristics such as schools, gender, ethnic groups, suburbs and regions depending on the latent attribute being measured. Examinees in one group may share characteristics as a group, leading to correlation within the group, thus violating the assumption of independence (Wang, Jiao & He, 2011) as differences among the group behaviour can reflect differences among the respondents from different groups and some similarities among respondents from the same cluster. As a result, it is important to control for such clustering when modeling the ability of persons when it exists since respondents and items may be nested in respondent and item clusters respectively (Santos, Azevedo & Bolfarie, 2012).

Item response theory (IRT) models are one of the most common methods used to measure psychometric ability of individuals taking a test. These are stochastic, mathematical models relating the probability of observing a certain result on a test item

to individual proficiency parameters and stimulus / item parameters. The standard unidimensional IRT models require that an individual's responses to different item be independent (local item independence) and that the responses of individuals to a test item are independent (local person independence) (Embretson & Reise, 2000, Lord & Novick, 1968) and these assumptions are usually violated when persons taking the test are from clustered samples and when items are categorised according to some sub-stimulus of the latent attribute being measured. Item clusters have been defined as tests within a test and coined as "*testlet*" (Bradlow, et al., 1999; Wainer & Kiely, 1987) or *"test bundles"* (Rosenbaum, 1988). Items within testlets are locally dependent because they are associated with the same stimulus and respondents in the same cluster are likely to give related responses to items or to items in a specific testlet. Modelling clustered data under a model that assumes independent observations is inappropriate because observations on examinees from the same cluster may be more homogeneous than observations on examinees from other clusters (Zenisky, Hambelton & Sireci, 2002; Wang, Jiao & He, 2011; Jiao & Zhang, 2014). The challenge for testlet developers is not to eliminate item and respondent clusters but rather to find a proper solution so that such local dependence does not negatively affect the reliability and validity of inference from the test.

Different models have been derived by different scholars to cater for person and item clustering. Some scholars who have studied person group effects studied the concept of differential item functioning (DIF), that is, when the item difficulty parameters vary from one group to another. Some scholars (eg Reckase, 2009) have treated item groups as the violation of unidimensionality assumption and hence applied multidimensional models where each item cluster is assumed to measure a unique latent ability, making the individual ability parameter to be a vector rather than a scalar (Lee et al., 2001; Weiner & Thissen, 1996). Other scholars have treated these sub-categories of items as sub-tests within tests (testlets) and derived models popularly known as testlet models

(Bradlow et al., 1999; Paek et al., 2009; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer & Wang, 2000; Wang & Wilson, 2005). The proponents of testlet models argue that existence of testlets in a test, if not accounted for, can reduce accuracy of parameter estimates, resulting in inaccurate inferences. An investigation by Lee, Dunbar and Frisbie (2001) showed a need for some parameters to model testlet effects.

To account for person clustering in psychometric tests, Rost (1990), Fox and Glas (2001), Bock and Zimowski (1997) and Rijmen (2008) proposed mixture and multilevel IRT models where the person clusters are factored into the model. They argue that failure to incorporate person clustering effects in a model may jeopardize measurement, leading to biased parameter estimates and the underlying distribution might be met in separate population groups but not on the entire population (Wilson, De Boeck & Carstensen, 2008). Hartig et al. (2008) also argue that failure to account for person clusters may negatively impact on the distribution of person ability parameters if assumed to be random effects.

Much of the researches done have dealt with these random effects separately. However, Jiao et al. (2012) proposed the multilevel testlet model for dual local dependence for binary response items while Jiao and Zhang (2014) derived a multilevel testlet model for polytomous data to account for dual dependence caused by person and item clusters by extending the Generalised Partial Credit testlet model. These researchers have considered the group membership to be manifest, determined through cluster sampling and supplied as model data in estimation. However, there could be some testing procedures where the actual groups are not known and have to be inferred from the data. Moreover, literature has shown mis-specification of the distribution of random effects to have severe consequences on properties of estimators (Grilli & Rampichni, 2009; Heargety & Kurland, 2001; Verbeke & Lesaffire, 1996).

1.1.1 Conceptual framework

This study assumes grouped respondents taking a test comprising of items that are grouped according to sub-content to measure a common trait. However, unlike earlier studies where respondents have looked at item clusters having an effect on an individual, the current study assumes the item cluster to have an effect on the entire respondent cluster, that is, there is interaction between items and persons groups as illustrated in Figure 1.1.

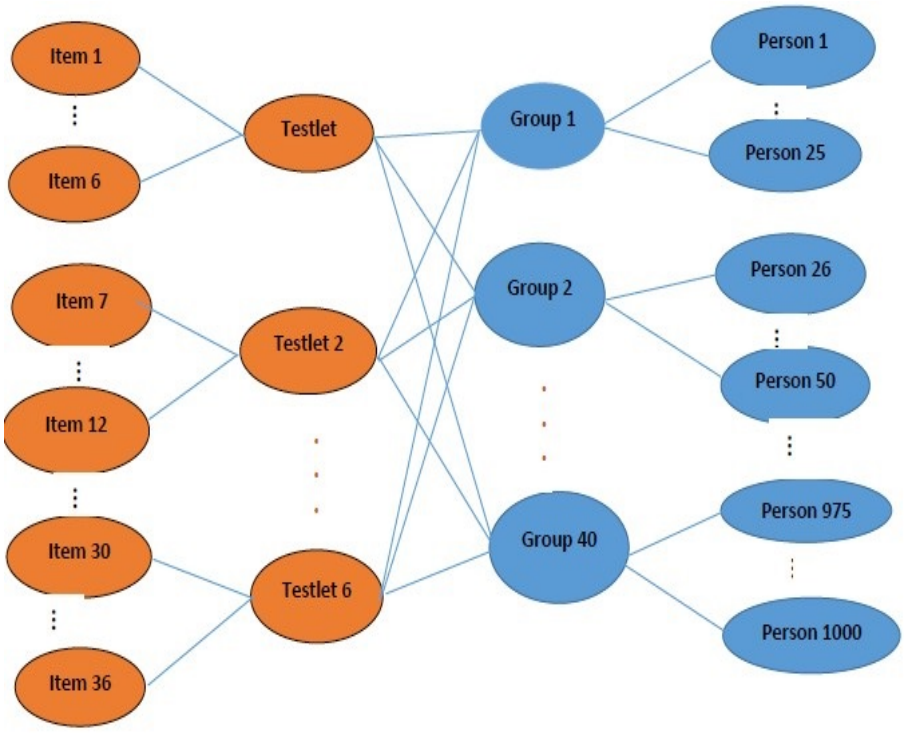


Figure 1.1: Illustration of the framework for the Hierarchical Dual Dependency Model

According to the diagrammatic illustration in Figure 1.1, the respondents are categorised into person clusters while the items are clustered into testlets. There is a possibility of an interaction existing between person groups and items clusters as shown in the illustration. This research is based upon such a framework and statistical theory will be extended and applied to model such relationships.

1.1.2 Statement of problem

IRT models are commonly used in educational, psychological, food security, behaviour science and health related researches to measure the latent ability of individuals from their responses to a set of multiple choice type of items. Most IRT models are based on the assumption of local independence of items and individuals answering the test. However, in reality, most psychometric tests comprise of either persons sampled from clusters such as gender, economic groups and schools or items that can be categorized according to content and measuring a single common stimuli of the latent variable under study, possibly introducing dependency among test takers and among test items. There is a possibility of having tests where both items and persons are clustered. However, researches in literature have mainly accounted for LID and LPD separately and very few studies have modelled LID and LPD effects simultaneously. Available dual dependency models have considered item difficulty parameters varying for person group and by item clusters. In addition, scholars who have modeled dual dependence have assumed no interaction between testlets and subject clusters. However, there is a possibility of interaction between such where members of one cluster perform better in one cluster of items than individuals in others clusters. In summary, the model can be extended to allow for group and testlet interaction.

Standard IRT models and existing IRT software are based on the assumption of normality for true ability distribution. However some researchers, (Duncan & McEcheran, 2008; Reise & Yu, 1990; DeMars, 2003; De Ayala & Sava-Bolesta, 1999; Luo, 2018) argue that there might be some trait distributions that are not necessarily normally distributed. Research has shown that mis-specifying the distribution of random effects may bias parameter estimates, thereby leading to wrong inference (Grilli & Rampichini, 2009; Verbeke & Lesaffire, 1996). Researchers who assessed the effects of modelling tests with non-normal ability traits using normal trait models (DeMars, 2003; Reise & Yu, 1990) have observed biased results with high root means square errors (RMSE) for

some distributions.

Factors that have been identified to affect the accuracy and stability of parameters are sample size (He & Wheadon, 2012, Zhang, 2010), number of items in the tests (Nord, 1968; Zhang, 2010) and number of response options (Lazano, Garcia-Cueto & Muniz, 2008; Lee & Paek, 2014; Weng, 2004; Preston & Colman, 2000). The current research proposes a Bayesian non-parametric multilevel model catering for item and respondents clusters without necessarily assuming restrictive normal assumptions for latent trait by employing the Dirichlet Process Mixture of normal distributions for the ability parameters. The group membership are assumed latent and inferred from the data by use of the "stick breaking" process. The stability of model parameter estimates were evaluated as the sample size, number of items per testlet and response options vary. The IRT modelling technique has mainly been utilised in educational sectors although it has potential for application in other fields such as the food security and health sectors where the technique had not been fully utilised for measurement.

1.2 Objectives of study

1.2.1 Main objective of study

The main research aim is to come up with a Bayesian hierarchical non-parametric testlet model that accounts for dual clustering. The model assumes Dirichlet Process Mixture ability parameters and determine the number of groups and group membership from clustered data with possible interaction between group and person clusters.

1.2.2 Specific objectives of study

1. To develop a non-parametric Bayesian hierarchical random effects dual dependence IRT model assuming Dirichlet Process Mixture ability parameters.

2. To assess the effects of ignoring item and person clusters in IRT modeling by comparing the proposed model with the Generalised Partial Credit Model, Rasch

testlet model and multilevel mixture models in terms of parameter stability, bias, errors of estimation, model fitness and total error variance.

3. To evaluate the stability of parameter estimates for the proposed models as the sample size, testlet size, number of response category options and magnitudes of item and person local dependency increases.

4. To compare the estimation prowess of the proposed model and other models for non-normal ability distributions

5. To determine the effects of ignoring the random slope in ability and threshold parameter estimation

6. To evaluate the ability of the proposed model to estimate person and item parameters for mixed-item tests with different number of response categories

7. To apply the developed model in the estimation of food insecurity for Windhoek urban households

1.3 Significance of study

Psychometric tests usually comprise of respondents sampled from subject classes such as geographical locations, schools, ethnic groups and gender depending on the latent attribute under consideration. In addition, many tests comprise of subsections within the test, linked in terms of content or a subsection of the stimulus being measured. The standard IRT models assume item and person independence. However, there might be dependency between subjects and items if the aforesaid circumstances exist. According to literature, failure to account for person and item dependency effects may result in erroneous inference. Furthermore, research has shown that there might exist ability parameters that may not necessarily be normally distributed. Literature has shown that the Dirichlet Process Mixture (DPM) can explain virtually all distributions by mixing distributions from the same family. Considering consequences of incorrectly

specified random-effects, the non-parametric Dirichlet Process Mixture ability parameters may have a better fit to the data and may give estimates that are more robust when compared to parametric models.

Moreover, given the repercussions of neglecting local dependence in IRT models cited in literature, the proposed hierarchical dual dependence may result in unbiased, accurate and precise parameter estimates. In addition, the proposed multilevel hierarchical non-parametric model allows for relative comparison of individuals and group abilities in terms of the latent variable. The DPM process proposed for latent trait can be used where latent groups are not known *a priori* and hence will be determined from the data and this is usually the case in research. This study contributes to existing literature by investigating the consequences of ignoring item and person clustering for latent groups determined by the stick-breaking process and comparing the model to models completely ignoring clustering and the parametric dual as the sample size, testlet size and dependency effects vary. The study also looks at the effect of mis-specifying the distribution of ability parameters on item and person parameter estimation.

1.4 Structure of report

Chapter 1 presented the overview, conceptual framework, background to item response theory modelling, orientation of the study and the statement of research problem and research objectives. Chapter 2 details the item response theory model developments in addressing item, person and dual dependency, a short review of the basic Bayesian modelling and the Bayesian non-parametric modelling including the Dirichlet Process, estimation methods, methods of assessing convergence and techniques for comparing the goodness of fit of Bayesian model. Chapter 3 highlights the development of the proposed non-parametric Bayesian Dirichlet Process mixture IRT model. Chapter 4 of the project gives the application and performance of the proposed model to simulated data compared to other models previously developed, looking at effects of ignoring

dual clustering when modelling clustered respondents given a test comprising of clustered items and Chapter 5 assesses the effects of samples and group sizes on ignored item and person dependence effects. Chapter 6 looks at the effects of ignoring dual dependence effects as the testlet length and number of categories changes and the ability of the models to estimate mixed-items tests. Chapter 7 looks at the consequences of mis-specifying the ability and slope distributions when items and persons are dependent. In Chapter 8 the model is applied to food security data to estimate a single food (in)security measure for food access, dietary diversity and food availability where sections on the Food and Nutrition Technical Assistance (FANTA) based questionnaire measuring household food insecurity assessment index, household dietary diversity score and months of (un)availability of adequate food are considered as testlet within a test, measuring sub-stimuli of the main latent variable, food insecurity. Chapter 9 concludes and recommends further studies.

Chapter 2

Literature Review

2.1 Local dependence in psychometric tests

Local independence assumption is one of the underlying assumptions for item response theory models. Local independence has two facets, local item independence and local person independence and is obtained when the relationship among items or persons is characterised by the IRT model (Embreston & Reise, 2000). Local item and person independence simplifies mathematical computation and the probability of a complete item response matrix across all items for all persons in a test can be expressed as the product of the probability of a response to every item by every person as defined by Reckase (2009) as in equation 2.1.1.

$$p(U|\theta) = \prod_{j=1}^{J} \prod_{i=1}^{I} p(u_{ij}|\theta_j) \tag{2.1.1}$$

where θ_j is the ability parameter for person j, for $j = 1, ..., J$ and u_{ij} is person j's response to item i, $i = 1, ..., I$.

2.1.1 Local item independence

Local item independence means that the response of a subject to one item will be independent of his/her response to another item, conditional on latent ability, that is, if items are locally independent, they will not be correlated after conditioning on θ_j (DeMars, 2010). When the assumption of local item independence has been met, the probability of person j, $j = 1, ..., J$ with ability parameter $\mathbf{\Theta}$'s vector of responses to multiple items, $p(U = u|\theta)$ can be expressed in terms of the product of a response to

each item, $p(u_i|\theta_j)$ by the person is shown in equation 2.1.2 (Reckase, 2009; Le, 2013).

$$p(U|\theta) = \prod_{i=1}^{I} p(u_{ij}|\theta_j) \tag{2.1.2}$$

Local item independence could be violated when there are content clusters within the test, leading to local item dependency (LID). The information that is contained in an item within a cluster needs to be discounted and the extend to which it needs to be discounted depends on the level of LID (Ip, Smits & De Boeck, 2009).

2.1.2 Local person independence

Local person independence implies that one person's response to one item does not affect another person's response to the same item (Jiao & Zhang, 2014). The person independence assumption is violated due to person clustering such as paired and nested samples. The assumption of local independence is satisfied when the relationships among all items and all persons are fully characterized by an IRT model (Embretson & Reise, 2000) and mathematically this implies that the probability of a complete item response matrix across all items for all persons in a test can be expressed as the product of the probability of a response to every item by every person (Reckase, 2009). The probability of a response vector to a specific item i by multiple persons with ability parameters θ_j in the vector $\mathbf{\Theta}$ is as given in equation 2.1.3.

$$p(U = u_i|\theta) = \prod_{j=1}^{J} p(u_{ij}|\theta_j) \tag{2.1.3}$$

2.1.3 Dual dependency in item response data

Dual dependency in IRT exists when both item and person independence have been violated, that is, when items are clustered based on their shared contents and stimulus or wording, and respondents are sampled from clustered samples such as schools, gender and ethnic groups. In such circumstances, the respondents for subjects from one cluster can be related as respondents that are nested within a hierarchy tend to perform in a more similar way than respondents in other clusters. For example, students taught by one teacher (in one class or school) tend to give similar responses to

test questions than those students in other classes or schools and households from the same ethnic/cultural/communal groups tend to consume similar types of foods than houses from other groups.

2.2 The Rasch model and its extensions

The IRT offers a set of mathematical techniques for modeling latent traits, posing forms of representing the relationship between the likelihood of an individual responding correctly to a given item, his latent trait level,"ability", and characteristics (parameters) of the items in the field under study. The relationship between examinee's ability and the probability that examinee endorses a particular response is expressed by the Item Characteristic Curve (ICC) which is usually modeled on a logistic or normal cumulative distribution function and is considered to be a monotonically increasing function. The one-parameter logistic (1PL) IRT measurement model commonly known as the Rasch model (Rasch, 1960) provides a theoretical base and a set of statistical tools to assess the suitability of a set of binary /dichotomous test items for scale construction, creates a scale from the items, and compares performance of a scale in various populations and test contexts. The binary response Rasch (1PL) model shown in equation 2.2.4.

$$P(U_{ij} = 1) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \tag{2.2.4}$$

where 1 indicates correct response and 0 indicate incorrect response, θ_j is the ability (latent trait) of subject j, b_i is the item difficulty parameter for item i. The probability of an item response to an item depends on the person's ability and item difficulty. The Rasch model assume unidimensionality of latent traits, monotonicity of the item response function, local independence of persons and items, homogeneity of items and absence of differential item functioning (DIF). The Rasch model also makes a strong assumption that each item is equally discriminating, implying that the item location can be clearly interpreted as "severity" and a person scoring higher displays more severity levels and the relative severity of items is identical for all persons (Raudenbush et al., 2003).

The model has been extended to a 2-parameter logistic (2PL) model which assumes a random slope where the item ability to discriminate respondents according to their latent ability varies at random. In order to maintain the increasing monotonicity of the ICC, the slope parameter (a_i) is set to be positive, where a_i is the slope of the curve when $\theta = b_i$. Larger values of a_i indicate better discrimination ability for item i and items depicting better discrimination prowess provide more information about the individual abilities. Rasch proposed a mixed model for polytomous data in 1961 which was a predecessor for a family of IRT models for polytomous data including the Partial Credit Model (PCM:Masters, 1982) which assumes equal item discrimination ability. The PCM is quite popular in assessing contexts due to its parsimonious nature. Because the PCM allows for a relatively small number of estimates per set of items, sample sizes as small as 300 return stable item and trait estimation (de Ayala, 2009).

The PCM has been extended to a Generalised Partial Credit Model (GPCM: Muraki, 1992) by assuming a random discrimination / slope. A normal prior is almost always assumed for the ability priors in the item response models while the gamma, log normal or truncated normal distributions are usually assumed for the discrimination parameter so that it is kept positive. The GPCM for polytomous responses is illustrated in equation 2.2.5.

$$P(u_{ij}|\theta) = \frac{exp(\sum_{k=1}^{K}[a_i(\theta_j - b_{ik})])}{1 + \sum_{r=1}^{R}(exp(\sum_{k=1}^{K}[a_i(\theta_j - b_{ik})]))} \tag{2.2.5}$$

where step is denoted by $r = 1, 2, ..., R$ and $k = 1, 2, ..., K$ represents the score category.

However, as highlighted in section above, the LID assumptions can be violated if the respondents are nested within the sub-populations they were sampled from and if the test items are sub-tests within a test, referred to as "*testlets*" or "*test sets*" (Bradlow et al., 1999; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Paek et al., 2009).

2.3 Models for local item dependency

Several researches have been conducted to study the effects of testlets and group dependence on ability and item parameter recovery, ICC, test information and test reliability separately. How to score and analyse tests involving sub-tests within a test has been an important area of research (Kogar & Keleciolu, 2017) as it violates one of the important assumptions of IRT modeling. Several models have been proposed to deal with testlets within a test. If two or more items are correlated, test developers could choose to drop one of the items from the test, combine the items into a single item and use polytomous response models or use a testlet model that allows groups of items to be measured as secondary dimensions (DeMars, 2010). As such, testlet data have been modelled with score-based polytomous IRT model such as the graded response model (GRM: Samejima, 1969), rating scale model (RSM: Andrich, 1978), or the partial credit model (PCM: Masters, 1982) where testlet are converted to polytomous items (Thissen et al., 1989; Wainer & Lewis, 1990), polytomous logistic regression, where each testlet with a set of m questions is treated as an item with a total score m, using the bi-factor model (DeMars, 2006; Kogar & Kelecioglu, 2017).

Bradlow et al. (1999) argue that the polytomous item approach has limitations in that some amount of information about item response patterns is lost since the score is aggregated to the testlet level and their application to polytomous response items is compromised. In addition, combining items into a polytomous item would be easier for binary items but would be rather less practical for polytomous items. Testlet response model scholars proposed incorporating the testlet structure (Bradlow et al., 1999; Rosenbaum, 1988; Waner & Kiely, 1987; Paek et al., 2009) into the model thereby preserving the exact item response patterns. Other models for analysing testlets are bi-factor models (De Mars, 2006; Cho, Cohen & Kim, 2014; Kogar & Keleciolu, 2017) and the testlet effect model (Rijmen, 2009) which is a restricted state of the bi-factor model. Zhang (2010) propounded that the use of polytomous models is efficient for

small testlet and non-adaptive tests and are more stable than the testlet based models when a large number of testlets is included in a test. However, the polytomous and bi-factor models are beyond the scope of this study. Some of the advancement in testlet based models used to handle item dependence are given in sections below.

2.3.1 The Rasch testlet model

Proponents of models to handle testlet effects argue that the existence of testlets in a test can reduce the accuracy of parameters estimation, resulting in inaccurate inferences. Testlet models include random effect parameters added to model the local dependency among items within the same testlet. Wainer and Wang (2000) proposed a testlet model and Wang and Wilson (2005) proposed a Rasch testlet model which combine features of the testlet model and Rasch model making it more desirable as the Rasch model has observable sufficient statistics for model parameters, relatively small sample size requirement and no distributional assumptions for the parameters is necessary as parameters are considered fixed. The Rasch testlet model, proposed by Wang and Wilson (2005) is illustrated in equation 2.3.6.

$$P(U_{ij} = 1) = \frac{e^{(\theta_j - b_i + \gamma_{jd(i)})}}{1 + e^{(\theta_j - b_i + \gamma_{jd(i)})}} \tag{2.3.6}$$

With the parameters as in equation 2.2.4 and $\gamma_{jd(i)}$ being the testlet d effect on person j. Wang, Bradlow and Wainer (2002) and Wainer et al. (2000) extended the Rasch model to the Rasch testlet model using the Bayesian approach whereas Wang and Wilson (2005) used the non-Bayesian method. These approaches assumed the person ability and testlet effect parameters to be independent and normally distributed parameters. Paek et al. (2009) extended the Rasch testlet model by considering the violation of independence of person parameters and testlet effect by suggesting a bivariate normal distribution for the person and item cluster parameters. Their model, the Multivariate Normal Extended Testlet Model showed equal or slightly better performance with regard to averaged root mean square errors (RMSE) compared to the

regular testlet model.

The researchers argue that the existence of testlets in a test can reduce the accuracy of estimation of model parameters. They discovered that in absence of LID to low LID, estimates of ICC observed from simulated data and estimates were very similar while the presence of medium to high levels of LID, the estimated ICC became slightly steeper than the true ICCs (Reese, 1995). Moreover, studies have shown that the estimation of ability parameters is not comprised much by the LID effects (Zhang, 2010; Jiao, Wang & He, 2013) although the standard errors are underestimated (Yen, 1993; Jiao & Zhang, 2014; Jiao et al., 2012; Eckes, 2014; Zhang, 2010; Wainer, 1995) leading to overestimation of the precision with which the proficiency levels are estimated (Ravand, 2015).

2.4 Models for local person dependency

Multilevel models have been designed to handle inter-dependencies among data points (Bryk & Raudenbush, 1992). Inter-dependencies among a set of items depicting local person dependence (LPD) can pause some problems when neglected. The heterogeneity between person groups can be taken into account by assuming that some of the model parameters follow some random distribution over the population of clusters (Rijmen et al., 2003), and hence the parameters are random variables while the models are random effects or mixed effects models where random units represent cluster effects. Mixed effects models are a collection of statistical tools that are well suited for analysing clustered data. Adams, Wilson and Wu (1997), Kamata (2001), Mislevy and Bock (1989) have treated IRT models as logistic mixed models while other scholars (Goldeinstein, 2001; Raudenbush & Bryk, 2002) have treated the models as hierarchical multilevel models. Differential Item functioning (DIF) models have been proposed to handle person groupings in IRT modelling. Adams, Wilson and Wu (1997) argued that the use of respondent level variables can lead to precision in the estimation of the

item and person parameters.

2.4.1 Multi-level IRT Models

Mislevy and Bock (1989) proposed a multilevel modelling framework where group and individual effects are combined in a hierarchical IRT model while Fox and Glas (2001) proposed a hierarchical model which entails multilevel regression model on the latent proficiency allowing for predictions on the individual level and group level. Fox (2013) provided the multi-stage IRT model for handling students drawn from different schools and the first level is given in equation 2.4.7.

$$P(U_{pji} = 1 : \theta_{pj}, a_i, b_i) = \Phi(a_i(\theta_{pj} - b_i)) \tag{2.4.7}$$

where $\Phi(.)$ represents the cumulative normal distribution function, a_i denotes the discrimination parameter, b_i the difficulty parameter while the latent variable θ_{pj} represents the students' ability, and U_{pji} denote the response of student p in school j to item i. The study concluded that multiple response modeling is useful in school effectiveness in detecting differences in response and abilities within and between schools, accounting for the nested structure of response and abilities. Leckie and Goldstein (2001) and Raudenbush et al. (2003) argue that if respondents are nested within the social settings, the IRT model should be multilevel to study the variation and co-variation of the latent variable at each level and multilevel or random effects models are key for analysis of hierarchical data including school effectiveness (Grilli & Rampichini, 2009).

2.4.2 Multiple Group IRT

Bock and Zimowski (1997) presented a multiple group IRT model (MGM) that included group-specific population distributions to handle the clustering of respondents into manifest groups that are known before hand, making use of a frequentist based non-parametric approach for estimation. Their model allows inferences to be made with respect to each of the sampled groups but not to some higher-level population across all groups. Their multiple group model has an additional set of population parameters they termed "multiple population parameters", characterising the latent population

distributions. They assumed the group-specific latent abilities to follow different symmetric normal distributions. However, Azevedo, Andrade, and Fox (2012) observed lack of normality in the latent traits for at least one of the person clusters, which might mean that the distribution is asymmetric, heavy tailed, or having kurtosis that deviates from the normal distribution (Santos, Avezedo & Bolfarine 2012). Azevedo, Andre and Fox explored the potentials of Markov Chain Monte Carlo (MCMC) estimation procedure and Bayesian fit tools for the MGM using a Gibbs sampling and the Metropolis-Hastings with the Gibbs sampling for non-conjugate priors. Fox and Glas (2001) assumes the population distribution represents a population of groups while-Bock and Zimowski framework focused on sampled groups.

Research has shown that latent traits departing from normality can lead to biased estimates and violation of such an assumptions has a negative effects on the accuracy of the estimates (Heargerty & Kurland, 2001, Grilli & Rampichni, 2000). To this effects, Santos et al. (2012) extended Bock and Zimowski (1997) model by modeling the ability parameters through group-specific skew-normal distributions under the centred parameterisation. Azevedo, Andrade and Fox (2012) generalized the multiple group IRT model following a Bayesian approach with item response functions allowed to be skewed probit, logit or log-log and multiple group ability parameters that can be represented by normal, skewed normal, student's t or skewed student t or any finite mixture of normal distributions by parameterising $l = 1, ...L$ mixtures of different response functions based on different cumulative distribution function $h = 1, ..., H$ with possibly different latent trait distributions across item groups. Their model is represented in equation 2.4.8.

$$P(U_{pji} = 1 : \theta_{pj}, \xi_{\mathbf{i}}, \omega) = \sum_{l=1}^{L} \prod_{h=1}^{H} F_{lh}(\eta_{pj}, \xi_{\mathbf{i}}, \omega)) \tag{2.4.8}$$

where $\theta_{pi}|\eta_{\mathbf{j}} \sim G(\eta_j)$, the cumulative distribution function F_{ih} has parameters ω and $G(\eta_{\mathbf{j}})$ represents a continuous population distribution function with parameters η_j for group j.

2.4.3 The Mixture IRT Models

The multilevel and multi-group models assume data were sampled from manifest group known before hand. On the other hand, the Mixture Rasch Model (MRM: Rost, 1990), mixture partial credit model (Rost, 1991) and the Mixture IRT Model (Bolt, Cohen & Wollack, 2001; Cho, Cohen & Kim, 2013; Choi, 2014) are latent models based on the assumption that respondents are drawn from unobservable latent clusters which cannot be observed directly and the Rasch model does not hold for the entire population but holds within these latent sub-populations of individuals. Members of the same latent class will have responses that have the same ordering of item difficulties while other classes will have responses that reflect a different ordering of item difficulties (Embreston, 2007). The latent class will have its own difficulty parameters for each of the item and the probability of a (correct) response for an individual is a mixture of the item response functions for each of the latent classes and the probability that the individual is in that latent class. The probability of a correct response will be a weighted sum of a correct response. The MRM (equation 2.4.9) does not require any assumptions about the ability distribution within the classes. It entails a large number of model parameters which can be reduced by imposing restrictions on the latent score distribution.

$$P(U_{jgi} = 1|\theta_{jg}, b_{ig}) = \sum_{g=1}^{G} \pi_g \frac{exp(a_{ig}(\theta_{jg} - b_{ig}))}{1 + exp(a_{ig}(\theta_{jg} - b_{ig}))} \tag{2.4.9}$$

with $\theta_{jg} = \mu_g, \sigma_g^2 \sim N(\mu_g; \sigma_g^2)$, where θ_{jg} is the ability of examinee j in latent class g, a_{ig} and b_{ig} are class specific item discrimination and difficulty parameters respectively and μ_g and σ_g^2 are class specific mean and variance respectively.

$$P(X = x|c) = \pi_{r|c} \frac{exp(\sum_{i=1}^{k} x_i \beta_{ic})}{\gamma exp(\beta_c)} \tag{2.4.10}$$

where r is the sum of raw scores, c is the latent class, π_c is the class size parameter, $\pi_{r|c}$ is the latent score probability, β_{ic} is the class specific item parameter. The number of components c, is not a model parameter but must be specified *a priori* or estimated by comparing the model fit under different number of classes (Rost, 1990). The rationale of

group membership is that it can be considered as a multinomial distribution consisting of a series of single draws on the g categories with mixture probabilities π_g (Congdon, 2003; McLachlan & Peel, 2000).

2.4.4 The Linear Logistic Test Model (LLTM)

The Linear Logistic Test Model (LLTM) proposed by Fischer (1973) and its various extensions (Rijmen & De Boeck, 2002) allows the direct assessment of the impact of item attributes on the difficulty of an item. Item attributes can provide more meaningful interpretation than can item-difficulty parameters (Ip, Smits & De Boeck, 2009). Further extensions on LLTM to the latent regression LLTM (Hartig et al., 2008; Desjardins & Bulut, 2018) the latent regression Rasch model (Zwinderman, 1991) and the Exploratory Item Response Theory Model (EIRM) for binary responses (Wilson, De Boeck & Carstensen, 2008) and for polytomous responses (Stanke & Bulut, 2019) and the locally dependent linear logistic test model (LID-LLTM: Ip, Smits & De Boeck, 2009) where both item and person parameters include explanatory properties to handle item and person clusters and their interactions. They reiterated that if person parameters are considered random, then it may have undesirable consequences if certain person properties are not taken into account where the assumed distribution no longer applies to the entire group but to subsets of persons who share common properties. Their studies concluded that ignoring LID may lead to biased estimates of item parameters and biased ability estimates and their associated standard errors.

2.5 Dual dependence models

Previous research has mainly dealt with LID and LPD separately. However, Jiao et al. (2012) proposed the multilevel testlet model for dual local dependence for dichotomous responses while Jiao and Zhang (2014) derived a multilevel testlet model for polytomous data to cater for dual dependence caused by person and item clusters by extending the GPCM. They compared their polytomous response "multilevel testlet model for testlet-based assessments with complex sampling designs" with the testlet PCM, the

multilevel PCM and the PCM in terms of parameter estimates. Their results indicated that ignoring item clustering effect produced higher total errors but did not have much impact on ability parameter estimates. On the other hand, ignoring person clustering effects yielded higher total errors in ability parameter estimates but did not have much effect on item parameter estimates. Moreover, their findings revealed that ignoring both clustering effects reduced the accuracy of item and ability parameter estimates. Their multilevel testlet model is illustrated in equation 2.5.11.

$$P_{jtigk}(X_{ji} = x/a_i; d_{ik}; \theta_{jg}; \delta_g; \gamma_{jt(i)}) = \frac{exp[\sum_{s=0}^{x} a_i(\theta_{jg} + \delta_g + \gamma_{jt(i)} - d_{ik})]}{\sum_{k=0}^{K} exp[\sum_{s=0}^{x} a_i(\theta_{jg} + \delta_g + \gamma_{jt(i)} - d_{ik})]} \tag{2.5.11}$$

where P_{jtigk} is the probability of person j with person-specific ability θ_{jg} in group g with a group effect of δ_g getting a score of X on item i with a step difficulty of d_{ik} and a discrimination parameter of a_i in item group t with an item group effect of $\gamma_{jt(i)}$. Person j is nested within group g. X represents any score from 0 to x. However, their model did not take into consideration the possibility of interaction between item and person clusters, giving a gap in knowledge for further research.

The person and item parameters have been assumed to be either fixed (Fox et al., 2006) or normally distributed random variables (Jiao & Zhang, 2014; Zhang, 2010)). On the other hand, a constant discrimination parameter is usually considered for model identifiability or the log-normal, truncated normal and gamma distributions are usually considered if the slope is assumed to be random. However, parametric models present quite strong assumptions about the mixing distributions, which might result in erroneous inference if the distribution is mis-specified (Grilli & Rampichni, 2009; Heargety & Kurland, 2001; Verbeke & Lesaffire, 1996), leading to severe consequences on properties of estimators. Non-parametric and semi-parametric models and models involving normal mixtures have been proposed to address the problem of making restrictive parametric assumptions on the random effects.

A review study conducted by Zhang (2010) has identified sample size, number of testlets within a test, testlet size, to be among factors that could influence testlet analysis in research designs. Mixed conclusions have been drawn on the effects of variation of these attributes on parameter estimation and test information and reliability. Some scholars argue that better psychometric properties are attained for fewer category options (see Lee, 2012) while others argue that scale reliability increases with response category options (Muniz et al., 2005; Lazano et al., 2008). In addition, some studies in literature assessed the effects of varying such factors when either items or persons are independent. Although items in a test may depict different contributions in categorising respondents according to their proficiency levels, researchers usually set item slopes to be invariant across items as a constraint to ensure model identifiability. The effects of varying such attributes in the presents of dual dependence effects may be worth investigating.

2.6 Non-parametric and semi-parametric models

The normality assumption for the trait distribution may not be met for the entire population, but in groups of respondents in the population clustered according to their proficiency levels. Rost (1990) proposed mixtures of Rasch models by mixing various logistic curves where the examinees are divided into subgroups based on membership attributes that cannot be observed directly and referred to such groups as latent classes or components. Rost's method requires that the number of latent clusters be known in advance and this can be determined using the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). He claimed that the values of item attributes differ in each cluster. After the probabilities of latent class membership have been estimated, obtaining the sums of the item response probabilities of each latent cluster membership is equivalent to mixing different ICC patterns. However, neither the latent classes nor the group membership maybe known prior to analysis and may

need to be inferred from the data. Sethumaran (1994) proved that any distribution function can be expressed by mixing distributions that belong to the same family of distributions, implying that ability distribution for clustered populations be expressed using the Dirichlet Process Mixture (DPM) priors.

On the other hand, Duncan and McEachern (2008) suggested two non-parametric IRT models by applying the DP priors to the entire curve and not only individual parameters and a parametric model with DP ability parameters. However, they applied their DP mixtures on 2PL and 3PL logistic models, and not hierarchical models. No estimate of difficulty or ability was estimated at component level and the component was for person clusters and not item testlets. Duncan and McEachern found that the non-parametric curve model produces significantly different item characteristic curves for a few of the items and that the corresponding ability estimates also change substantially for some individuals.

2.7 The Dirichlet Process Priors

In hierarchical models, there are questions of sensitivity of inferences to assumed distributions such as normal and gamma distributions for higher stage priors. The distribution of higher order priors for random effects are often uncertain and failure to account for such uncertainty may jeopardize the precision attached to posterior inferences (Congdon, 2006). Congdon also went on to say that inference can be affected by multi-modality in random effects due to inconsistencies with the assumed higher level priors. Instead of assuming a normal distribution for the ability parameter θ_j, the Dirichlet Process (DP: Ferguson, 1973) approach lets the form of higher stage density G itself be uncertain (West et al., 1994) and offers an approach which avoids parameter assumptions and is less impeded about the number of classes (Congdon, 2003).

The DP is a finite dimensional generalisation of the Dirichlet distribution to continuous spaces (Cho, Cohen & Kim, 2014) that can be used to set a prior on unknown distributions (Gelman et al., 2003). There must exist a random probability measure G on $(\Omega, \mathscr{B})$, such that for any partition $B_1, ...B_M$ on the support of G_0 the vector of probabilities $\{G(B_1), ..., G(B_M)\}$ follows a Dirichlet distribution with parameter vector $\{\alpha G_0(B_1), ..., \alpha G_0(B_M)\}$.

The probability measure G is referred to as the DP if $G \sim DP(\alpha G_0)$, where $\alpha > 0$ is a scalar precision or concentration parameter governing the concentration of prior for G about the mean G_0, where G_0 a baseline probability measure also on $(\Omega, \mathscr{B})$, the prior expectation of G. As α increases, the precision or concentration around the baseline prior increases, whereas small α tends to result in relatively large deviations from the form assumed by G_0 and will produce DP(αG_0) a distribution with a single possible value when $\alpha \to 0$ (Chen, 2015). The case $\alpha \to \infty$ means DP prior becomes nearly continuous (Chen, 2015) and hence become equivalent to a parametric model with G_0 known (Congdon, 2006), creating a distribution with potentially an infinite number of components, making it suitable for mixture models with an unspecified number of mixing components. The baseline G_0 is commonly chosen to correspond to a parametric model such as a Gaussian (Gelman et al., 2003). The definition of the DP and properties of the Dirichlet distribution imply that:

$$G(B) \sim Beta(\alpha G_0(B), \alpha(1 - G_0(B)) \text{ for all } B \in \mathscr{B} \tag{2.7.12}$$

so that the marginal random probability assigned to any subset B of the support distribution is simply a Beta distribution.

Suppose the vector y_j , $j = 1, ..., J$ is drawn from a distribution with unknown parameters θ_j, ψ_j

$$f(y_j|\theta_j, \psi_j)$$

and suppose there is greater uncertainty about the prior for parameters θ_j than for

parameters ψ_j (Escobar & West, 1998). One may adopt a DP prior for the θ_j, but a conventional parametric prior for ψ_i. Under the DP prior, the baseline prior G_0 is assumed from which candidate values for θ_i are drawn. So instead of a prior $\theta_j \sim G(\theta_j|\gamma)$ with G a known density and γ a hyperparameter, the uncertainty about the form of the prior is represented by introducing an extra step in the hierarchical specification:

$$\theta_j|G \sim G \tag{2.7.13}$$

$$G|\alpha, \gamma \sim DP(\alpha, G_0) \tag{2.7.14}$$

where G_0 has hyper-parameters γ.

Equation 2.7.13 states that the parameter θ arises from a distribution G but G itself arises from a distribution of distributions (equation 2.7.14), that is, it is a Dirichlet Process. The baseline distribution G_0 then serves as the initial guess of the distribution of θ and the concentration parameter α determines the *a priori* confidence in G_0. G refers to the measure and its distribution. Figure 2.1 gives an illustration of the Dirichlet Process.

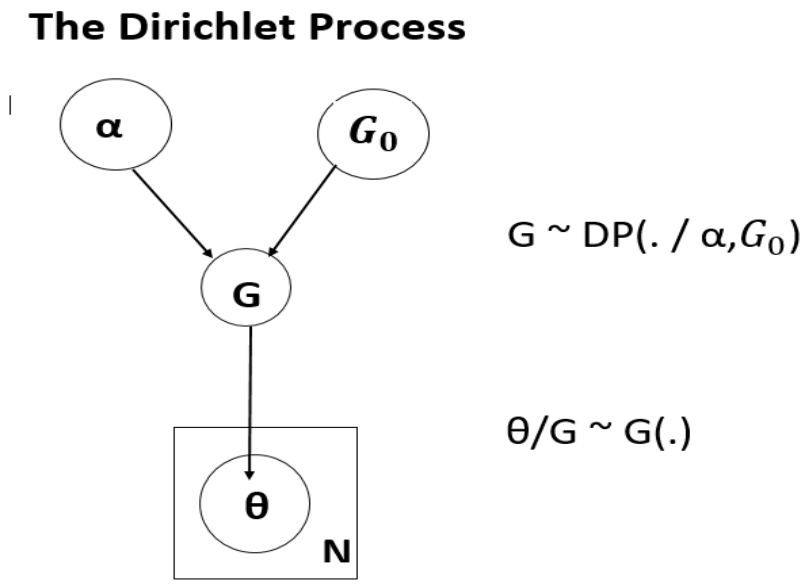


Figure 2.1: Diagrammatical illustration of the Dirichlet Process

The Dirichlet distribution is a conjugate prior for the multinomial distribution (Cho,

Cohen, Kim, 2011), thus making it an appropriate distribution for modelling the membership components of latent groups in mixture models which are themselves considered as multinomial distributions consisting of series of draws from the g categories (Congdon, 2003; McLachlan & Peel, 2000). However, the realisations from the DP are discrete distributions, hence $G \sim DP(\alpha G_0)$ implies that G will be atomic, with non-zero weights only on a set of atoms and will not have a continuous density on the real line. Despite all that, the DP has been useful in developing flexible models for a wide variety of problems. Contrary to this, Muthukumarana (2010) considers this to be an advantage, attributing it to the DP clustering ability, the clustering mechanism is implicitly carried out in the model framework and the number of component clusters need not be specified in advance as the DP process can determine the number of components (Miyazaki & Hoshino, 2009).

2.8 Representation of the DP Mixture model

The DP is currently one of the most popular Bayesian non-parametric models and can be described as a distribution over probability distributions, implying that the DP is a distribution and a random variable drawn from it will be another probability distribution (Blei et al., 2006). There are several ways to implement a DP prior. Following Sethuraman (1994), one way to generate the DP prior is to regard the θ_j as iid with density function $q(.)$ which is an infinite mixture of point masses or continuous densities (Ohlssen et al., 2007). This is also known as the 'constructive definition' of the Dirichlet Process (Walker et al., 1999). If G_0 consists of a continuous density f, then the DP forms a mixture of continuous densities (equation 2.8.15).

$$q(\theta_j) = \sum_{l=1}^{\infty} \pi_l f(\theta_j|\gamma) \tag{2.8.15}$$

where π is the component proportion. This structure is known as a mixed DP (Walker et al., 1999) and overcomes certain limitations of the original DP of Ferguson (1973). For example, a DP mixture with normal base densities would be as in equation 2.8.16.

$$q(\theta_j) = \sum_{l=1}^{\infty} \pi_l N(\theta_i|\mu_l, \psi_l). \tag{2.8.16}$$

The DP mixture models have often been treated as being potentially infinite dimensional by Ishwaran and Zarepour (2000, 2002) and Ishwaran and James (2001) demonstrated that a finite dimensional model with limited number of mixture components will accurately approximate an infinite dimensional modes if the number of mixing is large enough and thus suggested a finite dimensional DP with the dimension truncated at L components such that:

$$q(\theta_j) = \sum_{l=1}^{L} \pi_l N(\theta_j|\mu_l, \psi_l) \tag{2.8.17}$$

and $\sum_{l=1} \pi_l = 1$. This leads to an approximate or truncated DP which may be denoted:

$$\theta_j|G \sim G \tag{2.8.18}$$

$$G|L, \alpha, \gamma \sim TDP(\alpha, G_0). \tag{2.8.19}$$

If the distribution of G_0 has full support on the real line, then for every $\varepsilon > 0$, and distribution G^* on the real line, the prior probability that G lies in the neighbourhood $N_\varepsilon(G^*)$ is positive. This statement is true for any neighbourhood defined by any metric that generates the weak topology. The diagramatical illustration of the Dirichlet Process Mixture is highlighted in Figure 2.2.

There are several ways to implement the DP priors. The Polya characterisation of the DP (Escobar & West, 1995), the collapsed cluster sampling (Duncan & MacEchern, 2008). Ishwaran and James (2002) detail the usually close accuracy of this approximation to the infinite DP for typical α and L values. The most appropriate value θ_l^* for case j is then selected using a Dirichlet vector of length L with probabilities π_l for each value determined by the precision parameter α. The mixture weights π_j are constructed by 'stick-breaking' (Ishwaran & Zarepour, 2000), a constructive representation of the DP process which is useful in obtaining further insights into properties of the DP and as a stepping stone for generalisations.

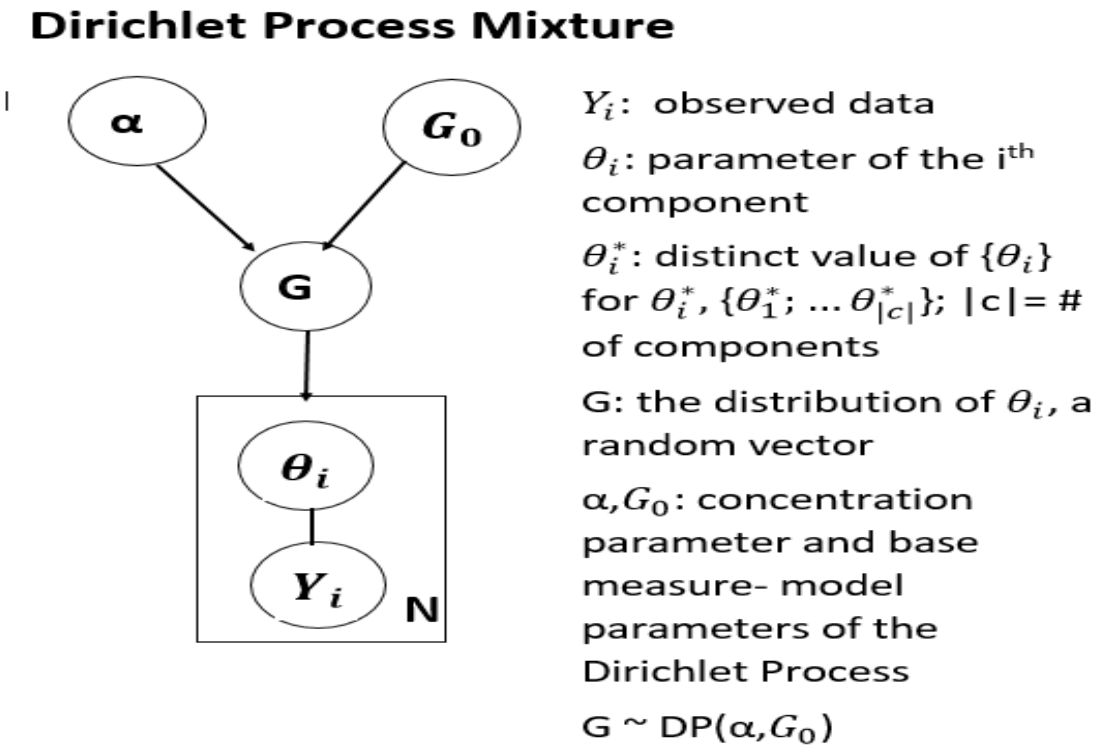


Figure 2.2: Diagrammatical illustration of the Dirichlet Process Mixture

2.8.1 The stick-breaking process

The stick breaking procedure starts with a stick of length 1 unit and at each step, we break off a piece of the remaining stick v_k and use the length of this piece as π_k. The stick-breaking allows one to induce $G \sim DP(\alpha G_0)$ by letting $V_L = 1$ and draw $L-1$ beta variables:

$$V_k \sim Be(1, \alpha) k = 1, ..., L$$

and set

$$\begin{aligned}
\pi_1 &= V_1 \\
\pi_2 &= V_2(1 - V_1) \\
\pi_3 &= V_3(1 - V_2)(1 - V_1) \\
&\vdots \quad \vdots \\
\pi_L &= V_L(1 - V_{L-1})(1 - V_{L-2})(1 - V_1)
\end{aligned} \tag{2.8.20}$$

Thus we set $V_L = 1$ to ensure that $\sum_{k=1}^{L} \pi_k = 1$ and draw $L-1$ beta variables independently drawn from, $V_k \sim Be(1, \alpha) k = 1, ..., L$.

2.9 Bayesian Estimation Methods

The Bayesian statistical technique in IRT is not new. Bradlow et al. (1999) proposed the Bayesian random effects IRT model while Wang et al. (2000) and Wainer et al. (2000) and Zhang (2010) applied the Bayesian approach to model the Rasch testlet models. In addition, Jiao and Zhang (2014) applied the Bayesian estimation technique to estimate the multilevel testlet model accounting for dual local item and person dependence effects. The non-parametric Bayesian approach was implemented by Karabatsos and Walker (2012) and Karabatsos (2015) to model the Bayesian non-parameteric model using a menu driven software developed by Karabatsos (2015).

In Bayesian statistical modeling, parameters are not viewed as fixed quantities but rather are random quantities that are assumed to emanate from probability distributions. Prior distributions need to be specified for each model parameter. Bayesian parameter estimation is based on the posterior distribution of the model parameters, which is proportional to the product of the likelihood and prior distribution, where the prior distribution quantifies uncertainty about the parameters before data are observed while the posterior distribution expresses the uncertainty about the parameter after observing the data. The objective of Bayesian inference is to maintain a full posterior probability distribution over a set of random variables.

Bayesian inference is based on the idea that one can update the prior knowledge about the parameters of interest given the information obtained from the observed data. For example, if $p(\theta)$ (prior distribution) represents prior knowledge about parameter vector $\mathbf{\Theta}$, which expresses all uncertainty before seeing the data, the data Y has a density function $p(y|\theta)$, the likelihood function which relates all variables into a "full probability model". Y is known, (observed data) so we should condition it, then the

posterior distribution becomes:

$$p(\theta|y) = \frac{p(y|\theta).p(\theta)}{p(y)} \propto p(y|\theta).p(\theta) \tag{2.9.21}$$

The posterior distribution $p(\theta|y)$ expresses all uncertainty about Θ after seeing the data and $p(y)$ is the inverse normalising constant given by

$$p(y) = \int P(y|\theta)p(\theta)d(\theta) \tag{2.9.22}$$

where θ can be a scalar or vector of parameters and y is the vector of observed data. Bayesian inference is based on the posterior distribution of the parameters of interest based on the entire posterior distribution (not point estimates) that maximise the likelihood of the parameters of interest. After drawing samples from the posterior distribution, the posterior mean, standard deviations and other summaries can be obtained and are more informative than point estimates (Burgos, 2011).

A Bayesian approach to multilevel IRT provides additional features as it supports a flexible way of incorporating prior knowledge to account for different sources of uncertainty, complex dependencies and other sources of information at separate level with subsequent inference possible at each of these levels from the posterior distribution of parameters, making it possible to handle sampling designs with complex dependency structures. Researcher continue to pursue studies on non-parametric IRT modeling which offers more flexible models for item response functions and provide better alternatives to model fit than the parametric approaches (Duncan & MacEachern, 2008) after Tsutakawa (1992) did early work on Bayesian non-parametric ability parameters. However, he only considered discrete distribution with a small, finite number of support points which cannot adequately approximate continuous models for ability parameters (Duncan & MacEachern, 2008).

2.9.1 Conjugate Priors

In Bayesian framework, a prior distribution needs to be specified for each of the model parameters. In simple models, integration in equation 2.9.22 can sometimes be avoided

by choosing conjugate priors such that the prior and posterior distributions are from the same family of distributions, making the integrals tractable. A prior distribution that is a member of distributional family D with parameter α is conjugate to the distribution $f(y|\theta)$ if the resulting posterior distribution $f(\theta|y)$ is also a member of the same distributional family. Therefore, if $\theta \sim D(\alpha)$, then $\theta|y \sim D(\tilde{\alpha})$, where α and $\tilde{\alpha}$ are the prior and posterior parameters of D respectively. Usually the posterior parameters are expressed as the weighted mean of the prior parameters and the maximum likelihood estimators.

Conjugate priors are mathematically convenient, but they do not always exist for all likelihoods. In more complex models, computation of integrals maybe be very difficult or sometimes impossible. In addition, non-conjugate priors are computationally challenging, but possible using Markov Chain Monte Carlo (MCMC).

2.9.2 Non-informative priors

If prior information about a parameter is known, it should be incorporated into the prior density. If no prior information is known about the parameter, non-informative priors with minimal influence on the parameter should be selected. When the prior distribution is non-informative, the range of uncertainty should be wider than the range of reasonable values of the parameters (Gelman & Hill, 2007). When the prior is reliable, the use of an informative prior with smaller uncertainty may facilitate the estimation of the posterior. When a non-informative prior is supplied, the posterior is estimated by relying mainly on the data (Jiao et al., 2012).

Various non-informative priors have been explored for variance parameters. Wang and Wilson (2005) and Bradlow et al. (1999) used the standard normal priors for θ, $\gamma_{nd(i)} \sim N(0, \sigma^2_{\gamma_{nd(i)}})$, $a_i \sim N(\mu_a, \sigma^2_{a_i})$, $b_i \sim N(\mu_b, \sigma^2_{b_i})$, $c_i \sim N(\mu_c, \sigma^2_{c_i})$, and $\mu_a \sim N(0, V_a)$, $\mu_b \sim N(0, V_b)$, $\mu_c \sim N(0, V_c)$, and $V_a^{-1}, V_b^{-1}, V_c^{-1} = 0$ reflecting lack of information about these parameters. They set all variance priors to the inverse $\chi^2_{g_z}$ is

the degrees of freedom they set to 0.05 for all distributions to reflect small amount of information.

Gelman and Hill (2007) reiterated that some non-informative prior distributions may unduly affect inferences when the number of groups is small and group level variability is close to 0 (zero) in multilevel models. They explored 3 different types of priors for variance estimation and concluded that the inverse-gamma ($\alpha = 0.001$, $\beta = 0.001$) prior distribution distorted the posterior distribution. On the other hand, the inverse-gamma ($\alpha = 1$, $\beta = 1$) prior distribution generally made the prior concentrate in the range (0.5, 5) and closely matched the prior distribution. In addition, they found the uniform distribution not to constrain the posterior inference. In general, they recommended a uniform prior with a wide range such as (0, 100) for the standard deviation. Gelman and Hill did not recommend the use of the inverse-gamma prior as non-informative prior but rather as a proper prior distribution when the group sample size is small and the group variance is near 0 (zero). The inverse-gamma prior is preferred by researchers due to its conditional conjugacy, resulting in cleaner mathematical properties (Jial et al., 2012). Jiao et al. (2012) experimented with all three options provided by Gelman and Hill (2007) ($\alpha = 0.001$, $\beta = 0.001$, $\alpha = 1$, $\beta = 1$, uniform(0,100)) as prior distributions for the variance parameter and the inverse-gamma ($\alpha = 1, \beta = 1$) was more suitable for their model. For their simulation data, the group variance was substantially different from 0 (zero).

2.10 Markov Chain Monte Carlo (MCMC)

In the case of complex posteriors, the integrals cannot be evaluated analytically and simulation procedures are often used to sample from the posterior distribution. The most widely used sampling methods are importance sampling and MCMC. For high dimensional problems, Markov Chains are recommended over importance sampling. A Markov Chain (MC) is a stochastic process in which future states are independent of

past states given the present state, that is, they depict the Markov property:

$$P(\theta^{(t+1)}|\theta^{(1)}, \theta^{(2)}, \theta^{(t)}) = P(\theta^{(t+1)}|\theta^{(t)}) \tag{2.10.23}$$

The MC is a bunch of draws from θ that is slightly dependent on the previous ones and it wanders around the parameter space, remembering only where it has been in the last state. MCMC is a class of methods in which one can simulate draws that are slightly dependent and are approximately from a distribution that has the posterior distribution as its equilibrium distribution (Muthukumarana, 2010). The draws are then used to calculate the quantities of interest for the posterior distribution. The underlying logic of MCMC sampling is that one can estimate any desired expectation by ergodic averages, that is, one can compute any statistic of a prior distribution from simulated samples from the distribution. The essential idea of iterative simulation is to draw values of a random variable x from a sequence of distributions that converge to the desired target distribution $f(x)$ of x as iterations continue. The MCMC procedure makes use of the prior information of the model parameters and the data to obtain the posterior estimates of the parameters.

$$E[f(s)]_P \approx \frac{1}{N}\sum_{i=1}^{N} f(s^{(i)}) \tag{2.10.24}$$

where P is the posterior distribution of interest, $f(s)$ is the desired expectation and $f(s^{(i)})$ is the i^{th} simulated sample from P. For example, the posterior mean:

$$E[x]_P = \frac{1}{N}\sum_{i=1}^{N} x^{(i)} \tag{2.10.25}$$

can be estimated. Some of the algorithms to implement the MCMC sampling technique are the Metropolis Hastings (M-H) and the Gibbs Sampling. The development of MCMC techniques has led to widespread use of DP priors across main domains and the most common use is for data clustering and identification of heterogeneous groups in populations.

2.10.1 The Metropolis-Hastings algorithm

Suppose $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(k-1)}$ are previously generated samples, the M-H procedure generates θ^* from the proposal density $q(\theta, \theta^{(k-1)})$ which may depend on $\theta^{(k-1)}$, the sampled

value is accepted with probability:

$$min\left(1, \frac{q(\theta^{(k-1)}, \theta^*)p(\theta^*|y))}{q(\theta^*, \theta^{(k-1)})p(\theta^{(k-1)}|y))}\right)$$

If θ^* is not accepted, then $\theta^{(k)} = \theta^{(k-1)}$. The rate at which the new values are accepted is called the acceptance rate. For large k, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(k)}$ is approximately realisation from the posterior distribution. The M-H algorithm requires an initial value $\theta^{(0)}$ in order to start simulation. Initial values far from the range covered by the posterior distribution often lead to chains that take long to converge to the equilibrium (posterior) distribution.

2.10.2 The Gibbs Sampling

The Gibbs Sampling (Geman & Geman, 1994; Draper, 2008) is a special case of the M-H algorithm which generates MC by sampling from the full conditional distribution indirectly without calculating the density by turning multivariate problems into a sequence of lower dimensional problems. The conditional distributions are called "full conditionals" as they condition on all the other parameters. Suppose we have random variables X_1, X_2, X_3 with a joint distribution $f(x_1, x_2, x_3)$, and we are interested in obtaining the characteristics of their marginal densities. The most natural way to obtain the marginal density $f(x_1)$ of X_1 by taking partial integrals:

$$f(x_1) = \int\int f(x_1, x_2, x_3)dx_2dx_3 \tag{2.10.26}$$

However, such an integral can be extremely difficult to perform either analytically or numerically. In such cases, the Gibbs Sampling provides an alternative method for obtaining the marginal distributions without having to calculate the density from integrals, through iterative sampling. The Gibbs Sampling generate a MC of random variables from the marginal distribution indirectly by sweeping through each variable or block of variables to sample from its conditional distribution with the remaining variables fixed to their current variables. X_1, X_2, X_3 are set to their initial values $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$ (often generated from the prior). At the i^{th} iteration, sample $x_1^{(i)} \sim$

$P(X_1 = x_1|X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)})$, sample $x_2^{(i)} \sim P(X_2 = x_2|X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)})$, and sample $x_3^{(i)} \sim P(X_3 = x_3|X_1 = x_1^{(i)}, X_2 = x_2^{(i)})$. The process continues until the distribution of the $x_k^{(i)}$ converges to a distribution similar to the posterior distribution $f(x_k)$ as $i \to \infty$. The advent of the Gibbs Sampling made estimation based on the DP tractable (Duncan & MacEachern, 2008). Albert (1992) proposed the Gibbs Sampler to fit the 3PL and 2PL respectively and to compile posterior summaries for the ability parameters while Patz and Junker (1999) suggested the use of the M-H within the Gibbs Sampling for the same model. Duncan and MacEachern (2008) implemented the Gibbs sampler that makes use of the Polya Urn scheme to obtain samples from the joint posterior distribution.

2.11 Model identification

A model is identified if its parameter values uniquely determine the probability distribution of the data and data uniquely determines the parameter values (Huang, 2005; Allman, Matias & Rhodes, 2009; de Araujo et al., 2009). Non-identifiability occurs when more than one set of parameter values yields the same value in the likelihood value and probability distribution given by the model (Cho, Cohen & Kim 2013), and is a particular problem for latent class models (such as IRT) as latent variables are not directly observed (de Araujo, 2009). Non-identifiability is usually caused by model mis-specification and the number of unique parameters exceeds the number of independent pieces of observed information and parameter estimates are not consistent as the likelihood function does not provide information to differentiate them.

In simulation studies, parameter identifiability can be assessed by checking whether we are able to recover the generating values. Non-identifiability can be rectified by constraining parameters (de Araujo et al., 2009) and placing restrictions are placed on the distributions that are mixed in non-parametric settings (Allman, Matias & Rhodes, 2009). IRT constraints include (1) fixing the mean θ or mean difficulty to 0 (zero) (2)

to allow the item and ability parameter to float and adjust the new floating parameter to a new defined value, and replace the model parameters with the new adjusted parameter to make the model identifiable (Gelman & Hill, 2007; Cho, Cohen & Kim, 2013). In multilevel IRT where the scale of the latent dimension is made up of several variance components, fixing one of these variances is not practical. Fox and Glas (2001) suggested imposing the identifying restrictions on the ability parameters and fixed the discrimination parameters to 1 and one difficulty parameter to 0. Jiao et al. (2012) and Jiao and Zhang (2014) constrained the mean item difficulty to 0 and the discrimination parameter to 1.

Model identification must be established before the model can be used for inference (Huang, 2005; Allman, Matias & Rhodes, 2009). However, identifiability constraints need not be imposed before any simulation takes place (Jasra, 2005) and constraints can be imposed before or after simulation with no adverse effect on simulation. Overfitting occurs when the wrong model fits the data better than the actual generating process. An overfitted model is a statistical model that contains unnecessarily more parameters than can be justified by the data. The kernel of overfitting is to have mistakenly extracted some of the residual variation (unexplained variation) as if that variation represented underlying model structure. To avoid over-parametrisation, we need to arbitrarily set to zero the effects of one level for each factor. The effects of the remaining levels then get the interpretation of differences relative to the base level.

2.12 Label switching

If exchangeable priors are placed upon parameters of a mixture model, then the resulting posterior distribution will be invariant to permutations in the labelling of the parameters (Stephens, 2000). As a result, the posterior distribution will be identical for each of the mixture components. Consequently, during MCMC simulation, the sample encounters the symmetries of the posterior distribution and the labels switches,

resulting in erroneous inference from the posterior ergodic summaries. Label switching significantly increases the effort required to produce satisfactory Bayesian analysis but it is a pre-requisite in an MCMC convergence and hence must be addressed (Jasra, 2005; Cho, Cohen & Kim 2013).

Label switching is potentially a serious problem in Bayesian estimation of mixture IRT models and it encompasses two types, across iterations within a single chain or when latent classes switch over replications (Cho, Cohen & Kim, 2013) or when the labels switches in multiple chains of a simulation, for example, when a high ability in one chain is a lower ability in another. If MCMC sampling takes place from an unconstrained prior with G groups, then the parameter space has $G!$ sub-spaces corresponding to different ways labelling the different states (Congdon, 2003) and the MCMC moves between these regimes in the parameter space. If these regimes are not sufficiently informed by the prior or the data, then label switching occurs.

Label switching can be detected when distinct jumps occur in the traces of the parameter and when the density of the parameter has multiple modes (Stephens, 2000). If the densities are unimodal then there is unique labelling and the presence of multiple modes signifies label switching in the parameter and interpretation of the posterior of the parameters is compromised. Solutions to the label switching include placing restrictive constraints on the priors distribution, relabelling algorithms, label invariant loss functions and random permutations sampling and maximum a posteriori (MAP) estimation.

2.13 Conclusion

The literature has shown the need to employ multilevel statistical modeling to take cognizance of item and person clustering effects in IRT modeling so as to obtain estimates of group specific parameters. Many scholars have handled local item and local person

dependency separately. However, in psychometric and biomedical studies, tests mainly comprise of testlets administered to persons sampled from clustered populations. In this respect, there is the need to come up with models that address dual dependency. Although much research conducted on the topic has treated these random effects as normally distributed random variables, in some circumstances these distribution may not hold, leading to biased estimates or increased errors in inference. As such, there is need to incorporate non-parametric models as these assume distributions other than the normal with its rigid parametric assumptions. In addition, sample size, testlet size, number of category options have been seen to significantly impact on parameter estimation and test reliability.

Chapter 3

Model Development

The current research proposes a Bayesian non-parametric approach to model dual dependency in item response functions considering the possibility of an interaction between group and person clusters. This will enable relative comparison of variance of latent traits for person categories and individual persons, providing a mechanism of accounting for correlation structures among the clustered observations. However, the model assumes no DIF, that is, the items have similar difficulty and discrimination ability per group and hence item parameters do not vary per group.

Since literature has shown that there might exist ability parameters that are not necessarily normally distributed, the research proposes a non-parametric model where the ability parameter is assumed to be Dirichlet Process Mixtures of normal distributions. The proposition is in line with Duncan and McEachern (2008) mixing who applied the Dirichlet Process Mixture ability parameters in the 2-parameter logistic (2Pl) IRT model. Sethumaran (1994) proved that any distribution can be expressed by mixing distributions that belong to the same family of distributions. This implies that any ability parameter distribution can be expressed as a mixture of normals using DP priors.

3.1 The Proposed Hierarchical Dirichlet Process Mixture Model

The proposed model is based on extending Jiao and Zhang (2014) multilevel testlet model for dual dependency in polytomous response items. They utilsed a constant discriminant parameter for identifiability and normal ability distribution. However, there are some ability parameters which may not necessarily be normally distributed. The Dirichlet Process Mixture allow the researcher to weaken the parametric assumptions to semi-parametric and non-parametric frameworks. The DP process prior supports strictly discrete distributions, providing a clustering mechanism thereby making it suitable for modeling clustered respondents. Moreover, clustering takes place as part of the model and observed data decides the clustering structure (Muthukumarana, 2010). The proposed model assuming DP ability parameter is represented as follows:

$$P(U_{jtig_jk}) = \sum_{l=1}^{L} g_l p(u_{jtigk}/g_j = l) \tag{3.1.1}$$

$$= \sum_{l=1}^{L} g_j \prod_{j}^{J} \prod_{i}^{I} P[u_{jtigk}/g_j = l] \tag{3.1.2}$$

$$P_{jtig_jk}(U_i = u/a_i; d_{tik}; \theta_{jg_j}; \delta_{g_j}; \gamma_{g_jt(i)}) = \frac{exp[\sum_{s=0}^{x} a_{it}(\theta_{jg_j} + \delta_{g_j} + \gamma_{g_jt(i)} - d_{ik})]}{\sum_{k=0}^{K} exp[\sum_{s=0}^{x} a_i(\theta_{jg_j} + \delta_{g_j} + \gamma_{gt(i)} - d_{ik}]} \tag{3.1.3}$$

$$\approx \Phi(a_i(\theta_{jg_j} + \delta_{g_j} + \gamma_{g_jt(i)} - d_{tik})$$

where P_{jtig_jk} represents the probability of person j, with person-specific ability θ_{jg_j}, belonging to person group component g_j with a group effect of δ_{g_j} getting a score of k on item i in testlet t with a step difficulty of d_{ik} and a discrimination parameter of a_i nested in testlet t with an item group effect of $\gamma_{g_jt(i)}$, the interaction between subject cluster g_j and testlet t. Person j is nested within group g_j. U represents any score

from 0 to K. The following distributions are assumed for the model parameters.

$$\begin{aligned} U_{jitg_jk} &= Categorical(P_{jtig_jk}) \\ \mathbf{a} &\sim LogN(\mu_{\mathbf{a}}, \sigma^2_{\mathbf{a}}) \\ \delta &\sim N(\mu_\delta, \sigma^2_\delta) \\ \gamma &\sim N(\mu_\gamma, \sigma^2_\gamma) \\ \mathbf{b} &\sim N(\mu_{\mathbf{b}}, \sigma^2_{\mathbf{b}}) \end{aligned} \tag{3.1.4}$$

For tests comprising of binary items, $U_{jitg_jk} \sim Ben(P_{jtig_jk})$. The log-normal distribution for a was selected in order to represent findings that are non-negative and typically near 1.0 and to make the ICC monotonically non-decreasing. In the absence of testlet effects, the model reduces to a multilevel model, in the absence of group effects, the model reduces to a testlet model while in the absence of both testlet and group effects, the model reduces to a GPCM.

3.1.1 The DP mixture ability parameter

The proposed non-parametric model assumes that the distribution of the latent ability θ_j, $j = 1 \dots J$, is a Dirichlet Process Mixture of normal distributions. We have J individuals on our data and each individual j belongs to a latent class or latent component g_j where $g_j \in 1 \dots G$, $G < J$. Each latent class has a mixing proportion π_g where $\sum_{g=1}^{L} \pi_g = 1$. The mixing proportion of each latent class is generated by a finite dimensional DP following Ishwaran and James (2001). Sethuraman (1994)'s stick breaking process (equation 2.8.20) was used to find the mixing proportions. The distribution of θ_j is given as follows:

- $(\theta_j/\mu_{\theta_j}, \sigma^2_{\theta_j}) \sim N(\mu_{\theta_j}, \sigma^2_{\theta_j})), (j = 1...J)$ with bivariate quantities $(\mu_{\theta_j}, \sigma^2_{\theta_j})$ drawn independent from a distribution $\mathrm{G}(\mu_{\theta_j}, \sigma^2_{\theta_j})$ where $G(.,.)$ is a bivariate DP, $G \sim DP(\alpha, G_0)$ with mean function $G_0(\mu_\theta, \sigma_\theta)$ and precision or total mass weight α. G_0 was taken to be the conjugate normal inverse gamma distribution such that:

- $\sigma_\theta \sim IG(1,1)$
- μ_θ was set to zero (0) for identifiability purposes

where IG and DP denotes an Inverse Gamma distribution and the Dirichlet Process respectively. The pairs $(\mu_{\theta_j}, \sigma^2_{\theta_j})$ concentrate on a set of $G \leq J$ distinct pairs (μ_g, v_g), $(g = 1, ..., G)$. The number of components, G is assumed to be unknown and determined from the data using the Stick-breaking process. This implies that the ability parameter is a Dirichlet Process Mixture of normals.

The group specific ability parameter δ_g is a Dirichlet Process such that:

$$\delta_g | G \sim G$$

$$G \sim DP(\alpha G_0)$$

where $G_0 \sim N(0, \sigma^2_{\delta_g})$; $\sigma_{\delta_g} \sim IG(1,1)$.

Following Jiao and Zhang (2014), person clustering effect, item clustering effect, person ability, item difficulty are assumed to be additive. However, contrary to Jiao and Zhang who assume that all effects are independent and mutual exclusive, the current model assumes an interaction between the cluster and item cluster effects, concurring with De Jong et al. (2008) model. Although the testlet effects in this study are assumed to be random, testlet effects can be assumed to be fixed (Beretvas & Walker, 2012) or constant over person. The random effects approach was selected because it deemed more appropriate in the presence of trait dependence (Wang & Wilson, 2005).

Unlike most mixture and multilevel models that assume item difficulty parameter to vary from one latent group to another, the proposed model assumes that item parameters are invariant over person groups, (cf Jiao & Zhang, 2014), that is, there is no differential item functioning (DIF). This means that the ICC for items are assumed to have the same shape although they might be at different levels on the continuum.

The model assumes rather the ability level differs from one group to another. For example, students in one class may have lower proficiency than students in another class, households in one community might be more food insecure than households in another community and a different ability distribution might hold in each cluster. The difficulty level of the questionnaire items remains invariant. The model assumes interaction between person clusters and item clusters.

Thus the proposed models is able to:

1. Detect the number of latent classes G
2. Identify the class g_j for each of the subjects j with greatest membership probability
3. Estimate group specific ability parameter δ_g
4. Estimate the person specific ability parameter θ_j
5. Detect the presence (and absence) of LID, LPD and dual dependence when present

3.2 Estimation of the parameters of the proposed model

The model was developed for estimating binary and polytomous items ($K \geq 2$). For a four category response ($K = 4$), the responses are scored 1, 2, 3 and 4. The model assumes that for successive response categories, the probability of selecting the k^{th} category over the $(k-1)^{th}$ category for item i follows a conditional probability governed by equation 3.2.5

$$\Phi_{jik} = \frac{P_{jtig_jk}}{P_{jtig_j(k-1)} + P_{jtig_jk}} \tag{3.2.5}$$

The probability of each response for person j in group g_j and testlet t for 4-response category items was estimated as follows:

$$s_1 = exp(\alpha(\theta_j + \delta_{g_j} + \gamma_{jt(i)} - b_{i1})) \tag{3.2.6}$$

$$s_2 = exp(\alpha(\theta_j + \delta_{g_j} + \gamma_{jt(i)} - b_{i1} - b_{i2}))$$

$$s_3 = exp(\alpha(\theta_j + \delta_{g_j} + \gamma_{jt(i)} - b_{i1} - b_{i2} - bi3))$$

$$P(U_{jitg_jk} = 1) = \frac{1}{1 + s_1 + s_2 + s_3}$$

$$P(U_{jitg_jk} = 2) = \frac{s_1}{1 + s_1 + s_2 + s_3}$$

$$P(U_{jitg_jk} = 3) = \frac{s_2}{1 + s_1 + s_2 + s_3}$$

$$P(U_{jitg_jk} = 4) = \frac{s_3}{1 + s_1 + s_2 + s_3}$$

$P(U_{jitg_{jk}} = 1, 2, 3, 4)$ is the probability of response score of 1, 2, 3 and 4 respectively for examinee j in group g responding to item i in testlet t. These expressions can be interpreted intuitively as though the person "passes through" each preceding response category before stopping at a response, thereby reflecting the person's standing on the latent variable continuum. The adjacent b_{jk} parameters represent the increment item difficulty parameters that the person has to step through in order to reach the next response category. The model estimated threshold parameter which is the sum of step and difficulty parameters and has the same interpretation as the difficulty parameter.

3.2.1 Data structure

The research considers a data structure in which all the J examinees receive a test of I polytomous response items with K_i category response options, where K_i is the number of response options for item i. Simulated and operational data are described by a fully observed $Y : (J \times I)$ matrix where entries y_{ji} are the responses of the j^{th} person to the i^{th} item with no missing values. Each individual j belongs to a latent class g_j where $g \in 1 \ldots G \leq J$ and each latent class g has a mixing proportion π_g where $\sum_{g=1}^{G} \pi_g = 1$. The mixing proportions of each latent class are generated by a finite dimensional DP via the stick-breaking procedure of Sethumaran (1994).

The I items are grouped into T, $(1 \leq T \leq I)$ mutually exclusive and exhaustive testlets. The testlet holding item i, $t(i)$ and the size of each testlet denoted by n_t, $(1 \leq n_t \leq T)$ with $t(1) = 1$ and $t(I) = T$. An independent item has $n_t = 1$ and if $T = I$, then each item is in its own testlet, and hence all items are conditionally independent. $\gamma_{gt(i)}$ is the group specific testlet effect and if $n_{t(i)} = 1$, then $\gamma_{gt(1))}$= 0 for all groups. The data was simulated in such a way that higher response order were more difficult to endorse than low response orders. Models involving missing data were not considered in all simulation studies. Data resembling testlets and clustered respondents for varying number of items, sample sizes, respondents clusters and testlet and clustering effects was simulated to address different objectives.

3.2.2 Prior distributions

The diagrammatical layout of the model parameters and their hyperparameters are given in Figure 3.1. (This is not the actual WinBUGS Doodle model).

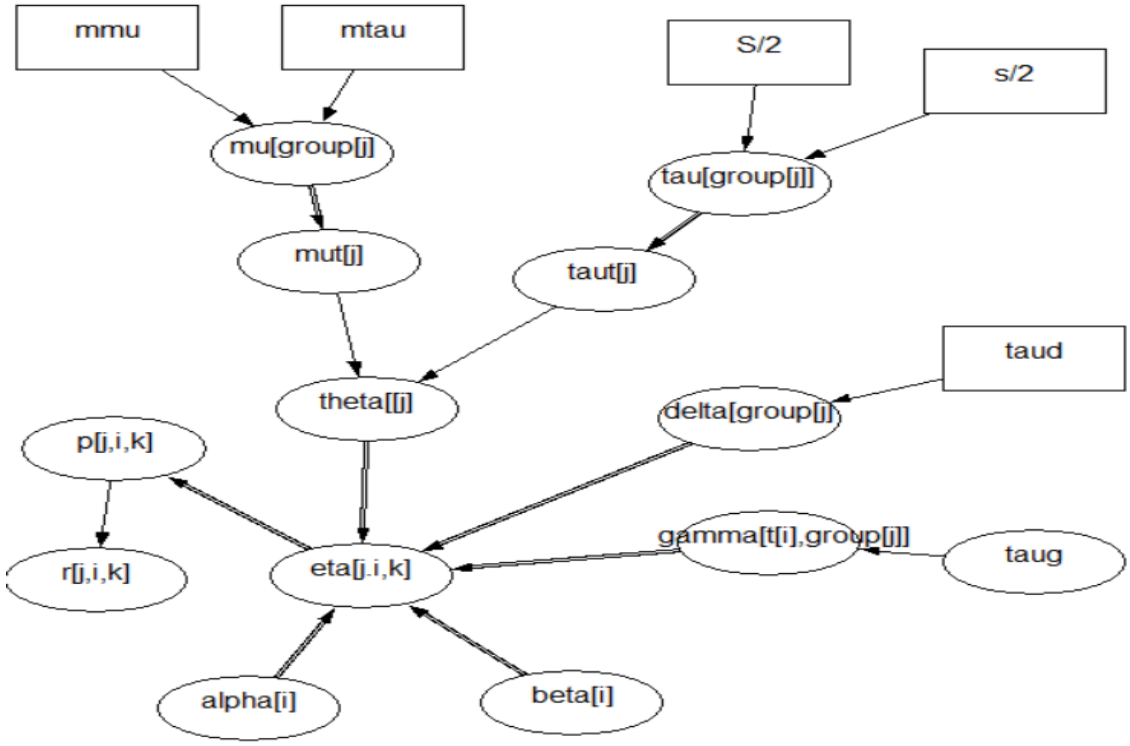


Figure 3.1: Diagrammatical illustration of the parameters and hyperparameters

The following prior and hyperprior distributions were utilised for the model parameters;

- $a_i \sim LN(0.2, 0.2)$
- $b_i \sim N(0, 1)$
- $\gamma_{gt} \sim N(0, \sigma^2_{\gamma_{gt}})$; $\sigma_{\gamma_{gt}} \sim IG(1, 1)$

where DP, IG and LN, denotes the Dirichlet Process priors, inverse-gamma and log-normal distributions respectively. The prior distribution for the ability variance, testlet variance, group variance and interaction variance were all set to the inverse-gamma(1,1) as the sample size was not small and the group variance was non-zero. The standard normal prior was selected for the difficulty parameter to provide rough bounds on the model parameters to make fitting procedures more stable (Cho et al., 2013).

The model differs from Jiao and Zhang (2014) simulation procedure in that the later had their model in a hierarchical manner while the proposed model was in the GPCM format. In addition, Jiao and Zhang model estimated the item difficulty and the step parameter. However, like the GPCM, our model is concerned with the threshold parameter, which is a linear sum of the item difficulty and step parameters. In this study, clustering takes place as part of the model framework and the data determines the clustering structure, hence the number of clusters need not be specified in advance.

3.2.3 Convergence Checks

To estimate model parameters and checking for chain convergence, four parallel chains with over-dispersed initial values were ran for each model. Practical convergence of the chain was checked by visualising when trace and history plots of the four sequences come together and where the distribution attains equilibrium / stationarity. If the model has converged, the trace or history plot will move around the mode of the distribution. The Brooks-Gelman-Rubin (BGR; Brooks & Gelman, 1998) statistics were used to assess convergence. The idea behind the BGR statistic is to use an Analysis of Variance (ANOVA) type diagnostic tool that compares within-chain and among-chain variability. If the chains converge to a similar posterior distribution, then

the between-chain variability will be small relative to the within-chain variability. The computation of the BGR is shown in equation 3.2.7

Take M widely dispersed starting points θ_m^0, $(m = 1, ..., M)$ and suppose that M parallel chains are run for $2n$ iterations. The first n iterations are discarded and regarded as burn-in. The M chains (θ_m^k) of length n produce means $\bar{\theta}_m = (1/n)\sum_{k=1}^{n} \theta_m^k$. The overall mean (across the chains) is $\bar{\theta} = (1/M)\sum_m \bar{\theta}_m$. As in classical ANOVA one calculates the within-chain and between-chain variability. Let

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2 \tag{3.2.7}$$

$$s_m^2 = \frac{1}{n}\sum_{k=1}^{n}(\theta_m^k - \bar{\theta}_m)^2$$

$$B = \frac{m}{M-1}\sum_{m=1}^{M}(\bar{\theta}_m - \bar{\theta})^2$$

The BGR analysis was done using BGR plots from WinBUGS / OpenBUGS / MultiBUGS. On the plot, the green lines give the width of the central 80% of the interval of pooled runs, the blue lines give the average widths of the 80% interval within individual runs while the red gives their ratio $R(= \textit{pooled/within})$. Values around 1 (one) indicate convergence with 1.1 considered acceptable by Gelman and Hill (2007). The square root of the BGR is the *potential scale reduction factor*. A requirement of a longer chain simulation is indicated by a scale reduction value greater than 1.0.

Auto-correlation functions were used to assess the rate to convergence. Posterior mean plots were used to check consistence of parameters and to assess if the plot exhibit elliptical shape. The converge of the chain was assessed for every parameter.

3.2.4 Label switching detection

Since label switching must be addressed before convergence diagnostics because it is a pre-requisite for convergence (Jasra et al., 2005), detection of label switch across iterations within a single MCMC chain was done by examining parameter traces for distinct jumps and when densities have multiple modes. If the estimated marginal posterior

densities are unimodal, then the sampling leads to unique labelling. If multiple modes exists for any of those densities, then label switching is present. Cross tabulations were done for the posterior modes and actual group membership. Label switching across chains was detected by having 4 initial values for the posterior and label switching was detected if the chains do not converge after attaining stationarity. The second form of label switching of latent class labels among replications was observed in the simulation study by comparing group membership with the data generating values for each chain to determine whether class labels had switched among replications.

3.2.5 Constraints for identifiability and label switching

Non-identifiability is overcome by assigning suitable prior distributions for the parameters, that is by setting appropriate restrictions that will make the model estimable. All models were identified by constraining the mean ability, threshold, testlet and group parameters to 0 (zero). In addition, one group membership and one testlet parameter was fixed to 0 (zero) while the ability parameter distribution was set to a zero mean normal distribution for parametric models and zero mean for the non-parametric model. The proposed model and the data generating process were concerned with modelling the interactions between the group and testlet and hence the multilevel GPCM and the testlet GPCM were adjusted accordingly.

3.2.6 Group membership recovery

The accuracy of parameter estimation for the proposed model is dependent on the accurate classification of examinees and classification accuracy is in-turn dependent on accurate parameter estimation (Cho et al., 2013). Recovery of class membership was assessed by comparing the simulated groups and group membership against the posterior mode for group membership. Since the naming of groups for the simulated and estimated groups were not necessarily the same, the proportion of estimated group membership was assessed by computing the proportion of respondents that belong to the same groups in both true and estimated groups. This was done after using the R-

package ***plyr*** to detect the sets of ordered pairs and their frequency.

3.3 Model selection procedures

The model fitness and ability to estimate and retain model parameters was determined using goodness of fit statistics, parameter and variance recovery, test reliability and test information.

3.3.1 Goodness of fit statistics

The goodness of fit statistics used for assessing the model fit are the Akaike Information Criterion (AIC: Akaike, 1974), the Bayesian Information Criterion (BIC: Schwarz, 1978) and the Deviance Information Criterion (DIC: Spielgelhalter et al., 2002) which deals with Bayesian posterior information and the Widely Applicable Information Criteria (WAIC, Watanabe, 2010). The DIC is composed of the Bayesian measure of fit or accuracy called the "posterior mean deviance" $\bar{D}$ and a penalty for model complexity p_D, the number of free parameters in the model. The BIC measures the trade-off between model fit and complexity of model. The AIC is a fined technique based on in-sample fit to estimate the likelihood of a model to predict/estimate future values. They are all methods of assessing model fit penalised for the number of estimated parameters in the model. The best fitting model is the one with the minimal AIC, BIC and DIC values. The AIC, BIC and DIC are computed as in equation 3.3.8, 3.3.9, and 3.3.10 respectively.

$$AIC = \bar{D} - 2p \tag{3.3.8}$$

$$BIC = \bar{D} + plogN \tag{3.3.9}$$

$$DIC = \bar{D_\xi} + p_D + 2p = D\hat{\xi} + 2p_D \tag{3.3.10}$$

where $\bar{D_\xi}$ is the posterior mean of the deviance, p is the number of parameters to be estimated by the model, N is the sample size and p_D denotes the difference between the posterior mean of the deviance and the deviance of the posterior mean $D\hat{\theta}$. The indices themselves are not very informative but the difference between the indices

for competing models are more useful with a difference of nine or greater providing stronger evidence to choose one model against the other (Anderson, 2008). The WAIC which the generalised version of the AIC onto the singular statistical model, reduces to equation 3.3.11 if the statistical model is realisable by and regular for a statistical model (Watanabe, 2013).

$$WAIC = AIC + o_p(\frac{1}{n}) \tag{3.3.11}$$

The benefits of using the proposed model were weighed against added complexity in the data analysis as Pett, Kim and Myung (2003) reiterated that the aim of model selection is not just maximum fit to the data but the model that best captures the characteristics or trends underlying the cognitive process of interest. In sum, the best model matches the purpose of study and can explain all the important features of the actual data without adding unnecessary complexity and without much loss of generality. As such, the model selection procedures employed take cognisance of the goodness of fit and model complexity in consideration. The p_D in the DIC was used as the effective number of parameters in the computation of the AIC and BIC statistics. Following Anderson (2008), the differences between a fit index (such as the BIC) was computed to estimate the expected Kullback-Leibler distance between the estimated best fitting model (model with the lowest index) and the i^{th} model to get a quantity:

$$\Delta_i = BIC_i - BIC_{min} \tag{3.3.12}$$

The likelihood of the i^{th} model given the data y becomes $L(g_i|y) \propto exp(-\Delta_i/2)$ which sets up the metric for assessing the relative strength of evidence between any competing models. Given the data and a set of Q competing models, the models were normalised to be a set of Akaike weights ω_i adding up to 1, where:

$$\omega_i = \frac{exp(-\Delta_i/2)}{\sum_{r=1}^{Q}(-\Delta_i/2)} \tag{3.3.13}$$

which denotes the probability that the model i is the expected best fitting model. The

evidence ratios between two models were computed as:

$$E_{ij} = \frac{L(g_i/y)}{L(g_j/y)} = \frac{\omega_i}{\omega_j} \tag{3.3.14}$$

and evidence ratios greater than 55 indicate that models differ significantly.

3.3.2 Parameter estimation and recovery

In this study, the model parameters estimated are person ability, item discrimination, item difficulty parameters, group-specific ability parameters, ability, group, testlet and group interaction variances and number of groups for the DPM dual model. The initial values for these parameters were generated from pre-specified parameter distributions. After initialising for the model parameters given response data, the MCMC data set was filled by sampling each parameter estimates from the joint distribution of the latest estimates of the parameters. When the Markov Chain reaches equilibrium, the burn-in process end and the monitoring process started. The results of after burn-in iterations were summarised to obtain model parameters estimates for inference.

The estimation ability for the models was compared in terms of the closeness of model estimates to observed (simulated) values in terms of Pearson's correlations between true and estimated values, bias, absolute bias (abias), standard errors (SE) and the root mean square errors (RMSE). Descriptive statistics, such as tabular summaries and graphical presentations were used to present these dependent variables. These criteria were used to assess the effects of manipulated testlet and group conditions on the item and person parameters. The ability of the model to detect the presence and absence of clustering was done by examination of testlet variance, ability variance and person clustering variances. Clustering effects were considered negligible for values less than 0.25. The bias, abias, SE and RMSE were computed as follows:

$$SE(\hat{\xi}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\xi}_r - \frac{\sum_{r=1}^{R}\hat{\xi}_r}{R}\right)^2} \tag{3.3.15}$$

$$Bias(\hat{\xi}) = \frac{\sum_{r=1}^{R}(\hat{\xi}_r - \xi)}{R} \tag{3.3.16}$$

$$Abias(\hat{\xi}) = \frac{\sum_{r=1}^{R} |(\hat{\xi}_r - \xi)|}{R} \tag{3.3.17}$$

The RMSE gives the average of the Euclidean distances between the observed (simulated) parameter values and the respective estimates.

$$RMSE(\hat{\xi}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\xi - \hat{\xi})^2} \tag{3.3.18}$$

where the bias, the SE and the RMSE represent the systematic error, the random error and the total error in model parameter estimates (Jiao & Zhang, 2014; Paek, 2009). ξ is the true model parameter, $\hat{\xi}_r$ is the estimated model parameter for the r^{th} replicate and R is the total number of replicates. The RMSE reflect both the sampling effect and probabilistic nature of IRT models and may be viewed as the real standard errors of the model estimates as suggested by Wright (1995). The rule of thumb employed for evaluating point forecast model is that the smaller the bias, SE and RMSE, the better the estimation ability of the model. The RMSE is the square root of the average of the parameter error variances, which is an index of precision of estimation. Under all research conditions, correlations of $r \geq 0.7$ were considered acceptable (Field, 2013) and the RMSE ≤ 0.40 and ≤ 0.25 in the ability and difficulty parameters respectively, were taken as criteria for considering a model as acceptable.

3.3.2.1 Inferential statistics on average dependent variables

The Monte Carlo simulation study is analogous of a statistical sampling experiment with a factorial experimental design and the number of replications being the number of times a simulation condition is repeated. In order to have adequate power of statistical tests, each simulation condition was replicated 10 times. The parameter recovery was assessed based on the bias, absolute bias, RMSE and standard errors of estimation. The general linear model was employed to compare the five dependent variables (bias, RMSE, abias, correlations and standard errors to assess the significance of the statistical effect of changing item and person dependency levels, samples / group size, testlet size/ number of test items, changing number of response category options and the effects ignoring the random slope to address the proposed research objectives. The

significance of the effect of independent factor levels on the dependent variables was determined using both the p-value and the effect size f (Cohen, 1988: equation 3.3.20). Comparison was also made based on the ratio between the average error variance over the sampling error variance. If the standard error estimates were accurate, then the ratio would be close to unity. If the ratio is larger than unity, then the standard error is overestimated. If the ratio is less than 1, then the standard error is underestimated (Wang & Chen, 2005).

3.3.2.2 Cohen effect sizes

The effect size (ES) can be treated as a parameter which takes on value 0 (zero) when the null hypothesis is true and some other specific non-zero value when the null hypothesis is false and hence it serves as an index of the degree of departure from the null hypothesis. The proportion of variance in the dependent variable explained by the factor:

$$\eta^2 = \frac{SS_{Between}(effect)}{SS_{Between} + SS_{Error}} = \frac{SS_{Between}(effect)}{SS_{Total}} \tag{3.3.19}$$

Cohen's effect (f) (Cohen, 1988 pp 280-288) is defined as:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \tag{3.3.20}$$

and quantified as small, medium and large using cut-off values of 0.10, 0.25 and 0.40 respectively. The effective size, contrary to the Fisherian formulation (Fisher, 1949) which posits the ES as 0 when the null hypothesis true and the ES as non-zero when the null is not true without further specification, the Cohen ES quantified the magnitude of the effect, thus providing a basis for statistical power.

3.3.3 Category characteristic curves

The category characteristic curves (CCC) for the non-parametric dual dependence model, the parametric model and the parametric models ignoring dual dependence effects were compared. The CCC shows the probability of each response category against the θ values that measure the location of the latent trait. The first line shows

the probability of selecting the first category response for persons of different ability levels. The second line shows the probability of selecting the second response category for persons of different ability levels etc. The points where adjacent category lines cross represent the transition from one category to the next. Thus respondents with low levels on the latent continuum are more likely to choose a lower category response. The ideal CCC is characterised by each response category being most likely to be selected for some segment of the latent continuum, with different segments corresponding to the hypothesised rank order of the response options. If this rank order is violated, the thresholds are said to be disordered.

3.3.4 Test reliability

The competing models were also evaluated in terms of the reliability of the measure. Test reliability refers to the degree to which a test is consistent and stable in measuring what it is intended to measure, that is, consistency within itself and across time. Reliability refers to dependability and consistency as well as the precision which enter into the measurement procedure (Anastasi, 1982). In classical test theory (CTT), the observed test score will be defined as:

$$X = T + E \tag{3.3.21}$$

where T is the true score and E is an error of measurement. This implies that their variances can be expressed as:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{3.3.22}$$

In which case test reliability will be defined as:

$$Reliability = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{3.3.23}$$

This could be transferred into the IRT framework as:

$$Reliability = 1 - \frac{\sigma_E^2}{\sigma_T^2} \tag{3.3.24}$$

where σ_T^2 is variance of expected a posteriori (EAP) scores.

Most examination experts suggested a minimum reliability estimate of 0.7 is desirable for routine classroom assessment (Royal, 2017). In CTT, reliability can be computed as the square of the correlation between the true and estimated scores. The true θ values were known since the data were simulated.

3.3.5 Spearman Brown prophecy formula

The reliability levels for models controlling for local dependence effects were compared to the GPCM for independent items and persons in terms of Spearman-Brown prophecy formula. The Spearman-Brown prophecy provides a rough estimate of how much the reliability of test scores would increase or decrease if the number of observations or items in a measurement instrument were increased or decreased. The Spearman-Brown formula for predicting the test length required for a dependence model to attain the reliability measured by the GPCM was calculated as:

$$m = \frac{\alpha^{model}(1 - \alpha^{GPCM})}{\alpha^{GPCM}(1 - \alpha^{model})} \tag{3.3.25}$$

where α^{model} = the test reliability estimate for the dependence model, α^{GPCM} = the reliability estimate for the GPCM model, m = the required length length for the dependence model to attain the GPCM model reliability.

3.3.6 Test information function

Another measure of model comparison was in terms of the test information estimated by the model. The test information function (TIF) is a measure of accuracy in ability estimates. TIF is equal to the inverse of the variability of estimates around the true value of the parameter (precision with which the parameter could be estimated). Higher values indicate accurate ability parameter estimates. Test information is calculated based on the summation of all the items information, that is, the TIF is estimated

$T(\theta) = \sum_i I_i(\theta)$ where $I_i(\theta)$ is the item information for item i and is defined as:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{3.3.26}$$

where $P_i(\theta)$ is the probability that an examinee with proficiency level θ will answer item i correctly, and $Q_i(\theta)$ is the probability that the examinee with proficiency level θ will answer item i incorrectly. $P_i'(\theta)$ is the first derivative of $P_i(\theta)$.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.3.27}$$

$$I(\theta) = \frac{1}{\sigma_\theta^2} \tag{3.3.28}$$

In IRT, the aim is to estimate the ability of an examinee (θ). Large information values imply that the ability level can be estimated with precision and all estimates will be reasonably close to their true values. Low information implies the ability cannot be estimated with precision and estimates will be widely scattered about the true values. The TIF estimates the ability parameter range well within an interval of the ability continuum. Outside this range the amount of information decreases rapidly, and the corresponding ability parameters are not well estimated. This means that the precision with which the ability parameter is estimated depends on where the examinee is located on the trait continuum. The peak of the curve shows where the measure yield the greatest information about the respondents. As shown in the derivation by Baker (2004), the higher the item/test information, the higher the item/test ability to discriminate between low and high proficiency level respondents.

3.4 Model estimation and statistical software

Several simulation studies were conducted to evaluate the performance of the proposed hierarchical non-parametric dual dependency model in a bid to address the factors that might affect the goodness of fit of the model. Data with different properties such as sample size, number of testlets, number of response categories, different levels of item and person dependency and varying number of response categories were simulated and the performance of the proposed model against available IRT models was assessed to

address the various research objectives. Ten (10) data sets were simulated for each condition in order to distinguish between sampling errors and estimation errors.

Data used for simulation studies were simulated in *R* version 3.2.5 (*R* Core Team, 2018) software and models were run in OpenBUGS (version 3.2.3 rev 1012: Lunn et al., 2009) and MultiBIGS 1.0 (Goudie, 2020) software for Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) techniques. The basic idea behind OpenBUGS and MultiBUGS is that the programmer needs to specify the statistical model, the prior distributions and the data, then they can determine the necessary sampling methods and Markov Chain calculations are done in the background and posterior summaries will be supplied for inference. The general linear modelling was conducted in SPSS version 25 (IBM Corp, 2017) while data manipulation and collation was done in Excel.

3.5 Application to operational data

The proposed model was applied to food insecurity measurement for urban households in Windhoek. A total of 835 households were interviewed using the African Food Security Urban Network (AFSUN) and Hungry Cities Partnership (HCP) Household Food Security Baseline Survey, which collects data on the experiential based Household Food Security Access Scale(HFIAS), the Household Dietary Diversity Score and the Months of Adequate Household Food Provision (MAHFP) score. Households surveyed in the 10 constituencies of Windhoek were identified using a two-stage sampling design. As a first step, primary sampling units (PSUs) were randomly selected with probability proportional to size. The PSUs were selected from a master frame developed and demarcated for the 2011 Population and Housing Census. Within the 10 constituencies, a total of 35 PSUs were selected covering the whole of Windhoek, and 25 households were systematically selected in each PSU. The HFIAS, HDDS and MAHFP questions were considered as 3 testlets within a main test intended to measure household food

access dimension.

3.6 Research Ethics

No human or animal subjects were directly used in this study as simulated and secondary data were used for theoretical statistical modeling. Moreover, permission for the use of secondary data was sought from the Hungry Cities Partnership (see Appendix D) and the data was handled in strict confidentiality. Above all, an ethical clearance was obtained from the University of Namibia Research Ethics Committee (approved under RESOLUTION: FPGSC-SCI/16/09/145).

Chapter 4

Assessing the effects of ignoring dual clustering in IRT models

4.1 Introduction

Studies on item parameter recovery in testlet effects concur that LID do not affect the estimation of difficulties or parameter estimates decreased slightly as LID levels increased (Reese, 1995) although high RMSE were reported (DeMars, 2006; Zhang & Jiao, 2014). However, high levels of positive LID are reported to result in overestimation of discriminant parameters (Reese, 1995) while negative correlation tends to underestimates discrimination ability (Tuerlinckx & De Boeck, 2001). In addition, correlations of true-estimated discriminant decreased as LID level increases (Reese, 1995), leading to the conclusion that high LID has devastating consequences on discriminant parameter estimation than difficulty and proficiency parameters. Theoretical analysis and empirical studies have shown that LID may result in overestimation of test reliability (Zhang, 2010; Wainer & Thissen, 1996; Zenisky, Hambleton & Sireci, 2002).

Although local dependence is undesirable in IRT modelling, there are good reasons for accounting for clusters within the population and for having testlets within a test. Many real life tasks require solving related problems, including content dependent items in a test. Testlets may increase construct validity. Data structures are often hierarchical in that individuals are often grouped into larger units consisting of a number of individuals (Raudenbush & Bryk, 1992; Stochet, 2005). For example, a group of

students that learn or study together may share certain group characteristics (Wang, Jiao & He, 2011) and the assumption of local person independence may be violated. De Boeck, Wilson and Carstensen (2008) reiterated that if person parameters are considered random, there might be undesirable consequences if certain person properties are not taken into account. For example, if a normal distribution is assumed for the random effects, the normal distribution may no longer apply to the entire group but only to clusters of person who share the same properties. As a result, the challenge for test developers is not to eliminate LID and LPD, but rather to properly model these dependencies (Zenisky et al., 2002).

Wang, Jiao and He (2011), Jiao and Zhang (2014) conducted simulation studies with 1000 respondents grouped into 40 clusters of 40 respondents each. Their results show that the estimation model, testlet effects and person dependence impacted on the accuracy of ability parameter estimates while only the estimation model and testlet dependencies have an effect on the estimation of item parameters. In addition, ignoring testlet effects resulted in higher total errors in item parameters while ignoring person dependencies magnifies the total errors in proficiency levels. Local dependency effects within testlet based tests varies from one test to another and hence the extent to which it affects model fit should also be examined. Levels of dependence considered for person dependence and local item dependency levels in literature are mainly with the range 0 to 1 (Zhang, 2010; Wang, 2002, 2005; Adams, Wilson & Wu, 1997; DeMars, 2006; Wang & Wilson 2006; Li, 2005; Zensiky, Hambleton & Sireci, 2002). In addition, Wang and Wilson (2005) argue that the variance of testlets in real life ranges from as small as almost 0 to as large as the variance of the latent trait, which is usually set to 1.0 in simulation studies. Jiao and Zhang (2014) and Jiao et al. (2012) considered dual dependence for 0.5 and 1 levels for both person and item dependence levels while Luo (2018) assessed the effects of small (0.25), medium (0.5) and large (1.0) testlet

effects and small (500 respondents), medium (1000 respondents) and large (2000 respondents) samples sizes on item parameter recovery. Furthermore, Jiao et al. (2011) used 10 replication in their simulation study while Jiao and Zhang (2014) applied 25 replications.

Studies on group dependence effects in IRT modelling usually assume the number of groups and group membership to be known in advance so that the group membership variable can be included as model data (eg Jiao & Zhang, 2014; Jiao et al., 2012). Those that used mixed models assume the number of groups to be known and the group membership to be unknown and hence employ the Dirichlet distribution to determine component membership (Cho, Cohen & Kim, 2011; Bolt, Cohen & Wollack, 2015). However, the number of groups in the data and group membership may not be known in advance and may need to be inferred from the data. In that case, the models in literature maybe rendered inadequate.

To assess the effects of ignoring clustering effects, the proposed non-parametric dual dependency model with the Dirichlet Process Mixture (DPM) ability was compared with the Generalised Partial Credit Model (GPCM) completely ignoring clustering effects, testlet GPCM model accounting for item dependency only, the multilevel GPCM model accounting for person clustering effects only and parametric dual dependency model for normally distributed ability parameters and logistic ICC in terms of model fit, bias, absolute bias (abias) root mean square errors (RMSE), error variances, precision and stability of parameters estimates. The simulation design is a fully crossed factorial design with three factors, LID and LPD each with three levels (0, none; 0.5, medium; 1, large) and calibration model with five levels. The comparison models are shown in Table 4.1.

The ability distribution was simulated from a mixture of normal distributions resulting

Table 4.1: Models for assessing effects of ignoring dual clustering in IRT modelling

Model	Description
Generalised Partial Credit Model	Ignoring both clustering effects, random discriminant parameters
Generalised Partial Credit Testlet Model	Accounting for testlet effects, random discrimination parameters
Multilevel GPCM	Accounting for person clustering effects, random discrimination parameters
Dual dependency GPCM	Parametric ICC and ability parameter, testlet and person cluster effects, random slope
Non-parametric dual dependence model	Testlet and person cluster effects, DPM ability parameter

in overall mean zero (0) and unity (1) variance. Equal component proportions were utilised. Mixture mean and variance were determined using the low of total expectation and the law of total variance equation 4.1.1 and equation 4.1.2 respectively.

$$\mu_\theta = \sum_{g=1}^{G} \pi_g \mu_{\theta_g} = 0 \tag{4.1.1}$$

$$\sigma_\theta^2 = \sum_{g=1}^{G} \pi_g \sigma_{\theta_g}^2 + \sum_{g=1}^{G} \pi_g \mu_{\theta_g}^2 - (\sum_{g=1}^{G} \pi_g \mu_{\theta_g})^2 = 1 \tag{4.1.2}$$

where $g = 1...G$ are ability clusters. μ_{θ_g} $\sigma_{\theta_g}^2$ are group specific mean and variances for group $g = 1...G$ respectively.

The effect of changing item and person dependency levels, testlet parameters $\gamma_{gt(i)}$, data were simulated from a normal distribution $N(0, \sigma_\gamma^2)$, by specifying three levels of testlet standard deviations as 0 (no dependency effects) 0.25 (small dependency effects) and 1 (large dependence effects). The person cluster variances were also simulated at three levels $\sigma_{\delta_g}^2 = 0$ (no group effects), 0.25 (small dependence effects) and 1 (large dependence effects). The 0 (zero) effect size were included as a control for baseline comparison and to confirm that models can adequately detect the presence and absence of testlet and person dependence effects. Nine local dependence conditions were simulated by considering all possible combinations of the person and item dependence effect as shown in Table 4.2.

4.2 Data simulation

To compare the effects of ignoring the dual clustering effects, data sets resembling testlet-based tests were simulated for nine different dependency conditions. The conditions considered are illustrated in Table 4.2.

Table 4.2: Conditions simulated for assessing the effects of ignoring item and person clustering in IRT modelling

Condition	Description
NoneNone	No item clustering effects ($\gamma_{gt(i)}$=0), no person clustering effects ($\sigma^2_{\delta_g} = 0$)
NoneSmall	No item clustering effects ($\gamma_{gt(i)}$=0), small person clustering effects ($\sigma^2_{\delta_g} = 0.25$)
NoneLarge	No item clustering effects ($\gamma_{gt(i)}$=0), large person clustering effects ($\sigma^2_{\delta_g} = 1$)
SmallNone	Small item clustering effects ($\gamma_{gt(i)}$=0.25), no person clustering effects ($\sigma^2_{\delta_g} = 0$)
SmallSmall	Small item clustering effects ($\gamma_{gt(i)}$=0.25), small person clustering effects ($\sigma^2_{\delta_g} = 0.25$)
SmallLarge	Small item clustering effects ($\gamma_{gt(i)}$=0.25), large person clustering effects ($\sigma^2_{\delta_g} = 1$)
LargeNone	Large item clustering effects ($\gamma_{gt(i)}$=1), no person clustering effects ($\sigma^2_{\delta_g} = 0$)
LargeSmall	Large item clustering effects ($\gamma_{gt(i)}$=1), small person clustering effects ($\sigma^2_{\delta_g} = 0.25$)
LargeLarge	Large item clustering effects ($\gamma_{gt(i)}$=1), large person clustering effects ($\sigma^2_{\delta_g} = 1$)

The discriminant parameter (slope) was simulated from a log-normal distribution i.e $a_i \sim LN(\mu_a = 0.2, \sigma^2_a = 0.2^2)$. Item location parameters were generated from a standard normal distribution. The item difficult parameters were generated form normal distribution. Testlet effects, person clustering effects and the effects of their interaction $b_{t(i)}$, δ_g, $\gamma_{gt(i)}$, were each generated from a normal distribution with 0 mean and standard deviations according to the conditions in Table 4.2.

For each of the joint levels of local item and person dependence, item responses were generated by incorporating true ability, item difficulty, group specific ability and LID parameters in equation 3.1.4. Ten (10) data sets with a common test structure were

simulated in R for $J = 1000$ respondents allocated to $G = 5$ person groups each with 200 respondents and $I = 36$ polytomous items, each with $K = 3$ response categories coded as 1, 2 and 3. The test items were grouped into $T = 6$ testlets of 6 items each. A sample size of 1000 was chosen because 1000 examinees were considered a minimum sample size required for IRT modelling (Nord, 1968; Resise & Yu, 1990) for estimating 3PL. The same set of population distribution for the ability, discriminant and difficulty parameters we simulated across all data sets.

To evaluate the suitability of the models in parameter recovery, correlations between the estimates and true values of $r \geq 0.7$ were considered acceptable (Field, 2013; Sahin & Amil, 2016) while the RMSE ≤ 0.40 in the ability parameter and RMSE ≤ 0.25 in the difficulty parameter were taken as evidence of precise estimation (cf Barnes & Wise, 1991; Hulin, Lissak, & Drasgow, 1982). The proposed dual models were used to detect LID and LPD in data sets by assessing the magnitude of testlet effects, group dependence and the effect of their interaction. Although there are no rules of thump for judging testlet (and group) effects parameters (Eckes & Baghaei, 2015) simulation studies have shown that testlet (and group) effects smaller than 0.25 are negligible (Wang, Bradlow & Wainer, 2002; Wang & Wilson, 2005; Zhang, 2010) and thus were also employed for the current study. Furthermore, the magnitude of testlet effect $\gamma_{gt(i)}$ was compared with the variance of the general ability parameter dimension. The higher the variance of testlet specific dimension compared to the general ability parameter variance, the more the local dependence the testlet has generated (Baghaei & Aryadoust, 2015; Baghaei & Ravand, 2016).

4.3 Results

4.3.1 Convergence checks

Convergence was checked based on multiple criteria to make sure that convergence was attained before the model parameter estimates were monitored. The Gelman-Rubin

R as modified by Brooks and Gelman (1998) was used. Four starting points were initialised and convergence was assessed by comparing within-chain (W) and between-chain (B) variability over the second half of those chains. The ratio $R = B/W$ was expected to be greater than one (1) if starting values were sufficiently different and would get closer to 1 as convergence approached. For the study, convergence was assumed for $R < 1.05$ following (Lunn et al., 2000). The R-statistic was close to 1 and smaller than 1.05 for all the competing models, indicating that all the models attained equilibrium.

Trace and history plots for the four starting points were employed to visualise where the sequences come together and where the distribution attains equilibrium / stationarity. The quantile plots, showed the running mean with the 95% confidence interval (CI). Results demonstrating the convergence of the MCMC sampler are shown in Appendix C.2. The auto-correlation function appear to dampen quickly, providing evidence of quick convergence of the Markov Chain (MC) to its equilibrium distribution and it would be appropriate to average the MC as if though the samples are independent. The parameter that depicted some auto-correlation were thinned 10 times after simulation. The trace, history and BGR plots indicated that all the parametric models had converged at 5000 burn-in iterations while the DP model converged after 10000 iterations (see Appendix C.2). The running means of the 95%CI for the four chains mixed well and reached equilibrium.

Based on these observations, 5000 iterations and 10000 iterations were discarded as burn-in for the parametric models and DP model respectively. Additional 5000 after-burn-in iterations were monitored and posterior inferences were estimated from these. No type I label switching was observed in WinBUGS although type II label switching was observed across each of the five chains (Cho, Cohen, Kim,2013). Post burn-in iterations were monitored for a posteriori summaries.

4.3.2 Goodness of fit statistics

The models were compared in terms of AIC, BIC and DIC fit indices and results for the first of 10 replications for each condition studied are shown in Table 4.3. The GPCM (true model) was selected as the best model in terms of fitness statistics for local independence conditions. Fit statistics for multilevel and GPCM models are equal except for DIC which is 10 units higher in the multilevel model. The DIC selected the multilevel model (true model) as the best fitting model in person dependence effects only while the AIC and BIC wrongly selected the testlet and dual models. In the presence of testlet effects only the dual model was selected as the best model ahead of the testlet model and in dual dependence effects, all fit statistics favoured the parametric dual model (true model) as the best fitting model, implying that the fit statistics accurately identified the true data generating model. In local independence, the non-parametric dual model performed slightly better than the parametric counter part. For data exhibiting dual effects, the testlet and dual models performed similarly with better fit while the GPCM and multilevel models also performed similarly with poorer fit. The fit statistics for independent person and items models and models accounting for person clustering effects are almost the same for all conditions while wide deviations were recorded for models with larger testlet effects.

According to the results shown in Table A2 and Table A3 (Appendix 1), in independence, all the fit statistics selected the GPCM model (the true model). In dual dependence effects, the differences between fit indices for the dual models and other models are greater that 10, implying that dual models are significantly better (see Anderson, 2008) than other models. In addition, there is strong evidence (smallest ratio of 33.333), implying that dual models are significantly superior in handing dual dependence effects than models accounting for LID and LPD effects only and those completely ignoring dual effects.

Table 4.3: Summary of fit statistics for assessing goodness of the models

Testlet	Group	Index	GPCM	Testlet	Multilevel	Dual	DualDP
None	None	AIC	48720	48750	48720	48750	48740
		BIC	49110	49110	49110	49150	49130
		DIC	49610	49710	49620	49820	49820
	Small	AIC	49750	49710	49760	49710	49710
		BIC	50140	50100	50150	50100	50080
		DIC	50650	50710	50640	50710	50710
	Large	AIC	43090	43040	43110	43080	42890
		BIC	43490	43430	43500	43470	43280
		DIC	43990	44050	43980	44140	44100
Small	None	AIC	50880	46450	50880	46380	46320
		BIC	51280	46840	51270	46770	46750
		DIC	51770	47520	51770	47440	47490
	Small	AIC	45440	42310	45390	42210	42290
		BIC	45790	42700	45790	42610	42650
		DIC	46300	43420	46270	43280	43320
	Large	AIC	45140	41650	45120	41540	41650
		BIC	45540	42040	45510	41930	42040
		DIC	46040	42720	45990	42590	42710
Large	None	AIC	56730	46320	56700	46190	46220
		BIC	57120	46710	57090	46580	46810
		DIC	57600	47410	57580	47270	47300
	Small	AIC	45640	36510	45620	36370	36410
		BIC	46030	36900	46010	36760	36820
		DIC	46500	37560	46480	37420	37480
	Large	AIC	49940	41020	49910	40890	41000
		BIC	50330	41420	50310	41290	41150
		DIC	50820	42120	50780	41960	42000

According to the results in Table A4 (Appendix 1), the ability parameter variance was well recovered by the true model in each simulation condition. In addition, the variance was generally well recovered by all models in local independence although it was slightly higher for dual models accounting for absent person and item clustering effects. For the GPCM model, the mean ability variance was not much impacted on by none and small item and person clustering effects, but overestimated by large person clustering effects and underestimated when large testlet effects were paired with none and small person clustering effects. The models ignoring LID (multilevel and GPCM) when present, underestimated the ability parameter variance while those failing to account for present LPD by the dual and testlet models while both dual and testlet models overestimated large interaction effects.

4.3.3 Group membership recovery

In local independence, all respondents were clustered in one modal group (100% group membership recovery) according to their ability levels, implying that no person dependence effects exist. However, in LID only, group parameter recovery was on average 97.5%, implying that LID might have a slight effect on ability parameter estimation and hence group parameter recovery. However, in group dependence effects only and in dual dependence effects, the average group recovery were at 85.8% and 82.7% respectively, suggesting that LID might have a slight effect on group recovery in addition to person clustering effects.

4.3.4 Category characteristic curves and information functions

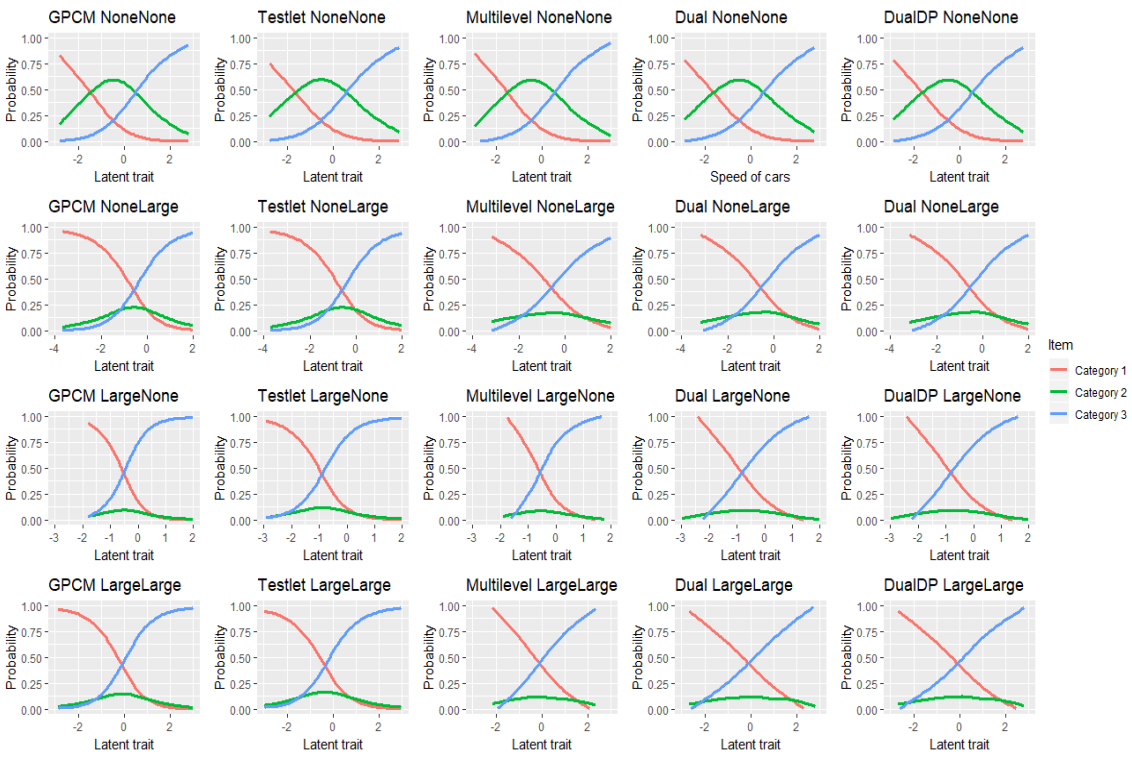


Figure 4.1: Category characteristic curves for the different models and dependence conditions

Figure 4.1 gives the category characteristic curves (CCCs) for the three response categories for the five calibration models under four of the simulated conditions. In independence, the thresholds (intersections between the category characteristic curves) are the same for all models, implying that the models optimally differentiate households at the same level on the ability continuum. The curve of the first score is a decreasing function of θ under all conditions and it cuts the 0.5 probability line at $b_{i1} = -2.5$ in the absence of dependence effects for all models. On the other hand, the highest score category is an increasing function of θ for all models under all conditions, and in the absence of dependence effects it crosses 0.5 probability line at the $b_{i3} = 0.5$, showing monotonicity, that is, higher category response levels were endorsed by respondents high on the ability continuum while the first category was mainly endorsed by households lower on the ability continuum.

In LPD only, the testlet and GPCM models have similar and steeper CCC showing higher discrimination ability while the multilevel and dual models CCCs depict a similar less steep pattern. The lower and upper thresholds for the GPCM and testlet are close to each other (both near zero (0)), implying that the two models optimally differentiate households whose proficiency levels are close to zero, and for a narrower proficiency range. On the other hand, the dual and multilevel models differentiate respondents for a wider proficiency range of $-2 \leq \theta \leq 1$, implying that the models are able to differentiate households of lower, medium and higher trait levels when compared to GPCM and testlet models. In LID only, the gap between lower and upper thresholds was narrower for GPCM and multilevel models ignoring LID than testlet and dual models accounting for LID in modelling.

However, in LPD and LID, the CCCs for all models show step disordering as the calibration from category 2 to category 3 is lower than the calibration from category 1 to category 2, and category 2 is not modal for all models. The thresholds are more

dis-ordered in LID than in LPD. However, the step disordering did not introduce the category disordering as the first category was mostly favoured by less proficient respondents while the third category was mostly endorsed by respondents high up on the trait continuum, implying that category definitions are not necessarily out of sequence. The slopes for GPCM and testlet models are generally steeper than multilevel and dual models across all dependence conditions, implying that although the models discriminate respondents at a narrower proficiency range when compared to the other models, their discrimination ability is higher than the other models. It is difficult to differentiate the CCCs for parametric and non-parametric dual models under all simulation conditions.

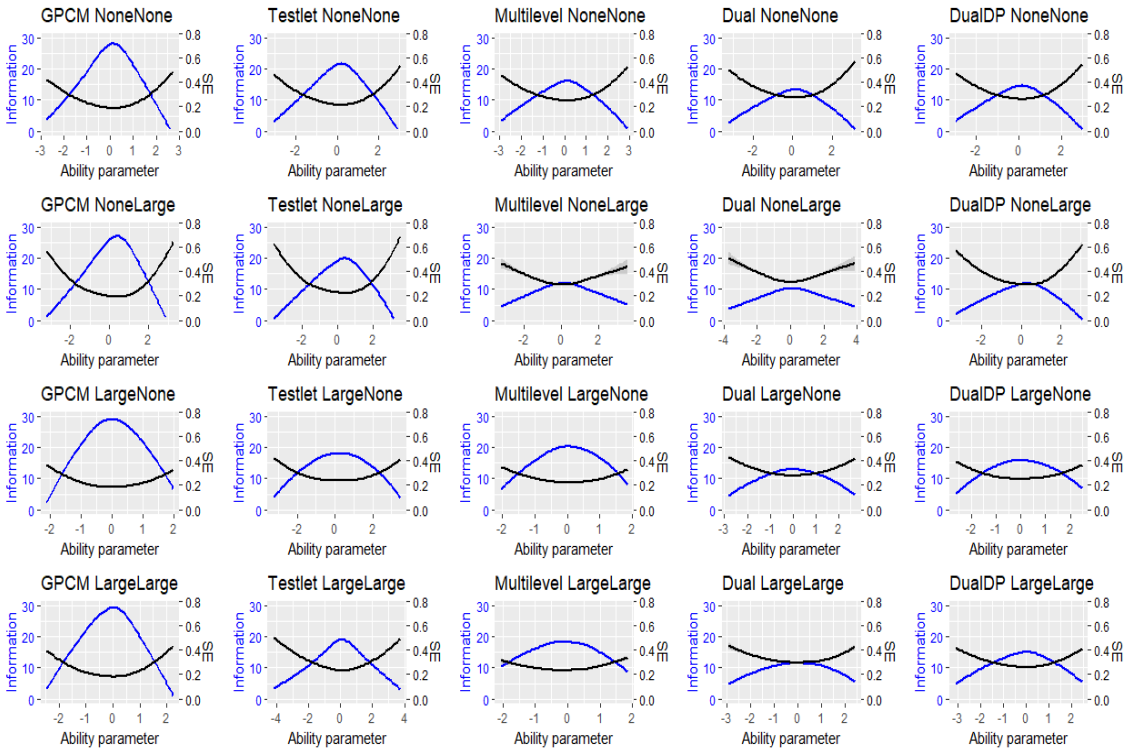


Figure 4.2: Test information for the different models and dependence conditions

Figure 4.2 shows the test information (blue line, values read from the x-axis) accounted for by the five models under comparison for four simulation conditions. For all estimation conditions, the highest amount of information is accounted for by the GPCM model, implying that the GPCM provided the best estimate for the proficiency. The

information in the testlet model is slightly lower than the GPCM in testlet effects while information in the multilevel model in group effects is even lower. The dual models accounted for the lowest test information than all models for all simulation conditions. The test information curve was narrower for GPCM across all simulation conditions and wider in multilevel and dual models in person dependence effects. In addition all models distributed the information uniformly across the proficiency range as the maximum amount of information was estimated at an ability of $\theta = 0.0$ for all models and all conditions except for the testlet models in large person effects where the maximum amount of information was recorded for $\theta = 0.5$.

In local independence, all models estimated the information for the ability range $-3 \leq \theta \leq 3$. Outside this range, the amount of information decreases rapidly and the corresponding abilities are not estimated well. However, in testlet effects, the GPCM model, although it estimated the highest amount of information, it did so for a narrower range of $-2 \leq \theta \leq 2$ whereas the models controlling for testlet effects estimated ability parameters in the range of $-4 \leq \theta \leq 4$. In short, the TIFs for the GPCM model are narrow and steep while the TIF for testlet and dual models are short and wide. The information function for dual models are too flat, implying that they overestimated the test information for respondents that are at the bottom of the proficiency scale and underestimated the information at the central values. The results shown suggest that the amount of information in the estimation of traits is acted upon more by testlet effects than group effects.

The test standard error curve (black line, values read from the right axis) represents the amount of measurement errors a test contains controlling for θ. In local independence, SE were lowest in the GPCM model, implying that there were less measurement errors in the test. However, there were lower measurement errors in the GPCM for all dependence conditions while in person dependence effects only, SE were largest in

multilevel model which control for group effects only. The results suggest that both item and person clustering effects result in overestimation of test information. The most peaked information curve in dual dependence effects was recorded in the GPCM model assuming independent items and persons.

4.3.5 Ability parameter recovery

The recovery of true model ability, difficulty and discriminant parameters was evaluated in terms of correlations between true parameters and parameter estimates, standard deviations, bias and mean square errors computed from 10 replicated data sets simulated for each condition, computed using equation 3.3.15 to 3.3.18. The descriptive statistics such as tabular summaries and graphical presentations were used to present these comparison variables. Plots showing the relationship between true and estimated parameter values were plotted and the spread and deviation from the slope of unit gradient was noted to assess the estimation ability of the model.

Table 4.4: Correlations between parameter estimates and true values for the ability parameters

Testlet effects	Group effect	GPCM	Testlet GPCM	Multilevel GPCM	Dual GPCM	Dual DP
None	None	0.98	0.97	0.97	0.96	0.95
	Small	0.92	0.92	0.96	0.96	0.96
	Large	0.71	0.72	0.95	0.95	0.95
Small	None	0.95	0.96	0.95	0.95	0.96
	Small	0.88	0.89	0.94	0.95	0.95
	Large	0.69	0.70	0.94	0.94	0.94
Large	None	0.94	0.97	0.95	0.95	0.95
	Small	0.83	0.85	0.92	0.93	0.94
	Large	0.69	0.70	0.93	0.94	0.94

The Pearson's correlation coefficients for the relationship between true ability parameter and their respective estimates are shown in Table 4.4. The correlations were calculated within each replication and averaged over replications. The estimates of ability parameters for dual and multilevel models incorporating group clustering effects were highly correlated with true ability parameters than the other models. The correlation

between true and estimated ability parameters for dual and multilevel models were about the same magnitude. This indicates that ignoring person clustering effects led to large differences in rank ordering of persons according to their latent traits. Correlations between ability parameter and their estimates from models accounting for person clustering effects (multilevel and dual models) were higher than correlations for testlet and GPCM models ignoring person clustering. The correlations for models ignoring person clustering effects were lowest for large effects, implying that ignoring person clustering impact negatively on the ordering of respondents according to their positions on the latent continuum. Item clustering effects have a minor effect on the rank-ordering of persons according to their ability levels.

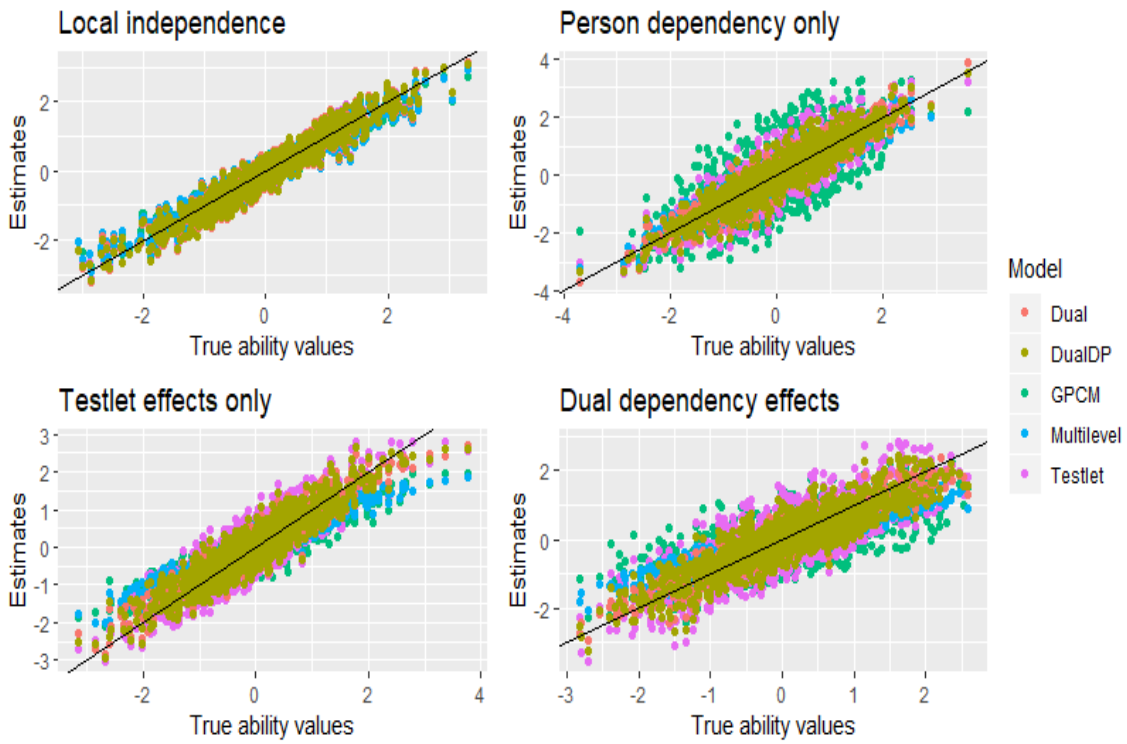


Figure 4.3: Plots of true ability parameters against estimates for item and person dependence conditions

Displayed in Figure 4.3 are scatter plots to show the relationship between true ability values and their respective estimates in LID and LPD. The plots in local independence show that for all models, parameter estimates were close to true parameter values as

points do not deviate much from the unit slope abline. In addition, large item clustering effects do not seem to impact much on rank-ordering of respondents on the latent continuum as the points for data with testlet effects only do not deviate much from true values. However, ignoring LPD significantly affected the ranking of respondents on the latent ability continuum as indicated by points for models ignoring person effects scattered far from the unity gradient line. In addition, in LPD effects, the GPCM and testlet models show *"outward"* biases, that is, the low ability estimates underestimate true values while high abilities are overestimated. On the other hand, in LID effects, the multilevel model seem to have *"inward"* biases where low true ability parameters are overestimated and high true ability values are underestimated.

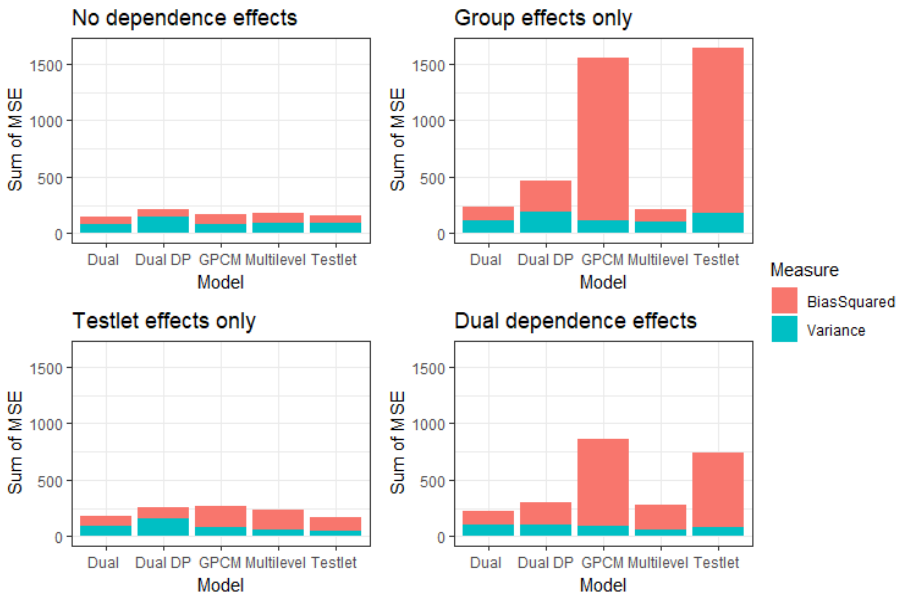


Figure 4.4: Random, systematic and total error in the ability parameter for changing person and item dependence conditions

The results in Figure 4.4 show sampling/systematic errors (biases) and random errors in ability estimates for the 5 competing models for 4 simulation conditions. If the data fits the model perfectly, the ratio of random and systematic errors should be close to unity (1). In local dependence, all models have minimum total errors in ability estimates. In group effects, ability parameter estimates are more biased in GPCM and

testlet models ignoring group effects. In LID, random errors were lowest in GPCM and multilevel models ignoring item dependence effects. However, testlet effects do not impact much on ability parameter bias although SEs seem to be lower in unaccounted for testlet effects.

In local independence conditions, all models have ratios between systematic and random errors close to unit, implying that the precision of ability parameter estimation is not overestimated or underestimated by any of the models. In person clustering effects, GPCM and testlet models ignoring the clustering effects have ratios between systematic and random errors way above unit, suggesting that the two models might be overestimating the precision in ability parameter estimation. On the other hand, in item clustering effects only, the GPCM and multilevel models ignoring item clustering effects seem to slightly overestimate the precision in ability parameter estimation as the ratio between sampling and random errors is slightly above unit.

Results in Table 4.5 SEs, bias and RMSE in estimation of ability parameters. From the results, lower SEs were recorded for GPCM and testlet GPCM models ignoring person clustering effects than the multilevel and dual models incorporating person clustering effects. SEs seem to be much affected by LID than LPD as lowest SE were recorded in GPCM and multilevel assuming local item independence. In addition, random errors in group effects are lower than random errors in person independence and the higher the group dependency effects, the lower the SEs in ability parameter estimation.

Although the ability parameter estimation bias had a near zero (0) mean for all models, ignoring person clustering effects resulted in more biased ability parameters as intervals/ranges in biases for GPCM and testlet models were wider in person clustering effects. Moreover, the standard deviations for the bias were higher for estimates from

Table 4.5: SE, Bias and RMSE in the ability parameters for different dependence conditions

		SE		Bias		RMSE	
Condition		Mean	SE	Mean	SE	Mean	SE
NoneNone	GPCM	0.23	0.00	0.01	0.01	0.31	0.00
	Testlet	0.24	0.00	0.02	0.01	0.32	0.00
	Multilevel	0.28	0.00	0.02	0.01	0.36	0.00
	Dual	0.30	0.00	0.00	0.01	0.38	0.00
	Dual DP	0.27	0.00	0.01	0.01	0.35	0.00
NoneSmall	GPCM	0.23	0.00	0.04	0.01	0.42	0.01
	Testlet	0.25	0.00	0.05	0.01	0.44	0.01
	Multilevel	0.29	0.00	0.04	0.01	0.40	0.00
	Dual	0.32	0.00	0.01	0.01	0.41	0.00
	Dual DP	0.31	0.00	0.02	0.01	0.41	0.00
NoneLarge	GPCM	0.29	0.00	0.02	0.03	0.88	0.02
	Testlet	0.30	0.00	0.01	0.03	0.85	0.02
	Multilevel	0.34	0.00	0.01	0.01	0.45	0.01
	Dual	0.36	0.00	0.00	0.01	0.48	0.01
	Dual DP	0.33	0.00	0.01	0.01	0.45	0.01
SmallNone	GPCM	0.22	0.00	0.02	0.01	0.36	0.01
	Testlet	0.25	0.00	0.02	0.02	0.49	0.01
	Multilevel	0.26	0.00	0.02	0.01	0.40	0.00
	Dual	0.30	0.00	0.01	0.01	0.42	0.00
	Dual DP	0.28	0.00	0.01	0.01	0.39	0.00
SmallSmall	GPCM	0.22	0.00	0.03	0.02	0.46	0.01
	Testlet	0.25	0.00	0.03	0.02	0.52	0.01
	Multilevel	0.30	0.00	0.03	0.01	0.43	0.00
	Dual	0.31	0.00	0.02	0.01	0.43	0.00
	Dual DP	0.28	0.00	0.01	0.01	0.32	0.00
SmallLarge	GPCM	0.23	0.00	-0.02	0.03	0.71	0.02
	Testlet	0.27	0.00	0.00	0.03	0.74	0.01
	Multilevel	0.27	0.00	-0.03	0.01	0.44	0.01
	Dual	0.31	0.00	-0.03	0.01	0.44	0.01
	Dual DP	0.29	0.00	-0.01	0.01	0.33	0.00
LargeNone	GPCM	0.18	0.00	-0.02	0.02	0.49	0.01
	Testlet	0.26	0.00	-0.02	0.01	0.45	0.01
	Multilevel	0.21	0.00	-0.02	0.02	0.47	0.01
	Dual	0.31	0.00	-0.02	0.01	0.42	0.00
	Dual DP	0.29	0.00	-0.01	0.01	0.38	0.00
LargeSmall	GPCM	0.18	0.00	0.03	0.02	0.53	0.01
	Testlet	0.29	0.00	0.00	0.02	0.70	0.01
	Multilevel	0.23	0.00	0.03	0.02	0.49	0.01
	Dual	0.34	0.00	0.04	0.01	0.47	0.00
	Dual DP	0.31	0.00	0.02	0.01	0.36	0.00
LargeLarge	GPCM	0.24	0.00	0.01	0.03	0.79	0.01
	Testlet	0.30	0.00	0.01	0.03	0.76	0.01
	Multilevel	0.28	0.00	0.05	0.01	0.49	0.01
	Dual	0.35	0.00	0.05	0.01	0.47	0.01
	Dual DP	0.33	0.00	0.01	0.02	0.64	0.01

the GPCM and testlet models ignoring person clustering, implying that these estimates are more biased when compared to models controlling for person clustering. To support this, the intervals are almost the same for all models in local item and person independence with the GPCM recording the narrowest interval and lowest standard errors for the bias.

In LID, SEs were lowest in GPCM and multilevel models while in LPD, SEs are lowest in GPCM and testlet models ignoring person effects. As the magnitude of LPD increased, SE increased across all models. Total errors (MSE, which reflect the actual variation of sample parameter estimates within replications for a specific simulation condition) in ability parameter estimates were slightly higher for GPCM and testlet models ignoring person clustering effects than for multilevel models controlling for respondent clusters. The MSE in ability parameters did not seem to be much affected by ignoring testlet effects. The estimation accuracy measured by bias, SE, MSE decreases as group effects increase. The results show that there are no wide magnitudes in the errors from the non-parametric and parametric dual models although the later performed slightly better in large dependence effects.

4.3.5.1 Inferential statistics on dependent variable

To evaluate the significance of effects of ignoring item and person clustering on ability parameter recovery, univariate three-way full factorial analysis of variance (ANOVA) was conducted for each of the error indices (bias, abias, SE, RMSE) as dependent variables and LPD (3 levels), LID (3 levels), and models (5 levels) as the model factors. The results show that bias in ability parameter estimates did not differ significantly across the conditions being studied as all the Cohen f effects sizes were negligible (< 0.10). The bias means for all models were close to zero. However, the absolute bias (abias), random error (SE) and total error (RMSE) in ability parameter estimation were significantly impacted by all factors and their interactions.

The ANOVA for abias has shown the main effects of group dependence and calibration model to be significant ($p < 0.05$) with effect sizes of $f = 0.19$ and $f = 0.26$ respectively. However, the main effects of testlet dependence were significant with a negligible effect size of $f = 0.08$. Two-way interaction between testlet and group effects ($f = 0.10$), testlet and model effects ($f = 0.10$), group and model effects ($f = 0.21$) and three way interaction between testlet, group and model effects all significantly affected the abias in ability estimates. For the main effect of testlet dependence, the Tukey *post-hoc* show that the lowest abias were recorded in the absence of testlet effects while highest were recorded for large testlet effects and all categories different significantly. There was no significant difference in abias for non-parametric and parametric dual and multilevel models and these abiases were significantly lower than abias for GPCM and testlet models which also did not differ significantly. Similar results were recorded for group effects. The smallest and not statistically significant abiases were recorded in all the models in local independence, dual model for all conditions, multilevel in group effects only and testlet model in item effects only, with the lowest being for GPCM for independent persons. The highest abias were recorded for testlet and GPCM models in large group effects.

The ANOVA for SEs in ability estimation has significant main effects ($p < 0.05$) for testlet, group and model factors with effect sizes of $f = 0.15$, $f = 0.29$ and $f = 0.58$ respectively. Tukey *post-hoc* for testlet effects show lowest SE for large LID effects although they did not differ significantly with SE for small LID effects. SE in item independence were significantly larger. Contrary to these finding, SE for group effects were significantly higher for large group effects and lowest in person independence. Lowest SEs were recorded in GPCM model and they differed significantly ($p < 0.05$) with testlet-based models. The random errors were higher for multilevel and dual models accounting for person effects. All two-way and three-way interactions between testlet effects, group effects and calibration models significantly affected SE in ability

parameter recovery. The *post-hoc* analysis for three-way interaction revealed lowest SEs for GPCM followed by the multilevel model in large to small testlet effects. Highest SEs were recorded for dual dependency models in both LID and LPD effects and multilevel model in person clustering effects only.

The calibration model, testlet and group effects had effect sizes of $f = 0.28$ (medium), $f = 0.15$ (small) and $f = 0.27$ (medium) on RMSE in ability parameters respectively, while the interaction between dependence conditions and calibration model was significant ($p < 0.05$) with an effect size of $f = 0.998$. According to the Tukey *post-hoc*, lowest errors were recorded in local independence, lowest being recorded for GPCM and testlet models. Highest total errors were recorded for models ignoring dual dependence (testlet and GPCM) when they are present. In conclusion, ignoring testlet effects led to lower SEs while ignoring group effects led to increased biases in ability parameters and ignoring both dependencies resulted in higher total errors.

4.3.6 Threshold parameter recovery

Table 4.6: True values and estimates correlations for the threshold parameters

Testlet	Group	GPCM	Testlet	Multilevel	Dual	Dual DP
None	None	0.99	0.99	0.99	0.96	0.96
	Small	0.99	0.99	0.99	0.98	0.97
	Large	0.99	0.99	0.99	0.98	0.97
Small	None	0.93	0.94	0.93	0.97	0.97
	Small	0.93	0.95	0.93	0.98	0.97
	Large	0.87	0.93	0.87	0.97	0.97
Large	None	0.85	0.97	0.85	0.96	0.96
	Small	0.84	0.96	0.86	0.96	0.96
	Large	0.85	0.95	0.85	0.97	0.97

The results in Table 4.6 show that the rank order of threshold parameters was in general well recovered by all models ($r > 0.7$) for all simulated conditions. Correlations between true threshold values and their respective estimates were high in all models in local independence with GPCM and multilevel models recording highest correlations. In addition, the rank ordering of items does not seem to be affected by LPD

as models not accounting for person effects still recorded high correlations in LPD. However, threshold parameter recovery was slightly affected by ignoring item clustering effects as correlations were lower for GPCM and multilevel GPCM in testlet effects.

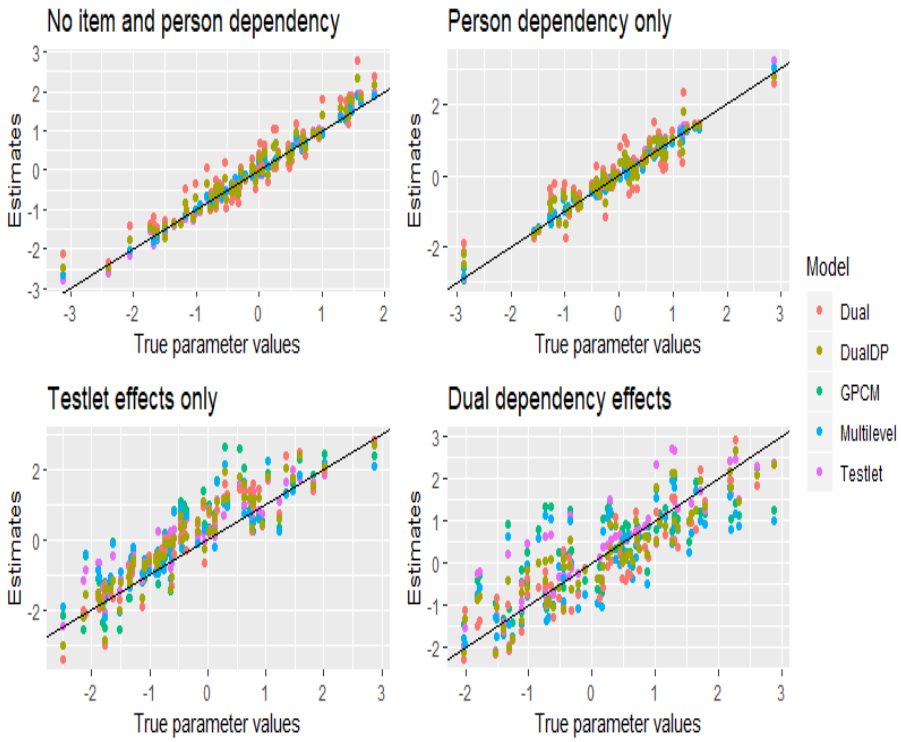


Figure 4.5: Plots of true vs estimated threshold parameters for local dependency conditions

Although the differences in fitness and ability parameters for all models in local independence are minimal, Figure 4.5 shows high deviance between estimates from dual and testlet models controlling for item dependence effects which do not exist, implying that mistakenly controlling for testlet effects when in fact they are absent distorts item difficulty levels. However, the step parameter estimates from GPCM and multilevel GPCM do not deviate much from their true values. In addition, the distribution of points below and above the line $x = y$ are almost the same, and thus the rank ordering of items on the difficulty continuum was not much affected in local independence. Contrary to this, ignoring testlet effects when they exist has negative consequences on item threshold levels. On the other hand, in item dependence, plots for testlet and dual

models are closest to the $(x = y)$ line while plots for GPCM and multilevel models are scattered further away from the line. In large person dependency effects only, plots for all models, including the dual models deviate from the $(x = y)$ line. This might imply that controlling for non-existent testlet effects may bias the step parameters. Points for multilevel and GPCM models ignoring testlet effects are close to the unity gradient $(x = y)$ line for local independence conditions.

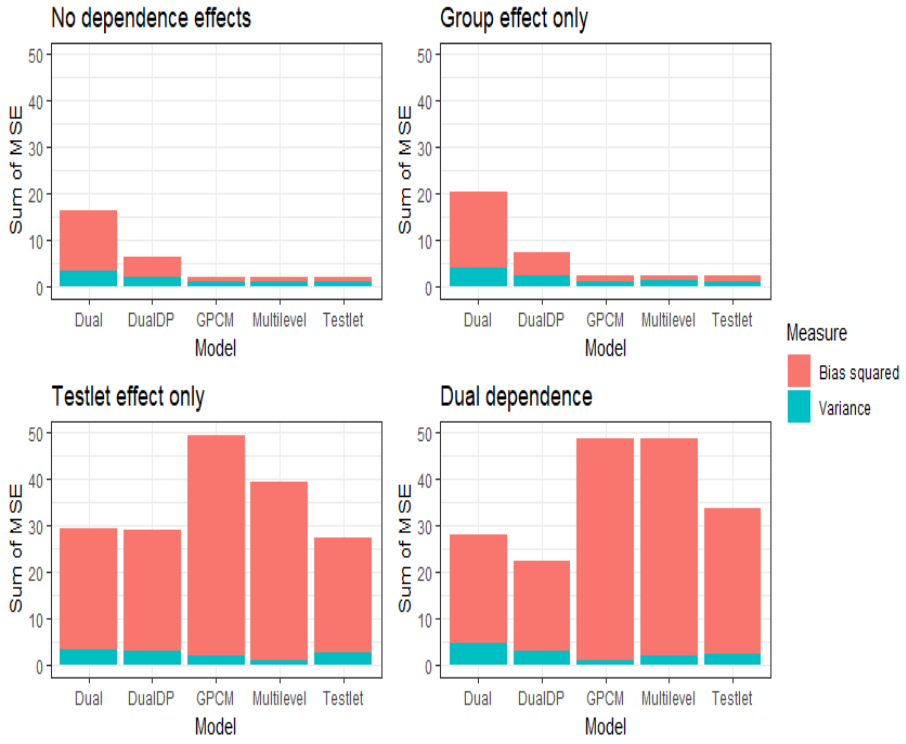


Figure 4.6: Random, systematic and total error in the threshold parameters for changing item and person dependency conditions

Figure 4.6 shows random, systematic and total errors in threshold estimates for different group and testlet dependence conditions. Random errors in threshold parameter were lowest in GPCM and multilevel models in LID while systematic errors were higher for dual models when they wrongly accounted for non-existent LID and LPD. However, systematic errors were highest for GPCM and multilevel models that ignore LID when it was present, resulting in higher total errors in threshold parameter estimates.

Table 4.7: Standard errors (SE), bias and Root Mean Square Errors (RMSE) in the threshold parameter estimation

Condition		Mean	SE	Mean	SE	Mean	SE
NoneNone	GPCM	0.12	0.04	0.02	0.12	0.16	0.07
	Testlet	0.12	0.03	0.04	0.13	0.16	0.07
	Multilevel	0.11	0.03	0.01	0.12	0.16	0.07
	Dual	0.21	0.04	0.17	0.39	0.22	0.12
	Dual DP	0.16	0.03	0.09	0.23	0.21	0.13
NoneSmall	GPCM	0.11	0.00	-0.10	0.07	0.19	0.04
	Testlet	0.18	0.01	-0.06	0.03	0.20	0.02
	Multilevel	0.13	0.00	-0.03	0.06	0.18	0.03
	Dual	0.20	0.00	0.04	0.03	0.22	0.01
	Dual DP	0.22	0.00	0.00	0.02	0.22	0.01
NoneLarge	GPCM	0.12	0.04	0.07	0.13	0.17	0.07
	Testlet	0.12	0.03	0.06	0.13	0.17	0.07
	Multilevel	0.14	0.03	0.01	0.12	0.18	0.06
	Dual	0.23	0.06	0.09	0.17	0.25	0.09
	Dual DP	0.18	0.04	0.05	0.16	0.23	0.11
SmallNone	GPCM	0.12	0.00	0.10	0.07	0.30	0.04
	Testlet	0.17	0.01	0.04	0.03	0.21	0.01
	Multilevel	0.13	0.00	0.12	0.07	0.31	0.04
	Dual	0.19	0.00	0.17	0.03	0.32	0.02
	Dual DP	0.19	0.01	0.19	0.04	0.23	0.04
SmallSmall	GPCM	0.12	0.04	0.06	0.25	0.27	0.15
	Testlet	0.14	0.04	0.04	0.20	0.19	0.07
	Multilevel	0.15	0.04	0.05	0.28	0.26	0.07
	Dual	0.20	0.04	0.03	0.20	0.24	0.12
	Dual DP	0.17	0.03	0.03	0.21	0.21	0.08
SmallLarge	GPCM	0.12	0.03	-0.02	0.26	0.36	0.09
	Testlet	0.15	0.03	-0.04	0.21	0.24	0.05
	Multilevel	0.16	0.02	0.02	0.29	0.33	0.09
	Dual	0.27	0.06	0.01	0.21	0.21	0.08
	Dual DP	0.21	0.03	0.01	0.22	0.25	0.07
LargeNone	GPCM	0.16	0.05	0.10	0.21	0.38	0.25
	Testlet	0.19	0.04	0.03	0.10	0.23	0.10
	Multilevel	0.12	0.02	0.09	0.23	0.38	0.19
	Dual	0.21	0.03	0.04	0.14	0.26	0.07
	Dual DP	0.20	0.03	0.06	0.15	0.25	0.08
LargeSmall	GPCM	0.14	0.08	0.10	0.27	0.41	0.12
	Testlet	0.19	0.07	0.03	0.22	0.23	0.07
	Multilevel	0.16	0.08	0.05	0.30	0.43	0.14
	Dual	0.27	0.07	-0.04	0.22	0.23	0.12
	Dual DP	0.23	0.06	-0.03	0.23	0.27	0.07
LargeLarge	GPCM	0.12	0.03	0.08	0.30	0.44	0.17
	Testlet	0.18	0.03	0.04	0.25	0.28	0.07
	Multilevel	0.17	0.02	-0.06	0.33	0.39	0.15
	Dual	0.25	0.04	0.03	0.25	0.29	0.12
	Dual DP	0.21	0.02	-0.03	0.26	0.25	0.04

According to Table 4.7 showing results for SEs, bias and RMSEs in threshold parameter estimation, in LID effects, SEs were similar and lower in GPCM and multilevel models ignoring the effects than testlet and dual modes controlling for LID whose SEs were also similar. Moreover, SEs were lower when there were no and small testlet effects than larger testlet effects. Although SEs do not seem to be affected by person dependence, they were higher in dual models and lower in GPCM and testlet models when LPD was present. The SEs were generally lower for GPCM and multilevel models assuming independent items than the models controlling for testlet effects.

The bias was high in dual models when both items and persons were independent, and lower for GPCM and testlet models, and were high in testlet and dual models when only items were independent. However, when items were dependent, testlet and dual models accounting for testlet effects recorded low bias while GPCM and multilevel models assuming independent items recorded higher bias.

In local independence, total errors were higher in dual models spuriously controlling for these dependencies. However, in LID, average RMSE were lower for dual and testlet models than for GPCM and multilevel models. The magnitude of differences in RMSE for these two sets of models increased as the LID levels increased.

4.3.6.1 Inferential statistics on dependent variable

The ANOVA results show that testlet effects, interaction between testlet and group effects and interaction between testlet and calibration model significantly affected bias ($p < 0.05$) in threshold parameter estimation with significant effects of $f = 0.22$, $f = 0.13$ and $f = 0.15$ respectively. Ignoring large testlet effects led to overestimation of threshold parameters as bias in LID was more for GPCM and multilevel models. The abias differed significantly for calibration models although the effects size of $f = 0.09$ was insignificant, group dependence (small effect size of $f = 0.18$), testlet effects (small effect size of $f = 0.20$), interaction between testlet and group effects (insignificant effect

size $f = 0.08$) and interaction between testlet effects and model (insignificant effect size $f = 0.013$). Further analysis has shown that abias was minimal for all models in item independence, followed by dual and testlet models in testlet effects and highest in GPCM and multilevel models that do not account for testlet effects. *Post-hoc* analysis on the interaction between testlet and group effects has also shown minimal bias in local independence, lower in group effects only and higher in testlet effects. The magnitude of bias increased when testlet effects were coupled with increasing group effects.

The SEs were significantly affected by LID, LPD and model factors each with a effect size $f = 0.34$, $f = 0.15$ and $f = 0.85$ respectively. In addition, there were significant interactions between LID and model factors. Further analysis on three-way interaction has shown that SEs were lowest in local independence for all models, for all testlet models in absence of testlet effects and for the GPCM and multilevel models in testlet effects. Higher SE were recorded for models that account for testlet effects in their presence.

All factors significantly affected RMSE in threshold parameter estimation. LID affected threshold parameter estimates with a medium effect of size $f = 0.36$ while the LPD affected the parameter with an effect of size $f = 0.17$. The calibration model and interaction between testlet and group effects were both significant though each has insignificant effects of $f = 0.09$. However, the interaction between LID and calibration model has a significant though small effect of size $f = 0.15$. Tukey's multiple comparison has shown that total errors was lowest for models with no dependence effects followed by the dual and testlet models when testlet effects were present. Highest RMSE were recorded for GPCM and multilevel models which do not cater for testlet effects when they were present.

4.3.7 Discriminant parameter

The recovery of discriminant parameters was also assessed using correlations between true and estimated values, systematic errors (bias), abias, random errors (SEs) as well as total errors (RMSE).

Table 4.8: True values and estimates correlations for the discriminant parameters

Testlet	Group	GPCM	Testlet	Multilevel	Dual	Dual DP
None	None	0.96	0.96	0.96	0.96	0.97
	Small	0.96	0.96	0.98	0.98	0.97
	Large	0.96	0.96	0.98	0.98	0.97
Small	None	0.89	0.95	0.87	0.96	0.95
	Small	0.86	0.94	0.85	0.96	0.96
	Large	0.85	0.94	0.84	0.96	0.94
Large	None	0.59	0.94	0.57	0.97	0.96
	Small	0.53	0.95	0.53	0.97	0.96
	Large	0.47	0.91	0.44	0.96	0.95

The true-estimated value correlations results in Table 4.8 show that the larger the testlet effect, the more the negative effects on discriminant parameter estimates rank ordering when ignored. Item discrimination ability is significantly lost by ignoring item clustering effects as average correlations between true discrimination ability and estimates as low as 0.38 and 0.47 (significantly lower than 0.7) were recorded for large testlet dependence effects. The correlation between true and estimated parameter values were slightly affected by person clustering effects as correlations are slightly lower in person independence.

The results in Figure 4.7 indicate that the discriminant parameter was in general well recovered by all models in local independence. However, the GPCM underestimated discrimination parameters in LID. The multilevel best recovered discriminant parameter in LID only. The results in these figures show that ignoring both item and person clustering effects, and accounting for absent testlet effects negatively affect item discrimination ability. To add more to this, except for dual models controlling for both

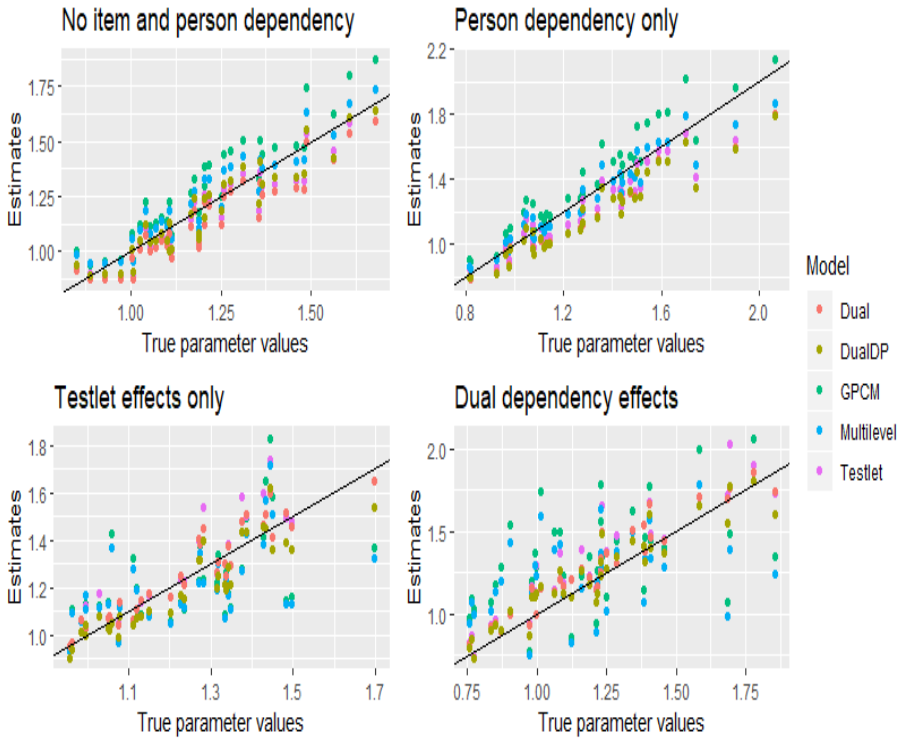


Figure 4.7: Plots of true vs estimated discriminant parameters for local dependency conditions

testlet and group effects, all the other models did not recover discrimination parameters well in both LID and LPD.

According to Figure 4.8, random errors were highest in the GPCM model across all simulation conditions. In local independence, dual models recorded highest systematic errors and hence highest total errors. In LPD only, multilevel models performed best, implying that LPD probably has an effect on discriminant parameter estimation. However, highest total errors were recorded in LID only, for models ignoring the effects.

4.3.7.1 Inferential statistics on dependent variables

According to ANOVA results, discriminant parameter estimation bias was affected by LID ($f = 0.16$), and model factors ($f = 0.39$) but not affected by person clustering effects. The bias was lower in dual models where the average was close to 0 (zero) and higher for testlet, GPCM and multilevel models. In addition, bias was minimal in

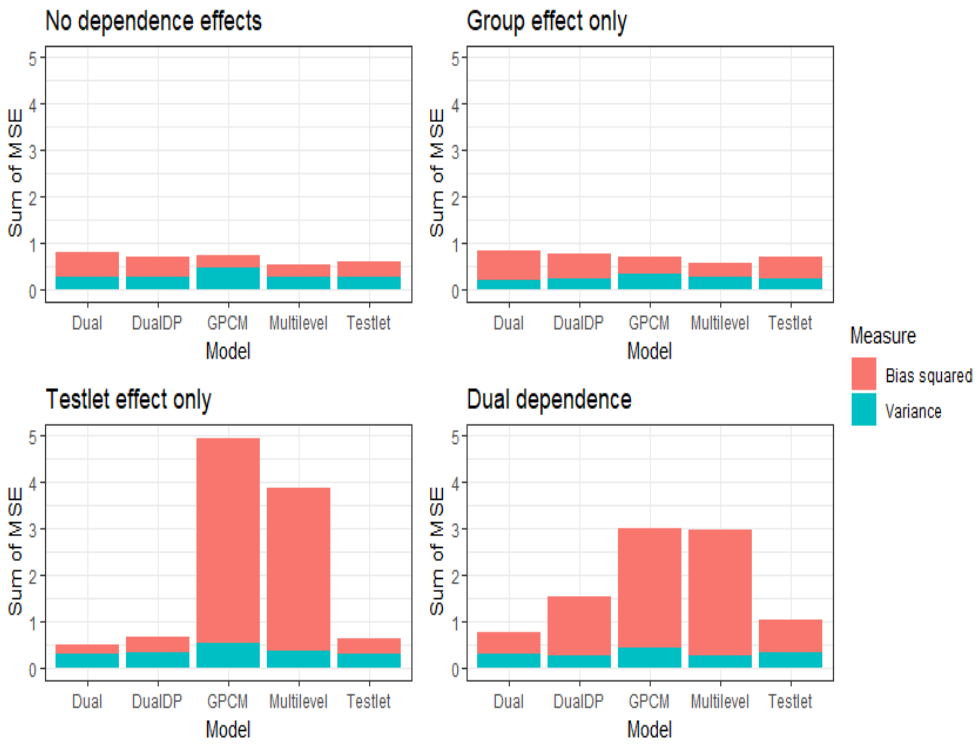


Figure 4.8: Random, systematic and total errors in the discriminant parameters local dependency conditions

item independence and increased as LID levels increase. The SEs in slope parameter were significantly influenced by LPD ($f = 0.15$), calibration ($f = 0.40$) and three-way interaction between LPD, LID and calibration model ($f = 0.16$). SEs were lowest in dual models in the absence of effects and highest in GPCM model for all estimation conditions.

The main effects for testlet ($f = 0.28$), calibration model ($f = 0.32$) were significant on discriminant parameter RMSE. The interactions between testlet and group effects, testlet and model effects, testlet, model and group effects were all significant with effect sizes of $f = 0.14$, $f = 0.14$ and $f = 0.17$ respectively. The Tukey multiple comparison analysis has shown that RMSE in discriminant parameters were lowest in item independence, followed by small effects and largest in large testlet effects. All pairs differed significantly. In addition, RMSE were lowest for both dual and testlet models and did not differ significantly and were higher for multilevel and GPCM models. The *post hoc* for interaction between group and testlet effects show lowest RMSE in item

Table 4.9: SE, Bias and RMSE in the discriminant parameters for different dependence conditions

		SE		**Bias**		**RMSE**	
Condition	Model	Mean	SE	Mean	SE	Mean	SE
NoneNone	GPCM	0.11	0.02	0.01	0.02	0.15	0.06
	Testlet	0.09	0.02	0.01	0.01	0.11	0.03
	Multilevel	0.09	0.02	0.01	0.01	0.12	0.03
	Dual	0.05	0.01	0.01	0.01	0.10	0.05
	Dual DP	0.07	0.01	0.01	0.01	0.10	0.03
NoneLarge	GPCM	0.12	0.04	0.02	0.02	0.17	0.07
	Testlet	0.09	0.03	0.01	0.02	0.13	0.06
	Multilevel	0.09	0.03	0.01	0.01	0.12	0.05
	Dual	0.08	0.02	0.02	0.03	0.16	0.08
	Dual DP	0.09	0.03	0.04	0.04	0.20	0.09
LargeNone	GPCM	0.11	0.02	0.03	0.04	0.19	0.09
	Testlet	0.09	0.02	0.01	0.02	0.12	0.06
	Multilevel	0.10	0.02	0.03	0.04	0.17	0.09
	Dual	0.09	0.02	0.00	0.01	0.11	0.03
	Dual DP	0.09	0.02	0.02	0.02	0.15	0.06
LargeLarge	GPCM	0.13	0.03	0.06	0.12	0.23	0.16
	Testlet	0.10	0.03	0.03	0.04	0.17	0.09
	Multilevel	0.09	0.02	0.09	0.11	0.27	0.15
	Dual	0.09	0.02	0.01	0.02	0.13	0.05
	Dual DP	0.09	0.02	0.02	0.03	0.14	0.07
SmallSmall	GPCM	0.12	0.03	0.05	0.08	0.23	0.12
	Testlet	0.10	0.03	0.02	0.02	0.16	0.05
	Multilevel	0.10	0.03	0.03	0.04	0.17	0.09
	Dual	0.10	0.03	0.01	0.02	0.13	0.06
	Dual DP	0.10	0.02	0.01	0.03	0.13	0.07
SmallLarge	GPCM	0.16	0.12	0.05	0.08	0.18	0.44
	Testlet	0.10	0.02	0.01	0.01	0.14	0.05
	Multilevel	0.09	0.02	0.02	0.05	0.16	0.10
	Dual	0.09	0.03	0.01	0.01	0.12	0.04
	Dual DP	0.09	0.03	0.02	0.03	0.14	0.07
LargeSmall	GPCM	0.16	0.12	0.05	0.08	0.18	0.34
	Testlet	0.11	0.04	0.02	0.03	0.16	0.07
	Multilevel	0.11	0.04	0.02	0.03	0.18	0.19
	Dual	0.10	0.04	0.01	0.01	0.14	0.05
	Dual DP	0.11	0.04	0.03	0.04	0.18	0.09

independence and in small testlet effects coupled with no, small and large LPD and were highest when large testlet effects were coupled with none, small and large group effects. The *post hoc* for three-way interaction between testlet and group effects and calibration models has shown that discriminant parameters RMSE were smallest in multilevel and testlet models in independence, for dual models in large testlet effects and were highest in GPCM and multilevel models in large testlet effects paired with

none, small and large group effects, implying that group effects have minimal effects on RMSE in discriminant estimates.

In summary, ignoring person clustering effects has more devastating effects on person ability parameters while ignoring item dependence has more negative effects on the difficulty and discriminant parameters.

4.3.8 Test reliability

Another measure of model comparison utilised in this study is the test reliability estimated by each model. The test information function (TIF) is a measure of accuracy in ability estimates. The estimated test reliability for each model for all dependence conditions are shown in Table 4.10.

Table 4.10: Test reliability for different dependence levels

LID	LPD	GPCM	Testlet	Multilevel	Dual	Dual DP
None	None	0.94	0.94	0.92	0.93	0.92
	Small	0.95	0.95	0.91	0.92	0.91
	Large	0.95	0.95	0.90	0.90	0.90
Small	None	0.91	0.96	0.72	0.90	0.90
	Small	0.94	0.94	0.89	0.92	0.91
	Large	0.96	0.95	0.87	0.90	0.91
Large	None	0.94	0.95	0.90	0.91	0.91
	Small	0.92	0.94	0.88	0.92	0.89
	Large	0.96	0.95	0.85	0.88	0.88

From the results in Table 4.10, for all simulated conditions, reliability estimates in GPCM and testlet models are higher than in models controlling for group effects. Highest test reliability values were noted for GPCM and testlet models in LPD and lowest in multilevel model in LPD, suggesting that reliability was affected more by group dependence effects than testlet effects.

4.3.8.1 Spearman-Brown prophecy

The test reliability of the local item and person dependence models were compared with the reliability of GPCM model (overestimated) to indicate the extent to which the tests length needs to be increased in order to reach the reliability it would have attained had items and persons be independent.

Table 4.11: Spearman-Brown prophecy for comparison with the GPCM model

LID	LPD	Testlet GPCM	Multilevel GPCM	Dual	Dual DP
None	None	0.94	1.46	1.30	1.47
	Small	1.02	1.73	1.69	1.72
	Large	0.88	2.44	2.25	2.51
Small	None	0.84	1.48	1.39	1.40
	Small	1.00	2.27	1.81	1.74
	Large	1.12	2.88	2.61	2.89
Large	None	0.75	1.63	1.80	1.88
	Small	0.73	1.95	1.51	1.39
	Large	1.02	3.23	2.51	2.65

The results from Table 4.11 show that as LPD and LID increase, the Spearman-Brown prophecy coefficient increases. This implies that when dependence levels increase, the test length must be increased in order for models accounting for these dependence effects to attain the reliability measured by independent items and persons models. For example, the test length for the parametric dual model would need to be increased by 1.30, 1.39 and 1.80 folds respectively in order to attain the test reliability measured by the GPCM model in the presence of none, small and large testlet effects when persons are independent. The test would need to be magnified by 1.28, 1.69 and 2.25 units respectively in none, small and large group effects when items are independent.

4.4 Discussion

The objective of this study was to assess the effects of ignoring local item and person dependency on item and person parameter estimates and their precision as well as the reliability and information provided by the test as determined by the models ignoring item and person clustering effects. The results have shown that AIC and BIC often

wrongly identified the correct model with the DIC being the fit statistics that correctly identified the data generating model, consistent with the findings from other studies (eg. Jiao & Zhang, 2014; Jiao et al., 2010, 2012). This is probably because BIC and AIC are computed from the data based on codes supplied while the DIC is inbuilt in the BUGS software employed for analysis.

In the presence of dual dependence effects, testlet and dual models were rated to have a better model fit ahead of multilevel and GPCM models probably because they are more complex or they have more parameters to be estimates compared to GPCM and multilevel models. However, this could imply that testlet effects have more undesirable consequences on model fit than person dependency effects, making models controlling for testlet having better model fit statistics. In support of this, higher fit indices were recorded in testlet effects only than group effects only. In the same view, dual and testlet models falsely accounting for testlet effects had the worst fit statistics when both persons and items are independent, again indicating that controlling for absent testlet effects negate model fitness while the same cannot be said about the multilevel model falsely accounting for absent group effects. Furthermore, fit statistics for models with no person and item clustering effects and models with person clustering effects are almost the same for all models while wide deviations were recorded for models with larger testlet effects, suggesting the argument that testlet effects have more negative impacts on model fitness than person clustering effects.

In local independence, the GPCM was the best fitting model, implying that it is not necessary to employ complex models accounting for item and person clustering effects if such are not present, hence there should be mechanisms to detect respondents and testlet clustering effects in IRT modelling. On the other hand, the multilevel performed better when persons were independent. Similar results on fitness statistics were recorded in studies conducted by Jiao and Zhang (2014), Min and He (2014),

the model-data fit and parameter statistics obtained from models dealing with testlet effects were shown to be better than the ones obtained from models ignoring testlet effects.

The thresholds in CCCs in LID and LPD are not ordered in the same way as the response categories, giving rise to a psychometric phenomenon termed "disordered thresholds" or "reversed deltas" (Stanke & Bulut, 2019). As the step calibration becomes more disordered, the central category becomes narrower. The disordered step calibrations may mean that the category definition is too narrow or that too many category options have been presented to respondents. Step disordering increases item discrimination and may indicate that the item is highly discriminating over a limited region of the variable but less informative in other regions and this high item discrimination is not synonymous with better function or being more informative. However, in the framework of polytomous IRT models such as the PCM (Masters, 1982), the mixture extension (Rost, 1991) and the GPCM (Muraki, 1992), there is no reason to assume why thresholds should be ordered (Wetzel & Carstein, 2014).

The occurrence of disordered thresholds in the current study maybe because of the existence of groups, concurring with Adams, Wilson and Wu (2012) who studied controversies in derivations of the PCM and reported that reversed threshold are merely a consequence of low frequency in the categories concerned and they do not affect the order of rating scales and reversed threshold often occur in subgroups of participants. As a results, the existence of reversed steps maybe because of existence of person subgroups and do not necessarily mean that the order of response categories is violated. This is supported in this study by monotonically increasing mean trait estimates for each category. According to Andrich (2006, 2011), ordered thresholds are relevant and central when categories are intended to form order. However, he goes on to say that the situation no longer holds in testlet effects where there is no reason for threshold

to be ordered. In fact, researchers argue that estimated parameter disorder is an indication of dependence among the underlying items (Andrich, 2006) or person (Adams, Wu & Wilson, 2012) and the more the dependence among items, the more the disorder (Andrich, 2011) and does not imply category disorder. Andrich (2006, 2011) and Adams et al. (2012)'s arguments probably explains why the results of this study depict disordered thresholds in item and person clustering effects. Similarly, software developers do not consider the disordered estimates as a violation of the intended order of response categories in items nor is it an indicator of model misfit nor is it an indicator of underlying category disorder (Masters, 1982). This probably explains why the threshold parameters in this study are disordered in item dependence but the CCC still depict monotonicity.

It can be inferred from the results that failure to account for both LID and LPD results in overestimation of the discrimination ability of the model as the CCC from the GPCM and testlet models are steeper in LPD and GPCM and multilevel models are similar in LID. The CCC for GPCM and testlet models are almost similar in dual effects, suggesting that the CCCs are more affected by group effects than item clustering effects.

In local independence, all the examinees were clustered in one posterior modal group, implying that the non-parametric dual model detected the absence of clustering effects well. However, on average, 97.5% of group membership was recovered in LID effects only, suggesting a slight effect of item dependence on group recovery. The recovery ability was weaker in LPD effects, increasingly so when the dependence levels increase. This resulted in the parametric dual model performing better than the non-parametric model when group effects were large.

The true ability parameter variance was on average well recovered by the true data generation model in each simulation, implying that in dual dependence effects, the dual model recovered the variance well. Ignoring LPD inflates the ability variance while ignoring LID underestimates ability variance and consequences increase with dependence levels. The effects of LID seem to be stronger as the GPCM underestimates the variance in dual effects. Similar results were observed in Jiao and Zhang (2014) where the dual dependency model recovered ability variance well but the same was overestimated by the testlet and underestimated by multilevel model. The testlet model overestimates the ability variances probably because it combines the ability and group variances that it does not control for. Jiao and Zhang (2014) made a similar argument, noting that ability variance estimates for the testlet was close to the sum of ability and group variances, implying that the ignored group variance is incorporated into the ability variance, resulting in overestimation of the variance by the testlet model. The multilevel model underestimated the variance probably because it splits the variance between ability and group. The underestimation of ability parameter variance by the multilevel model might be due to the testlet parameters averaging out part of the ability parameter variance. The impact of calibration model, LID and LPD on ability parameter variance recovery in line with Jiao et al. (2012) and Jiao and Zhang (2014).

The estimates of testlet and group interaction variances were close to true values for both dual and testlet models for none and small interaction effect while overestimated for large interaction effects. Large group variance was underestimated by the multilevel model while small group variance was overestimated by the dual models. In general, group and item cluster interaction variances were well recovered by the dual model, implying that the model can accurately detect the presence and absence of item and person clustering effects within the data, so that the best model can be utilised for estimation.

The results reported in the current study show that the levels of testlet effects did not much impact on the recovery of proficiency levels as correlations between true and estimated ability parameters were high. The results are consistent with other results in literature where despite observing large to huge testlet effects in their study researchers still observed high correlations between true values and estimated values (DeMars, 2006; Eckes, 2014; Min & He, 2014) for all prediction models or between parameter estimates of models independent items models and testlet based models (eg Baghaei & Ravand, 2016). However, the rank ordering of respondents by their traits was significantly affected by person clustering and dual dependence effects, implying that group dependence effects have a significant impact on ability parameter estimation than item dependence effects. As noted in literature, the existence of person clustering effects in a test may result in misleading results on vertical scaling (Jiao, Jin & Thum, 2010). Wang, Jiao and He (2011) intra-cluster correlation (group dependence effects) has little effect on ability parameters recovery (as determined by the correlations between the true and estimated parameters).

The average bias in estimation of ability, threshold and discrimination parameters was close to zero (0) for all models. This is probably because regardless the magnitude, bias usually cancel out when averaged across item since some deviation are negative and some are positive (DeMars, 2006). In addition, the means in person and item parameters were constrained to zero (0) for identifiability purposes. The bias in person parameters was not affected by LID and LPD and did not differ for calibration model, testlet effects, person effects and their interactions. Similar results were observed in earlier researches on testlet (Zhang, 2010; Jiao, Wang & He, 2013) where there was no evidence to indicate that the recovery of ability parameter was influenced by any of the factors under study. However, results from Jiao et al. (2012) and Jiao and Zhang (2014) showed the calibration model to significantly affect ability parameter estimation bias. The outward biases recorded in the current studies are similar to observations by

Zhang (2010) who reported high magnitudes of bias at the end of the ability continuum.

Although to a lesser extent than LPD, LID led to biased ability estimates as determined by the squared bias and variance (Figure 4.4), concurring with Paek et al. (2009) who reported largest bias in the target dimension (ability) when testlet effect was large. Overestimates in ability parameters were predicted when standard IRT models were used, ignoring local item dependency (Chang & Wang, 2010) although Kogar and Keleciolu (2017) discovered ability parameters estimated using standard IRT and testlet models to be close. However, their studies only handled LID and not LPD, which is shown in the current study to have more impact on ability parameter recovery than LID. A study by Wang, Jiao, Jin and Thum (2010) shows that the degree of group dependence could lead to biased person parameter estimation and misleading results in vertical scaling.

From the study findings, LID led to inward biases while LPD effects led to outward biases. However, these findings contradict Reese (1995, 1999) who reported minimal bias in ability parameters for zero to mid-level LID and underestimation of the scores of low ability test takers while scores of high ability test takers were overestimated. The underestimation of proficiency levels for low ability test takers and overestimation for high ability test takers probably explains why the average bias in LID (and LPD) is close to zero even for models showing high discrepancies on the scatter plot. In agreement with the current results, Jiao, Wang and He (2013) postulated that the testlet variance did not impact on the bias in ability parameter recovery. However, they also reported that the variability in ability parameter estimation bias increased as testlet effects increased and the pattern was consistent across all calibration models.

The absolute biases recorded in the presence of person effects were higher, rendering the factor to have statistically significant effects in agreement with Wang, Jiao and

He (2011), who concluded that the inter-class correlations (group effects), have not only statistically significant effect on the accuracy of the person ability estimation, but also large effect sizes. Jiao, Jin and Thum (2010) showed that the degrees of person dependence could lead to biased person parameter estimation and misleading results on vertical scaling. The effect of LID on absolute bias in the estimation of the ability parameter was significant according to the p-value, but with a negligible effect size, in line with Tuerlinckx and DeBoeck (2001) who reiterated that LID cause partiality in the estimation of the ability and item parameters.

The calibration model factor significantly affected the bias, with models ignoring person dependence effects recording more biases when the effects are present. Similar results were recorded by Jiao et al. (2012) where the effect of calibration model on ability parameter recovery was statistically significant. However, unlike in the current study where LPD and LID had significant effects although with negligible effect size, for their study no other factors significantly impacted on ability parameters.

All the factors affected ability parameter SEs, concurring with findings by Jiao and Zhang (2014). However, Jiao et al. (2012) study discovered that the model factor significantly impacted on random errors while all other factors had negligible effect sizes. Violation of local item independence assumptions led to underestimation of standard errors, concurring with other studies in literature (eg Zhang, 2010; Jiao & Zhang, 2014, Jiao et al., 2012; Yen, 1993; Wainer, 1995; Sireci, Wainer & Thissen, 1991; Kogar & Kelecioglu, Chang & Wang, 2010; Eckes, 2014) which also means the precision and test information would be overestimated (Ravand, 2015). The problem was exacerbated when testlet effects increased, also noted by Bradlow, Wainer and Wang, (1999). The higher SE recorded in testlet and dual models might have been due to increased number of parameters estimated in these models. Models ignoring LID effects recorded

lower SE while models ignoring LPD recorded lower SEs. Kogar and Keleciolu concluded that having lower values of errors for UIRT compared to models accounting for testlet effects maybe a sign of the existence of local dependency among items.

However, these results contradicted Jiao et al. (2012) findings that there was no consistent pattern in the models when the magnitude of LID and LPD were changed and recorded higher SE in multilevel and dual models for all dependence conditions. A possible explanation for dual and multilevel models having higher estimates in LPD is that they estimated both individual and group ability parameters separately, which probably increased the difficulty in separating the effect for individual persons and groups (Jiao et al., 2012). The same should also be the case for dual and testlet models having higher SE in item clusters. Similar findings in ability parameter bias and SE were recorded by Ravand (2015) and Baghaei and Ravand (2016) who reiterated that the estimates obtained for ability parameters were almost the same but SE were significantly different with the testlet response model (TRT) model having higher SEs.

As the total errors are a contribution of both random and systematic errors in parameter estimation, they present a better picture of the discrepancy between true and estimated parameter values. Because RMSE is the deviation around true values, it includes the effect of bias. The SE is the deviation around the expected value and does not included bias, thus the RMSE is the true measure of standard deviation. In the current study, both the RMSE and SE in the ability parameter were affected by all factors under study and their interaction. Ignoring person dependence effects by using independent-persons models has been shown in the current study, to inflate total errors in ability parameter estimation, as RMSE were generally lower for dual and multilevel models than for testlet and GPCM, leading to the conclusion that models considering local dependency provided better model-data compatibility than models that do not. These results are in agreement with results from other studies (eg Jiao et al., 2012;

Kogar & Kelecioglu, 2017; Jiao & Zhang, 2014; Zhang & Jiao, 2014). Contrary to these findings, Zhang (2010) found no evidence to conclude that any factor impacted on the RMSE in ability parameter estimation. However, controlling for dual dependence effects when persons and items are independent led to increased total errors, implying that controlling LPD spuriously can jeopardize the estimation of ability parameters, although total errors for independent items and persons are less than 0.40 and hence acceptable (Amil & Sahil, 2016; Barnes & Wise, 1991).

The total errors in ability parameter estimation were also significantly increased by LPD, consistent with previous studies (Zhang, 2010). Drester (2004) reported that when all items in a test were in testlets, the RMSE for ability parameters were higher when dependencies were ignored, which is the case in the current study, where all the test items belonged to testlets. The ability parameter estimates were poor in models that assumed independent persons when persons were clustered. However, although the DPM non-parametric model had estimates that were close to those from the parametric dual model, the parameters from the DP model are a little bit weak. These differences in ability parameter are related to the detection of latent group membership since latent classes are characterised by differences in the group specific ability parameters across latent classes. The analysis of group membership detection has shown that some individuals were wrongly classified by the DPM model. Overestimation of the individuals' scores increase the possibility of wrong classification of individuals based on their proficiency levels (DeMars, 2006; DeMars, 2012; Wang & Wilson, 2005; Yen, 1993) and this may lead to erroneous decisions. As such, it is important that correct modelling techniques be employed in order to make rightfully informed decisions.

The estimation accuracy measured by bias, SE, RMSE decrease as group effects increased. Group effects significantly affected the RMSE in ability parameter, accounting for most of the variance (as explained by the significant effect sizes) in the bias and

RMSE in estimation of ability parameters while LID has more impact on random errors. Despite the GPCM model recording smallest SE in ability parameters in LID, the same recorded highest bias and highest total errors in dual dependence effects. In addition, the ratio between SE and sampling errors was more than 1. This implies that the model overestimates the precision of ability parameter estimation in LID effects, resulting in biased estimates in group (and dual dependence) effects. The lower RMSE in the better fitting models was as expected that models that fit the data generating process better should have lower total errors of estimation.

Difficulty parameters did not seem to be mush impacted on by LID and LPD as all true and estimated value correlations are > 0.7 and hence are acceptable (Field, 2013; Amil & Sahin, 2016). This supports Wainer and Wang (2000) and Wainer et al. (2000) observation that item difficulties were still well estimated even when testlet effects were ignored. However, the fact that correlations were highest when there were no item and person clustering effects may mean that both item and person clustering negatively affect the rank-ordering of questions by their difficulty. Furthermore, the correlations for dual models were slightly lower in local independence, implying that controlling for effects that do not exist may negatively impact on the rank ordering of items on the difficulty continuum.

The study results have shown that controlling for absent testlet effects and ignoring present testlet effect overestimates the threshold parameters. The results are in agreement with findings of Jiao and Zhang (2014) whose studies reported overestimates in the threshold parameter, significantly affected by both item clustering and calibration model and their interaction effects with significant effect sizes. They also reported overestimated thresholds increasingly with increase in item dependence levels when independent items models were used for estimation. The current results contradict Wainer and Wang (2000), where difficulties were well recovered even when testlet

effects were ignored. However, this could be because their study was on the item location (difficulty) while the current study assessed the effects of the recovery of the step (threshold) parameter. Furthermore, Wang and Wilson observed the difficulties estimates from the independent items Rasch model to shrink slightly towards the mean compared to estimates from the Rasch testlet model.

The total errors in models controlling for testlet effects were much smaller than for models ignoring the effects. In line with results from Jiao et al. (2012), all factors significantly affected RMSE in threshold parameter estimation while Jiao and Zhang (2014) observed LID, model factors and their interaction to significantly affects total errors in estimation of thresholds. Total errors in models ignoring LID increased with the magnitude of item clustering effects. Similar results were recorded by other researchers in item difficulties (DeMars, 2006) who recorded higher RMSE for independent items models. However, even though independent-items models recorded lowest random errors, they recorded highest total errors in thresholds, implying that (1), bias is higher in these models, (2), the models overestimate the precision with which threshold parameters are estimated. The average RMSE in testlet and dual models are consistent with the expectation that better fitting models usually have lower total estimation errors. However total errors were highest in dual models in local items and person independence, implying that controlling for absent dependence effects can compromise item parameter estimation.

The discriminant parameters in GPCM and multilevel models in large testlet effects were very low, significantly less than 0.7 (see Field, 2013; Sahin & Amil, 2016) when compared to ability and threshold parameter correlations, implying that ignorance of testlet effects has more negative consequences on slope estimation than step and proficiency levels. Min and He (2014), Kogar and Kelecioglu (2017) also reported that the intercept parameters were less influenced by the model factor in LID than the slope

parameters predictions. The low correlation between true and estimated discriminant parameters are comparable with findings by Reese (1995) who recorded low correlation of about 0.26 in high levels of LID.

Bias in slope parameters was significantly affected by LID and model factors and was higher for multilevel and GPCM models, suggesting that ignorance of LID overestimates the discriminant parameter. The current studies are in agreement with Baghaei and Ravand (2016) who recorded smaller errors of estimation in the standard IRT models and higher discriminant parameter estimates compared to their TRT counterparts. In support of these project findings, Yen (1993) proclaimed that discrimination parameters may be biased in item dependence and consequently, ability parameters may be biased because scoring weights depend on the discrimination (Embreston & Reise, 2000). However, the current results contradict other studies where slopes were consistently underestimated by independent-items models increasingly with LID levels (DeMars, 2006). Similar to the current results where dual models had biased results in item independence, DeMars (2006) reported that using testlet effects models for data that has independent items led to small positive bias in the slope. Bradlow, Wainer and Wang (1999) conducted a simulation study for tests comprising of independent and testlet items and discovered that when testlet effects were not modelled, item discrimination were underestimated for testlet items and overestimated for independent items. The results contradict Luo (2018) who discovered bias not affected by the testlet effects but by sample sizes.

The higher discriminant parameter in GPCM model also reflected in steep CCCs and higher test information values for the model. The current results study are in agreement with findings of Min and He (2014) who reiterated that when testlets were analysed with independent-items models, there was a greater error in the slope parameter. The overestimation of the slope parameter increasingly with levels of LID are in agreement

with findings by Reese (1995) whose study discovered that the zero, low and medium LID levels led to overestimation of the discriminant parameter while high LID levels resulted in stronger overestimation of the slope parameter. However, controlling for absent testlet effects did not bias the estimation of the slope.

The study has shown that both model and LPD significantly affected random errors in discriminant parameters. Although highest errors were recorded for the GPCM ignoring both LID and LPD, the effects of LID inferred from the ANOVA model were not significant. Different results have been reported in literature. Some studies concluded that testlet effects significantly affected SE in discriminant parameter estimates and when testlet-based tests were analysed using independent items models, there was greater error in the estimates (Min & He, 2014; Luo, 2018). Highest SE were recorded for the large testlet effects, no significant differences between SE in medium and small testlet effects (Luo, 2018). On a different note, some studies (Kogar & Keleciolu, 2017; Chang & Wang, 2010; Baghaei & Ravand, 2016) discovered that the average error values of discrimination parameters obtained from models ignoring testlet effects were lower under all simulation conditions compared to those yielded by models accounting for testlet effects. However, despite recording greater errors in models controlling for testlet effects, Kogar and Keleciolu still recommended testlet models for accurate estimation of the slope.

Similar to findings in trait estimation, controlling for group effects significantly reduced SE in slope parameters and large group variances resulted in significantly smaller SE than large testlet variances. This might imply the overestimation of precision of slope measurement. In LID only, dual model spuriously controlling for absent group effects underestimated the precision of estimation of the slope parameter as the ratio between systematic and random errors was less than unity. This is in agreement with Schochet (2005) who reported that spuriously controlling for absent group effects can result in

underestimates of the precision with which the parameter is estimated.

The LID and model factors significantly affected total errors in slope estimation, with greater errors in models ignoring item effects. The findings are in agreement with other studies who reported testlet effects (Luo, 2018) and model effects (DeMars, 2006) significantly affected the RMSE in estimation of discriminant parameters. RMSE was significantly highest for large testlet effects and there was no significant differences between medium and small effects. The existence of testlets in a test has been reported to reduce the accuracy of estimation of model parameters (Wainer & Thissen 1996; Wainer & Wang 2000). Baghaei and Ravand (2016) concluded that the RMSE show that the discrimination parameter being estimated with the same precision across models while the discrimination parameter was estimated with higher precision for models catering for testlet effects. Similar results were reported in the current study with overestimation recorded in LPD instead. In addition, they reported the standard IRT models and testlet response theory (TRT) to perform equally well for parameter stability across different populations though the performance for the TRT model was slightly better. Further, the standard IRT models recorded smaller errors of estimation and higher parameter estimates compared to their TRT counterparts.

The current findings have shown that other than the underestimation of slope precision, group dependency has little effect on item parameter recovery, in agreement with Mislevy, (1987), Mislevy and Sheehan (1989), who have shown that the use of collateral information (person categories such as gender) will lead to smaller errors in estimation of item parameters and reduce RMSE on ability parameter estimates. However, Adams, Wilson and Wu (1997) discussed a two-level model incorporating the treatment of latent proficiency as a dependent variable in a regression model and argue that the inclusion of collateral variables has a negligible effect on accuracy of item parameters and does not lead to substantial decrease in the RMSE of ability predictions. However,

Adams, Wilson and Wu go on to say that although the collateral information plays a minor role in improving item parameter estimates, it can have a significant effect on expected *a posteriori* ability predictions.

In LID, the choice of model has a bigger impact on item parameter estimation than trait estimates. The threshold and discriminant parameter estimates were biased in LID. However, most studies (e.g Bradlow et al. 1999; Wainer & Wang, 2000; DeMars, 2006) reported accurate recovery of difficulties using the independent-items models though with higher RMSE and biased discriminant parameters. This is probably because the current study dealt on threshold (step) parameters which were reported by Jiao and Zhang (2014) to be biased with high RMSE in LID and not the location (difficulties) reported in most studies. Moreover, some of the studies referred to are on binary and not polytomous item responses. On the other hand, the estimates of item parameters were not much affected by LPD while the ability parameter estimates were biased with high RMSE. However, both dependence conditions led to underestimation of SEs and overestimation of precision, based on the ratio between systematic errors and random errors that is more than unity in value, in agreement with studies in literature (e.g Schochet, 2005; Hedges, 2004; Walsh, 1947).

The magnitude of LID does not seem to have an effect on the accuracy of ability parameter estimation, concurring with results from other scholars (e.g Zhang, 2010). However, the study only investigated 3 levels of LID (0, 0.25 and 1). Maybe different results could have been observed if higher levels of LID were considered.

Testlet data, if ignored by applying standard IRT models with conditional independence of items assumption, resulted in overestimation of proficiency estimates as well as bias in item difficulty and discrimination parameters. However, accounting for testlet effects that do not exist may negatively affect the rank ordering of items and persons

according to their difficulty and ability levels respectively. This suggests that there is need to effectively detect item and person dependence so as to account for their effects if they exist, or rather use standard models assuming conditional independence of items and testlets. There are no standard accepted rules of thumb for judging testlet effect parameters. However, Eckes and Baghaiei (2015) through simulation studies show that testlet effects smaller than 0.25 are negligible. Item difficulty $RMSEs \approx 0.25$ and person ability $RMSEs \approx 0.40$ have been taken as evidence of precise estimation in previous simulation studies (Barnes & Wise, 1991; Hulin, Lissak, & Drasgow, 1982).

The results show that the discriminant estimation SEs in LPD are lower than SE values in person independence, in agreement with findings of Jiao, Wang and He (2011). The main effects of group effects accounted for most of the variation in bias, SE and RMSE in the estimation of ability parameters although the test length played the most significant role in the correlations between true and estimated ability parameters. All models estimated the ability parameter satisfactorily in testlet effects only (cf DeMars, 2006) though the estimated test information and reliability were overestimated for the independent-items and persons models when testlet effects are indeed present. However, the current study only investigated a small range of testlet effects (0, 0.25 & 1). Larger values of LID are worthy further investigations. In addition, the study was based on Markov Chain Monte Carlo (MCMC) estimates only. The estimation of the proposed model using maximum likelihood based methods can be investigated.

The study results have shown that the CTT reliability computed by squaring the true-estimate correlations for ability parameter were affected by dual dependence effects where reliability was lower for models ignoring testlet effects (cf DeMars, 2006), lowest in the presence of person dependence effects. In the presence of LID, the standard error of θ were small for the GPCM and multilevel, (reliability is high), the examinees

weighted mostly in one θ group while in the presence of LPD, the θ values were narrower for the GPCM and testlet models ignoring group dependence effects while the standard errors were larger and reliability and information lower for models accounting these dependence effects.

The IRT test information curves show peaked values for GPCM and testlet models, implying that the GPCM and testlet models differentiate the examinees according to their proficiency levels better than other models increasingly better in large dual effects. However, these results contradict other findings since GPCM and testlet models have higher total errors and less correlations with true ability parameters leading to the conclusion that the two are misleadingly overestimating the reliability and test information levels. In addition, the ratio between sampling and systematic errors in the models suggest that the ability parameters precision is being overestimated. Despite the GPCM and testlet models recording the smallest SE in LID, the same model recorded highest total error in estimation of ability parameters in group dependence effects. This implies that ignoring testlet effects result in overestimation of precision of ability parameter estimation while ignoring group dependence led to biased results. Furthermore, despite the GPCM model having recorded the worst fit statistics when items are dependents, the model reported lowest SE, best reliability and test information, again suggesting that these psychometric properties are being overestimated. In fact, it could be that overestimation has caused the poorer model fit. However, as expected, the IRT based reliability was accurately estimated by multilevel, testlet and dual models that accurately estimated this decrease.

Controlling for testlet effects resulted in lower test information than ignoring them, in line with other results in literature (Baghaei, 2010; Zhang, 2010; Wang & Wilson, 2005; Wainer & Wang, 2000; Bradlow et al., 1999; Yen, 1993; Wainer & Thissen, 1996). The current reliability results are coherent with results from other studies where reliability

increase from around 0.91 in item independence to 0.92, 0.93 and 0.97 in low, medium and high LID respectively (Reese, 1995; Zhang, 2010). In addition, similar results for CTT and IRT based reliability in testlet effects were reported by DeMars (2006) where the independent-items models overestimated the IRT reliability while testlet models properly estimated the decrease in reliability.

The decrease in reliability for the multilevel model accounting for person effects, and the testlet model accounting for testlet effects as well as the dual models accounting for both dependence effects suggest that failure to account for both effects result in overestimation of reliability. However, most studies on test reliability in literature are on testlet based data. The current results are consistent with findings in other studies that ignoring person clustering effects can lead to serial overestimates of precision levels (Wang, Jiao & He, 2011; Schochet, 2005; Walsh, 1947; Hedges, 2004) increasingly as the heterogeneity across clusters increases (Walsh, 1947). Moreover, the results for items and persons have shown that mistakenly introducing absent clustering effects can lead to gross underestimates for precision levels, in agreement with Schochet (2005) who reported that falsely introducing spurious clusters may result in serious underestimation of precision. In summary, ignoring both LID and LPD effects result in underestimates of SE, thereby overestimating the precision of estimation of trait levels. The same conclusion can be inferred from high biases and low standard errors as well as Spearman-Brown prophecy which show that in ocal item and person dependence effects, the test has to be elongated in order for models accounting for these dependence effects to attain the reliability estimated by the GPCM assuming independent persons and items, increasingly so when the dependence levels increase.

The overestimation of test reliability for models ignoring LID may be due to the fact that when items show LID, they are correlated resulting in possibly stronger correlation with the total score. Test reliability will then be overestimated by excess correlation

among testlet items and thus the true reliability will be lower than the estimated reliability. However, while some researchers incorporated independent and testlet items in their tests (eg Zhang, 2010; Bradlow et al., 1999), this study is based on test items that belong to testlets only.

As indicated in earlier work by Swaminathan et al. (2003), sampling and random errors in IRT estimation can have important implications on the development and analysis of test data. The overestimation of the precision of ability parameter might result in inaccurate inference and false prediction (Eckes, 2014; Ip, 2000; Reese, 1995; Bradlow et al., 1999) premature termination of exam in computer adaptive testing where the criteria for termination is the standard error of ability parameter estimates (Wainer & Wang, 2000). In programmes targeted to identify beneficiaries based on trait levels, failure to account for items clusters or person clusters may lead to wrong beneficiaries being selected. There were no significant differences in the parametric and non-parametric dual models, especially in local independence and for smaller person clustering effects. However, margins were slightly higher for large dependence effects as the recovery of group membership was also lower.

4.5 Conclusion

Testing is evolving. Psychometric tests are usually constructed in units more than a single binary item where more than one option is supplied for each item, e.g Likert scale item. Such units are sometimes aggregations of items on a single stimuli. Testlets are frequently used in standardised tests for efficient usage of time by asking related questions as items with related content are more appropriate for real life and are regarded to measure high proficiency levels. As such, using testlets in large scale and standardised tests is inevitable. However, although testlets test are beneficial in measurement, they violate one of the most important assumptions, local item independence. In addition,

respondents are usually categorised according to their proficiency levels, making respondents in one cluster to have similar or correlated responses to items in one testlet and these clusters maybe latent, unknown before hand. The objective of examiners is not to do away with testlet tests and clustered respondents but rather to come up with measurement methods that will produce the best results in the presence of correlated items and examinees.

The CCC for GPCM and testlet models were steeper than the CCCs for dual and multilevel models accounting for person dependence effects, implying that the former have higher discrimination ability compared to the later. In accordance with the phenomenon that the higher the discrimination ability the better the θ values will be, the proficiency levels from the GPCM model assuming independent items and independent persons are better than those derived from models controlling for both forms of local dependence. In addition, the GPCM have better SEs, reliability and test information values compared to the dual models. However, despite all these positive factors in favour of the GPCM model, the correlations were lowest and the bias and total errors highest in the GPCM and testlet models in dual dependence effects, suggesting that the reliability, discriminant and test estimation parameters could have been overestimated. Furthermore, the results from the Spearman-Brown prophecy and the ratio between systematic and random errors greater than unity in value suggest the GPCM model overestimates test information, reliability and the precision of proficiency estimation.

In summary, the results have shown that in LID, testlet IRT models demonstrated good model fit, small bias and satisfactory accuracy in item parameter recovery while in LPD, multilevel models demonstrate prowess in ability parameter recovery. The dual models performed well in testlet, group or dual dependence effects. Group effects have the most influence on the estimation of ability parameters, accounting for more than half of the variances in bias, abias, SE and RMSE. In general, average indices

of bias, abias, SE and RMSE increases with group dependence effects. Controlling for testlet effects resulted in lower test information than ignoring them, in line with other studies in literature. However, the current study has shown that ignoring person group effects (and dual dependence effects) led to even more overestimation of the test information than ignoring them.

In local item and person independence, adding extra parameters where they were not needed over-capitalised on chance and increased the error variance. In general, the use of a more complex model when a less complex model was adequate led to higher RMSE but not biased results. To avoid the errors introduced by using an inappropriate, over-fitting model when items and persons are independent, the non-parametric dual model should be used first as a diagnosis tool before the appropriate model can be selected for estimation. The dependence effects should be deemed significant enough to warrant modelling if they are more than 0.25. If the data shows the existence of person groups that are not known in advance, the non-parametric model can then be used to detect the group memberships. However, given the ease and speed with which the parametric models can be run on open source software, in contrast with the DPM ,kmjnhbvcg1y90 model that requires more computational time, practitioners may use the parametric dual dependence model if the group effects are known or assumed.

Chapter 5

Effects of changing sample and group size on LID and LPD

5.1 Introduction

This chapter assessed the effects of changing properties of respondents on parameter recover. One of the major factors that affect the stability and accuracy of parameter estimates is sample size used to calibrate the items (He & Wheadon, 2012; Zhang, 2010), which stem from the probabilistic nature of IRT models. A substantial amount of simulation studies have looked at the influence of sample size on the estimation of IRT model parameters. The first study by Nord (1968) investigated the sample size and test length requirements for estimating item parameters accurately in a 3PL and concluded that a minimum of 50 items and 1000 examinees were required to estimate the discrimination parameter with high accuracy. After subsequent studies by other scholars, 1000 was considered to be the minimum sample size requirement for item parameter estimation in IRT modelling. However, later studies investigated the use of less than 1000 examinees on item parameter estimation and sample sizes between 200 and 500 were suggested as feasible sample sizes for IRT modelling.

Wang and Chen (2005) investigated how item parameter recovery, standard error of estimates and fit statistics are affected by sample size and test length in the Rasch Models (Rasch, 1960) and the Rating Scale Model (Andrich, 1979). DeMars (2003)

also studied the effects of sample size, test length, category size and ability distribution on parameter estimation for polytomous items using the Nominal Response Model (Bock, 1972) and suggested that sample size might depend on the number of parameters per item. Swaminathan (2003) studied the effects of small sample sizes on parameter accuracy for 2PL, 2PL and 3PL IRT models and concluded that item parameter estimation in small samples can only be accomplished through a Bayesian approach. Stone and Yamamoto (2004) studied the effects of sample size on parameters of the Rasch. Results from the studies generally concluded that the magnitude of variation between sample estimates decreases with increasing sample size. However, most of these studies were done in local independence. DeMars (2010) argued that the proficiency parameter θ may not be well estimated for short tests. Furthermore, Reise and Yu (1990) reiterated that at least 500 examinees are needed to ascertain respectable correlations and RMSE and about 1000 and 2000 examinees if structural parameter recovery is crucial.

Kogar and Kelecioglu (2017) conducted a study to assess the effects of testlet length and sample size on ability and item parameter estimation, considering sample of size 250, 500 and 1000 respondents. Luo (2018) assessed the effects of small (0.25), medium (0.5) and large (1.0) testlet effects and small (500 respondents), medium (1000 respondents) and large (2000 respondents) samples sizes on item parameter recovery and 30 test items to compare item parameter recovery in a 2PL for different estimation methods. Zhang (2010) studied the effect of sample size in testlet-based tests with samples of size 250, 500 and 1000 examinees and testlet effects of size 0.25, 0.5, 0.75 and 1. These researchers observed psychometric properties to improve with sample size increase.

An example of the "Eye tracking data" in Congdon (2006) and Lunn et al. (2012) used a maximum of $M = 19$ clusters for allocating a sample of size $n = 104$ cases using the

Dirichlet process mixture model for a Poisson Gamma mixture to model heterogeneity. They used a maximum of $M = 10$ clusters for the "Galaxy data" which has a prior belief of 6 clusters and $n = 82$ cases. Ishwaran and James (2002) found at least 5 to 6 clusters on the Galaxy data with the DP approach using an inverse gamma prior for the cluster variance and 4 clusters under the uniform prior. Congdon illustratively explained the determination of number of clusters (Congdon, 2003, pp 203) highlighting that taking $\varepsilon = 0.001$ and $\alpha \sim Unif(0.5, 10)$ will guarantee a maximum of $M = 50$ number of clusters, highlighting that M should reflect the nature of the data. Jiao and Zhang (2014) and Jiao, Wang and He (2011) used 1000 students grouped into 40 groups of 25 respondents each while Jiao et al. (2012) employed 1000 respondents grouped into 50 groups of 20 respondents each. However, their membership variable was fixed and supplied as data in the Bayesian model. Choi (2014) and Cho, Cohen and Kim (2013) used the Dirichlet distribution to allocate subjects into 4 groups.

The objective of this study is not to determine the minimum number of respondents required for parameter estimation but rather focuses on assessing the effects of sample size on the stability and accuracy of ability and item parameters in LID and LPD. As such, the study considered assessing the effects of sample size by making the number of items and response categories constant at 36 items (in 6 testlets) and 3 options respectively. Since as low as 350 examinees can be used for a test of 30 items, the researcher considered using a minimum of 400 respondents since the dual models have more parameters when compared to the 3PL and testlet models, using the Bayesian estimation criteria that has been noted to perform better for smaller samples (Swaminathan, 2003).

Experiments in section 4.2 were repeated for samples of size 400 (small), 1000 (medium) and 2000 (large) respondents for groups of size 5, 20 and 40, guided by the group sizes explored in literature, so that the model performance can be inferred for small to

relatively large cluster sizes. The design for the study thus becomes a four factor design with 3 local items effects (0, 0.25 and 1) × 3 local person dependency effect (0, 0.25, and 1) × 3 calibration samples sizes (400, 1000 and 2000 respondents) × 3 groups (5, 20 and 40) × 5 calibration models (GPCM, testlet GPCM, multilevel, dual GPCM and dual DP). However, because of the longer time required to run data for 20 and 40 clusters, the analysis for variant cluster sizes was only done for large person dependence effects only, compared to the control where persons are independent. The small person dependent effects were only compared for changing sample sizes with the number of groups fixed at 5.

5.2 Results

5.2.1 Goodness of fit statistics

According to the results in Chapter 4 (section 4.3.2) and according to literature (eg Cho et al., 2013; Jiao & Zhang, 2014) the results from the AIC (and BIC) are less accurate at detecting correct model. As such, model comparison in this (and following chapters) was done using the DIC produced by OpenBUGS/ MultiBUGS. The DIC fit statistics used to evaluate how well the data fits the models (Table 5.1) and the model with the lowest statistics was the best fitting model.

According to the goodness of fit statistics, in local independence, the GPCM model had better fit. In dependence of varying magnitude, the dual model outperformed all competing models across all sample sizes. In LID and LPD only, the non-parametric dual model was better or performed similar to the data generating model for smaller samples. However, for large group dependence effects the multilevel model performed better for all sample sizes. Large testlet effects resulted in a wider margin in fit statistics for models incorporating and those ignoring clustering effects and the margin of differences increased with sample sizes. Person clustering effects did not result in much differences in goodness of fit for models ignoring the effects and models accounting for

Table 5.1: DIC fit statistics for 400, 1000 and 2000 respondents

LID	LPD	Sample size	GPCM	Testlet	Multilevel	Dual	Dual DP
None	None	400	21160	21230	21150	21230	21150
		1000	49610	49710	49620	49820	49820
		2000	106200	106300	106400	106400	106300
	Small	400	18950	18990	18940	18730	18710
		1000	50650	50710	50640	50710	50710
		2000	105000	105100	105000	105000	105000
	Large	400	19620	19620	19570	19610	19500
		1000	43990	44050	43980	44140	44100
		2000	88240	88360	88170	88240	88240
Small	None	400	19740	18590	19730	18590	18500
		1000	51770	47520	51770	47440	47500
		2000	104800	98320	104800	98320	98240
	Small	400	18500	17660	18940	17610	17490
		1000	46300	43420	46270	43280	43320
		2000	97130	91000	97110	90820	90750
	Large	400	18240	17320	18230	17620	17510
		1000	46040	42720	45990	42590	42710
		2000	89930	84320	89670	84010	83990
Large	None	400	23690	19590	23700	19450	19380
		1000	57600	47410	57580	47270	47300
		2000	120200	95710	120200	95380	95280
	Small	400	19800	16410	19800	16360	16300
		1000	46500	37560	46480	37420	37480
		2000	86320	79400	86100	79310	79290
	Large	400	19460	15620	19640	15540	15450
		1000	50820	42120	50780	41960	42000
		2000	90720	73640	90570	73200	73110

them as the sample size increased.

5.2.2 Variance recovery

According to results in Table A6 and Table A7 (Appendix 1), for independent items and persons, the ability parameter variance estimates were generally well recovered for all calibration models and approach their true values as the sample size increases. In LPD only, the ability parameter variances increases with sample size, becoming overestimated for the testlet and GPCM models. Although the overestimation seem to be increasing with sample size, it is highest for a sample of 2000 respondents and lowest for 1000 respondents (not 400 respondents as expected). In LID only, the GPCM and multilevel ability parameter variances were slightly underestimated for small testlet effects and well underestimated for large testlet effects and they changed slightly as

the sample size increases. In addition, large testlet effects led to underestimation of ability variances which increased with sample size in item independence. When large testlet effects were combined with small and large person clustering effects, the ability variances were reduced in magnitude. However, the testlet and dual models recover the variance fairly well in LID and the variance increases with sample size. In large person effects, the multilevel underestimated the ability parameter variances across all sample sizes while the GPCM overestimated for large samples and underestimated for small the sample of 400 respondents. The dual model recovered the variances fairly well across sample sizes while the testlet model overestimated the variances increasingly with sample sizes. However, the overestimation was minimal when clustering effects were coupled with item effects, the variance in ability parameter estimates increase with sample size (group size) for all models.

The testlet and dual models detected the absence of group by testlet interaction effects and the performance (effects approaching 0) as the sample size increases. The non-parametric and parametric dual and multilevel models recovered the absence of group dependence effects fairly well, recovery improving with increase in sample size. On the other hand, the interaction effect variances in the testlet and dual models decrease with sample size.

5.2.3 Ability parameter recovery for changing group ad sample sizes

The effects of sample size on ability and item parameter recovery in the presence of item and person clustering effects was assessed using the Pearson's product moment correlation coefficients for correlations between true and estimated values, systematic/sampling errors (biases), absolute biases (abias), random/standard errors (SE) and root mean square errors (RMSE). The correlations between the true and estimated ability parameter values for varying group and samples sizes are given in Table A1. From the results in Table A1, the effect of group size is not very clear as the trend

is not the same across sample sizes, suggesting the possibility of interaction between sample size and group size. However, the correlations are high for ungrouped respondents and generally low in models ignoring person effects for all group and samples sizes.

Further analysis on the errors of estimation in the ability parameters in Table A8 and the analysis of variance ANOVA has that total errors decrease with sample sizes for ungrouped respondents and for models controlling for group effects while the errors increase with sample sizes for models ignoring person dependents effects. In addition, the errors of estimation increase as the number of groups increase across all models and it became more difficult for the non-parametric model to detect respondent groups as the number of simulated groups increased from 5 to 40. However, the dual and multilevel models performed significantly better across all samples and group size and the non-parametric model, although the detection of groups was becoming increasingly difficulty as the number on groups increased, it still performed better than the independent persons models. The computation time required for 20 and 40 groups was significantly higher. As a results, further analysis on the effects of increasing sample sizes in the presents of dual dependence effects was done for 5 groups only.

The analysis for interaction between group and sample size has shown that in group effects only, the errors increase with the number of groups in all models for 400 respondents with the 20 groups recording the worst errors for the GPCM and tesltet models assuming independent persons. Twenty groups recorded lowest errors for 1000 respondents across all models while errors increased with the number of groups for samples of size 2000. For dual dependence effects, errors increased with group size for dual and multilevel models for 400 and 1000 respondents although the errors were generally low while 20 groups were best and worst for 400 and 1000 respondents respectively, for 400 and 1000 samples. Errors decrease as the number of groups increase across all models for 2000 respondents.

Table 5.2: True-estimated ability correlations for 400, 1000 and 2000 examinees

Sample size	Testlet	Group	GPCM	Testlet	Multilevel	Dual	Dual DP
400	None	None	0.98	0.97	0.96	0.96	0.96
		Small	0.87	0.87	0.94	0.94	0.94
		Large	0.72	0.73	0.92	0.92	0.92
	Small	None	0.95	0.94	0.96	0.96	0.96
		Small	0.87	0.90	0.95	0.95	0.95
		Large	0.78	0.79	0.94	0.95	0.94
	Large	None	0.97	0.98	0.94	0.93	0.92
		Small	0.81	0.76	0.92	0.92	0.92
		Large	0.68	0.76	0.90	0.92	0.91
1000	None	None	0.98	0.97	0.97	0.96	0.97
		Small	0.92	0.92	0.96	0.96	0.96
		Large	0.71	0.72	0.95	0.95	0.95
	Small	None	0.95	0.96	0.95	0.95	0.95
		Small	0.88	0.89	0.94	0.95	0.95
		Large	0.69	0.70	0.94	0.94	0.94
	Large	None	0.97	0.97	0.95	0.95	0.95
		Small	0.83	0.81	0.92	0.93	0.94
		Large	0.69	0.76	0.93	0.94	0.94
2000	None	None	0.97	0.97	0.96	0.96	0.96
		Small	0.87	0.88	0.95	0.95	0.95
		Large	0.72	0.72	0.92	0.92	0.92
	Small	None	0.95	0.90	0.96	0.96	0.96
		Small	0.87	0.90	0.95	0.95	0.95
		Large	0.78	0.79	0.94	0.95	0.94
	Large	None	0.97	0.97	0.96	0.95	0.94
		Small	0.81	0.85	0.96	0.96	0.94
		Large	0.68	0.70	0.98	0.97	0.95

The ability correlations for 400, 1000 and 2000 respondent samples for 5 groups are shown in Table 5.2. It can be inferred from the results that the ranking of respondents according to their their were well recovered by all models ($r > 0.7$) across sample sizes, except for the GPCM and testlet model in large LPD effects. The correlations do not seem to be affected by sample size but were rather affected by LPD, as correlations for each model and each condition remained almost constant across sample sizes.

Figure 5.1 depicts the ratio of systematic errors to model standard errors (making up the total error of estimation) for different sample sizes. According to the results, in local independence, SE and total errors decreased as the sample size increased.

Bias in models accounting for LPD decreased with increasing sample size. On the other hand, for models ignoring LPD, ability parameters estimates were biased and the bias was magnified by sample size increase. However, testlet effects do not seem to affect ability parameter recovery and the effect of sample size on such is not very clear from the graphs. There was an approximately 1:1 ratio between the squared bias and the standard errors. In LPD and dual effects, bias increased as sample size (and hence group size) increased for GPCM and testlet models ignoring LPD. Bias in ability parameter recovery does not seem to be affected by the testlet effects.

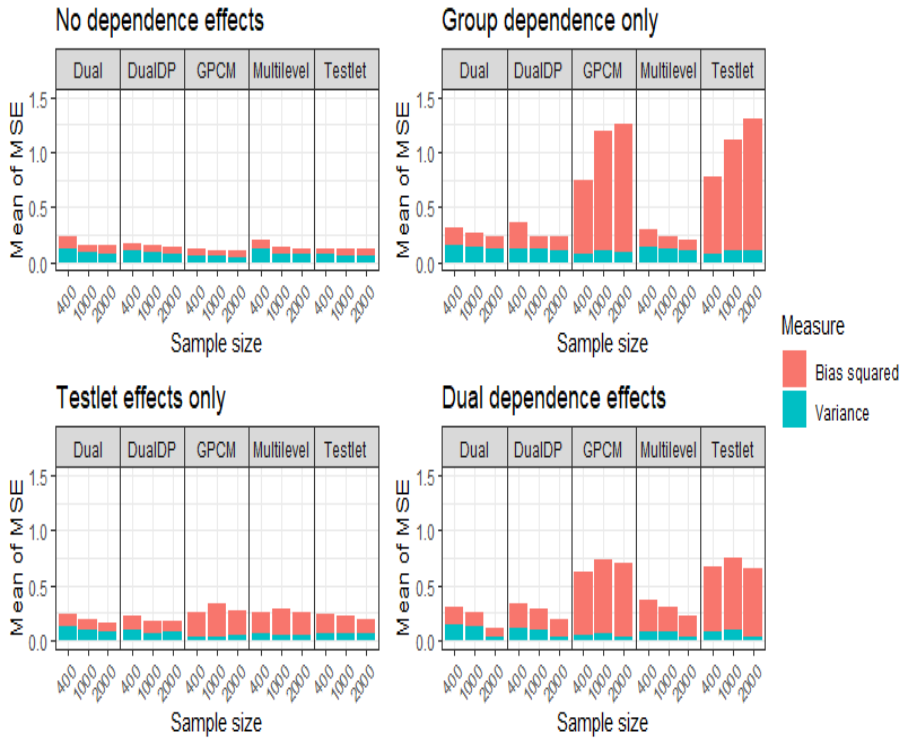


Figure 5.1: Random, systematic and total errors in the ability parameters for different sample sizes

The ratio between variance and bias for models controlling for person dependence effects (dual and multilevel models) was close to 1:1 and there was no significant effect of sample size on total variance. However, the total error for controlled testlet effects (dual and testlet models) decreased with increase in sample size. In dual dependence, the effect of sample size on bias and SE and hence total errors, was not very clear. In person and item clustering effects, the ratio between the error variance and the

sampling error is more than unit, and the ratio is higher for person effects than item effects, suggesting that both person and item dependence effects underestimated SEs, thereby overestimating the precision of trait measurement. The ratio between systematic errors (bias) and random errors (SE), increased with sample size, implying that the overestimation of precision of ability measurement increased with sample size.

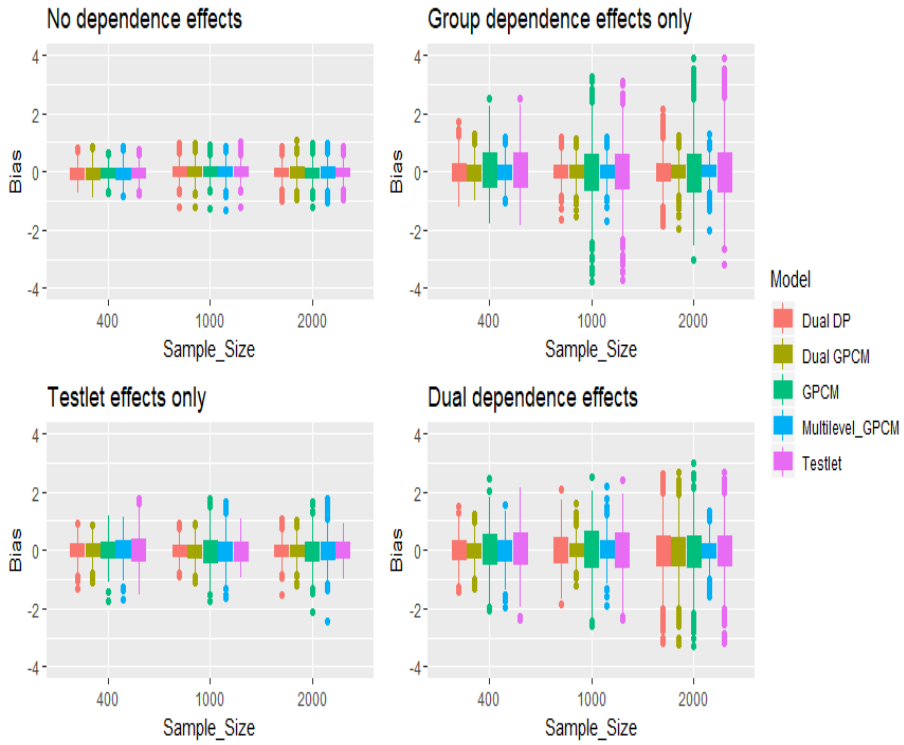


Figure 5.2: Bias in the ability parameters for varying sample sizes

The results in Figure 5.2 give bias recorded in ability parameter recovery for 400, 1000 and 2000 examinees. All models recovered the ability parameter fairly well in LID and LPD across sample sizes as bias box plots for all models are centred around the 0 axis for all sample sizes. In group dependence only, bias bands for GPCM and testlet models ignoring group effects were wider although the width appeared to be constant across sample sizes. In addition, GPCM and testlet models recorded more outliers. Bias did not seem to be much affected by testlet effects although the bands for all models were wider than in local independence. In dual dependence, wider bias widths were recorded for GPCM and testlet models not controlling for group effects, suggesting that group

effects impact more on ability parameter recovery than item clusters.

5.2.3.1 ANOVA

To assess the effects of the number of respondents on ignoring local person and item dependence in ability, threshold and discriminant parameter recovery, four factors (model, LID, LPD and sample size) fully-crossed factorial ANOVA was used for analysis using the General Linear Model in SPSS. Both the p-value and Cohen effect size f (Cohen, 1988) were employed to quantify the magnitude of effect if any. For the ability parameter, the ANOVA results with bias as the dependent variable have shown sample size and its interactions with other factors to be statistically insignificant with negligible effect size. However, analysis with the absolute bias (abias) as the response variable has shown sample size, LID and their two-way interactions with model, and LPD to have negligible effects. However, LPD and model factors were significant with small effects sizes of $f = 0.14$ and $f = 0.19$ respectively. In addition, three-way interaction between sample size, LPD and model factors and four-way interaction between sample size, LID, LPD and model had significant but small effects of $f = 0.11$ and $f = 0.17$ respectively. The Tukey *post hoc* analysis showed that the average abias in the 400 respondent sample was lower but did not differ significantly with abias in the 1000 respondent sample. On the other hand, the bias for 2000 examinees was significantly higher than the 400 and 1000 respondent samples. The bias was smaller for tests with no LID, followed by small LID and highest for large LID. Similar results were obtained for LPD. For the model factor, abias were lowest in dual and multilevel models and significantly higher in testlet and GPCM models. The interaction between LID, LPD, calibration model and calibration sample size had shown no significant differences in abias for independent items and persons for all models and samples sizes although biases were lowest for GPCM (and testlet) model for large sample sizes. Conversely, abias were highest in GPCM and testlet models for large person clustering effects, more so when they are coupled with large testlet effects, for large samples sizes of 2000 respondents. The biases for the same were significantly lower for 400 respondents.

All the factors and their interaction effects were significant on random errors in ability parameter recovery. Sample size significantly affected the SE, ($f = 0.21$) while the model factor has a large effect of size $f = 0.45$. The interaction between sample size and testlet effects has a significant effects of $f = 0.19$ while the interaction of sample size and person dependence had an insignificant effect of $f = 0.06$. In addition, the interaction of sample size with model had a significant effect of size $f = 0.23$. All three-way and four-way interactions of sample size with other model factors had medium effects on the random error in the ability parameter estimation. Tukey *post hoc* show that the SEs were significantly lower for 2000 respondent tests and increases as the sample size decreased to 400 respondents. Four-way interaction between calibration model, sample size, LID and LPD show that the SE were lowest for the GPCM (followed by the multilevel) in the presence of large testlet effects, for large sample sizes, more so when large testlet effects were coupled with large person effects. The SE for the same combination increased when sample sizes decreased. The SEs were largest in the dual models in the presence of testlet effects, for small sample sizes. Large testlet effects recorded the lowest standard errors and SE were highest in the absence of LID. For LPD, lower SE were recorded in person independence and highest for large person dependence.

ANOVA for RMSE in ability parameter estimates has shown that although sample size has significant effects according the p-value, the effect size of $f = 0.03$ is statistically negligible. However, the calibration model and LPD significantly affected total errors which were lowest for 1000 respondents and highest for 400 respondents and larger group effects. There was a significant interaction between model and group effects with the GPCM and testlet models recording significantly higher total errors according the *post hoc* results. The four-way interaction between LID, LPD, model and sample size has shown that although there were no significant differences in total errors

for all models and sample sizes for independent items and persons, total errors were lowest in GPCM model (followed by testlet and multilevel models) and these errors decreased with increase in sample size. For dual dependence cases, total errors were lowest for dual (and multilevel) models for larger samples and these errors increased with sample size decrease. The errors were highest in testlet and dual models in LPD for large sample sizes.

5.2.4 Threshold parameter recovery for 400, 1000 and 2000 respondents

Table 5.3 shows average correlations for the 10 simulations for each factor combination. From the results, correlations for the threshold parameter recovery improved with increase in sample size for most dependence conditions. The model ranks for each condition were maintained as the sample size increases. Changes were only noticed in magnitude which increased with sample size. When both items and persons are independent, high correlations were recorded for all models for 400 respondents. However, in item clustering effects, GPCM and multilevel models have low correlations (< 0.7). As the sample size increases to 1000 and 2000, high correlations (> 0.7) were recorded for all models and for all conditions.

The results in Figure 5.3 show that in local independence, SE in all models decreased with sample size. Total errors were lowest in GPCM, multilevel and testlet models. However, biases and hence total errors in dual models falsely accounting for non-existent dual effects increased with sample size. Although the impact of falsely accounting for testlet effects in group effects only is not very clear, biases and total errors were high in dual models. There was a 1:1 ratio between variance and squared bias in the true model (the multilevel model) and total errors decreased with sample size. Bias in dual, GPCM and multilevel models in LID only were high although the distinct effect of sample size cannot be determined from the graph. However, the true model

Table 5.3: True-estimated thresholds correlations for 400, 1000 and 2000 examinees

Sample size	Testlet	Group	GPCM	Testlet	Multilevel	Dual	DualDP
400	None	None	0.99	0.99	0.99	0.95	0.97
		Small	0.99	0.99	0.99	0.95	0.97
		Large	0.99	0.99	0.99	0.95	0.97
	Small	None	0.93	0.94	0.93	0.95	0.97
		Small	0.82	0.92	0.82	0.93	0.94
		Large	0.81	0.81	0.81	0.96	0.97
	Large	None	0.88	0.98	0.88	0.94	0.95
		Small	0.83	0.92	0.83	0.94	0.95
		Large	0.81	0.88	0.81	0.93	0.94
1000	None	None	0.99	0.99	0.99	0.96	0.96
		Small	0.99	0.99	0.99	0.96	0.96
		Large	0.99	0.99	0.99	0.98	0.97
	Small	None	0.93	0.94	0.93	0.97	0.97
		Small	0.93	0.95	0.93	0.98	0.97
		Large	0.87	0.93	0.87	0.97	0.97
	Large	None	0.85	0.97	0.85	0.96	0.96
		Small	0.84	0.96	0.86	0.96	0.96
		Large	0.85	0.95	0.85	0.97	0.97
2000	None	None	1.00	1.00	1.00	0.99	0.99
		Small	1.00	1.00	1.00	0.96	0.98
		Large	1.00	1.00	1.00	0.99	0.99
	Small	None	0.94	0.95	0.93	0.98	0.97
		Small	0.95	0.95	0.95	0.98	0.98
		Large	0.83	0.89	0.83	0.96	0.92
	Large	None	0.83	0.97	0.82	0.98	0.97
		Small	0.89	0.97	0.89	0.98	0.98
		Large	0.92	0.94	0.93	0.94	0.93

(testlet) has low variances and biases depicting a 1:1 ratio and total errors seemed to be decreasing with increasing sample size. In dual dependence, total errors were generally high across all models although they were lowest in both dual models followed by the testlet model, and highest in GPCM and multilevel models ignoring testlet effects.

Bias in local independence was minimal for all models (Figure 5.4). The effect of sample size on threshold parameter recovery was not clear though bias range seemed to decrease with increase in sample size. For group effects only, GPCM and testlet models underestimated the threshold parameter for 400 respondents as the ratio between error variances and sampling errors was more than unit. However, the true model

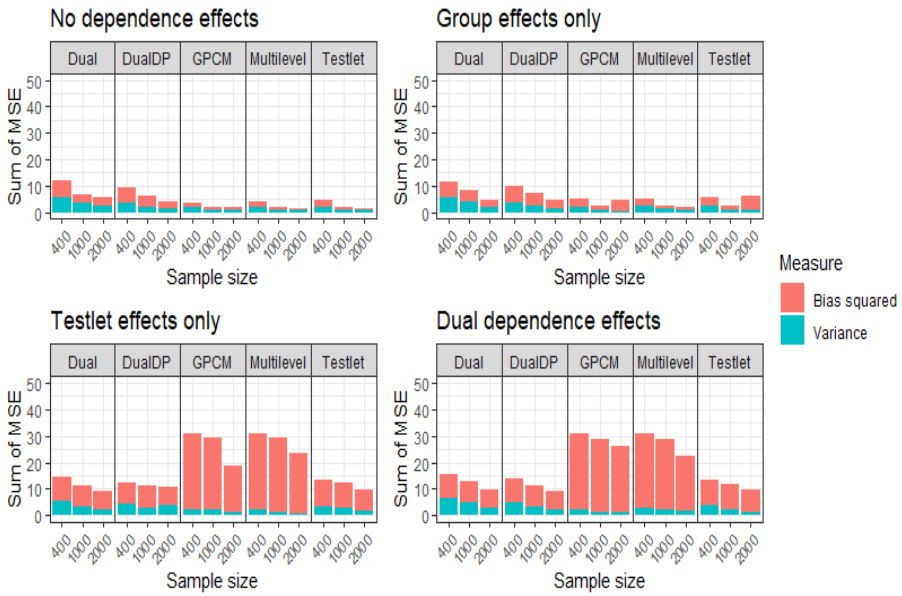


Figure 5.3: Random, systematic and total errors in the threshold parameters for 400, 1000 and 2000 respondents

(multilevel) recovered the threshold parameter well across all sample sizes. Group effects do not seem to impact on threshold parameter estimation much since narrow bias bands were reported for all models across sample sizes. However, in testlet and dual dependence effects, wider bands were reported for all models and the range was widest for GPCM and multilevel models not accounting for testlet effects, suggesting that threshold parameter recovery is affected more by testlet effects than group effects.

5.2.4.1 ANOVA

ANOVA analysis shows that sample size has a significant on bias, ($p < 0.05$) although the Cohen effect size of $f = 0.07$ is negligible. The interaction between sample size and LID had significant effects although interaction between sample size and LPD did not significantly affect bias. All conditions and their interactions significantly affected abias in threshold parameters. However, only the testlet effects, group, estimation model, and interaction between testlet and model factors had significant effects of 0.25, 0.17, 0.16, and 0.17 respectively. The rest had insignificant effects.

The bias were lowest in independence and higher for small and large effects which do

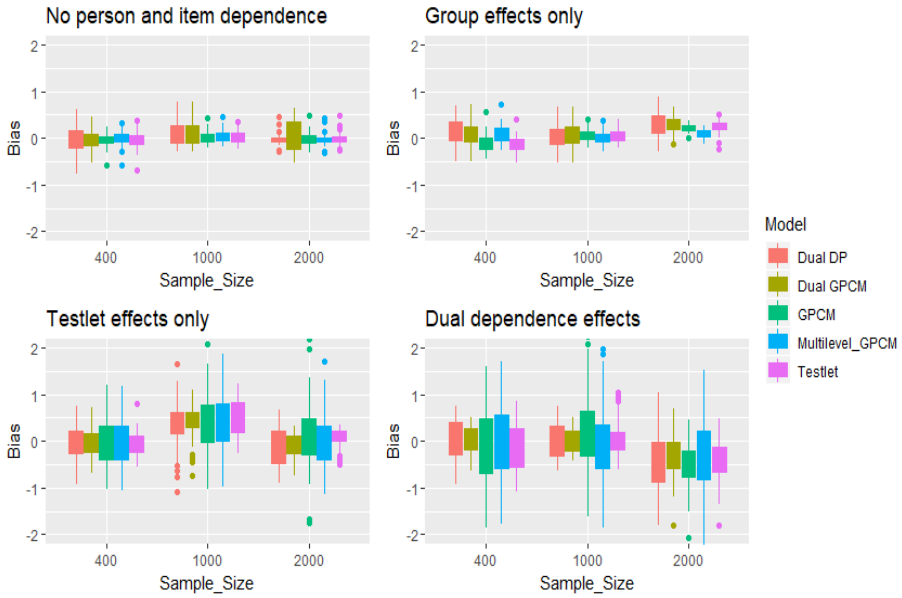


Figure 5.4: Bias in the threshold parameters for 400, 1000 and 2000 sample sizes

not differ significant for both LID and LPD. Bias was lowest for 2000 respondents and there was no significant difference between bias for 2000 and 1000 respondents. However, bias was significantly higher for 400 respondents, implying that smaller samples overestimated the threshold parameter. There was no significant difference between threshold parameter bias for dual and testlet models which were significantly lower than bias in GPCM and multilevel models ignoring testlet effects. The dual models underestimated the parameter while GPCM and multilevel overestimated the parameter. This means that bias was highest in smaller samples for models not accounting for testlet effects and lower in larger samples for models accounting for testlet effects.

All factors significantly affected threshold parameter SEs. Sample size and model had a large effect sizes of $f = 0.80$ and $f = 0.92$ respectively. The interaction between sample size and group ($f = 0.48$), interaction between sample size and testlet ($f = 0.11$) and interaction between calibration model and sample size ($f = 0.15$) had significant effects on SE in threshold parameters. The SE decreased with increase in sample size and thus were largest for 400 respondents and lowest for 2000 respondents. There was a significant interaction between sample size and LID ($f = 0.12$). *Post hoc* results

show that SEs were largest in large and small LID effects, for 400 respondents and were lowest for 2000 respondents in small and large LID. This indicated that testlet effects increased SEs, more so for small samples. Furthermore, the interaction between sample size and group effects significantly affected random errors in threshold parameters ($f = 0.14$). Errors were largest in group effects and for small sample sizes. In addition, SEs were highest for dual models controlling for both LID and LPD for small sample size (400) and lowest for models ignoring LID effects for large samples (2000 respondents).

All factors and their interactions significantly affected RMSE in threshold parameters. Sample size had an effect of size $f = 0.11$. The interaction between LID and sample size had an effect size of $f = 0.13$ and the interaction between sample size and group effects had an effect size of $f = 0.16$ and model and sample size factors had an effect of size $f = 0.12$. As the sample size increased, the discrepancy between estimates and real data decreased as signified by lowest total errors being recorded for samples of 2000 respondents and highest for 400 examinees. For independent items and persons, RMSEs were generally low for all models, becoming smaller as the sample size increased and lowest RMSE were in GPCM model. In dual dependence effects, RMSEs were lower in dual models for large samples and higher in the GPCM model for large samples, implying that RMSE for models ignoring dual dependence effects increased with sample size while RMSE in models controlling for dual effects decreased as sample size increased.

5.2.5 Discriminant parameter recovery for 400, 1000 and 2000 respondents

According to results in Table 5.4, correlations between true and discriminant parameter estimates increased with sample size, implying that the ranking of items according to their ability to categorise respondents according to their latent abilities was well recovered for larger samples. Similar trends were noted across correlations for different

Table 5.4: True-estimated discriminants for 400, 1000 and 2000 respondents

Sample size	Testlet	Group	GPCM	Testlet1	Multilevel	Dual	DualDP
400	None	None	0.92	0.92	0.92	0.91	0.92
		Small	0.90	0.90	0.90	0.90	0.90
		Large	0.87	0.86	0.87	0.87	0.87
	Small	None	0.63	0.73	0.63	0.73	0.72
		Small	0.89	0.92	0.90	0.93	0.92
		Large	0.77	0.86	0.76	0.86	0.84
	Large	None	0.49	0.81	0.49	0.82	0.79
		Small	0.49	0.82	0.48	0.85	0.62
		Large	0.56	0.70	0.55	0.75	0.77
1000	None	None	0.96	0.96	0.96	0.96	0.97
		Small	0.96	0.96	0.98	0.98	0.97
		Large	0.96	0.96	0.98	0.98	0.98
	Small	None	0.89	0.95	0.87	0.96	0.97
		Small	0.86	0.94	0.85	0.96	0.95
		Large	0.85	0.94	0.84	0.96	0.96
	Large	None	0.59	0.94	0.57	0.97	0.96
		Small	0.53	0.90	0.53	0.97	0.96
		Large	0.47	0.93	0.44	0.96	0.95
2000	None	None	0.95	0.95	0.95	0.95	0.95
		Small	0.96	0.96	0.96	0.96	0.96
		Large	0.98	0.97	0.98	0.98	0.98
	Small	None	0.91	0.98	0.91	0.98	0.97
		Small	0.80	0.91	0.79	0.95	0.91
		Large	0.94	0.96	0.94	0.98	0.98
	Large	None	0.42	0.97	0.41	0.98	0.95
		Small	0.53	0.96	0.54	0.97	0.94
		Large	0.67	0.99	0.68	0.99	0.98

dependence conditions across sample sizes.

Figure 5.5 shows random and systematic errors in discriminant parameters for 400, 1000 and 2000 respondents. In local independence, SE in discriminant parameters are almost equal across all models although they were higher in the GPCM model. In addition, all models depict approximately 1:1 ratio between the systematic and random errors as expected and the ratio was constant when sample size increased. However, total errors decreased with sample size increase in all models. Similar results were observed in LPD only although systematic errors in dual models falsely accounting for absent testlet effects seemed to increase with sample size. In LID only, dual and testlet models had a 1:1 ratio between random and systematic errors and total errors decreased

with increase in sample size. However, systematic errors were high for models not accounting for testlet effects although the effect of sample size is not obvious from the graph. There is an approximately 1:1 ratio between error variance and squared bias in dual models (true models) in dual dependence. Total errors in dual models decreased with increase in sample size. However, systematic errors were highest in GPCM and multilevel models ignoring LID.

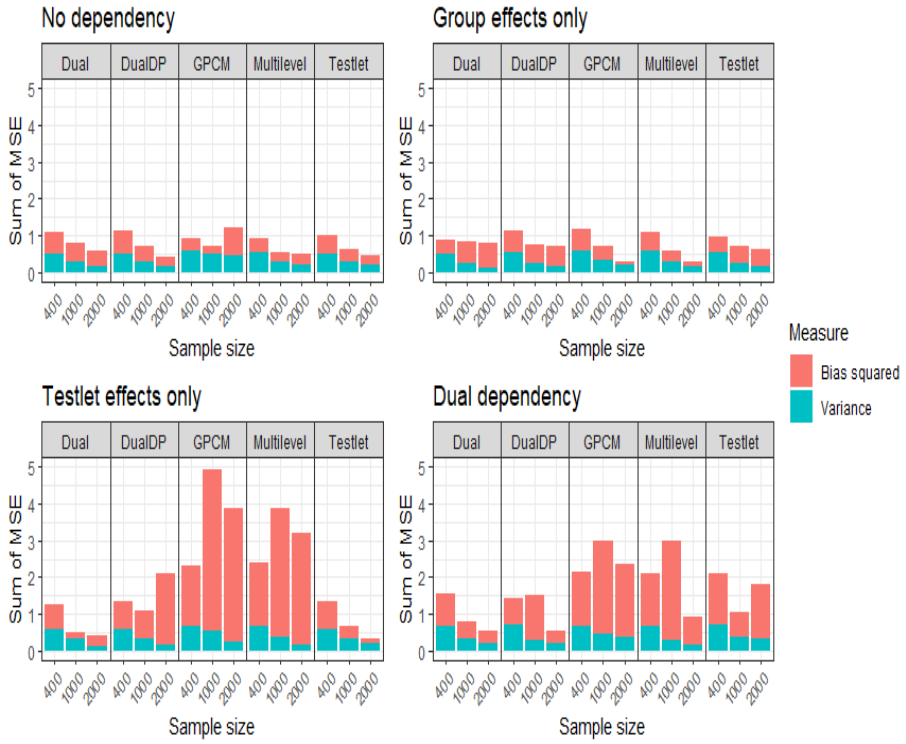


Figure 5.5: Random, systematic and total errors in the discriminant parameter for 400, 1000 and 2000 respondents

From the results in Figure 5.6, in local independence, the dual parametric model underestimated the discriminant parameter for small samples and the underestimation disappeared as the sample size increased. On the other hand, the GPCM model overestimated the discriminant parameter as shown by the box above the zero (0) y-axis line and the overestimation increased with sample size. The multilevel and testlet model seemed to recover the discriminant parameter fairly well. In LPD only, GPCM and multilevel models seemed to recover the discriminant parameter well while the

dual model underestimated the parameter. The range was narrow for all models in local independence and in LPD only. In testlet effects, the bias range was wider for GPCM and testlet models ignoring item clustering effects and narrower for dual and testlet models controlling for testlet effects. In dual dependence, dual models recovered discriminant parameters well, followed by the testlet model, as indicated by narrower bands along the zero axis line. The range for discriminant parameter bias in multilevel and GPCM models were wider.

Figure 5.6: Bias in the discriminant parameters for changing sample sizes

5.2.5.1 ANOVA

The ANOVA to compare bias in discriminant parameters show all factors and their interactions to had significant effects except sample size ($f = 0.08$). There was a significant interaction between sample size and testlet, group and model effects with effect sizes of 0.15, 0.13 and 0.12 respectively. In addition, all three-way and four-way interactions between sample size and other model factors significantly affected discriminant parameter bias. The interaction between sample size and testlet effects was significant with models ignoring testlet recording highest bias for samples of 1000 respondents and bias in all models decreased with increase in sample size in item independence.

In addition, bias for models falsely accounting for absent testlet effects were slightly higher than independent items models. The Tukey *post hoc* show lowest bias in independent items models when items were indeed independent and higher in dual and testlet models falsely accounting for absent LID effects and the bias was higher for larger samples. Overall, bias were highest in independent items models when items were not independent and more so for large samples of 1000 and 2000 respondents. Similar results were recorded for absolute bias.

Sample size and calibration model significantly affected the random errors in the estimation of discrimination parameters with large effect of $f = 0.77$ and $f = 1.36$ respectively. In addition, the interaction between sample size and testlet effect, group effects and calibration model significantly affected the random errors with effect sizes of $f = 0.14$, $f = 0.15$ and $f = 0.15$ respectively. The SE was lowest in sample of size 2000 and differed significantly with samples of 1000 and 400 (highest). The SE across all sample sizes differed significantly with each other. The SE were highest in the GPCM and dual models and increased as group and testlet effects increase. In testlet and dual dependence effects, SE were lowest in the dual models followed by the testlet model for 2000 respondents and highest in the GPCM followed by the multilevel model for samples of size 400. Regardless of the condition and model, SE decreased with sample size increase and were highest in independent-items models.

ANOVA with RMSE as the dependent factor also show the calibration model and calibration sample size to significantly affected the SE in discriminant parameter estimation ($f = 0.27$ and $f = 0.20$ respectively) in addition to testlet and group effects (see section 4.3.7). In addition, all interactions between sample size and other model factors had small to medium effects on the RMSE in discriminant parameter recovery. The Tukey *post hoc* show that smallest total errors were recorded for samples of size 2000 which differed significantly with RMSE for 1000 and 400 respondents. In

addition, the errors for 400 examinee samples were significantly higher than errors in samples of 1000 and 2000 respondents. The RMSE were lowest for 2000 examinees in the testlet model when only items are dependent, in the dual model for dual dependence and testlet effects and multilevel model when items are independent. The RMSE were highest in independent items models when items were not independent, increasingly with increase in sample size.

5.2.6 Test reliability for 400, 1000 and 2000 respondents tests

Table 5.5 shows the changes in test reliability as the sample size increased. From the results, it can be concluded that the test reliability increased with increase in sample size across all models. As noted in section 4.3.8 above, the reliability levels were highest in GPCM and testlet models in dual dependence effects.

The results in Table 5.5 give the average reliability estimates for the competing models and dependence conditions as the sample size increases. The reliability levels are higher in the GPCM and testlet models and lower in the parametric and non-parametric dual and multilevel models accounting for person dependence effects, lowest in dual effects. The reliability levels for all dependence conditions seem to increase as the sample size increases.

From the results in Table 5.6, the Spearman-Brown prophecy coefficients were lower for the testlet model, an indication that there is little difference between the reliability coefficients for the two models. However, Spearman-Brown values were higher for multilevel and dual models and the values decreased as the sample size increased. For example, in both LID and LPD, the test would need to be elongated by 2.71 folds for the reliability of the parametric dual model to be equal to the reliability value for GPCM model. For 1000 and 2000 respondents, the test would need to be magnified by 2.51 and 1.98 folds respectively. This implies that the test reliability overestimation by the GPCM model decreased as the sample size increased.

Table 5.5: Test reliability for 400, 1000 and 2000 respondents

		GPCM	Testlet	Multilevel	Dual	DualDP
400	NoneNone	0.95	0.95	0.90	0.90	0.89
	NoneSmall	0.95	0.95	0.88	0.88	0.86
	NoneLarge	0.96	0.95	0.85	0.86	0.85
	SmallNone	0.94	0.94	0.88	0.88	0.88
	SmallSmall	0.95	0.94	0.83	0.86	0.86
	SmallLarge	0.96	0.95	0.84	0.84	0.84
	LargeNone	0.91	0.94	0.84	0.86	0.85
	LargeSmall	0.93	0.94	0.83	0.86	0.85
	LargeLarge	0.94	0.94	0.83	0.84	0.84
1000	NoneNone	0.94	0.94	0.92	0.93	0.92
	NoneSmall	0.95	0.95	0.92	0.91	0.90
	NoneLarge	0.95	0.95	0.90	0.90	0.90
	SmallNone	0.94	0.94	0.92	0.91	0.91
	SmallSmall	0.94	0.95	0.88	0.91	0.91
	SmallLarge	0.95	0.95	0.88	0.89	0.85
	LargeNone	0.94	0.95	0.90	0.91	0.91
	LargeSmall	0.95	0.94	0.86	0.89	0.88
	LargeLarge	0.95	0.95	0.85	0.88	0.88
2000	NoneNone	0.94	0.94	0.93	0.93	0.93
	NoneSmall	0.95	0.95	0.93	0.92	0.92
	NoneLarge	0.96	0.93	0.91	0.91	0.91
	SmallNone	0.94	0.94	0.92	0.93	0.92
	SmallSmall	0.94	0.95	0.90	0.92	0.91
	SmallLarge	0.97	0.96	0.89	0.90	0.90
	LargeNone	0.94	0.96	0.89	0.90	0.89
	LargeSmall	0.95	0.94	0.84	0.92	0.91
	LargeLarge	0.96	0.94	0.92	0.96	0.96

5.3 Discussion

The models maintained their ranking according to goodness of fit statistics as the sample size varied. However, in item or person dependence effects only, the dual model was better or performed similar to the data generating model for smaller samples. However, for large LPD, the multilevel performed better from small to larger samples. The group membership identification for the non-parametric model improved with sample size (group size) and so did the recovery of ability parameter. However, longer test length and larger sample sizes improved group membership identification (Reise & Yu, 1990). The results have that more groups were favorable for large sample sizes and smaller groups were favourable for smaller sample sizes.

Table 5.6: Spearman's Brown prophecy for 400, 1000 and 2000 respondents

		Testlet	Multilevel	Dual	DualDP
400	NoneNone	1.10	2.25	2.21	2.43
	NoneSmall	1.18	2.95	2.94	3.39
	NoneLarge	1.16	3.69	3.59	3.89
	SmallNone	1.12	2.14	2.13	2.16
	SmallSmall	1.18	3.95	3.09	3.15
	SmallLarge	1.21	3.94	3.98	3.98
	LargeNone	0.67	1.97	1.66	1.77
	LargeSmall	0.81	2.78	2.31	2.45
	LargeLarge	1.01	3.25	2.74	2.81
1000	NoneNone	0.94	1.46	1.30	1.47
	NoneSmall	0.96	1.61	1.78	1.93
	NoneLarge	0.88	2.44	2.25	2.51
	SmallNone	0.99	1.41	1.49	1.57
	SmallSmall	0.93	2.23	1.68	1.65
	SmallLarge	1.12	2.79	2.53	3.80
	LargeNone	0.75	1.63	1.80	1.88
	LargeSmall	1.20	2.98	2.43	2.70
	LargeLarge	1.02	3.23	2.51	2.65
2000	NoneNone	1.05	1.27	1.22	1.32
	NoneSmall	1.04	1.38	1.53	1.70
	NoneLarge	1.96	2.60	2.61	2.85
	SmallNone	0.91	1.23	1.16	1.37
	SmallSmall	0.85	1.92	1.39	1.65
	SmallLarge	1.18	3.41	3.15	3.27
	LargeNone	0.66	1.84	1.68	1.87
	LargeSmall	1.14	3.66	1.67	1.93
	LargeLarge	1.03	2.15	1.98	1.97

The effect of sample size on the estimation of ability parameter variance in local independence was as expected where the estimates approached true parameter value as the sample size increased. However, the effect of sample size on ability variance in dual dependence was not very certain as the variance was lowest for 1000 respondents. This is probably because the testlet effects led to underestimation of the variance increasingly with sample size while person effects led to overestimation of the same as sample size increased. Probably the pooling effects make it difficult to determine the actual effect of increasing sample size on ability variance estimation in dual dependence effects.

The correlations between true and estimated ability parameters were not much affected by sample size, but were significantly affected by LPD effects. These results contradict results from Reise and Yu (1999) who observed a pronounced effect of sample size on

ability parameter recovery with lower correlations for smaller (250 respondent samples) and higher correlations of 0.95 for larger samples of 2000 respondents and suggested that 1000 respondents were required to maintain the true correlation of 0.90. The difference in these results could be that Reise and Yu considered smaller sample sizes (250 respondents) than the minimum considered in this study (400 respondents). However, even in the Reise and Yu (1990) study, average correlations between the true and estimated ability parameters were above 0.90 for all sample sizes, implying that the linear ordering of persons on the ability continuum was retained while the RMSE statistic show that sample size is not a major factor in determining the θ recovery.

Bias in ability parameters was not affected by sample sizes probably because (1) for all sample sizes, the mean ability parameter was constrained to 0 for identifiability and (2), positive and negative biases usually cancel out regardless of magnitude. The results concur with Reise and Yu (1990) who also observed bias in ability parameters not to be affected by sample size. However, in local independence, abias in ability parameters decreased as sample size increased but increased with sample size when LPD effects were not accounted for. The results local independence are in agreement with other studies (e.g Wang & Wilson, 2005) where bias (in ability and item parameters) was lower for larger sample sizes. On the other hand, abias when LPD effects are ignored worsens when the sample size increased, concurring with Zhang (2010) who observation that as the sample size increased, the discrepancy between the real data set and the model estimates increased and the polytomous and Rasch testlet models offered an advantage over the standard IRT models as they avoid underestimation of SE and better ability parameters in small testlet situations.

As expected, the study results show SE for all conditions decreased with sample size and hence the precision of measurement increased with increase in sample size. This is due to the fact that increasing the sample size tend to reduce the sampling error,

making the sample statistic less variable. However, the SE were lowest for larger samples for models ignoring dual dependence and testlet effects. It was also observed that for independent persons models, bias increased with sample size. Thus, the decrease in SE and increase in biases with sample size imply that the overestimation of precision of ability parameter estimation is worsened by increasing sample size. The SEs were largest in dual models for testlet-based tests, for small sample sizes and these were accurately measured. However, making a sample large cannot correct a methodological problem that produce bias. That is why although the GPCM model recorded the lowest SEs dual dependence effects, the same recorded bias increasingly with sample size.

The study show RMSE to be significantly affected by the main effects of sample size although there was significant interaction between sample size and dependency levels and model factors. In the absence of person clustering effects and when person clustering effects are accounted for, the total errors decrease with sample size, in support of earlier researchers (Reise & Yu, 1990; Wang & Wilson, 2005) who reported a decrease in the RMSE for increasing sample sizes in their study which has independent polytomous items. However, the presence of person effects, RMSE in independent-persons models increase with sample size. As a result, the ANOVA results show negligible effects of sample size on RMSE in trait parameters because the RMSE increase in independent person models and decrease in models accounting for person effects with increase in sample size cancel out. The ability parameter estimates did not change significantly due to changes in sample sizes. This is probably because the samples sizes considered in this study are already relatively high. Reise and Yu concluded that at least 500 examinees are needed for adequate calibration under the graded response model.

Threshold correlations have been observed to increase with sample size, in line with

observations by Reise and Yu (1990) who observed lower correlations in difficulty parameters for smaller samples and increased as the sample size increased. The GPCM and multilevel models assuming independent items, have poorer correlations for small samples of 400 respondents, implying that the models compromise the true rank ordering of items by their difficulty levels for smaller sample sizes. Group effects have little impact on item parameter recovery. Sample size significantly affected threshold bias where small samples overestimates the thresholds especially for models ignoring testlet effects. On the other hand, bias were lower in larger samples when items are independent and when item effects are accounted for. Other researchers also observed that sample size significantly affected bias, SE and RMSE in difficulty estimates (Luo, 2018; Cho, Cohen & Kim, 2013). According to Luo, increasing the sample size from 500 to 1000 and 2000 generally seemed to improve the quality of estimation of both difficulty and discriminant parameters.

The thresholds SEs decreased as sample size increases and were largest for dual models in LID for smaller samples and lowest in models ignoring testlet effects for large samples. Similar to ability parameters, ignoring testlet effects underestimates SE in threshold parameters, thus overestimating the precision with which it is measured. The findings support earlier researchers (Wang & Chen, 2005; Embrestson & Reise, 2000; Ra, 2010) who reported decrease in SE in estimation of item (difficulty) parameters with increase in sample size and consequently increasing the precision of its measurement. Sample size significantly affected RMSE in thresholds estimation. In local independence, total errors decreased with increase in size of sample in dual models. Several other researchers who studied with independent and testlet items also recorded a decrease in total errors in estimation of item difficulty parameters as the number of examinees increased (Reise & Yu; 1990; Cho, Cohen & Kim, 2013; Wang & Wilson, 2005). However, Choi (2014) observations that test length and sample size did not appear to affect the recovery of item parameters are contradictory. On the other

hand, in testlet and dual dependence effects, the RMSE in independent items models increased with sample sizes, implying that ignoring testlet effects result in erroneous estimation of thresholds increasingly with sample size.

The correlations in discrimination parameters increased with samples size especially when items are independent, in agreement with Reise and Yu (1990) who reported the true-estimated correlations for the discriminant parameters to be highly influenced by the sample size, becoming better as the sample size increased. In addition, bias discriminant parameter estimation significantly decreased as sample size increased. These systematic errors were lowest in GPCM and multilevel models when items were independent and higher in testlet and dual models, an indication that falsely accounting for absent testlet effects bias slope estimates. However, in LID, bias were highest in independent items models for large sample sizes, implying that ignoring LID bias the slope more for larger samples.

Similar results were reported by Luo (2018) who discovered sample size to significantly affect bias in estimation of discriminant parameters and bias was lowest for 2000 respondents followed by 1000 and was highest for 500 examinees. However, their study contradict the current findings in that testlet effects did not affect bias in discriminant parameters. This could be because Luo considered testlet variances of 0.25, 0.5 and 1, while the current study included a 0 testlet variance control where independent items models performed significantly better. It can be inferred from the current results that independent items models are the best when items are indeed independent and in LID, testlet models should be employed for analysis if item discrimination ability is to be retained. When items are independent, SE are higher in the GPCM. When items are dependent, SE are lower in testlet and dual models, further decreasing with sample size increase. However, SE were high in independent-items models for small samples. So regardless of the model and simulation condition, SE decrease with sample size

increase. Similar results were reported by other researchers (eg Luo, 2018).

When items are independent or their dependence is accounted for, RMSE decrease when the sample size is increased, in line with other findings (DeMars, 2003; He & Wheadon, 2012; Wang, Bradlow & Wainer, 2002; Ra, 2010; Luo, 2018; Wollack et al., 2002) who reported high RMSE in smaller samples and RMSE estimates decreased to acceptable levels as the sample size increases. In the same vein, Kogar and Kelecioughu (2017) claimed that the slope parameter is more sensitive to model and sample size changes and the error values of item parameters for different models got closer and generate similar values when the sample size gets large. However, ignorance of item dependence effects result in high RMSE, increasingly with sample size. Contrary to the current findings, Reise and Yu (1990) observed the RMSE in the discriminant parameter were not greatly affected by sample size although they followed the same pattern as the correlations which they reported to be highly influenced by the sample size. The RMSE in item parameters were smaller for larger samples probably because when the sample is large, the number of scores within each response category are increased, thereby providing more information on which to estimate the item parameters.

The current results have shown for independent person and items and when testlet and group effects are accounted for, both item and person parameters become better. The same conclusion was reached by other testlet-based test researchers (eg Zhang, 2010) who concluded that sample size was influential with better results obtained for larger samples. However, when local dependence effects are ignored, the average systematic errors and total errors worsen although the rank ordering of items and persons improved with samples size. According to Atilgan (2013), item parameter estimation error due to small samples did not result in poor person parameter estimation.

Although several researchers reported a significant overestimation of test reliability by

independent items models compared to testlet-based models in testlet effects (Zhang, 2010), the current study discovered no significant differences between the reliability estimates from PCM and testlet models and the magnitude was maintained across different sample sizes. This is probably because although the GPCM underestimated SEs in ability parameters, the models also underestimated the posterior ability parameter variance in testlet effects and the underestimation increased as the sample size increased. In addition, the current study reported the extent of overestimation of test reliability by the GPCM model to decrease with increasing sample sizes as denoted by decreasing Spearman-Brown prophecy values. This is probably because the underestimation of the posterior ability parameter variance in independent items models decreased with increasing sample size. The testlet and GPCM models have high reliability values in LPD because they both overestimated the ability variance increasingly with sample size. Contrary to the current findings, Zhang (2010) reported the standard Rasch model ignoring testlet effects overestimated test reliability increasingly with sample size.

Although Atilgan (2013) study on reliability revealed that a sample size of 400 or greater make the reliability estimates robust and stable and increasing the sample size further does not necessarily make a significant contribution to the reliability estimates, the reliability estimates in the current study changed for samples of at least 400. This is probably because Atilgan used the generalizability coefficients and the index of dependability estimated using the multivariate generalisability theory model.

5.4 Conclusion

The current results have shown sample size to influence model parameter recovery as it significantly affected bias, SE and RMSE in estimation of item and person parameters in local dependence effects. Increasing sample from 500 to 1000 and 2000 generally

improved the quality of estimates in local independence and when the dependence conditions are accounted for. In local item and person dependency, increasing sample size magnifies the errors of ignorance of the consequences on item and person parameters respectively, implying that using a smaller sample would be better for inference. This is probably the rationale behind the argument by some studies in literature that the effective sample size from cluster sampling maybe reduced due to dependence among individuals within clusters, which could lead to more errors in parameter estimation (Fox & Glas, 2001; Kamata, 2001). However, some researchers still argue that cluster sampling requires a larger sample size to achieve accuracy in model parameter estimation compared to that based on simple random sampling.

The study findings suggest that relatively large sample sizes are required under group based experimental designs and the required sample sizes increase substantially as the clustering effects increase. However, in the event that local dependence effects are not accounted for, the consequences may increase with sample size. The reduction in SE with increase in sample size is due to the fact that increasing the sample size tend to make the sample statistic less variable. However, making a sample large cannot correct a methodological problems that produce bias. That is why although the GPCM model recorded the lowest SEs in the presence of dual dependence effects, the same recorded bias increasingly with sample size. This simply means that the parameter estimates from the independent items and persons model were less variables, but highly biased, increasingly so when the sample size increased.

Conversely, introducing spurious sources of clustering can lead to serious underestimates of precision level (Schochet, 2005). Thus either researcher must clearly specify the sources of clustering under the design if known, or the assumptions underlying them or use independent items and persons models if no such dependence conditions exist. In addition, researchers and test developers must use methods that can detect

the presence of clustering effects. If the clusters and the number of clusters are not known prior to analysis, then the use of the models that can detect the latent clusters is encouraged. This makes the proposed Dirichlet Process Mixture dual model handy for modelling in such circumstances.

In summary, increasing sample sizes resulted in reduction in estimation for independent items and persons and when dependence effects are accounted for. However, increasing sample size increased the overestimation of precision of ability parameter measurement when item clustering effects are ignored and increase in estimation errors in ability and item parameters when LPD and LID are ignored respectively. Increasing sample / group sizes improved the group identification for the non-parametric dual model, thereby boosting the proficiency recovery ability.

Chapter 6

Effects of changing the characteristics of the test items

6.1 Introduction

The proposed model was evaluated when testlet and item characteristics were varied. Testlet length (number of items), and number of response categories per item have been noted to affect the goodness of fit and model parameter recovery. In addition, the proposed model is expected to model tests comprising of items with varying number of response categories. This chapter assessed the effects of varying the testlet length, number of response categories and modeling mixed items models. According to He and Wheadon (2012), parameter stability is affected by the sample size, number of categories in items and the distribution of category scales within the items.

6.1.1 Effects of changing the test length and testlet size

There is a large volume of published literature on the effects of test length on parameter estimation in IRT based testing. Testlets are usually set from 5 to 10, or more (Wang & Wilson 2005, Wainer & Wang 2000) and small to medium items per testlet. Testlets of size 2-4 were applied by Ip, Smits and De Boeck (2009) and DeMars (2006). Zhang (2010) investigated three different models on testlet type data and small (3) and medium (5) testlet sizes. Wang and Wilson (2005) considered 4 and 8 testlet tests for a sample size of 200 and 500 respondents and testlet effects ranging from small to large (0.25, 0.50, 0.75 and 1.00). Bradlow, Wainer and Wang (1999) compared tests of 5 and 10 items per testlet for 3 testlet conditions ($\sigma_\gamma^2 = \frac{1}{2},\ 1$ and 3) and a control

condition with all items independent, equivalent to $\sigma_\gamma^2 = 0$. Their studies concluded that estimating testlet data with standard models cause (a) being partial in the determination of item difficulty ability, (b) overestimation of the discrimination parameter, (c) overestimation of individual scores and (d) overestimation of the test information and reliability. However, very few studies looked at ability and item parameter recovery, test reliability and information as testlet size and local dependence levels change in magnitude.

6.1.2 Effects of changing the number of response categories

Previous findings on the relationship between number of response options and reliability have been inconsistent. Some researchers concluded that the number of categories has no effect on scale reliability (eg Dawes, 2008) while others reiterate that the scale reliability increases with more response options per item (Muniz, Garcia-Cueto & Lazano, 2005; Lazano et al., 2008). The majority of studies that attempted to discover the optimum number of response alternatives using the scale reliability criterion have concluded that the optimum number of alternatives for ensuring an appropriate level of reliability is four (Lazano et al., 2008), with response categories between 4 and 6 popular (Lee & Paek, 2014; Weng, 2004). Weng (2004) propounded that item homogeneity may be a plausible explanation to the differences in effects of the number of response categories among studies. Preston and Colman (2000) argue that scales with more category options tend to show better item discrimination than those with fewer response options. However, Muniz et al. (2005) found that the benefits of having a higher number of scale points tend to reach a plateau beyond 4 points while Neummann and Neumann (1981) argued that the benefits climax at 5 points and no further improvement in the psychometric properties will be recorded if the scale is increased beyond that point. Some schools of thought recommend fewer scale points (see Lee, 2012) arguing that when more options are provided, respondents perceive the difference between adjacent categories to be smaller, which can introduce inconsistent responses.

Lee argues that response styles or systematic errors can be easily introduced if a greater number of response alternatives is provided. Lee and Paek (2014) recorded the most deteriorating performance in the transition from 3 to 2 categories while Lazano et al. (2008) recorded better performance for 2 options when compared to 3 options.

Very few studies have looked at the effect of changing the number of category options on parameter estimation and recovery but rather have assessed the impact of varying category options on scale reliability and validity. Most of the studies determined the optimal number of response categories which most concur to lie between 4 and 7. The fact that questionnaire validity (in local dependence conditions) changes according to the number of response options (DeMars, 2006; Lazano et al., 2008; Muniz et al., 2005; He & Wheadon, 2012) calls for a need for further research on the effects of number of category options on parameter recovery, scale reliability and test information in dual local dependence effects.

6.1.3 Modelling mixed tests in the presence of LID and LPD

Although most polytomous item response theory models are usually used for items with the same number of category options, some models can be utilised for estimation even when the number of response categories for test items differs. Wang and Wilson (2005) proposed the Rasch testlet models that can be used for both dual and polytomous items in testlet based tests. They manipulated the type of items (dichotomous only, mixture of dichotomous and polytomous and polytomous only), number of items per testlet, sample size and testlet effects of 0.25, 0.5, 0.75 and 1, representing small to large effect sizes. The test comprising of a mixture of testlet and polytomous items had 24 three-point polytomous items in 4 or eights testlet and 20 dichotomous items in 2 or 4 testlets (10 or 5 items each). They discovered that item and person parameters and random testlet effects could be accurately recovered under all simulation conditions. Bradlow, Wainer and Wang (1999) proposed a parametric Bayesian model which involves a modification of the standard item response theory (IRT) model that explicitly

account for the nesting of items within the same testlet and can be applied to more generally to multiple choice questions that are comprised of mixtures of independent and testlet items. Wang, Bradlow and Wainer (2002) extended the 2PL testlet model to a more general Bayesian model that can accommodate mixed formats containing both dichotomous and polytomous items. They manipulated the number of items per testlet (3, 6, 9) and the number of categories. They ran 5 replications per condition. The consideration of mixed items tests is important for practical and theoretical considerations as such tests do exist in measurement theory.

The current chapter comprises of two separate studies aimed and assessing the effects of varying test characteristics in the presence of local item dependence (LID) and local person dependence (LPD). The first study evaluates the effects of varying test(let) size and the number of response options in testlet items while the second study assesses the effects of having mixed items in the same test.

6.2 Effects of changing the test(let) length and the number of category options

In the current study, the effects of ignoring the presence of local person and item dependence on item and person parameter estimation, variance and group recovery and test reliability were evaluated for varying testlet sizes, number of response options. Although the study is not aimed at optimising the number of response option but rather the effects of changing category options on item, ability parameter recovery and test information and reliability, tests of items with 3 (odd), 4 (even) and 5 (odd) category responses per item were considered, guided by the prevailing literature, where four response options have been set either as minimum or optimum, leading the researcher to include slightly lower and slightly higher options. Furthermore, the 3, 4 and 5 category options were considered mainly because of the limitations caused by the model complexity and the computational power required for the proposed Dirichlet process

model to compile and converge. Testlets with 60 items grouped into 6 testlets of 10 items each, 36 items grouped into 6 testlets of 6 items, and 18 items grouped into 6 testlets of 3 items each were considered for comparison. The study is a five factor fully-crossed factorial design with 2 (LID levels) × 2 (LPD levels) × 5 (calibration models) × 3 testlet sizes (3, 6, 10) × 3 category options (3, 4, 5), resulting in 135 simulation conditions. Because studies have considered sample size to be influential on the psychometric properties, the sample size for this study was fixed at 1000 respondents based on Reise and Yu (1990) argument that a sample of at least 500 cases is needed for reasonable parameter recovery of the graded response model with 5 category response items.

6.3 Results on the effects of changing the testlet length and category options

The calibration models were compared based on the test reliability and information they provide as well as their ability to predict person proficiency and item attributes. Graphical, tabular and inferential comparisons were done. Correlation between the true simulated parameter values and estimates and errors of measurement were utilised to assess the estimation ability of the models.

6.3.1 Goodness of fit of the calibration models

The results in Table 6.1 give the DIC fit statistics as the number of items per testlet and category options per time change. For independent items and persons, multilevel and GPCM models had better model fit according to the DIC values compared to testlet and dual models controlling for LID and the trend was maintained as the testlet size and category response options changed. The multilevel and GPCM models had better fit across all category options when only persons were dependent. However, in testlet and dual effects, the testlet and dual models had significantly lower DICs than independent-items models across all options. The fit statistics for GPCM and multilevel models were similar and higher than the dual and testlet models which were

also similar. The trend in fit statistics did not change as the number of items changed, implying that the performances of models are robust to changes in item attributes. However, for 4 category options, the testlet and dual models outperformed the multilevel model.

Table 6.1: DIC fit statistics for changing testlet sizes and category options

Items	Categories	Condition	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	NoneNone	26890	21310	26880	26590	26500
		NoneLarge	23120	23200	23070	23150	23100
		LargeNone	29780	23840	29760	23640	23650
		LargeLarge	26700	21310	26640	21150	21430
	4	NoneNone	33260	33280	33260	33290	33300
		NoneLarge	25620	25720	25610	25640	25700
		LargeNone	27130	22220	27110	22120	22200
		LargeLarge	27720	21700	27680	21560	22150
	5	NoneNone	31490	31560	31500	31540	31520
		NoneLarge	26520	26750	26460	26530	26530
		LargeNone	31360	23770	31360	23720	24150
		LargeLarge	30280	23000	30160	22710	23750
6	3	NoneNone	49610	49710	49620	49820	49820
		NoneLarge	43990	44050	43980	44140	44100
		LargeNone	57600	47410	57580	47270	47300
		LargeLarge	50820	42120	50780	41960	42000
	4	NoneNone	58370	58440	58520	58330	58320
		NoneLarge	52020	52060	51890	51960	51990
		LargeNone	63650	49320	63400	48970	49000
		LargeLarge	59930	47680	59930	47280	47300
	5	NoneNone	62240	62300	62110	62160	62180
		NoneLarge	56880	56770	56730	56810	56820
		LargeNone	69330	52720	69120	52360	52360
		LargeLarge	59500	44900	59280	44470	44480
10	3	NoneNone	77980	78100	77980	78070	78010
		NoneLarge	71890	72050	71890	71960	72020
		LargeNone	85860	70220	85850	70070	71000
		LargeLarge	78520	64610	78520	64480	66410
	4	NoneNone	87580	87670	87580	87690	87610
		NoneLarge	89010	89120	89000	89100	89000
		LargeNone	97780	76950	97770	76860	76950
		LargeLarge	91430	70930	91410	70630	72540
	5	NoneNone	102100	102200	102100	102200	102100
		NoneLarge	93880	94080	93830	93940	94040
		LargeNone	112500	84840	112500	84540	88950
		LargeLarge	97490	73610	97490	73590	73940

6.3.2 Variance recovery

According to the results in Table 25 (Appendix 1), in local independence, ability parameter variances were close to their true values and variances approached their true values as the testlet size increases and number of categories for all models. The GPCM and multilevel models for independent items recorded lower variances when compared to testlet and dual models falsely accounting for absent testlet effects. However, variances were lower for 4 category tests than for 3 and 5 category tests. Multilevel and GPCM models recorded ability variances that were almost constant across all testlet sizes. In LID and dual dependence effects, independent items models underestimated the ability parameter variance increasingly with the number of items per testlet and response categories. However, the decrease was not very clear for tests with 4 response category items. The variances were slightly overestimated by the testlet model and well recovered by dual model but becoming underestimated as the number of items and response options increases.

In group effects only, the ability variances were overestimated by testlet and GPCM models ignoring person dependency effects and the overestimation increased with testlet size. However, the effects of increase in number of response categories was not very clear as the variances were lower for 4 category response options than for 3 and 5 responses. The variances were well recovered by the multilevel model and slightly higher for the dual model and decreased for 5 categories across all testlet sizes. In local independence, the variances in ability, group effects and interaction between testlets and person group did not appear to be significantly affected by increasing the number of response categories. However, the testlet model seemed to overestimate the ability variance in LPD increasingly with testlet size while the underestimation of ability variances by independent items model in LID was maintained across testlet sizes.

The multilevel and dual models detected the absence of group dependence effects while

dual and testlet models detected the absence of item dependence effects for 3 response category tests and becoming smaller as both the number of items and response options increased. The multilevel models generally underestimated the ability variance for all conditions but overestimated the group variances when they were present. In addition, testlet-based models detected the presence of group and testlet interactions when they were present. In group effects only, the multilevel and dual models recover the group variances well. However, in LPD only, the multilevel recovered group variance well while the dual models overestimated the group variances across all testlet sizes and response categories sizes.

6.3.3 Ability parameter recovery

The prowess of the models to estimate the proficiency levels in LID and LPD effects, for varying testlet length and number of response categories were assessed in terms of the ability to retain the rank ordering of respondents according to their traits, systematic, random and total errors. Descriptive and inferential statistics were used for model comparison.

Table 6.2 shows the Pearson's correlation coefficients between true and estimated ability parameters under different dependence condition for changing test(let) sizes and category options. The results show that in local independence and as the both test(let) length and number of category options increased, the correlation between true and estimated parameters increased for all calibration models. When person dependence effects were accounted for, correlations between the true-estimate for ability parameters increased as the number of items and response categories increased. However, the effects of number of categories on correlations when dependence conditions were not accounted for was not very clear, but seem to be increasing. The correlations for models ignoring LPD effects decreased as the number of response options increased for 3 items tests but the effect of increasing response options was not very clear for 6 and

Table 6.2: True-estimated abilities correlations for changing items and categories

Items	Categories	LID	LPD	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	None	None	0.95	0.95	0.93	0.94	0.96
		None	Large	0.76	0.78	0.92	0.93	0.93
		Large	None	0.87	0.77	0.92	0.93	0.92
		Large	Large	0.68	0.75	0.92	0.92	0.90
	4	None	None	0.96	0.96	0.95	0.95	0.94
		None	Large	0.73	0.75	0.93	0.93	0.91
		Large	None	0.86	0.84	0.92	0.94	0.91
		Large	Large	0.66	0.79	0.92	0.93	0.90
	5	None	None	0.97	0.97	0.97	0.97	0.95
		None	Large	0.61	0.66	0.94	0.94	0.90
		Large	None	0.87	0.80	0.93	0.93	0.91
		Large	Large	0.60	0.67	0.92	0.94	0.88
6	3	None	None	0.97	0.97	0.97	0.97	0.94
		None	Large	0.71	0.72	0.95	0.95	0.95
		Large	None	0.85	0.92	0.94	0.95	0.95
		Large	Large	0.66	0.70	0.94	0.95	0.95
	4	None	None	0.98	0.98	0.97	0.97	0.96
		None	Large	0.63	0.65	0.94	0.95	0.92
		Large	None	0.88	0.77	0.94	0.95	0.95
		Large	Large	0.64	0.72	0.95	0.96	0.95
	5	None	None	0.99	0.99	0.98	0.98	0.94
		None	Large	0.73	0.75	0.96	0.96	0.94
		Large	None	0.88	0.80	0.95	0.96	0.95
		Large	Large	0.68	0.66	0.95	0.96	0.90
10	3	None	None	0.99	0.98	0.03	0.97	0.95
		None	Large	0.70	0.72	0.96	0.96	0.93
		Large	None	0.90	0.79	0.95	0.95	0.98
		Large	Large	0.67	0.64	0.94	0.96	0.92
	4	None	None	0.99	0.99	0.97	0.97	0.97
		None	Large	0.74	0.76	0.97	0.97	0.98
		Large	None	0.92	0.81	0.96	0.96	0.96
		Large	Large	0.73	0.76	0.95	0.95	0.95
	5	None	None	0.99	0.99	0.98	0.98	0.97
		None	Large	0.70	0.72	0.97	0.97	0.94
		Large	None	0.90	0.90	0.96	0.97	0.96
		Large	Large	0.69	0.69	0.69	0.97	0.94

10 items tests. The correlations in LID only, although slightly lower for GPCM model, are generally high across all models, increasing as the number of items and response options increased.

The random, systematic and total errors in the ability parameter recovery when the number of items per testlet increases are shown in Figure 6.1. According to the results,

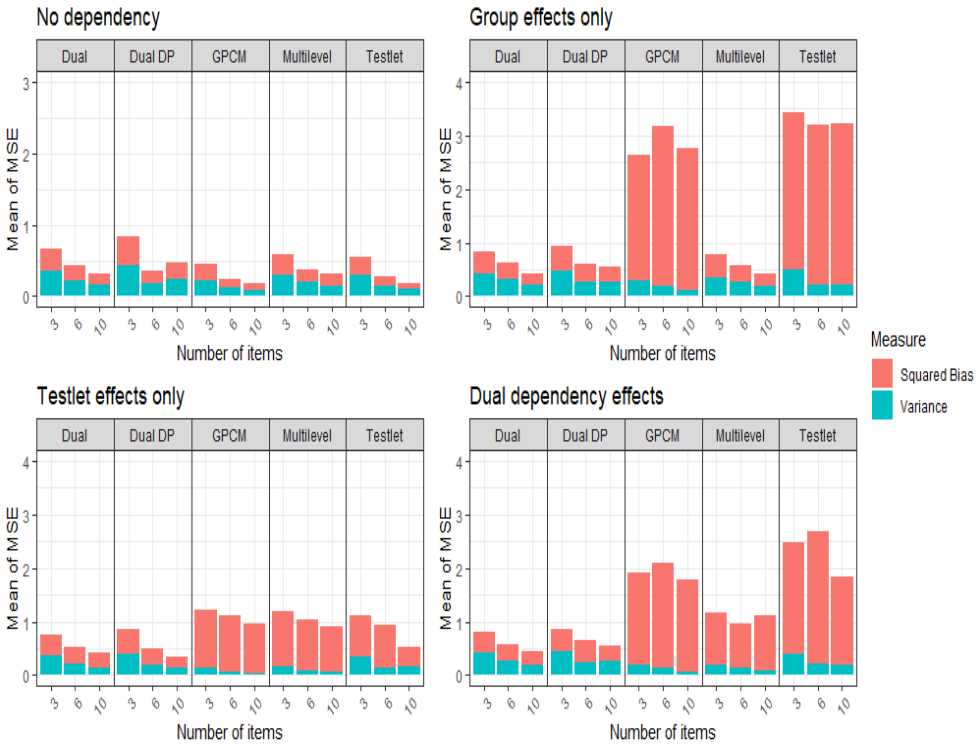


Figure 6.1: Random, systematic and total error in the ability parameters changing test(let) sizes

in the absence of LID and LPD, the total, systematic and random errors decreased as the testlet size increases. Systematic errors were high for the GPCM and testlet models not controlling for group effects and the effect of testlet size is not very clear from the graph. Testlet effects do not seem to affect the ability parameter recovery although the errors in ability parameter estimation across all models decreased as the testlet sizes increased. In the presence of dual dependence effects, the total errors in the GPCM and testlet were generally high. The effect of varying testlet length in the presence of person dependence effects is not very clear as the models ignoring person dependence effects recorded highest errors in testlet of 6 items than tests of 3 and 10 items per testlet.

In LID and LPD, all models recorded low systematic and random errors across for 3, 4 and 5 response categories and total errors decreased when the number of response categories in items increased (Figure 6.2). Failure to account for LPD effects resulted in larger systematic errors which decreased as the number of response categories increased while failure to account for testlet effects slightly affected the systematic errors

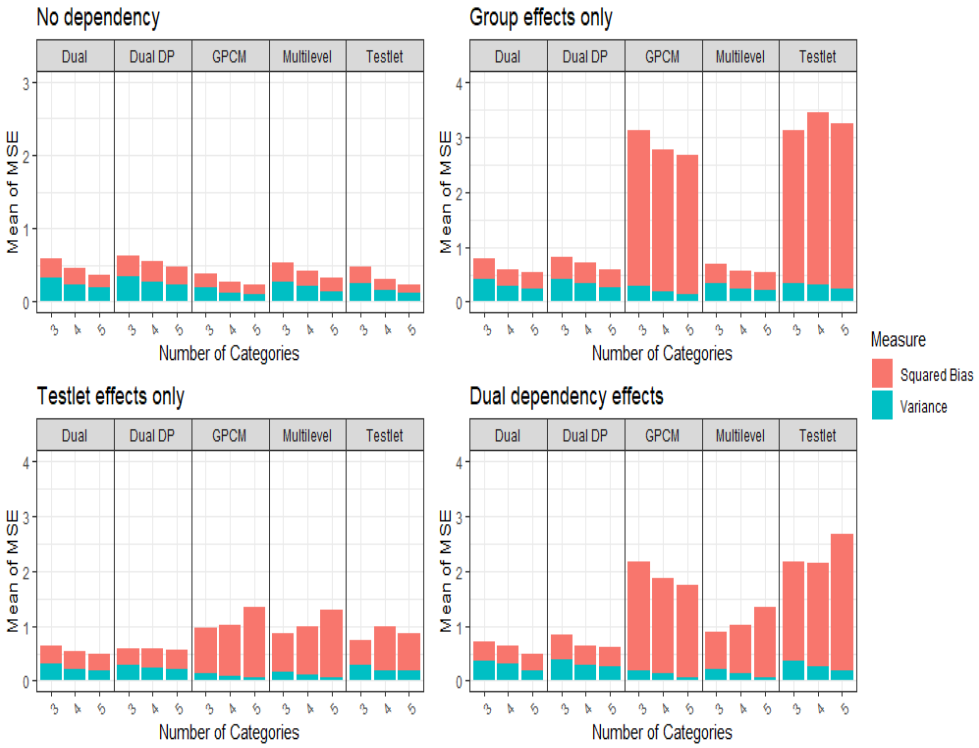


Figure 6.2: Random, systematic and total errors for ability parameters for changing number of response categories

which were highest for tests with 5 response category items. In LID only, total errors in GPCM and multilevel models increased as the number of items increased. In dual dependence effects, systematic errors were higher in GPCM and testlet models and seemed to be decreasing as the number of category options increased. Although the multilevel model recorded lower errors in dual effects, the errors increased as the number of category options increased. From the results, both RMSE and SE seemed to be affected by varying number of response categories in items. A model that fits the data well will have a unit ratio between systematic and random errors. It can be concluded from the figure that most of the models, real standard errors were higher than model standard errors and the ratio increased with the number of response categories. The RMSE generally increased with increasing number of categories in an item.

In local item and person independence, bias in ability parameter estimation were low for all models and narrow bands were maintained across testlet sizes (Figure 6.3) although they seemed to be decreasing as the number of items per testlet and number of category options per item increased. However, in group dependence, bias were high in GPCM

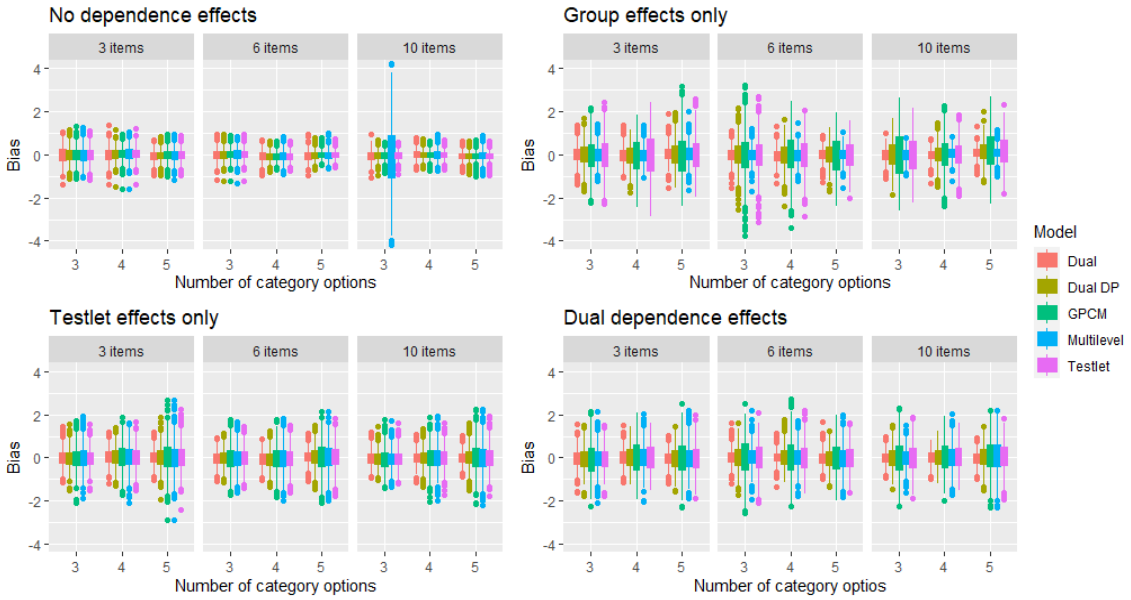


Figure 6.3: Bias in the ability parameters for changing test(let) sizes

and testlet models ignoring LPD and seemed to be increasing with testlet size and number of category options for 3 and 6 items per testlet tests but not very clear for 10 items. The errors were lower in multilevel models across all item and category sizes. Bias in ability estimation was slightly affected by LID and similar patterns were recorded across different testlet and category sizes, making it difficult to detect the actual effects of variant testlet and category size on ability parameter recovery if any. The dual models recovered person parameters with the least total error across testlet and category sizes. In LPD effects, bias ranges were wider in testlet and GPCM models and remained almost constant as the number of response categories increased for all dependence conditions.

6.3.3.1 ANOVA

To evaluate the effects of varying test(let) length and number of response categories on parameter recovery, a crossed 2 (LID conditions) $\times$ 2 (LPD conditions) $\times$ 3 (test(let) sizes)$\times$ 3 (category sizes) $\times$ 5 (calibration models) factorial design was used for inference. Although the model had item dependency, group effects and calibration model as factors, analysis was mainly for the main effects of testlet size and number of response options and their interaction with other factors since the other factors have been analysed in Chapter 4. Bias, absolute bias (abias), random and total errors were

used as dependent variables of GLM. From the ANOVA results, although testlet size and number of response options and their interactions with other model factors had significant p-values < 0.05 for both bias and abias, all the Cohen effects were negligible. Tukey *post hoc* analysis showed no significant difference between abias for tests with 5 and 3 response categories and abias was significantly lower than test with 4 response categories items. Comparisons for interaction between group and model effects shows that abias was highest in group and dual dependence effects and for testlet and GPCM models ignoring LPD, highest when 10 item tests were coupled with 4 response categories and lowest for all models in local independence for 5 response categories.

Both testlet size and number of response category options significantly effected SEs in ability parameter estimation with effect sizes of $f = 0.28$ and $f = 0.28$ respectively. In addition, LID, LPD and calibration model were also significant. The two-way interaction between testlet size and LPD and model factors were significant with small effect sizes of $f = 0.12$ and $f = 0.13$ respectively. However, two-way interaction between testlet length and LID were insignificant. Three-way interaction effects for testlet size, LID and model and three-way interactions between testlet size, model and LPD were significant with small effect sizes of $f = 0.12$ each. The Tukey *post hoc* for main effects of testlet size show that SEs were smallest for longer testlets of 10 items each and were highest in testlets and means for all testlet sizes differed significantly. The *post hoc* analysis shows that SEs were lowest for tests with 5 response categories items and highest for items with 3 response categories items, implying that random errors decreased as the number of response categories increased.

The *post hoc* for interaction between testlet size and LID indicated that smallest SEs were recorded for 10 items for larger LID. The interaction between number of categories and LID, LPD and calibration model were significant (p-value < 0.05) with effects of $f = 0.07$ (negligible), $f = 0.14$ (small) and $f = 0.14$ (small) respectively. The SE were

lowest in GPCM and multilevel models in LID and dual effects, and they decreased as number categories increased. Highest SEs were recorded in dual models in testlet and dual effects, for fewer response categories. SEs were highest for 10 items, highest for independent. Contrary to these results, the *post hoc* for group and testlet effects show that SEs were lowest for longer tests of 10 items per testlet in item independence and largest for shorter tests in large LID. Testlet size and model interaction indicated lowest SEs in GPCM and multilevel models ignoring LID and for 10 items for 5 and 4 categories and were highest in dual models for shorter tests of 3 testlet items. *Post hoc* for three-way interaction between LPD, LID and testlet size show lowest SEs in LID only and in local independence for longer tests of 10 items per testlet and highest in dual dependence and large LPD effects for shorter tests of 3 items per testlet coupled with 3 and 4 category options. SEs were lowest for GPCM in local independence or when LID effects were ignored, for longer tests and were higher for models controlling for LID and models ignoring LPD for testlets of 3 items.

The main effects of changing testlet size and number of response options on RMSE in ability parameter estimates were statistically significant with small effect sizes of $f = 0.13$ and $f = 0.12$ respectively. The Tukey multiple comparison analysis showed that RMSE were smallest for longer testlets of 10 items per testlet and largest for tests with 3 items per testlet. The RMSE decreased with the increase in number of response options, highest for 3 categories and lowest for 5 categories. Two-way interaction between testlet effects and testlet sizes and group effects and testlet size show that RMSE were smallest in item independence, for 10 items per testlet and largest in large LID for 3 items testlets. However, the ordering for LPD was that all tests with no dependence effects had smaller RMSE, ordered according to testlet size. Three and four-way interactions between testlet size and model, LID and LPD were statistically significant according to p-values although all effect sizes were less insignificant (< 0.10). Four-way interaction between LID, LPD, model and testlet size show smallest RMSE in local

independence for GPCM model for longer tests of 10 and 6 items per testlet and largest for dual large person effects, for the GPCM model for 3 items testlets, implying that ignoring dual and person dependence in ability estimation has more negative effects on total errors for shorter testlets.

The RMSE in local independence decreased with increase in number of categories, lower for GPCM and testlet and slightly higher for dual and multilevel models, lower for 3 response categories. In local independence, RMSE are lower in dual and multilevel, higher in the testlet and GPCM models and increase as the number of categories increase. Lowest RMSE were recorded in local independence for all models, for large tests of 10 items per testlet and 5 category options. Highest RMSE were recorded in the GPCM and multilevel models for 3 response options across all testlet sizes, suggesting that the number of categories impacted more on total errors in proficiency estimation than testlet size.

6.3.4 Threshold parameter recovery

The ability of the models to predict item step parameters for variant item and category options size was evaluated using the true and estimated parameter correlations as well as the estimation errors. Both descriptive and inferential statistical methods were utilised for model comparison.

The results in Table 6.3 show that in local independence, threshold parameters were well recovered by all models across testlet and category sizes although correlations were slightly lower in dual models falsely accounting for absent LID. Correlations appeared to increase with testlet length when items are independent and when item dependence are controlled for and seem to decrease when LID was omitted in modelling. However, in testlet effects, correlations for testlet and dual models increased with number of response categories while correlations for GPCM and multilevel models were lowest for four categories, making it difficult to determine the effects of varying category options.

Table 6.3: Thresholds correlations for changing items and category options

Items	Categories	LID	LPD	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	None	None	0.99	0.99	0.99	0.93	0.96
		None	Large	1.00	1.00	1.00	0.94	0.95
		Large	None	0.62	0.71	0.62	0.85	0.87
		Large	Large	0.72	0.83	0.72	0.77	0.79
	4	None	None	0.99	0.99	0.99	0.98	0.98
		None	Large	1.00	0.99	0.99	0.94	0.95
		Large	None	0.80	0.87	0.78	0.86	0.89
		Large	Large	0.67	0.71	0.67	0.83	0.91
	5	None	None	0.99	0.98	0.9	0.96	0.96
		None	Large	0.97	0.97	0.98	0.94	0.94
		Large	None	0.73	0.81	0.73	0.82	0.88
		Large	Large	0.63	0.73	0.62	0.80	0.96
6	3	None	None	0.99	0.99	0.99	0.93	0.98
		None	Large	0.99	0.99	0.99	0.87	0.96
		Large	None	0.84	0.94	0.85	0.93	0.92
		Large	Large	0.75	0.85	0.75	0.90	0.90
	4	None	None	0.99	0.99	0.99	0.99	0.99
		None	Large	0.99	0.99	0.99	0.99	0.99
		Large	None	0.65	0.82	0.65	0.94	0.82
		Large	Large	0.72	0.91	0.72	0.96	0.90
	5	None	None	0.99	0.99	0.99	0.95	0.95
		None	Large	0.99	0.99	0.99	0.98	0.98
		Large	None	0.72	0.93	0.72	0.97	0.90
		Large	Large	0.76	0.96	0.76	0.97	0.91
10	3	None	None	1.00	1.00	0.99	0.97	0.96
		None	Large	1.00	0.99	1.00	0.93	0.93
		Large	None	0.74	0.77	0.74	0.87	0.87
		Large	Large	0.66	0.70	0.66	0.93	0.94
	4	None	None	0.99	0.99	0.99	0.95	0.95
		None	Large	0.99	0.99	0.99	0.98	0.98
		Large	None	0.76	0.93	0.77	0.94	0.93
		Large	Large	0.61	0.65	0.61	0.93	0.93
	5	None	None	0.99	0.98	0.98	0.94	0.93
		None	Large	0.96	0.95	0.99	0.98	0.97
		Large	None	0.66	0.73	0.65	0.95	0.94
		Large	Large	0.71	0.89	0.93	0.93	0.92

Results in Figure 6.4 show that in local independence, errors were generally low and seem to be decreasing as number of items increased across all models. Higher systematic errors were recorded in dual models spuriously accounting for absent dependence effects and seemed to be increasing with testlet size, implying that controlling for LID when items are independent may bias the step parameters. Errors were highest in 3

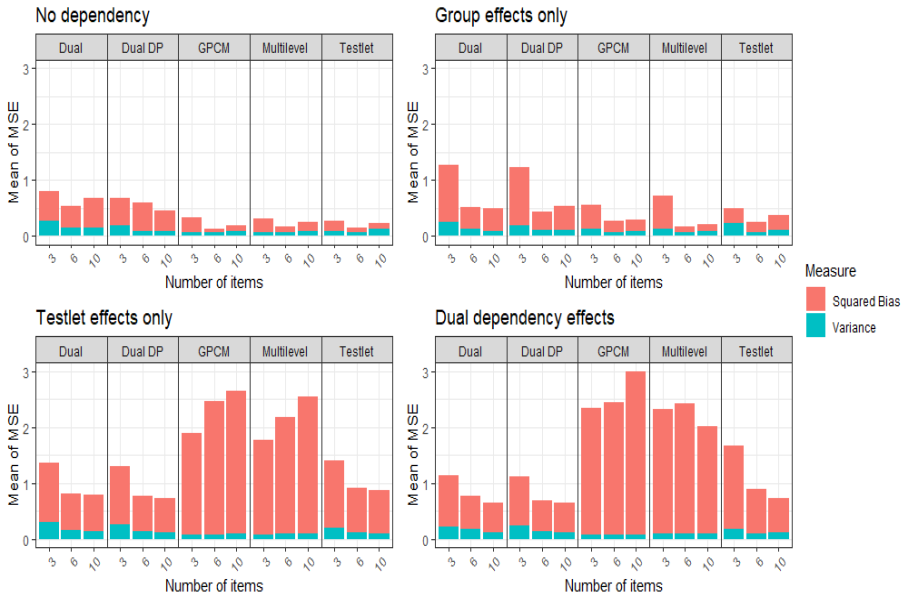


Figure 6.4: Random, systematic and total errors in the thresholds for changing test(let) sizes

items per testlet tests and seem to be decreasing as number of items increase. Failure to control for LID increased bias and total errors increasingly with testlet size.

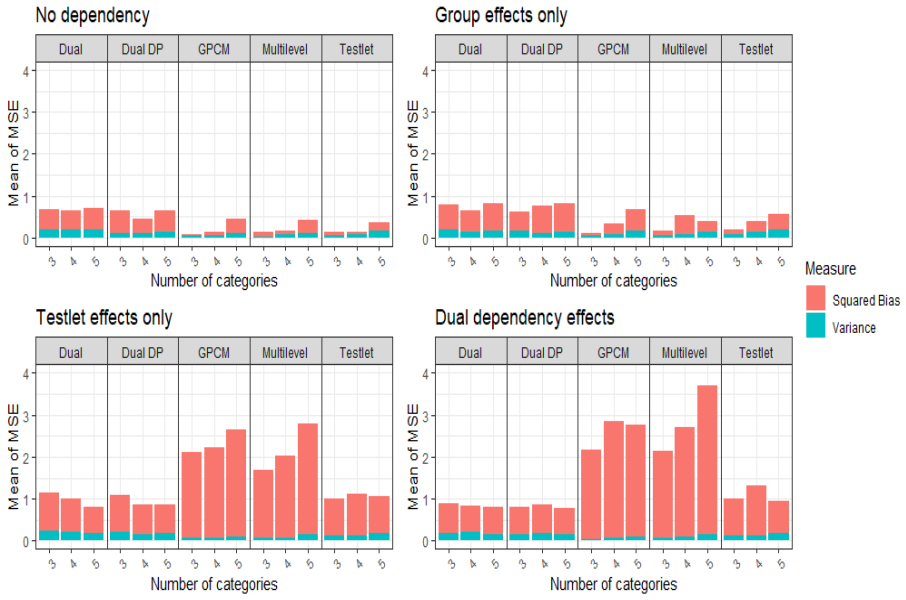


Figure 6.5: Plots of true threshold parameters against estimates for item and person dependency conditions

The ratio of error variance over sampling variance estimates were plotted for each dependence condition in Figure 6.5. The results show that in independence, total errors although very small, seem to be increasing with the number of categories for testlet,

GPCM and multilevel models. Systematic errors were highest in dual models spuriously accounting for absent effects and errors seemed to increasing with number of categories. In person dependence effects, errors seemed to increase with number of categories for GPCM and testlet models ignoring LPD while the effects of increasing category options is not very clear for dual and multilevel models. In item and dual dependence effects, systematic errors were generally high and increased with number of categories in GPCM and multilevel models. However, effects of increasing categories when dependence conditions were accounted for is not clear.

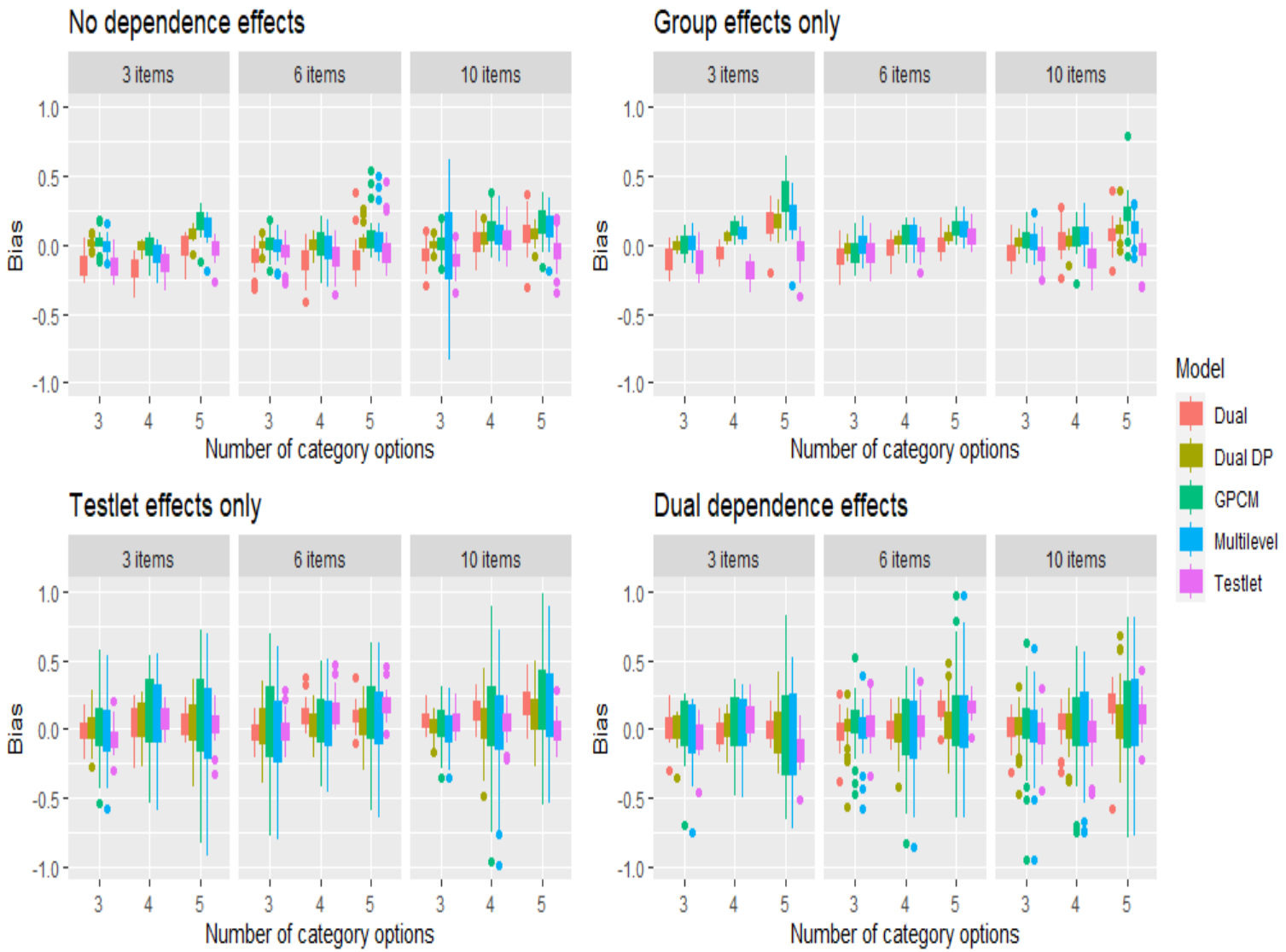


Figure 6.6: Bias in the threshold parameters for changing category options

From the results in Figure 6.6, in local independence, threshold bias were low across all models although slightly higher for dual models falsely accounting for absent LID. Group effects do not seem to impact on threshold parameters significantly although bias appear to be lower for 10 items than 3 items per testlet tests. In addition, medians for biases were slightly lower for 3 items testlets and higher for 6 and 10 items. Similar pattern were recorded in LID only and medians were lower than zero (0), an indication that true values were underestimated. The medians increased to values more than zero (0) for 10 items testlets. However, testlet size did not seem to significantly affect the range of biases. In dual dependence, threshold parameters for 3 items were underestimated and the underestimation was cleared by increasing testlet size to 6 and 10 items. In item independence, bias was low in all models across all response categories but were slightly higher in dual models. Modes ignoring LID recorded high threshold bias and the trend was maintained as categorisation increased. Thresholds seem to have been underestimated by all models for 4 categories.

6.3.4.1 ANOVA

The GLM shows all factors not to have significant impact on bias in threshold parameter estimation. However, main effects of testlet length were significant on abias with an effect size of $f = 0.17$ while main effects of number of categories had significant p-value but insignificant Cohen effects. All two-way interaction between testlet length and model, testlet length and group effects and testlet length and LID were significant with small effect sizes of $f = 0.13$, $f = 0.15$ and $f = 0.10$ respectively. Three-way interaction between between testlet length, LPD and LID, testlet length, group and model effects and four-way interaction had significant p-values and effect sizes of $f = 0.16$ (small effect), $f = 0.10$ (small effect) and $f = 0.03$ (insignificant effect) respectively. The *post hoc* analysis show that bias were smallest (negative) for large LID for 3 items per testlet tests, followed by tests with no testlet effects, and largest (positive) for LID for 6 and 10 items per testlet, an indication that in LID, threshold parameters were underestimated for smaller testlet tests.

Multiple comparisons for the effects of interaction between LPD and testlet size show that bias was lowest (negative) for tests of 3 items per testlet, and the tests with 10 items per testlet having the highest (positive), arrangement for each testlet size arranged in order of the magnitude of LPD, higher in large LPD and lower in the absence of LPD. For the *post hoc* for the interaction effects between calibration model and testlet length, the lowest bias was recorded for GPCM and multilevel models for shorter testlets of 3 items and largest for GPCM for tests of 10 items per testlet. This implies that ignoring LID leads to underestimation of threshold parameters for shorter testlets and overestimation of the same for longer testlets. Bias were smaller (negative) in dual dependence and large LID for shorter tests of 3 items testlets and highest in LID and dual dependence for 6 and 10 items per testlet tests, indicating that LID affect threshold bias more than LPD. In addition, LID led to underestimation of threshold parameters for shorter tests and overestimation for longer tests, and the underestimation and overestimation were worst for GPCM and multilevel ignoring testlet effects. Similar results were recorded for abias and abias was less for longer tests of 10 items per testlet and higher for shorter tests of 3 items per testlet.

All two-way to four-way interaction effects between number of response categories other models factors were significant (p-value < 0.05) but with negligible effects sizes less than 0.05. The Tukey *post hoc* comparison shows that lowest abias were recorded for 5 response categories and were significantly lower than abias in 3 and 4 categories. Lowest abias were recorded in local independence and for controlled testlet effects and higher for models ignoring LID. The multiple comparison for four-way interaction effects show lowest bias in GPCM and testlet models in independence, bias decreased as number of categories increased. In dual and testlet effects, biases were lower in dual models, decreasing with increasing number of options although 3 options performed better than 4 options. The bias were highest in person, item and dual effects for 3 and

4 category options, worst for 4 categories.

The main effects of testlet size were significant on thresholds SE with a small effect size of $f = 0.17$ while the p-value for main effects of number of categories had an insignificant f. Model, group and testlet factors were significant on SE. Two-way interaction effects between number of categories and LID, LPD and model had significant effects (p-value < 0.05) on SE with negligible to small effect sizes of $f = 0.04$, $f = 0.14$ and $f = 0.03$ respectively. All three-way and two-way interactions between testlet size and model, LID and LPD had significant effect sizes except for two-way interaction between testlet size and LID and three-way interaction between testlet size, LID and LPD. Model effects were highly significant in aggregating SE threshold parameter estimation with a high effect size of $f = 0.41$.

According to the Tukey *post hoc*, SE were lowest for longer testlets of 10 items each, with no significant differences between 6 and 3 items per testlet. Lowest SE were recorded in independence for longer tests of 10 and 3 items per testlet and were highest in large LID for 3 items per testlet. SE were highest for shorter tests of 3 items per testlet in person independence while they were lowest for longer tests of 6 items. The 5-way interaction between model, LID, LPD, testlet size and categorisation show lowest SEs in multilevel and dual models for items of 10 and 6 items and 3 categories followed by 4 categories. SEs were highest in GPCM model in LID and the ordering of items and categories was not clear according to *post hoc* analysis. Lowest SE were for tests with 5 response categories and they differed significantly from SE from tests with 4 and 3 response categories (p-value < 0.05). In addition, SE were lower for GPCM which differed significantly from multilevel and testlet models and highest for dual model (p-value < 0.05). However, although categorisation had a significant ANOVA p-value < 0.05 on total errors and a negligibly low ($f = 0.06$) Cohen f. Group effects had a low effect of $f = 0.00$ while testlet and model effects had small effects of size

$f = 0.15$ and $f = 0.12$ respectively. Five-way interaction showed the lowest SE in GPCM and multilevel models for large tests of size 6 and 10 testlet items and fewer categories of size 3 and 4 as well as for dual and testlet models in item in LID. SEs were highest in GPCM model in LID for small tests of size 3 and smaller response options of size 3.

The ANOVA analysis for total errors in thresholds estimation had significant effects for testlet length ($f = 0.15$). However, although the two-way, three-way and four-way interaction effects between testlet length and other factors were significant (p-value $<$ 0.05), the Cohen effect sizes were all negligible. The multiple comparison for interaction between testlet length and LID show that total errors were significantly lower in item independence for testlets with 5 and 6 items, followed by large LID for 5 and 6 items per testlet while they were highest for shorter tests of 3 items per testlet, both in the presence and absence of testlet effects. For the *post hoc* for interaction between LPD and testlet length, lower total errors were recorded in the absence of LPD, lowest for 10 items per testlet conditions, followed by 6 items per testlet tests. RMSE were higher in large dependence effects, and again arranged in order of testlet length, with the highest being recorded for shorter tests of 3 items testlets. Furthermore, total errors were lowest for GPCM, multilevel and testlet models in local independence for longer tests of 10 followed by 6 items per testlet while errors were highest for GPCM and multilevel in dual dependence for 10 items per testlet and 6 items per testlet tests. The errors were also high in large LID only, for GPCM and multilevel models ignoring the effects, highest for longer tests with 10 and 6 items per testlet. This implies that when LID is not controlled, RMSE increased with testlet size but in the absence of LID and when the condition is accounted for, RMSE decreased as testlet size increased.

Two and three-way interaction between LID and calibration model had small but significant effects on RMSE. Total errors were lowest for 3 category items and they

differed significantly with total errors in for 4 and 5 categories. In addition, total errors were lowest in testlet models and differed significantly with all other models (p-value < 0.05). However, the total errors in dual models did nod differ significantly but were significantly lower than errors in GPCM and multilevel models. T otal errors in GPCM and multilevel models ignoring LID were highest and they did not differ significantly.

6.3.5 Discriminant parameter recovery

The results in this section show the effects of changing number items per testlet and number of response categories in test items on recovery of item discriminant parameters. The results for correlations between true and estimated discriminants in Table 6.4 show that in local independence, correlations decreased as number of categories increased. In group effects only, the rank ordering of items according to discrimination ability was not much affected. In dual dependence, multilevel and GPCM models had generally low correlations, lowest for 4 categories.

Table 6.4, show that in local independence, correlations decreased as number of items increased across all models. In item dependence only, GPCM and multilevel models failed to retain item discrimination ability increasingly with number of categories, more so for shorter tests. The effects of testlet length is not very clear for independent items as discrimination abilities were maintained. However, correlations for models ignoring testlets effects seem to decrease as testlet size increase while correlations for models accounting for LID increase slightly with testlet length.

Total errors in discriminant parameters in independence are generally low for all models and seemed to be decreasing as number of items per testlets increased (Figure 6.7). Similar results were observed in group effects only. However, in testlet and dual dependence effect, high systematic errors were recorded in GPCM and multilevel models ignoring testlet effects and seem to be increasing as number of items within testlets

Table 6.4: True/estimated discriminant correlations for changing items and categories

Items	Categories	LID	LPD	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	None	None	0.96	0.95	0.96	0.95	0.95
		None	Large	0.92	0.85	0.91	0.90	0.90
		Large	None	0.25	0.88	0.25	0.92	0.91
		Large	Large	0.60	0.89	0.92	0.92	0.93
	4	None	None	0.94	0.93	0.93	0.93	0.93
		None	Large	0.97	0.94	0.96	0.96	0.95
		Large	None	0.15	0.87	0.14	0.80	0.90
		Large	Large	0.39	0.84	0.39	0.90	0.90
	5	None	None	0.92	0.93	0.92	0.92	0.92
		None	Large	0.84	0.81	0.75	0.88	0.90
		Large	None	-0.04	0.89	-0.03	0.95	0.90
		Large	Large	0.30	0.84	0.40	0.93	0.80
6	3	None	None	0.94	0.94	0.94	0.94	0.94
		None	Large	0.94	0.93	0.94	0.93	0.93
		Large	None	0.49	0.94	0.46	0.97	0.96
		Large	Large	0.60	0.91	0.58	0.94	0.85
	4	None	None	0.94	0.94	0.94	0.94	0.94
		None	Large	0.93	0.94	0.93	0.94	0.94
		Large	None	0.41	0.94	0.38	0.96	0.87
		Large	Large	0.09	0.88	0.09	0.91	0.64
	5	None	None	0.89	0.88	0.89	0.88	0.88
		None	Large	0.97	0.96	0.97	0.96	0.96
		Large	None	0.41	0.94	0.40	0.95	0.81
		Large	Large	0.48	0.95	0.47	0.96	0.78
10	3	None	None	0.94	0.94	-0.04	0.94	0.94
		None	Large	0.95	0.95	0.95	0.96	0.95
		Large	None	0.63	0.91	0.63	0.93	0.94
		Large	Large	0.55	0.90	0.56	0.93	0.93
	4	None	None	0.90	0.90	0.89	0.90	0.91
		None	Large	0.95	0.95	0.95	0.95	0.95
		Large	None	0.18	0.93	0.18	0.94	0.94
		Large	Large	0.35	0.91	0.35	0.94	0.94
	5	None	None	0.93	0.93	0.92	0.92	0.93
		None	Large	0.91	0.92	0.95	0.93	0.93
		Large	None	0.34	0.93	0.33	0.93	0.93
		Large	Large	0.37	0.91	0.90	0.87	0.90

increased.

In independence and LPD effects only, total errors were minimal in all models and increased slightly as number of categories increased (Figure 6.8). In LID only, total errors were highest for multilevel and GPCM models ignoring testlet effects and decreased as number of response categories increased. However, errors in dual and testlet models

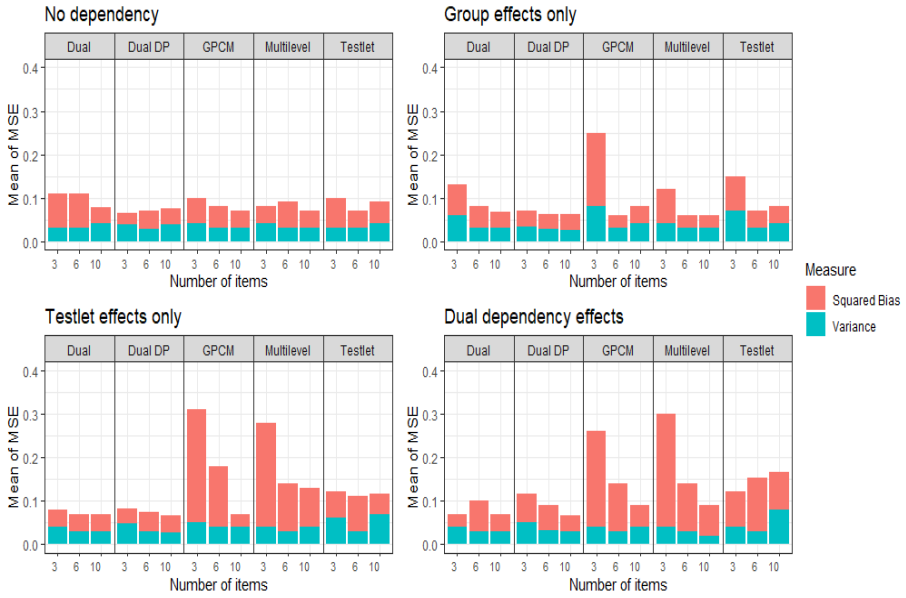


Figure 6.7: Bias, SE and MSE in discriminant parameters for different test(let) sizes

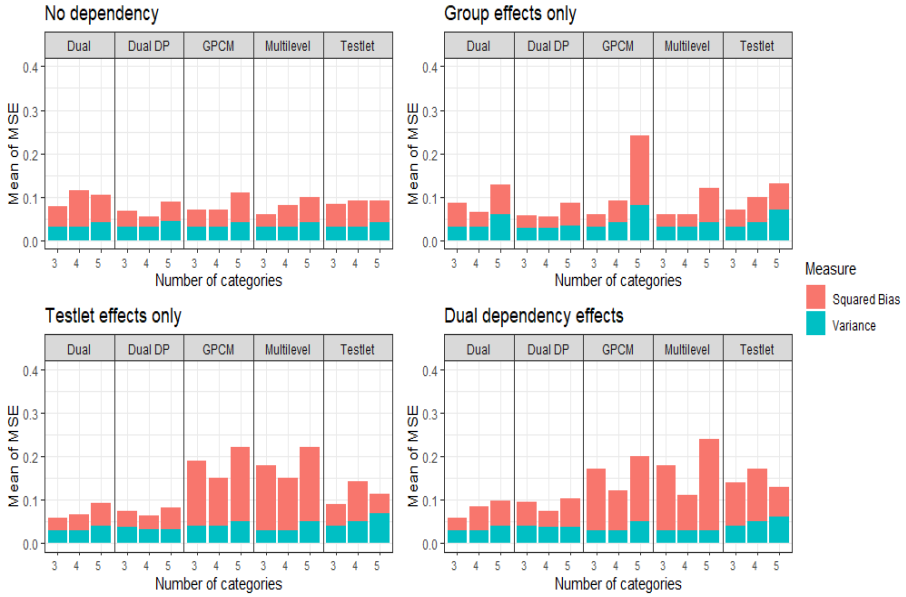


Figure 6.8: Bias, SE and MSE in discriminant parameters for changing no. of categories controlling for item dependence increased with number of categories. Moreover, in dual effects, errors in discriminant parameters increased as number of categories increased from 3 to 5.

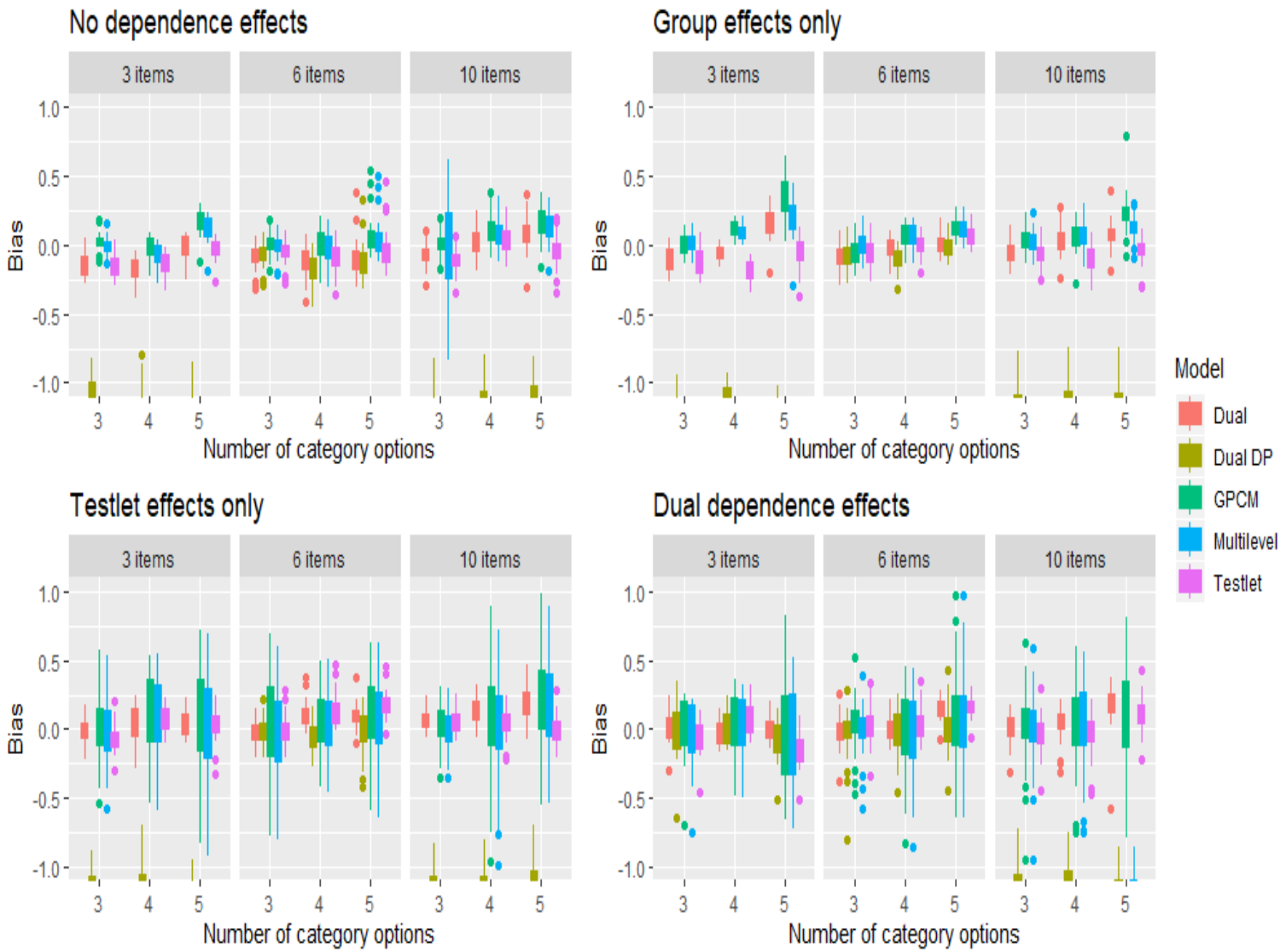


Figure 6.9: Bias in the discriminant parameters for changing category options

Figure 6.9 shows that in local independence, GPCM and multilevel models seemed to recover discriminant parameters well and bands were narrower for 6 and 10 items per testlets. The discriminant estimates were becoming lower than their true values as testlet size increased and the trend was noted across all models. In local independence, dual and testlet models underestimated the discriminant parameter for all testlet sizes. In group effects only, GPCM and multilevel models overestimated discriminant parameters in 3 items testlets while dual and testlet models underestimated the discriminant parameter for large LID. In LID only, dual and testlet models overestimated discriminant parameters for 3 items testlets while GPCM and multilevel models parameter estimates were more biased. The effect of testlet size on the range of bias in discriminant parameters did not seem to be significant. The results show that ignoring LID affected the discriminant parameter recovery more than LPD effects.

In local independence, dual and testlet models underestimated discriminant parameter while multilevel and GPCM recover the parameter well (Figure 6.9). In LPD only, dual and testlet models underestimated discriminant parameters for fewer categories and recovered well for tests with 5 options. On the other hand, the multilevel and GPCM overestimated the parameter as number of categories increased. In testlet and dual dependence effects, dual and testlet models recorded narrower and zero median biases although the testlet model overestimated the parameter when response options increased. Multilevel and GPCM models ignoring testlet effects had wider although zero (0) median bands.

6.3.5.1 ANOVA

The GLM with systematic errors (bias), abias, random errors (SE) and total errors (RMSE) as dependent variables was employed to evaluate the models in terms of their ability to retain the properties of discriminant parameters when testlet size and categorisation increased. The factors considered for modelling are the calibration model, testlet effects, group effects, testlet length and number of response options. However,

the analysis in this section concentrated on the effects of changing testlet size and category options and their interactions with other factors in the model. Main effects of testlet length and number of options on discriminant parameter bias were significant with effect sizes of $f = 0.18$ and $f = 0.23$ respectively. All the other model factors did not significantly affect bias in recovery of discriminant parameter.

The interaction between testlet length and other factors had significant p-values but only the interaction between testlet size and person dependence had a significant effect ($f = 0.10$) on bias. However, interaction between testlet length, calibration model, and testlet effects had significant effects on abias. The smallest biases were recorded for 10 item per testlet tests, followed by 3 items per testlet test while highest bias were recorded for 6 items per testlet. The *post hoc* analysis for interactions did not lead to conclusive results on the effects of changing the testlet size on bias. Number of categories and group dependence had negligible effects on abias of $f = 0.06$ each. Model and testlet factors had significant effects of $f = 0.22$ each. More negative biases were recorded for 4 category options, indicating slope underestimation of slope, while higher biases were recorded in 3 and 5 response categories, across all models.

Increasing testlet length significantly affected random errors in discriminant parameter recovery ($f = 0.18$). However, number of category options had insignificant main effects ($f = 0.06$). Smallest SEs were recorded for longer tests of 10 items per testlet while highest SEs were recorded for 3 items per testlet, all different significantly. The interaction effects between testlet length and LPD and model factors were significant although the interaction between testlet length and testlet effects was not statistically significant. The random errors were smallest in large LPD effects for shorter tests, and were highest in person independence for longer tests, suggesting that the random errors might be underestimated in the presence of person dependence effects especially when testlet based tests are shorter. The lowest random effects were recorded for models with

4 response categories items tests and they differed significantly between the SE for the 3 and 5 response categories items tests, which in turn did not differ significantly. The lowest standard errors were recorded in dual models and multilevel models accounting for group dependence effects while highest SEs were recorded in GPCM and testlet models ignoring group effects.

The effect of changing testlet sizes on total errors in discriminant parameter estimation had significant p-value but negligible Cohen f while variant number of response categories was significant ($f = 0.12$). Model and testlet factors were significant, resulting in significant interactions between testlet length and testlet effects. Total errors were lowest in item independence for longer testlet (10 items) while they were highest in large item effects for shorter tests of 3 item testlet. The effects of changing testlet length in group effects is not very clear as the pattern is not sequential. However, smallest total errors in discriminant parameter recovery were recorded for testlet and dual models controlling for LID across all test sizes and the order of arrangement was such that small errors were recorded for longer tests.

Interaction between number of categories, group and model effects was significant on RMSE with an effect size of $f = 0.13$ while four-way interaction between number of categories, calibration model, testlet and group effects were significant with medium effects ($f = 0.33$). Lowest total errors were recorded for tests with 3 categories items and did not differ significantly with errors in 4 categories per item. However, 5 categories items tests recorded significantly higher total errors. Lowest RMSE were recorded in dual and testlet models and did not differ significantly, while RMSE were higher in GPCM and multilevel models which in turn did not differ significantly.

6.3.6 Test reliability

The models were compared in terms of test reliability as the number of testlet length and category options varied. Correlations in ability parameters have been noted to

slightly increase as testlet length increased. This implied that the classical testlet reliability determined by squaring correlations between true and estimated ability parameters would have been increased by increasing testlet length. The correlations would be very low ($r < 0.7$) for models failing to account for local person effects.

Table 6.5: Test reliability for different testlets size and category options

Items	Categories	Condition	GPCM	Testlet	Multilevel	Dual	DualDP
hline 3	3	NoneNone	0.90	0.90	0.87	0.87	0.87
		NoneLarge	0.93	0.92	0.85	0.86	0.86
		LargeNone	0.86	0.90	0.79	0.85	0.85
		LargeLarge	0.88	0.90	0.78	0.84	0.84
	4	NoneNone	0.93	0.93	0.90	0.90	0.90
		NoneLarge	0.94	0.86	0.87	0.86	0.87
		LargeNone	0.88	0.91	0.80	0.86	0.86
		LargeLarge	0.92	0.92	0.82	0.87	0.88
	5	NoneNone	0.94	0.94	0.92	0.92	0.90
		NoneLarge	0.96	0.95	0.89	0.89	0.89
		LargeNone	0.90	0.95	0.85	0.91	0.90
		LargeLarge	0.93	0.95	0.82	0.88	0.89
6	3	NoneNone	0.94	0.94	0.92	0.93	0.92
		NoneLarge	0.95	0.95	0.88	0.89	0.90
		LargeNone	0.94	0.95	0.90	0.89	0.91
		LargeLarge	0.92	0.95	0.82	0.88	0.90
	4	NoneNone	0.96	0.96	0.93	0.94	0.93
		NoneLarge	0.97	0.97	0.92	0.91	0.93
		LargeNone	0.95	0.96	0.93	0.92	0.92
		LargeLarge	0.96	0.97	0.89	0.92	0.93
	5	NoneNone	0.97	0.97	0.95	0.95	0.94
		NoneLarge	0.98	0.98	0.93	0.93	0.93
		LargeNone	0.96	0.97	0.90	0.93	0.93
		LargeLarge	0.98	0.98	0.95	0.95	0.94
10	3	NoneNone	0.97	0.96	0.92	0.94	0.93
		NoneLarge	0.98	0.97	0.93	0.93	0.93
		LargeNone	0.95	0.96	0.91	0.93	0.93
		LargeLarge	0.97	0.97	0.90	0.92	0.92
	4	NoneNone	0.98	0.97	0.95	0.95	0.95
		NoneLarge	0.98	0.97	0.94	0.94	0.94
		LargeNone	0.96	0.97	0.92	0.95	0.93
		LargeLarge	0.97	0.97	0.91	0.94	0.92
	5	NoneNone	0.98	0.97	0.96	0.96	0.95
		NoneLarge	0.98	0.98	0.94	0.95	0.95
		LargeNone	0.96	0.95	0.93	0.95	0.94
		LargeLarge	0.98	0.98	0.97	0.95	0.95

Table 6.5 shows the changes in IRT test reliability as the testlet length varied. The

reliability increased with testlet size for all models across all dependence conditions. However, the reliability values were generally low in models accounting for LPD, more so in the multilevel models controlling for LPD only. Test reliability coefficients were highest for GPCM and testlet models ignoring LPD and were lowest in multilevel and dual models controlling for person effects. In addition, test reliability coefficients generally increased as the number of categories increased for all models and for all dependence conditions. Reliability coefficients were highest in person effects.

6.3.6.1 Spearman-Brown prophecy for changing test(let) size and categories

Table 6.6 illustrates changes in Spearman-Brown coefficients, giving an indication of the magnitude by which the test length for each model, for different dependence condition, has to be elongated for it to attain the reliability they would have measured in local independence conditions. As expected, Spearman-Brown prophecy coefficients in Spearman-Brown prophecy values are highest in multilevel model than dual and testlet models controlling for testlet effects. In addition, the coefficients increased with the number of categories across all models and conditions, an indication that tests with more category options have to be longer in order to attain the reliability estimated by the GPCM model than tests with fewer response category options.

6.4 Evaluation of modelling mixed items tests in LID and LPD

The study objective was to investigate the effects of changing number of response categories per item on item and ability parameter recovery, test information and test reliability as the LID and LPD changes, including binary items. Although the proposed models is for polytomous items, items with two categories are clearly not polytomous but were considered since it is not uncommon to find a test comprising of a combination of binary and polytomous items of varying response category options. The study evaluated the ability of the proposed non-parametric model's capacity to retain person and item parameters for tests with binary and polytomous items (10 four-point items,

Table 6.6: Spearman-Brown prophecy against the GPCM model

Items	Categories	Condition	Testlet	Multilevel	Dual	DualDP
3	3	NoneNone	1.03	1.29	1.26	1.34
		NoneLarge	1.15	2.19	2.15	2.16
		LargeNone	0.67	1.66	1.11	1.08
		LargeLarge	0.87	2.17	1.42	1.40
	4	NoneNone	1.02	1.40	1.37	1.48
		NoneLarge	2.80	2.49	2.78	2.34
		LargeNone	0.74	1.79	1.16	1.19
		LargeLarge	0.98	2.47	1.70	1.57
	5	NoneNone	1.01	1.47	1.34	1.74
		NoneLarge	1.30	2.90	2.97	2.97
		LargeNone	0.53	1.57	0.95	1.00
		LargeLarge	0.71	2.93	1.86	1.64
6	3	NoneNone	1.02	1.51	1.35	1.36
		NoneLarge	0.85	2.32	2.27	2.11
		LargeNone	0.75	1.61	1.82	1.55
		LargeLarge	0.57	2.51	1.53	1.30
	4	NoneNone	1.04	1.77	1.65	1.68
		NoneLarge	1.01	2.95	3.28	2.64
		LargeNone	0.67	1.24	1.58	1.43
		LargeLarge	0.74	2.57	1.98	1.92
	5	NoneNone	1.10	1.97	2.06	1.64
		NoneLarge	1.21	4.06	3.86	2.76
		LargeNone	0.77	2.93	2.00	1.46
		LargeLarge	1.00	2.39	2.46	1.95
10	3	NoneNone	1.22	2.82	1.96	2.43
		NoneLarge	1.35	3.48	3.71	3.69
		LargeNone	0.66	1.80	1.36	1.43
		LargeLarge	0.96	3.18	2.62	2.81
	4	NoneNone	1.07	2.21	2.13	2.58
		NoneLarge	1.62	3.11	3.11	3.13
		LargeNone	0.75	2.21	1.49	1.81
		LargeLarge	1.04	3.05	2.04	2.81
	5	NoneNone	1.42	2.04	2.23	2.58
		NoneLarge	1.51	3.56	3.39	2.58
		LargeNone	1.24	2.06	1.25	1.53
		LargeLarge	0.86	1.21	2.38	2.58

10 three-point items and 10 binary items) in local dependence effects. The design is a three factor factorial design with 2 local item dependence condition (0 and 1) $\times$ 2 local person dependence condition (0 and 1) $\times$ 5 models (GPCM, testlet, multilevel, parametric dual and non-parametric dual).

Table 6.7: DICs for testlets with different response category items

LID	LPD	GPCM	Testlet	Multilevel	Dual	DualDP
None	None	11360	11540	11430	11390	11370
	Large	16270	16400	16380	16330	16280
Large	None	16150	16120	16150	16130	16120
	Large	18630	18850	18920	18810	18800

6.4.1 Goodness of fit statistics

The fit statistics in Table 6.7 behaved in the same way as in polytomous items only, where the data generation models were selected as the best fitting model. The fit statistics for the non-parametric dual are lower than the parametric dual for all conditions.

Table 6.8: variance parameters for models for different categories testlets

Model	Ability	Group	Interaction	Model	Ability	Group	Interaction
NoneNone				NoneLarge			
GPCM	0.92			GPCM	0.93		
Testlet	1.28		0.09	Testlet	1.62		0.10
Multilevel	1.04	0.04		Multilevel	0.98	1.14	
Dual	1.33	0.07	0.10	Dual	1.01	0.92	0.10
DualDP	1.29*	0.09	0.12	DualDP	1.08	1.11	0.10
LargeNone				LargeLarge			
GPCM	0.95			GPCM	1.07		1.15
Testlet	1.76		0.10	Testlet	1.49		
Multilevel	1.02	0.13		Multilevel	0.51	0.83	
Dual	1.02	0.08	0.10	Dual	0.93	0.98	1.12
DualDP	0.99	0.08	0.11	DualDP	0.97	0.99	1.18

The variances in Table 6.8 behaved in similar pattern as in the polytomous items only, that is, overestimated in GPCM and testlet models in LPD and underestimated in GPCM and multilevel in the presence of LID. Absence of LID and LPD was well detected by the testlet effects models and LPD was well detected by the person effects models.

6.4.2 Parameter recovery

Table 6.9 shows the correlation in the estimation of ability, difficulty parameters for tests of mixed items testlets. The correlations were lower than when all items are

Table 6.9: Ability, threshold and discriminant parameter correlations

Parameter	Condition	GPCM	Testlet	Multilevel	Dual	Dual DP
Ability	NoneNone	0.88	0.88	0.87	0.87	0.88
	NoneLarge	0.64	0.65	0.85	0.85	0.84
	LargeNone	0.65	0.65	0.85	0.85	0.84
	LargeLarge	0.51	0.51	0.79	0.78	0.79
Difficulty	NoneNone	0.81	0.82	0.81	0.88	0.85
	NoneLarge	0.76	0.78	0.78	0.78	0.78
	LargeNone	0.87	0.87	0.87	0.83	0.85
	LargeLarge	0.87	0.85	0.85	0.83	0.85
Discriminant	NoneNone	0.51	0.52	0.51	0.58	0.55
	NoneLarge	0.46	0.48	0.48	0.48	0.48
	LargeNone	0.57	0.57	0.57	0.53	0.55
	LargeLarge	0.55	0.55	0.55	0.53	0.55

independent with polytomous items of the similar number of response options. In LID only, correlations in the ability parameter and threshold parameters were high in all models. In LPD, ability correlations were low in the independent person models. However, the discriminant parameters were generally low across models and conditions.

Table 6.10: SE, Bias and RMSE in ability estimates for testlets with different response options

Condition	GPCM		Testlet		Multilevel		Dual		Dual DP	
SD	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
NoneNone	0.42	0.01	0.63	0.02	0.83	0.02	0.93	0.02	0.88	0.02
NoneLarge	0.34	0.01	0.46	0.01	0.55	0.01	0.56	0.01	0.55	0.01
LargeNone	0.38	0.01	0.50	0.01	0.60	0.01	0.56	0.01	0.58	0.01
LargeLarge	0.35	0.01	0.42	0.01	0.43	0.01	0.44	0.01	0.43	0.01
Bias										
NoneNone	-0.26	0.03	-0.20	0.06	-0.21	0.06	-0.24	0.06	-0.27	0.07
NoneLarge	-0.23	0.04	-0.24	0.05	-0.05	0.04	-0.06	0.04	-0.07	0.04
LargeNone	-0.18	0.04	-0.11	0.06	-0.02	0.04	-0.02	0.04	-0.07	0.04
LargeLarge	0.17	0.05	0.11	0.05	0.03	0.02	0.01	0.02	0.01	0.04
RMSE										
NoneNone	1.03	0.02	1.73	0.03	1.94	0.02	2.33	0.03	2.06	0.02
NoneLarge	1.01	0.02	1.54	0.03	1.09	0.02	1.10	0.02	1.27	0.02
LargeNone	1.14	0.02	1.61	0.03	1.14	0.01	1.13	0.02	1.32	0.02
LargeLarge	1.23	0.03	1.47	0.03	0.72	0.01	0.75	0.01	1.00	2.00

According to the results in Table 6.10, standard errors and total were generally higher than their counterparts for tests with equal number of category options for all items and testlets. The GPCM reported low SE (< 0.40) in group and item effects than all

local dependence models. The ability traits measured by the GPCM and testlet models were generally biased across all conditions while the dual and multilevel models were less biased in local dependence effects. However, lowest total errors were recorded in the GPCM model when compared to local dependence models.

Table 6.11: SE, Bias and RMSE in the threshold estimates for different category options tests

	GPCM		Testlet		Multilevel		Dual		Dual DP	
Condition	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
NoneNone	0.18	0.01	0.23	0.01	0.27	0.01	0.29	0.01	0.28	0.01
NoneLarge	0.18	0.01	0.21	0.01	0.21	0.01	0.27	0.01	0.24	0.01
LargeNone	0.19	0.01	0.23	0.01	0.25	0.01	0.27	0.01	0.26	0.01
LargeLarge	0.22	0.01	0.25	0.01	0.24	0.01	0.29	0.01	0.27	0.01
Bias										
NoneNone	0.10	0.17	0.10	0.23	0.05	0.23	0.04	0.17	0.13	0.16
NoneLarge	0.07	0.20	0.06	0.23	-0.05	0.23	-0.05	0.17	-0.06	0.16
LargeNone	0.08	0.17	0.08	0.20	0.01	0.20	0.03	0.20	-0.02	0.21
LargeLarge	0.05	0.17	-0.05	0.18	0.06	0.18	0.04	0.18	0.04	0.18
RMSE										
NoneNone	0.21	0.17	0.25	0.23	0.28	0.23	0.30	0.17	0.31	0.16
NoneLarge	0.19	0.20	0.25	0.23	0.25	0.23	0.31	0.17	0.28	0.16
LargeNone	0.20	0.17	0.27	0.20	0.28	0.20	0.30	0.20	0.29	0.21
LargeLarge	0.23	0.17	0.28	0.18	0.27	0.18	0.33	0.18	0.30	0.18

According to the results in Table 6.11, the SE in thresholds were generally low in GPCM and testlet models across all dependence conditions. Bias was higher for local independence conditions for all models. The RMSE were lower in the GPCM model when compared to local dependence models.

According to the results in Table 6.12, bias in discriminant estimates were generally high for all models, implying that the models were failing to accurately retain the discrimination ability of the test items if the number of response categories per testlet was variant.

Table 6.12: SE, Bias and RMSE in the slope estimates for different category options tests

		SE		Bias		RMSE	
Condition	Model	Mean	SE	Mean	SE	Mean	SE
NoneNone	GPCM	0.72	0.12	2.46	0.40	2.62	0.41
	Testlet	0.20	0.02	0.54	0.12	0.74	0.09
	Multilevel	0.20	0.02	0.50	0.12	0.71	0.08
	Dual	0.18	0.02	0.32	0.11	0.61	0.07
	Dual DP	0.19	0.02	0.52	0.12	0.72	0.09
NoneLarge	GPCM	0.52	0.09	2.31	0.47	2.46	0.46
	Testlet	0.18	0.02	0.59	0.17	0.84	0.14
	Multilevel	0.18	0.02	0.57	0.17	0.83	0.14
	Dual	0.17	0.02	0.55	0.17	0.82	0.13
	Dual DP	0.18	0.02	0.58	0.17	0.84	0.14
LargeNone	GPCM	0.33	0.05	1.24	0.27	1.47	0.24
	Testlet	0.15	0.02	0.33	0.14	0.68	0.10
	Multilevel	0.15	0.01	0.32	0.14	0.68	0.09
	Dual	0.15	0.02	0.31	0.14	0.67	0.09
	Dual DP	0.15	0.02	0.33	0.14	0.68	0.09
LargeLarge	GPCM	0.70	0.17	2.24	0.67	2.70	0.65
	Testlet	0.18	0.03	0.43	0.21	0.97	0.15
	Multilevel	0.18	0.03	0.45	0.22	0.99	0.15
	Dual	0.18	0.03	0.41	0.21	0.96	0.14
	Dual DP	0.18	0.03	0.44	0.22	0.94	0.15

6.5 Discussion

The models maintained their ordering in term of goodness of fit as the testlet size and number of categories vary. The overestimation of ability variance by independent persons models in LPD escalates as testlet length and number of categories increased. This is probably because when persons are clustered, elongating the test result in more variant responses and this will inflate variance of ability parameter estimates. Although the effect of varying testlet length was not distinct on independent items models, it would have been expected that the underestimation of variances testlet effects be magnified by increasing the testlet length. The multilevel model underestimates variances increasingly with number of categories. This is because more category options result in higher variance in response options, probably increasing the testlet effects in the presence of LID, which will in turn result in underestimation of the ability variance.

The percent of correct identification of latent ability classes was better for longer tests

of 10 items per testlet, thus improving ability recovery prowess of the non-parametric dual model. The findings are in agreement with other studies (Cho, Cohen & Kim, 2013; Reise & Yu, 1990) that also observed longer test length and larger sample sizes improved group membership identification. This is probably because increasing the testlet length provides many response options that will make it easier to categories respondents according to their ability levels than when response options are fewer.

In the current study, ability correlations increase with test(let) length, in agreement with earlier researchers (Wang, Jiao & He, 2011; Lee & Paek, 2014). According to Wang et al. (2011), test length played the most significant role in the correlations between true and estimated ability parameters. Increasing testlet length provides more response options with which to categories respondents according to their ability levels. As a results, the rate of categorising respondents according to their true ability levels improves with testlet length. In addition, the psychometric properties tend to be better with more response categories. The correlation between true parameters and estimates increased when the number of response category increased, in agreement with the results of Lee and Paek (2014) who reported ability correlations, validity and reliability to be better when number of categories increased. They also reported a small differences for 4, 5 and 6 categories, indicating a plateau from 4 points and beyond.

Bias in ability parameter estimation decreases when testlet length increases when persons are independent or when person clustering effects are accounted for, but increase when group effects are omitted in the model. In general, average indices of bias, abias, SE and RMSE decreases as test(let) size increases. However, among all the manipulated variables, LID have the most impact on ability parameter estimation, in agreement with Wang, Jiao, and He (2011) who also observed the main effects of person clusters to account for most of the variation in the bias, SE and RMSE in the estimation of ability parameters. Bias is lowered by the ability to detect true ability groups necessitated

by provision of more response options by increasing the testlet length. This will in turn lower the random and total errors in ability estimation. However, providing more responses in LPD increases the variance of ability traits. Thus erroneous inference is escalated when the person dependence are not controlled for in estimation. The testlet effects did not significantly affect bias in ability parameters, concurring with Zhang (2010) who observed no obvious pattern between bias and testlet size.

The bias in estimation of trait levels generally decreased with increasing category options across all simulation conditions except for GPCM and multilevel models in local item dependence effects only and for the multilevel model in dual effects. This suggest that ignoring LID increases errors of estimation increasingly with number of category options. This is in agreement with Weng (2004) who propounded that item homogeneity may be a plausible explanation to differences in the effects of number of response categories among studies. However, person effects do not cause such an increase as the errors still decrease as the number of categories increase. This is probably because increasing the number of categories makes responses more variant, escalating the dependency among testlet items, thereby increasing the consequences of ignoring local item dependence effects on ability measurement. On the other hand, the more diverse responses when coupled with group dependence effects make it more easier to classify respondents according to their traits and group membership detection improves, thus the psychometric properties improve when category options are increased in group dependence effects.

The SE have been observed to decrease with increase in testlet length across all simulation conditions, and are significantly lower in GPCM and multilevel for longer tests, suggesting that the overestimation of precision of ability parameter estimation worsens as the testlet length increases. The SE were highest in dual and testlet models for shorter tests. The decrease in bias and SE when the testlet size increases imply that

the ratio of systematic errors to random errors also decrease, indicating a reduction in the overestimation ability parameter estimates precision, test information and reliability decrease with increase in test length.

The SE in proficiency estimation in the current study decrease as number of category options increase. This is probably because the variability among responses increase as the number of options increase. As a result, ability parameter estimates will be retained more precisely, and this will reduce SEs, which is a measure of deviation of the estimates about the sample mean. In local item dependence effects, this is likely to magnify the between testlet variance, thereby increasing the consequences of neglecting testlet effects, which include underestimation of the SE in trait estimation and overestimation of the precision of trait measurement, trait information and trait reliability. Similarly, total errors in ability parameter estimates decrease as the number of response categories increase.

Generally, RMSE were lower for longer tests in dual models and highest in the GPCM for longer tests group dependence effects. This implies that ignorance of group dependence effects worsens inference for longer tests as the errors of estimation by employing a wrong estimation methods will be magnified by increasing the testlet size. Studies in literature also recorded smaller RMSE (in the ability and item parameters) for larger testlet than for small testlets for all sample sizes (Wang & Wilson, 2005).

Correlation between estimated and true thresholds increased with testlet length especially when items are independent and items dependence effects are accounted for, while group effects do not seem to significantly affect threshold parameter correlations. The ordering of models according to their performance remained invariant as testlet length increases, implying that the models are robust to testlet sizes. The current results are in agreement with Jiao, Wang and He (2011) who observed that as test

length increased, the correlation coefficients increased across. In addition, DeMars (2006) reported that the mean squared correlations between true and estimated traits increased with test length. Jiao et al. (2011) also reported group effects to have little impact on item parameter recovery. In addition, correlations in difficulty parameters increase with number of response options. Neumann and Neumann (1981) observed similar findings where the Pearson's correlations generally increase with the number of category options and 7-10 categories have similar results.

From the study results, bias increased with decrease in testlet size when LID is omitted in modelling as thresholds are underestimated in shorter tests and slightly overestimated in longer tests, in agreement with earlier studies (Wang & Chen, 2005; Royal; 2017) who reported that although the estimates of item parameters were biased, the magnitude of biases were trivial in long tests. However, contrary to the current findings, DeMars (2003) noted that increasing the number of items has little effect on item parameter recovery. The increase in bias with the number of response categories concur with Neumann and Neumann (1981) who reported deviations of actual averages from the theoretical means increases as the number of response choices increase.

The SE have been observed to decrease with testlet length especially when LID effects are ignored, aligning with Kogar and Kelecioglu (2017) who noted that increasing the number of items decreased the average errors in parameter estimation. SE in threshold parameters were lowest when testlet effects are ignored and higher when they are accounted for, but generally decrease as the number of categories increase.

The total errors in item step and discriminant parameters were lower for independent items and when LID is accounted for, and decrease with testlet length. However, total errors increase with testlet length when LID effects are not accounted for. When testlet effects were not accounted for, total errors in estimation of item parameters increased

with testlet length. This is probably because when the test increases in length, the correlation among responses for items in the same testlet is enhanced and the consequences of omission in estimation are escalated. In support of this view, Weiner and Thisses, (1996) argue that the larger the testlet from which the test is constructed, the more correlation there is among items and the lower the reliability (Wainer & Thissen, 1996).

The increase in total errors with test length when LID effects are ignored tally with DeMars (2006), who noted that increasing the number of items has little impact on item parameter estimation but increases the error variance of parameter estimation. Contrary to the current observations, Lee (1997) and Wollack and Cohen (1997) found item parameter recovery to worsen as test length increased. However, they looked primarily at longer tests comprising of at least 40 items. Other studies (eg DeMars, 2006; Wang & Wilson, 2005; Cho, Cohen & Kim, 2011) reported lower RMSE for longer test(lets) than for shorter test(lets). On the other hand, Wollack et al. (2002) observed that for a fixed number of examinees, pattern of test length was not so clear as there was a drastic increase in RMSE when the items increased from 10 to 30 items but the RMSE from 20 to 30 items did not change.

In agreement with Rijmen et al. (2003), current study favours longer tests in local independence conditions and when local dependence conditions are accounted for, for more accurate results. Omission of local dependence bias parameters, increase total errors of estimation and provide misleading precision of measurement and standard errors. However, Neumann and Neumann (1981) are of the opinion that shorter tests appeared too crude and usually resulted in more favourable conditions Neumann and Neumann (1981). The study findings are in agreement with findings by Kogar and Kelecioglu (2017) who observed that the number of items and sample size most influence the discriminant parameter but less influential on the location parameter. By adding a testlet in a test, the error averages of the parameters generally increase.

Correlations in the discriminant parameters increase with the number of response options. Neumann and Neumann (1981) observed similar findings where the Pearson's correlations generally increase with the number of category options and 7-10 categories have similar results. Preston and Colman (2000) argue that scales with more category options tend to show better item discrimination than those with fewer response options. This implies that the ability of the measure to separate respondents according to their trait levels improves as the number of categories increases. This is in agreement with the test reliability that was reported to increase as the number of response options increases. This difference can be attributed to the variability in responses as the number of categories increases.

The bias in estimation of discriminant parameters were overestimated when number of options increased, in agreement with earlier researchers who noted that scales with greater number of response options tends to show better item discrimination than those with fewer response options (Lee & Paek, 2014; Preston & Colman; 2000), arguing that respondents' ability to discriminate between adjacent categories will be sharper for more category options than when options are fewer. This is probably because variances between responses tend to increase when options increase, that is, examinees are more likely to have different response patterns, which will make it easier to disaggregate them according to their proficiency levels. However, many response options may also result in adjacent alternatives being almost similar, making respondents having challenges in differentiating them.

However, the bias in the discriminant parameter were more negative for 4 category options, implying that the slope was underestimated and were higher for 3 and 5 category options. Higher biases for the testlet and GPCM models. The change in pattern of bias observed for 4 point scales could be because the 4 point scale lack the neutral

midpoint which is favoured by respondents who don not want to commit themselves. However, according to Neumann and Neumann (1981) 3 point scales are not so desired on short tests as the midpoint provides a convenient choice for hesitant respondents. They observed five point scales to be more favourable.

Random errors in discriminant parameters decreased with increase in testlet size and were smallest when person clustering effects are ignored, implying that failure to account for local person dependence inflates the estimation of precision of slope parameter. Discriminant parameter SEs were lowest for 4 category options and significantly higher for 3 and 5 options. The SE were lowest in the dual and testlet models when then falsely accounting for absent group effects. This implies that accounting for absent LPD underestimates slope SEs, thereby inflating the precision of its estimation.

The RMSE, which reflects the sampling effect and probabilistic nature of IRT models and may be viewed as the real standard errors of estimation. The findings observed for RMSE in item parameters (difficulty and discriminant) are in agreement with other studies (He & Wheadon, 2012; DeMars, 2003) that reported the mean RMSE for each category measure to generally increases with number of categories in items. The increase in mean RMSE for items with more category options reflect the notion that higher category items have more model parameters than lower category items and that any respondent will respond with only one of the category options regardless of the number of category options the item has. As a results, item with more category options will result in responses that have higher variation than lower category items. Consequently, the mean error of measurement (bias) increase with response category options. Low RMSE for the 3 category case coincides with Garcia-Cueto et al. (2002) who argue that when simulated data are used, variability is low for the case of 3 categories probably due to the use of the normal distribution for traits, the majority of scores accumulate in the central category, resulting in a lower variability.

In the current study, both the Classical Test Theory (CTT) reliability based on the mean squared correlations between true and estimated traits and IRT reliability increase as the testlet length increase, in line with other findings reported in literature (Royal, 2017; DeMars, 2006). This finding goes in line with expectation as it is more reliable to estimate ability levels of respondents based on their responses on several items than a few, as this reduces chance responses. However, contrary to these findings, Zhang (2010) observed no evident pattern on the association between test reliability and testlet size.

The results show reliability increase with number of response categories and this trend was maintained for various local dependence conditions. Similar results were recorded by Lazano et al. (2008), Lee and Paek (2014) whose studies observed an increase both reliability and validity of the rating scale as number of response categories increased. Their study concluded that the optimum number of response categories is between 4 and 7 with fewer than 4 categories resulting in lower reliability and validity and scarcely increase when the number of response categories exceed 7. Neumann and Neumann (1981) also observed an increase in test information as the number of items increased and argued that if fewer response points are used, some information might be lost although the scale might be less ambiguous to respondents, while a finely graded scale may prevent the potential loss of information, but might be beyond the examinees' ability to discriminate. However, the greater the categorisation, the more the tests needs to be lengthened in order to attain the reliability it would have attained had the items been independent. Similarly, Lazano et al. (2008) argued that the greater the categorisation, the greater the loss of information and in turn the greater the attenuation of the relationship between items.

As expected, Spearman-Brown indices decreased as the test length increased, implying

that the reliability estimates from the models controlling for local dependence become increasingly closer to the reliability measured by the GPCM as the testlet length increases. The coefficients are highest in the multilevel model because the model underestimates the posterior mean for ability variance in testlet and dual dependence effects, an aspect attributed to possibility of the ignored testlet effects averaging out part of the ability (Zhang & Jiao, 2014).

Lee (2012) argue that provision of a greater number of alternatives for respondents' choice may introduce response patterns and hence systematic errors will be increased. In line with his argument, the better reliability recorded in tests with more response category options may be due systematic method variance and not necessarily due to traits. In addition, reliability and validity will be positively associated with the amount of item variance, which tends to be greater when more scale points are used (Lee & Paek, 2014). In the case of 3 category options, majority of response are concentrated in the central category, thus resulting in decreased variability (Lazano et al., 2008), which will in-turn affect all evaluation criteria, including reliability coefficients. Thus the study results contradict findings from studies that report better performance for tests with fewer response options. However, it is not clear what fewer relate to as usually negative consequences of increasing the number of response options were reported for options more than those considered in the current study.

Although the reliability increased with increase in number of response alternatives, there was a sharp increase from 3 to 4 options and the rate of increase from 4 to 5 category options decreased. This is probably because 3 and five category options include a middle option which is usually favoured by respondents who do not want to provide committed answers, thus reducing the variability of the responses to items and hence scale reliability. Probably the results are due to the fact that in the case of 3 category options, majority of response are concentrated in the central category, thus resulting

in decreased variability (Lazano et al., 2008), which will in-turn affect all evaluation criteria, including reliability coefficients.

Some scholars believe that the benefits of having a higher number of options climax at 4 response points, while some reported a number between 4 and 7, arguing that psychometric properties do not further improve beyond 7 options. This is probably the reason why the increase in reliability from 4 to 5 category options was not as sharp as the increase from 3 to 4 response options. However, fewer response options were studied and hence the optimum number of response options favoured by the proposed model cannot be determined and could be a subject for further research. In addition to the psychometric properties, it should be born in mind that respondents prefer formats with large number of response options as this permits them to express their view points more clearly. However, related research has shown that there is a plateau on the response options, above which psychometric properties will not improve further.

Some studies have recommended fewer categories options for better psychometric properties, arguing that when more options are provided, the examinees may perceive the differences between adjacent categories to be smaller, and this can also lead to inconsistencies. However, such researchers usually use up to 9 or 10 options and thus the options considered in this study are their examples of fewer category options. Lee (2012) argue that response styles or systematic errors can be easily introduced if a higher number of category options is supplied. However, such researchers conducted their studies using operational data and such patterns in responses may not apply to simulated data. Further research is recommended using operational data.

The GPCM and multilevel models have lower SEs and higher biases while dual models have higher SEs and lower biases, resulting in all models have higher total errors than when all items have the same number of response category options. Although total

errors of estimation are above the acceptable level (> 0.40: Amil & Sahil, 2016) ability correlations in multilevel and dual models across all conditions and GPCM and testlet models when person are independent are generally high (> 0.7: Field, 2013) and hence if the study or program objective is to measure the proficiency level of respondents when items have different category options, the models can be employed without much loss of generality.

The models can be used to estimate the ability of respondents when testlets have items of different response category options as well as retain the difficulty level of the test items. However, the models are failing to recover the discrimination ability of the items. When a test is composed of heterogeneous item types, (items with varying number of response categories), a reasonable number of responses to items with highest category response options must be ensured in order to maintain the sampling error to acceptable levels. The study has clearly shown that when sampling errors associated with category measures for items with higher number of categories are generally larger than those items with low number of categories. Although total errors are higher than when response options are the same across items, models can be used to estimate traits when the number of response options per item differ.

6.6 Conclusion

In summary, when items and persons are independent or when their dependence effects are accounted for, better psychometric properties were recorded for longer tests with more category options. However, when local dependence effects are ignored, the repercussions worsen as test length and categorisation increase. LPD impacted on the accuracy of person ability estimation most while LID affected items parameters most. As shown in a number of previous studies, ignoring LID resulted in overestimation of precision of ability parameter measurement and the problem exacerbated when

testlet length increased. The negative effects of testlet length which under normal circumstances of local independence of items and persons and when local dependence is accounted for, results in improved estimation, implying that it is important that the best model be selected for estimation. This highlights the need to detect the presence of LID and LPD before estimation, making the proposed non-parametric dual model handy as it can detect local item dependence and local person dependence even when the number of groups and group membership are not known prior to estimation.

It is apparent that dual effects be accounted for in IRT trait measurement especially if longer tests with greater response options are to be used. Thresholds bias generally increase as the number of response options, more so when testlet effects are ignored. The systematic and hence total errors in the slope increase as the number of response options increased for all simulation conditions and all models except in the GPCM in local item effects only where the errors decrease as the number of response options increase. There is no significant differences noted in item parameter estimates for the parametric and non-parametric dual models.

The reliability levels increased with increase in number of categories although Spearman-Brown prophecy values indicate that tests with greater categorisation need to be increased in length with higher magnitudes in order to attain the reliability had items and person been independent, than items with fewer number of category options. However, the current study was only on assessing the behaviour of psychometric properties in local dependence effects when the number of response options are variant. The optimum number of categories and the optimum sample size and test length are subjects that require further further research. Although the psychometric properties are weaker than when the number of response options vary for items in different testlets, the proposed model can be utilised to effectively estimate the proficiency levels of respondents when items in different testlets have different response options.

The main goal of the current study was to evaluate the effects of varying response category choices on parameter recovery (variances, correlations, bias, absolute bias, standard errors and root mean square errors) and the test information and reliability. Studies in literature make diverse recommendation on the number of response categories for polytomous items with ordered response categories (Likert-type, rating-scale-type). Some researchers argue that more category options are associated with better psychometric properties such as reliability and correlations while some vie for fewer categories for better psychometric properties. Reliability estimates are positively associated with the amount of item variance, which tend to increase when more options are provided.

Chapter 7

Effects of mis-specifying the distributional properties of the parameters

7.1 Introduction

7.1.1 Effects of mis-specifying the ability parameter distribution

Parameter recovery studies have shown that there is item and person parameter estimation error when the true ability parameter distribution, $g(\theta)$ is non-normal and a normal distribution is specified (Woods & Lin, 2009; Woods & Thissen, 2006). Flexibility in model specification for $g(\theta)$ is important, and the IRT model should be able to accommodate non-normal ability parameter distributions while also maintaining accuracy in parameter estimation and a level of parsimony. Current practices in many statistical software assume normally distributed traits by default. However, there are many studies where the population of test scores may have approximately normal distributions but non-normal distributions for the latent trait, (Lord & Novick, 1968).

Generally, parsimony has always been a major consideration in Statistics and especially in psychometric modelling where involvement of many parameters is usually associated with non-identifiability issues. Therefore, flexible yet parsimonious options to parameter estimation have always been considered, including the use of models assuming normality assumption for ability traits. However, the advent of high processing statistical software is now providing room for more complicated statistical tools as less and

less time is required to attain convergence.

The most commonly used specification for $g(\theta)$ is the standard normal distribution or a normal distribution with prior distributions assigned to mean and variance. According to Woods (2006), the normal distribution is easier to work with and is likely to be a reasonable approximation for many latent variables in psychometric testing. In addition, many IRT software are based on the assumption of normality for latent traits. However, the distribution of theta is unobservable and may not be normally distributed. The non-parametric approaches to model non-normal ability distributions includes the Bock-Aitkin (Bock & Aitkin, 1981) or empirical histogram solution, Dirichlet distribution and the Dirichlet Process to estimate the distributional parameters that characterise a discrete $g(\theta)$ using latent class probability estimates. Casabianca and Lewis (2015) focused on a semi-parametric model for $g(\theta)$ via log-linear smoothing, Duncan and McEachern (2008) employed the Dirichlet Process prior for $g(\theta)$ in a 2-parameter logistic model (2PL). They concluded that corresponding ability estimates change substantially for some individuals.

Several studies have been conducted to assess the effects of mis-specifying the ability distribution by applying a normal ability distribution for non-normal ability distributions. Reise and Yu (1990), De Mars (2003), De Ayala and Sava-Bolesta (1999) and Luo (2018) simulated data from the uniform, skewed and normal distributions and assessed the effects of applying normal ability distribution on ability parameter recovery while Woods and Lin (2009) and Casabianca and Lewis (2015) simulated data from the normal, skewed and bimodal distributions and evaluated the effects of mis-specifying the distribution by using parametric models assuming the trait distributions to be normal on the recovery of latent trait model parameters. Their model treats the rating thresholds as random parameters that are subject to the mixture and has stick-breaking mixture weights that are covariate dependent allowing the rating

category to vary across items and examinees. In addition, the distribution of category thresholds vary flexibly as a function of the covariates. The studies concluded that the mis-specification of ability trait distribution compromise the recovery of person and item parameters. However, their studies were conducted on simulated data that meets the local person and item independence assumptions.

Woods (2006) argued that the extent to which the appropriate $g(\theta)$ is influenced by violations of other assumptions such as conditional independence is not known and should be evaluated. The objective of the current study was to assess the effects of mis-specification of the trait distribution in local item and person dependence effects by applying normal distribution models on traits that are non-normal. The proposed non-parametric model was weighted against standard normal trait models in the estimation of parameters for test whose true latent distribution were not necessarily normal, for varying local dependence conditions.

7.1.2 Effects of ignoring the stochastic nature of the discriminant parameters

Several distributions have been utilised as priors distributions for the discriminant parameter, including the uniform distribution, the truncated normal, the log-normal and the gamma distribution. One of the important considerations in choosing the right model is whether the item discrimination parameters should be random and vary across items or whether they should be assumed to be equally discriminating as in the Rasch type of models. The item discrimination parameter tells us how steep the slope is, or how rapidly the probability of endorsing a response category changes at the item difficulty level. The steeper, the slope, the stronger the relationship between the ability and a (correct) response. If the item is not well discriminating, the probability will not increase much with increasing levels of the trait and the item will not help measure the examinees levels on the ability continuum, making the item redundant. Thus, the parameter enables the evaluation of how well an item performs in terms of its relevance

or contribution for measuring the underlying construct targeted by the questionnaire / test, possible redundancy of the items relative to other items in the scale. An item's relationship with the underlying construct is reflected in the magnitude of the discrimination parameter. Both item relevance (discrimination) and item location (difficulty) are important features in defining the best items for a particular study population.

Generally, parsimony has always been a major consideration in statistics and especially in psychometric modelling where involvement of many parameters is usually associated with non-identifiability issues (Haberman, 2005). Therefore, flexible yet parsimonious options to parameter estimation have always been considered. Despite the highlighted importance of the item discrimination parameter in measurement tests, IRT researchers usually set the discriminant parameter invariant across all items in order to ensure identifiability of the model. Miyazaki and Hoshino (2009) proposed a semi-parametric model with a Dirichlet Process IRT in order to ensure identifiability and their model performed better than existing parametric and non-parametric models. Jiao and Zhang (2014) ensured identifiability of their dual dependence multilevel model by setting the discriminant parameter invariant across all items. On the other hand, Duncan and MacEachern set the discriminant parameter to be positive so as to make the item characteristic curve (ICC) monotonically increasing. In addition to aiding in model identifiability, models assuming equal discrimination are parsimonious and require less computational power (Embretson & Reise, 2000).

However, few studies have been conducted to assess the effects of making the slope invariant when it is indeed random. Lee (2013) fitted and compared a constrained graded response model (GRM) with equal slopes and an unconstrained GRM with random slopes using the likelihood ratio test and the Kullback-Leibler approach. In a constrained model, all items are assumed to provide the same information about the respondent's ability. In the unconstrained model, with distinct a_i, $i = 1...I$, the items

might carry different amounts of information about the proficiency. Lee reported no significant difference between the constrained and unconstrained model for a sample of 1000 respondents. Their results for 500 respondents showed that the constrained model might have some deficiencies although further analysis attributed the differences to sample size. Smyth (2015) illustrated the effects of ignoring a random slope in the graded response model. According to the likelihood traction test, the unconstrained model with random slope explained the data significantly better than the constrained model.

According to Embretson and Reise (2000), two response patterns with the same total score can have different trait levels as succeeding on highly discriminating items and failing on poorly discriminating level items lead to higher trait level estimates. According to them, it is often the case that a model with random slope is a more accurate reflection of the data generating process. In addition, the choice of constant or random slope has implications on other parameters in the models (Embretson & Reise, 2000). As a result, the current study was aimed at assessing the effects of assuming a random slope on the ability and difficulties recovery, test information and reliability as well as the ability to detect clustering effects within a test. The sample size was set at 1000 respondents guided by Lee (2013) study. The data simulated in section 4.2 (with a stochastic slope) was re-run using similar models with a constant, equal slope ($a_i = 1, i = 1...I$), resulting in a four-factor factorial design with 2 LID levels (0, 1) $\times$ 2 LPD levels (0, 1) $\times$ 2 slope status (constant, random) $\times$ 5 calibration models (GPCM, testlet, multilevel, parametric dual, dual DP). The models under comparison are shown in Table 7.1.

Table 7.1: Models for assessing effects of ignoring dual clustering in IRT modeling

Model	Description
Partial Credit Model (PCM)	Ignoring both the item and person cluster effects, constant discriminant parameters
Generalised Partial Credit Model	Ignoring both clustering effects, random discriminant parameters
Partial Credit Testlet Model	Catering for testlet effects, constant discrimination parameters
Generalised Partial Credit Testlet Model	Testlet effects, random discrimination parameters
Multilevel PCM	Catering for person clustering effects, constant discrimination parameters
Multilevel GPCM	Catering for person clustering effects, random discrimination parameters
Dual dependency PCM	Testlet and person cluster effects, constant slope
Dual dependency GPCM	Testlet and person cluster effects, random slope
Non-parametric dual dependence model 1	Testlet and person cluster effects, constant slope
Non-parametric dual dependence model 2	Testlet and person cluster effects, random slope

7.2 Study design

In the current study, some explicit distributions of the trait θ, $g(\theta)$ were assumed. Because the distribution of θ may not be normal, the proposed model relaxes this assumption and estimates the latent distribution by employing mixtures of normal distributions with components membership for the mixed distributions determined through the stick-breaking Dirichlet Process. The model was weighed against similar parametric dual, GPCM, testlet and multilevel models in terms of error variances and stability and accuracy of parameters when trait distributions were non-normal. A Markov Chain simulation study to compare parameter recovery when the ability distribution was mis-specified in local item dependence (LID) and local person dependence (LPD) effects.

The distribution of ability parameters were simulated from uniform, normal, bimodal and skewed distributions. The distributions were scaled to have a zero (0) mean and unit (1) variance (cf DeMars, 2003). The uniform distribution was in the $[-\sqrt{3}, \sqrt{3}]$

range, the skewed distribution was a transformation of a beta distribution with $\alpha = 1.25$ and $\beta = 10$, multiplied by 11.4 after subtracting 0.11 so that the mean and standard deviation would be zero (0) and one (1) respectively. The skewed distribution was then truncated on the left side at -1.23 and was positively skewed with a long tail in the right hand side. The normal ability traits were simulated from the standard normal distribution while the bimodal distribution was created as a mixture of two normal distributions with the parameters $\mu_1 = -0.705$, $\sigma_1^2 = 0.254$, $\mu_2 = 1.058$, $\sigma_2^2 = 0.254$ with mixing probabilities of 0.6 and 0.4 respectively. The data were simulated from distribution with the same mean and ability variance so as to allow easy comparison of parameters estimated to evaluate the effects of mis-specifying the ability distribution.

Manipulated factors include the latent distribution (normal, skewed, uniform, bimodal), testlet and group effects (each with none and large levels), and five calibration models (GPCM, testlet, multilevel, dual, dual DP) resulting in a fully crossed study with 135 simulation conditions. Other than the latent traits, all the other models parameters were simulated as in section 4.2. Ten (10) data sets with 1000 respondents categorised into 5 groups of 200 respondents were simulated for each condition. Testlet affects, item difficulty parameters and the item step parameters were simulated from the standard normal as in Section 3.2.2.

7.3 Results on changing the distributional properties of the ability parameters

To assess the effects of using normal trait models for traits that are non-normal in dual dependence effects, the normal trait models and the non-parametric dual model were compared in terms of goodness of fit and errors of estimation, variance recovery as well as test reliability.

Table 7.2: DIC statistics for models mis-specifying the ability parameter distribution

LID	LPD	Distribution	GPCM	Testlet	Multilevel	Dual	DualDP
None	None	Bimodal	50930	50970	50930	50960	50930
		Skewed	48200	48270	48180	48260	48200
		Uniform	52330	52420	52320	52390	52380
		Normal	49710	49710	49610	49820	49820
	Large	Bimodal	44660	44720	44630	44660	44620
		Skewed	44770	44850	44730	44830	44400
		Uniform	46730	46780	46680	46740	46710
		Normal	43990	44050	43980	44140	44100
Large	None	Bimodal	52900	43410	52880	43220	43240
		Skewed	54100	42990	54080	42890	42920
		Uniform	49730	40410	49710	40340	40360
		Normal	57600	47410	57580	47270	46220
	Large	Bimodal	44180	36640	44110	36440	36440
		Skewed	45120	37640	45110	37440	37460
		Uniform	49130	41150	49080	40990	40940
		Normal	50820	42120	50780	41960	42000

7.3.1 Goodness of fit statistics

The results in Table 7.2 show that the models maintained their performance across different ability distributions for all dependency conditions. In local independence conditions, the skewed distribution had the lowest deviance information criteria (DIC) values in all models, followed by the normal distribution. In person dependence only, the normal distribution (correctly specified) had the lowest fit statistics. However, in dual dependence effects, the normal ability distribution recorded higher DIC. Models ignoring local item dependence effects when present recorded the highest statistics across all distributions. In LID effects only, the uniform model was best fitting.

7.4 Variance recovery

According to results in Table A20 (Appendix 1), all ability variances are closer to their true values for the skewed and normal distributions although the variance for the GPCM is lower in the skewed and closer to true values in the bimodal and normal distributions. The dual models recovered the variances well across all distributions in person dependence effects while the testlet overestimated the variances across all

Table 7.3: Average correlations between true trait values and posterior means

Condition	Distribution	GPCM	Testlet	Multilevel	Dual	DualDP
NoneNone	Bimodal	0.97	0.97	0.96	0.96	0.97
	Normal	0.98	0.97	0.97	0.97	0.97
	Skewed	0.97	0.97	0.96	0.96	0.96
	Uniform	0.97	0.97	0.96	0.96	0.97
NoneLarge	Bimodal	0.70	0.72	0.95	0.95	0.91
	Normal	0.71	0.72	0.95	0.95	0.95
	Skewed	0.66	0.68	0.94	0.93	0.92
	Uniform	0.66	0.68	0.94	0.94	0.92
LargeNone	Bimodal	0.87	0.86	0.95	0.95	0.95
	Normal	0.94	0.97	0.94	0.95	0.95
	Skewed	0.86	0.78	0.93	0.95	0.95
	Uniform	0.87	0.82	0.93	0.94	0.94
LargeLarge	Bimodal	0.59	0.58	0.92	0.93	0.92
	Normal	0.69	0.70	0.94	0.95	0.94
	Skewed	0.60	0.70	0.93	0.94	0.93
	Uniform	0.65	0.70	0.94	0.95	0.91

distributions. The variances for skewed and normal distributions were underestimated more than the other models in LID effects. However, the variances did not differ much across distributions but rather across dependence conditions as in Chapter 4. The testlet and dual models detected the presence and absence of testlet effects well while the dual and multilevel detected the presence and absence of group dependence fairly well across all ability distributions.

7.5 Ability parameter recovery

The recovery of true ability parameter values across trait distributions and dependence levels were compared by using errors of estimation and correlations between true and estimated parameters. Descriptive and inferential statistics were employed for comparison.

In local independence, the GPCM and testlet models exhibited highest correlations of 0.97 for all distributions (Table 7.3). Correlations were generally high for all models in all distributions. In LPD only, correlations were slightly lower for skewed and uniform distributions more so in independent persons models. In LID only, correlations were high (> 0.85) in all distributions across models. In large dual effects, correlations were

lower in the bimodal and skewed distribution compared to corresponding values for other distributions for the same models. There isn't much difference in correlations for the non-parametric and non-parametric dual models for all dependence conditions and models.

Table A21 (Appendix 1) illustrates systematic, random and total errors in estimation of abilities using normal trait distribution models and the non-parametric trait distribution dual model. When local independence was attained, SEs were lowest in the GPCM followed by testlet model. It is difficult to draw conclusive decisions on the effects of trait distribution mis-specification on random errors based on the tabulated means as models behaved different across dependence conditions and trait distributions.

In local independence, bias were slightly higher (positive) for bimodal distribution for all models and lower (negative) for skewed and uniform distributions. However, all the means were close to zero. In LID only, the bias means were close to 0 for the bimodal and skewed distributions and slightly negative for normal and uniform distributions across all models. In dual effects, bias were slightly positive in the bimodal, normal and skewed distributions and slightly negative for the uniform distribution. However, tabulated means were not conclusive on the effects of trait distribution mis-specification as they were all close to 0 in all distributions and models.

Total errors in local independence were all acceptable (< 0.40) for all models and distributions except for the dual model in the uniform distribution where a posterior average of 0.43 was recorded. In LID, the RMSE were relatively low in all models across distribution although they were generally above 0.40. In person clustering effects, the RMSE were largest in GPCM and testlet models. However, the tabulated means were not conclusive on the effects of distribution mis-specification on total errors in ability parameter estimation. The parametric and non-parametric dual models behaved

similarly across all distributions and conditions.

7.5.0.1 ANOVA

To evaluate the effects of mis-specifying the trait distribution on the ability parameter recovery, a crossed 2 (LID conditions) $\times$ 2 (LPD conditions) $\times$ 4 (ability distributions) $\times$ 5 (calibration models) factorial design was used for inference. Univariate four-way ANOVA with RMSE, SE, bias and abias as dependent variables were conducted to evaluate whether any of the observed differences across simulation conditions are statistically significant. In addition to the Tukey *post-hoc* test, the Cohen's f was used as a measure of effect size to quantify the magnitude of the statistically significant differences.

The distribution factor did not significantly affect both bias and absolute bias in ability parameter estimation ($f = 0.0$ and $f = 0.031$ respectively), implying that bias for different ability distributions did not differ significantly. Neither the random errors ($f = 0.08$) and total errors $f = 0.00$ were significantly affected by the ability distribution, implying that using normal ability traits to estimate skewed, bimodal, uniform and ability traits did not affects the estimation of ability parameters significantly. However, the two-way interaction between group effects and ability distribution significantly affected the random errors in ability estimation with a small effect of $f = 0.11$. The *post hoc* for the interaction effects show low SE in person independence, lowest in skewed, followed by normal and highest in the uniform distribution. In group effects, SEs were lowest in the skewed and highest in the normal distribution. No significant difference was noted between non-parametric and parametric models. However, significant difference in estimate properties were noted for dependence conditions as highlighted in Chapter 4.

Table 7.4: Average correlations between true threshold values and posterior means

Condition	Distribution	GPCM	Testlet	Multilevel	Dual	DualDP
NoneNone	Bimodal	0.99	0.99	0.99	0.93	0.96
	Normal	0.99	0.99	0.99	0.96	0.96
	Skewed	0.91	0.91	0.92	0.91	0.91
	Uniform	1.00	1.00	1.00	0.94	0.93
NoneLarge	Bimodal	1.00	0.99	1.00	0.96	0.98
	Normal	0.99	0.99	0.99	0.98	0.97
	Skewed	0.97	0.96	0.97	0.97	0.97
	Uniform	1.00	1.00	1.00	0.96	0.98
LargeNone	Bimodal	0.82	0.96	0.82	0.96	0.89
	Normal	0.85	0.97	0.85	0.96	0.96
	Skewed	0.83	0.89	0.83	0.92	0.76
	Uniform	0.86	0.82	0.86	0.87	0.90
LargeLarge	Bimodal	0.89	0.96	0.88	0.98	0.97
	Normal	0.85	0.95	0.85	0.97	0.97
	Skewed	0.85	0.91	0.84	0.95	0.96
	Uniform	0.85	0.94	0.86	0.95	0.94

7.6 Threshold parameter recovery

Models were compared in their ability to retain the item difficulty levels for different ability distributions and local dependence conditions by comparing simulated values and estimated values in terms of correlations and errors of estimation.

The correlations between true and estimated step parameters when trait distributions have been incorrectly specified in local dependence effects, are shown in Table 7.4. Correlations for independent items are high in uniform and normal distributions and low in skewed distribution. In LID, correlations in GPCM and multilevel models are lower across all models. In the presents of LID, correlations are higher in the normal distribution (correctly specified) and low in mis-specified models. Correlations in the parametric and non-parametric dual models are high across all dependence conditions and ability distributions and do not differ significantly.

7.6.0.1 ANOVA

The ANOVA results show that the effect of trait ability distribution on threshold estimation bias is not statistically significant with a Cohen $f = 0.08$. However, absolute bias was significantly affected by distribution differences, $f = 0.27$, with lowest bias

being recorded in the uniform and skewed distributions followed by the normal and highest in the bimodal distributions. The two-way interaction between ability distribution and testlet effects was significant with effect size of $f = 0.15$ while three-way interaction between distribution, testlet effects and model factors was significant with an effect size of $f = 0.15$. The ability distribution factor significantly affected SEs in estimation of threshold parameters with a large effect size of $f = 0.55$. The SE were lowest in the skewed distribution for all models and highest in the dual models for the uniform and bimodal distribution in local dependence effects. Total errors in estimation of thresholds were affected by the ability distribution with an effect size of $f = 0.32$. The interactions between ability distribution and other factors were significant with small but significant effects sizes. The RMSE were lowest in the skewed distribution in person dependence effects for dual and multilevel models and highest in multilevel and GPCM models in bimodal and uniform models.

7.7 Discriminant parameter recovery

The ability of the calibration model to retain item discrimination ability for misspecified ability distributions in local dependence effects were assessed by comparing estimated and simulated slopes in terms of correlations, random, systematic and total errors of estimation.

Correlation in discriminant parameters (Table 7.5) were lower in the skewed distribution when local independence was met. In person clustering effects only, correlations were slightly low in the bimodal distribution for all models. In item effects, correlations were lower in the skewed and bimodal models. There was no major difference between the correlations for dual models.The results in Table A23 (Appendix 1) show bias, SEs and RMSE in estimation of discriminant parameters across different trait distributions. Bias was generally low in all models and distributions and slightly high in the skewed distribution for the GPCM model. In person effects only, bias were generally high in

Table 7.5: Average correlations between true slope values and posterior means

Condition	Distribution	GPCM	Testlet1	Multilevel	Dual	DualDP
NoneNone	Bimodal	0.95	0.95	0.95	0.95	0.95
	Normal	0.96	0.96	0.96	0.96	0.96
	Skewed	0.91	0.91	0.92	0.91	0.91
	Uniform	0.96	0.96	0.96	0.96	0.96
NoneLarge	Bimodal	0.94	0.94	0.94	0.93	0.94
	Normal	0.96	0.96	0.96	0.96	0.96
	Skewed	0.97	0.96	0.97	0.97	0.97
	Uniform	0.96	0.96	0.96	0.96	0.96
LargeNone	Bimodal	0.60	0.91	0.64	0.95	0.95
	Normal	0.59	0.96	0.57	0.95	0.96
	Skewed	0.33	0.89	0.33	0.92	0.92
	Uniform	0.62	0.88	0.63	0.89	0.85
LargeLarge	Bimodal	0.45	0.84	0.44	0.92	0.92
	Normal	0.47	0.93	0.44	0.96	0.95
	Skewed	0.45	0.81	0.44	0.96	0.95
	Uniform	0.51	0.91	0.51	0.93	0.92

the GPCM across distributions with lowest bias being recorded in the normal distribution. Bias in the multilevel model in person effects only were generally close to zero (0) although slightly high in the uniform distribution. Bias in models controlling for testlet effects were negative across all distributions. In dual dependence effects, bias were high across all models and highest in the bimodal distribution. In local independence, larger SEs were recorded in the skewed distribution for all models. There did not appear to be noticeable differences in LPD only while larger SE were recorded in the skewed distribution for models ignoring person effects. In dual dependence, larger SEs were reported in the bimodal distribution for the GPCM model.

7.7.0.1 ANOVA

The GLM results have shown distribution factor and its two-way interaction with group and model to have significant effects of $f = 0.10$, $f = 0.19$ and $f = 0.11$ respectively. Three-way interaction between group, model and ability distribution and four-way interaction between testlet effects, group, model and ability distribution to had significant effects of 0.13 and 0.12 respectively. Multiple comparison for main effects of ability distribution show that bias was lowest in the normal distribution and did not differ significantly from the uniform distribution, while higher in the bimodal and

skewed distributions. Highest bias were recorded for the GPCM and multilevel model in the skewed and bimodal distributions when both item and person independence were violated. The effects of differences in ability distributions on SE were significant ($f = 0.18$). Two-way interactions between ability distribution and group dependence were significant with effects of 0.15 and 0.11 respectively. However, the four-way interaction between the four independent factors was not significant ($f = 0.04$). SEs were lowest in the normal and uniform with no significant difference and significantly higher in the skewed distribution. RMSE were significantly affected by ability distribution ($f = 12$). The four-way interaction between model factors was significantly with a small effects size of 0.12. RMSE were lowest in the dual model for the normally distributed traits and highest in the GPCM and multilevel models in dual dependence for skewed, normal and bimodal distributions. No major differences were reported in the estimation errors for the parametric and non-parametric dual models for all dependence conditions and ability distributions.

7.8 Results on the distributional properties of the discriminant parameter

To see the effects of ignoring the random slope, constrained and unconstrained slope models were compared in terms of goodness of fit, their ability to recover ability, testlet and group variances as well as their estimation prowess assessed by comparing the errors incurred in retaining true simulated parameters. Descriptive and inferential statistics were employed for analysis.

7.8.1 Goodness of fit

Table 7.6 reports the goodness of fit parameters for the constrained slope and unconstrained slope models. Although the unconstrained model had an additional set of parameters to estimate, the fit statistics were slightly lower than their constrained counterparts. However, the rank ordering of the models according to their goodness of fit remained the same for constant and random slope models for all dependence

Table 7.6: Model fit statistics for random and constant slope

Slope status	Condition	GPCM	Testlet	Multilevel	Dual	Dual DP
Constant Slope	NoneNone	50010	51120	50030	50110	50090
	NoneLarge	44520	445101	44260	44480	44480
	LargeNone	57720	47720	57700	47560	47560
	LargeLarge	51190	42720	51170	42550	42490
Random Slope	NoneNone	49710	49710	49620	49820	49820
	NoneLarge	43990	44050	43980	44140	44900
	LargeNone	57600	47410	57580	47270	47300
	LargeLarge	50820	42120	50780	41960	42000

conditions.

Table 7.7: Ability, group and interaction variances for constant and random slope

	Constant slope			Random slope		
Model	Ability	Group	Interaction	Ability	Group	Interaction
NoneNone						
GPCM	1.32	-	-	0.90	-	-
Testlet GPCM	1.34	-	0.03	1.10	-	0.03
Multilevel GPCM	1.33	0.10	-	0.94	0.09	-
Dual Parametric	1.36	0.11	0.03	1.20	0.10	0.03
Dual Dirichlet	1.39	0.11	0.04	1.14	0.12	0.03
NoneLarge						
GPCM	2.74	-	-	1.60	-	-
Testlet GPCM	2.61	-	0.05	1.93	-	0.04
Multilevel GPCM	1.44	1.32	-	0.97	0.90	-
Dual Parametric	1.48	1.46	0.04	1.17	1.12	0.03
Dual Dirichlet	1.43	1.41	0.04	1.06	1.08	0.04
LargeNone						
GPCM	0.72	-	-	0.50	-	-
Testlet GPCM	2.21	-	1.98	1.44	-	1.31
Multilevel GPCM	0.63	0.15	-	0.45	0.13	-
Dual Parametric	1.33	0.17	2.51	0.89	0.14	1.67
Dual Dirichlet	1.37	0.19	2.47	0.89	0.13	1.54
LargeLarge						
GPCM	1.21	-	-	0.74	-	-
Testlet GPCM	3.07	-	2.37	1.94	-	1.50
Multilevel GPCM	0.62	0.65	-	0.43	0.48	-
Dual Parametric	1.27	3.01	2.61	0.89	1.76	1.75
Dual Dirichlet	1.32	2.41	2.39	1.06	1.76	1.64

Results in Table 7.7 compare the ability, testlet, group and interaction variances for models with constant and random slope. The results show that for all dependency conditions, models ignoring the random slope had higher variances than their counterparts. However, all the models detected the presence and absence of dependence

effects well.

7.8.2 Ability parameter recovery

Table 7.8: Average correlations between true values and estimates for ability parameter for constrained and unconstrained slope

Slope status	Condition	GPCM	Testlet	Multilevel	Dual	Dual DP
Constant Slope	NoneNone	0.97	0.97	0.97	0.96	0.95
	NoneLarge	0.72	0.71	0.95	0.95	0.95
	LargeNone	0.90	0.94	0.95	0.95	0.95
	LargeLarge	0.69	0.67	0.93	0.93	0.91
Random Slope	NoneNone	0.98	0.97	0.97	0.96	0.95
	NoneLarge	0.71	0.72	0.95	0.95	0.95
	LargeNone	0.94	0.97	0.95	0.95	0.95
	LargeLarge	0.69	0.70	0.93	0.94	0.94

From the results shown in Table 7.9, ignoring the random nature of the slope and assuming it to be constant instead did not affect the rank order of the respondents according to their proficiency on the ability continuum. Models behaved similarly with a constant and with a random slope. The correlations for parametric and on-parametric dual dependency models did not differ significantly.

According to the results in Figure 7.1, SEs are lower for models accounting for random slopes compared to their counterparts assuming constant slopes. Total errors in models ignoring random slope are higher across all dependency conditions. However, in LID+ only, systematic errors for GPCM and multilevel GPCM models are higher than their counterparts assuming a constant, unity slopes.

Figure 7.2 shows the plots of true parameter values against their respective estimates, for models with constant and random slope. It is very difficulty to determine the effects of ignoring the random slope as the graphs appear similar for all dependence conditions. However, the bands for graphs assuming a constant slope seem to be slightly wider than the models with a stochastic slope.

Figure 7.1: Random, systematic and total errors in ability parameters for constant and random slope

7.8.2.1 ANOVA

To evaluate the effects of ignoring the random slope on ability parameter recovery, a crossed 2 (LID conditions) × 2 (LPD conditions) × 2 (random and constant slope) × 5 (calibration models) factorial design was used for inference. All factors did not significantly affect bias and abias in estimation of ability parameter, including the state of the slope as their f values were close to zero (0). Except for two-way interaction between slope and model factors and three-way interaction between slope, group effects and model factors with significant effects of size $f = 0.10$ and $f = 0.16$ respectively, all the other interaction effects had effect sizes less than 0.10. The *post hoc* for the effects of interaction between group effects, model and the status of slope show lowest bias in person independence for dual and multilevel models, with models having a random slope outperforming models with constant slopes. On the other hand, bias were highest in testlet and GPCM models in LPD, random effects performing better than constant effects. The analysis on interaction between group, testlet, model effects and

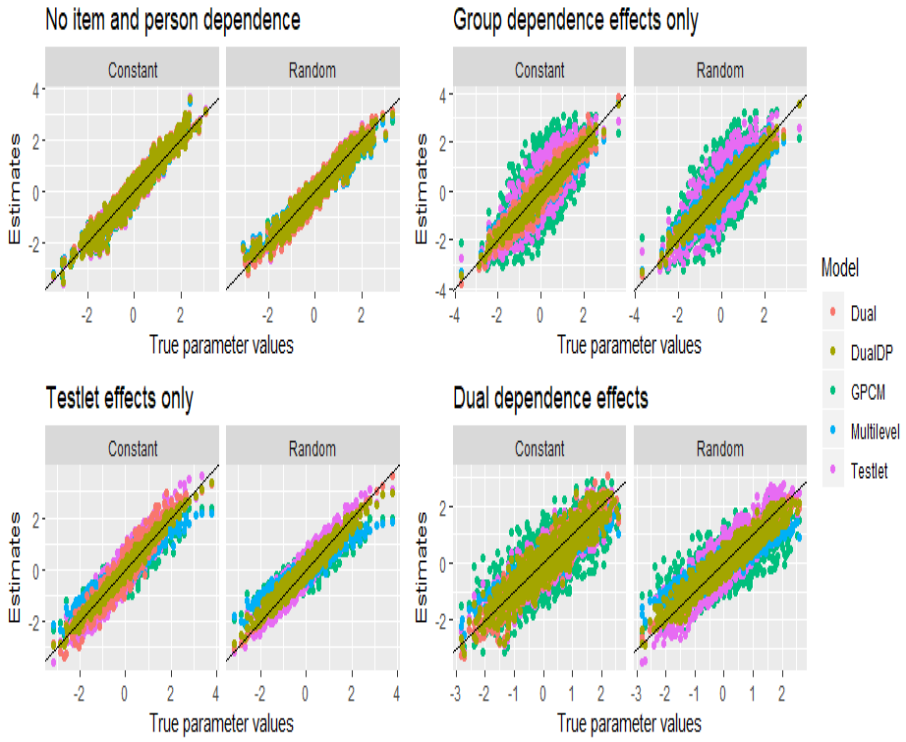


Figure 7.2: Bias in the ability parameter for constant and random slope

the status of slope had resulted in lower biases in testlet and GPCM models for local independence, again models with random slopes having lower bias. Bias were higher in dual and multilevel models, models with random slopes having lower biases.

The status of discriminant parameters significantly impacted on the random errors (SE) in the estimation of ability parameters with a large effect size of $f = 0.47$, just behind the calibration model effects of $f = 0.74$. The effects of interaction between slope status and testlet dependence was significant with large effects of size $f = 0.61$. All the other interactions between slope status and the other model factors were significant according to the p-values but Cohen's effect sizes were negligible (less than 0.10). Despite the slope status and calibration model having large effects on the SE, effects of their interaction was negligible ($f = 0.07$). The *post hoc* were done based on the p-value and not the effect sizes. The SE were significantly lower for random slopes compared to constant slopes. For the effects of interaction between slope and testlet effects, the SE were lower for random slope in the presence of large testlet

effects, followed by random slope in the absence of testlet effects. However, for the interaction between group effects and the slope status, lowest SE were recorded in the absence of group effects and for random slope followed by random slope in the absence of group dependence effects. The largest were recorded for large group effects and constant slope. The SE were lowest for the GPCM and testlet models when the slope was random, in the presence of dual dependence effects and largest in the dual model for constant slopes in the presence of dual dependence effects.

The slope status did not significantly affect the total errors (RMSE) in ability parameter estimation. However, the two, three and four-way interaction between slope status and other model factors were significant with significant effect sizes except for the interaction between slope status and testlet effects which had a negligible effect size. The RMSE were lowest for the GPCM and testlet models in the absence of LID and LPD for random slopes followed by the same for constant slopes and were highest in the GPCM and testlet models in the presence of group dependence effects for constant slopes.

7.8.3 Recovery of the threshold parameter

To assess the effects of constraining the slope when dependence effects are not accounted for, the ability of the calibration models to retain the item step parameters was compared using the correlations and estimation errors.

Table 7.9: Average correlations between true and estimated for thresholds for constrained and unconstrained slope

Constant Slope	Condition	GPCM	Testlet	Multilevel	Dual	Dual DP
	NoneNone	0.97	0.97	0.97	0.97	0.97
	NoneLarge	0.97	0.98	0.98	0.95	0.96
	LargeNone	0.85	0.96	0.85	0.94	0.94
	LargeLarge	0.84	0.94	0.84	0.93	0.93
Random Slope	NoneNone	0.99	0.99	0.99	0.96	0.97
	NoneLarge	0.99	0.99	0.99	0.98	0.97
	LargeNone	0.85	0.97	0.85	0.96	0.96
	LargeLarge	0.85	0.95	0.85	0.97	0.97

Table 7.9 compares true and estimated values correlations in threshold parameter for models with constant and stochastic slopes. For all dependency conditions, the correlations for models assuming a constant slope are lower than those for the models incorporating a random slope, implying that replacing a stochastic slope with a constant one affects the rank ordering of items according to their difficulty levels. The difference between correlations in constrained and unconstrained models is more distinct for the parametric and non-parametric, with the two models behaving similarly.

Figure 7.3: Random, systematic and total errors in the threshold parameters for constant and random slope

The results in Figure 7.3 show that ignoring a random slope increases systematic and total errors in threshold parameter estimation for all models and dependency conditions.

According to the results in Figure 7.4, there does not seem to be much differences in the plots for true values against estimates for the threshold parameter for respective

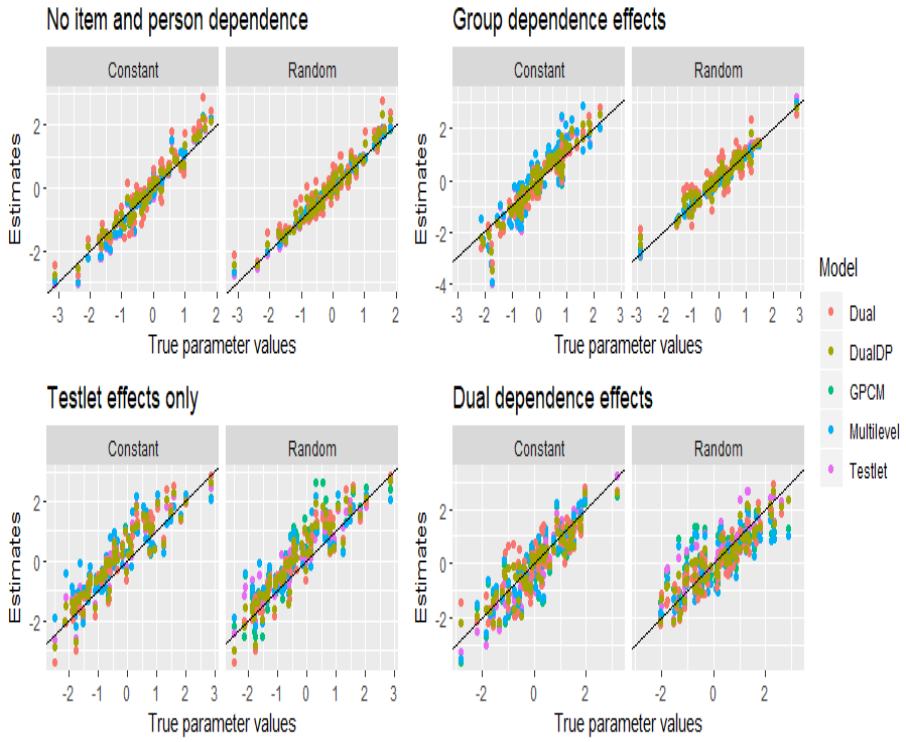


Figure 7.4: Bias in the threshold parameter for constant and random slope constrained and unconstrained models. However, models accounting for the random slope seem to have a slightly narrower band than their counterparts ignoring the random nature of the simulated slope. There does not seem to be significant differences in the results from dual and testlet models controlling for LID and similar results were recorded for independent items models.

7.8.3.1 ANOVA results

The use of a constant in place of a random slope did not significant affect bias and abias in the estimation of thresholds as both Cohen's effect f values were negligible. The interaction between slope status and testlet effects was significant with an $f = 0.12$ while three-way interaction between slope, testlet and group dependency had a small effect of size $f = 0.11$. All the other interaction effects had none to negligible effects. Bias was small for independent items condition for random slope followed by independent items for constant slope and largest for large testlet effects in unconstrained models. A similar pattern was observed for group effects except that bias were largest when large group effects were coupled with constant slope. In addition, bias was lower

in local independence and in group effects only for random slope followed by same conditions for constant slope and highest in testlet and dual dependence for random slope. Lowest biases were recorded for the GPCM and testlet model in independence or in group effects only for random slope while highest were recorded in GPCM and multilevel model in large testlet effects and for random slopes.

The effects of slope status was minimal while group and testlet effects had significant impact on threshold parameter SE. All interaction effects between status of slope and other model parameters had significant and negligible effects although they were all significant according to the p-value criteria. The SEs were lowest in LID, with the random slope having lower SEs than constant slope and the highest were recorded for large testlet effects and random slope. For group dependence effects, lowest SE were recorded in person independence, with the random slope having lower SE than constant slope. However, for group effects, SE were significantly higher when large group effects were coupled with constant slope. Lowest were recorded in GPCM and multilevel models while higher SE were recorded in the dual models in large testlet effects and higher for random slope although the difference between random and constant slopes were not statistically significant.

The slope had negligible main effects on total errors in threshold parameter recovery although the interaction between slope status and model and three-way interactions involving calibration model had significant but small effects.

7.8.4 Test reliability

There does not seem to be much difference in test reliability coefficients for models assuming a constant slope and those with a stochastic slope (Table 7.10). However, test reliability coefficients in models constraining the slope are slightly higher in of local dependence effects, with the margin of differences wider in group effects. The reliability estimates or dual models are not very different across all simulation conditions

Table 7.10: Test reliability for constant and random slope

Condition	Slope	GPCM	Testlet	Multilevel	Dual	Dual DP
NoneNone	Random	0.94	0.94	0.92	0.93	0.92
	Constant	0.94	0.94	0.92	0.92	0.92
NoneLarge	Random	0.95	0.95	0.88	0.89	0.88
	Constant	0.97	0.96	0.93	0.93	0.92
LargeNone	Random	0.94	0.95	0.90	0.89	0.88
	Constant	0.92	0.95	0.88	0.91	0.91
LargeLarge	Random	0.92	0.95	0.85	0.88	0.88
	Constant	0.94	0.96	0.86	0.89	0.89

Table 7.11: Spearman-Brown prophecy

Condition	Slope	Testlet	Multilevel	Dual	Dual DP
NoneNone	Random	0.94	1.46	1.30	1.47
	Constant	1.06	1.53	1.49	1.45
NoneLarge	Random	0.88	2.44	2.25	2.51
	Constant	1.40	2.62	2.91	3.02
LargeNone	Random	0.71	1.57	1.75	1.88
	Constant	0.56	1.50	1.17	1.13
LargeLarge	Random	0.58	2.01	1.56	1.65
	Constant	0.61	2.45	2.00	1.92

7.8.4.1 Spearman-Brown prophecy

The Spearman-Brown prophecy coefficients in Table 7.11 are slightly higher for constrained slope models when compared to their unconstrained slope counterparts. This implies that longer tests are desired for constrained models in order to increase their reliability levels to values they would have been if local independence assumptions were not violated. The reliability measures for the parametric dual model assuming known group membership and the non-parametric that estimated the group membership from the data are almost the same.

7.9 Discussion

Although a researcher may hypothesise about the shape of $g(\theta)$, for any particular θ, the distribution is not known in advance. Researches on the effect of using parametric and non-parametric methods for estimation when the ability distributions are not

normally distributed have concluded that non-parametric estimation provides greater flexibility than does the normal parametric form for $g(\theta)$. However, most of the studies were conducted in local independence conditions. The current study compared parametric and non-parametric models for estimation when ability distributions are mis-specified in local dependence effects.

High correlations in ability parameter for all models in local independence conditions, implying that mis-specification of the trait distribution did not much affect the ranking of respondents based on their traits when local independence is not violated. The effects of mis-specifying the trait distribution on the ordering of respondents according to their traits in the presence of person and dual dependence is worst in the bimodal and skewed models than the normal (correctly specified) and uniform distributions. Better correlations in the uniform distribution were also observed by Reise and Yu (1990). However, Reise and Yu (1990) recorded lower correlation of 0.90 in the normal and skewed distributions compared to the 0.97 recorded for the GPCM in the current study. Mis-specification of the ability distribution has more negative consequences when person clustering effects are ignored as lower correlations are recorded in the other distributions than when the traits are normally distributed.

In the current study, none of the main effects of factors in the model significantly affected the bias, SE and RMSE in ability parameter recovery. This is probably because the ability parameters for all distributions were simulated from zero mean and unity standard deviation conditions. However, the two-way interaction between theta distribution and group effects was significant with SE lowest in the skewed and highest in the normal theta model when local person independence is violated. However, Reise and Yu (1990) reported significantly high correlations and worst average RMSE in the uniform distribution when compared to the normal and skewed. Simulation studies have shown that results from normal IRT models can be non-trivially biased when the

true population distribution for θ is non-normal.

The correlations in the threshold parameter were highest in the uniform distribution and lowest in the skewed theta condition. The threshold bias are lower in uniform and skewed distributions for the GPCM model and higher in normal and bimodal distributions. However, contrary to the current findings, Woods (2006) reported near zero bias and lower correlations in the normal - normal case.

SE are lowest in the skewed and highest in the uniform and bimodal distributions in dual models. Total errors are lowest in the uniform distribution in person clustering effects in the dual and multilevel models and highest in multilevel and GPCM models for the bimodal and normal theta distributions. Highest correlations and lowest RMSE in the uniform true θ conditions were also reported by Reise and Yu (1990). They reported equal higher RMSE and lower correlations for the normal and skewed distributions.

The correlations between true and estimated discriminant parameter values were generally low in the skewed and bimodal distributions and higher in the uniform and normal distributions. It is expected that accuracy should be high in the normal case because there is an exact match between the true and assumed $g(\theta)$. Similar results were reported Reise and Yu (1990) who observed the uniform true theta (θ) conditions to be slightly superior on average, followed by the normal and lastly the skewed distribution.

Bias in the slope was higher in the bimodal and skewed θ and lower in normal and uniform ability distributions, in line with DeMars (2003) who concluded that the pattern of data set with a uniform trait distribution were similar to those from the standard normal distribution traits. However, like in other studies, bias did not differ significantly according to the ability distribution (DeMars, 2003; Reise & Yu, 1990). Woods

(2006) observed bias and RMSE for item parameters in the skewed-normal and non-parametric cases to be larger than those for the normal-normal. However, the bias and RMSE for the non-parametric models were lower than those for the mis-specified normal models.

The SE in slope estimation was largest in the skewed distribution for models ignoring person clustering effects and lowest in the normal and uniform for models accounting for person effects. The total errors in the discriminant estimation were lowest in the dual models for normal and uniform distributions and higher in the GPCM and multilevel models for skewed distributions. Reise and Yu (1990) also observed uniform ability parameter distribution to be superior on the estimation of RMSE in the discriminant parameter, with the normal and skewed distribution behaving similarly higher results. The current results concur with DeMars (2003) who observed that the error variance in the slope was higher for skewed distribution of the ability parameter than the normal and uniform distribution and Woods (2006) who observed lower RMSE for the normal-normal case than the skewed-normal case.

It appears that the uniform condition resulted in the worst average RMSE in the ability for local independence conditions despite it having given the best results in the estimation of the difficulty and discriminant parameters, an indication that the item parameters were best estimated in the uniform condition, in agreement with observation by Reise and Yu (1990). The higher RMSE in the uniform true θ conditions may imply that there were more extreme values estimates as compared to the other ability distributions.

Casabianca and Lewis (2015) recorded little errors in the normal case. For the negatively skewed distribution, non-parametric or semi-parametric approaches for $g(\theta)$ may be preferred over the fixed normal models as they properly model deviations from

normality in the characterisation of $g(\theta)$ thereby yielding more precise item parameter estimates. Non-parametric procedures yielded better item parameter estimates in terms of RMSE. However, the problematic part about the flexibile non-parametric approaches is that they involve estimation procedures requiring probably many additional parameters and this may lead to estimation and identifiability problems (Haberman, 2005).

The non-parametric dual model provides an approximation of the latent variable distribution without a parametric constraint and provide person and item parameter estimates form the mis-specified normal distribution. The non-parametric model was not really advantageous in the estimation of abilities in the presence of person effects compared to its parametric counterpart as expected because the non-parametric model had to detect group membership variables with measurement errors while the data generating groups are considered known *a priori* for the parametric dual model. This is probably why the current results do not tally with results in literature where non-parametric models were observed to perform better when trait distributions were not necessarily normally distributed.

The fit statistics are lower for models with a random slope probably because the model have more parameters to be estimated compared to the model with a constrained slope. However, this could simply be an indication of better model fit as results in Chapter 4 of this project had shown that the GPCM model with lowest number of parameters was selected as the best fitting model when local independence is not violated. Higher ability variances were recorded in the restricted model although both detected the absence of local dependence effects well. The higher variances reported for the constrained model are probably because the variance for the omitted discriminant parameters are incorporated into the ability parameter variance.

Employing a constant slope when it is stochastic did not affect the rank ordering of respondents according to their proficiency significantly. If the reason for analysis is to estimate respondents' ability or ranking the respondents according to their proficiency levels, then applying a constrained model to ensure identifiability does not affect the results. The difference between parametric and non-parametric dual models is not very significant for constrained and unconstrained models and they both recovered the ability and item parameters well for all dependence conditions. Ignoring the random nature of the slope has been shown to bias the ability parameter estimates when person clustering effects are not accounted for, implying that failure to account for group dependence effects has more negative consequences in the estimation of trait levels when a stochastic slope is constrained to invariance.

The lower SE recorded in GPCM and testlet models for random slope and large item dependency effects and higher SE in large testlet effects and constant slope suggest that the underestimation of precision of the ability parameter estimation (and hence the overestimation of test reliability, testlet information and model discrimination ability) was minimal when the discriminant parameter was assumed constant than when it is random. RMSE were highest in the GPCM and testlet models in the presents of group effects, an indication that negative effects of ignorance of group dependence effects are exacerbated by ignoring the random nature of the slope.

The rank ordering of items according to their step parameters was compromised when the random slope was constrained to be invariant. If the objectives of study or programme intervention includes the determination of test items that are more difficult to the examinees, then the inference might be erroneous. However, although ignorance of random slope distorts the rank ordering of items by their difficulty levels, local items independence still play a major role in item parameter estimation. The bias in threshold parameter estimation was not significantly affected by making the slope constant

although highest absolute biases were recorded in the independent items models for random slopes. This means that ignoring the random slope worsens the effects of LID on item parameter estimates.

Ignoring a random discrimination parameter significantly affected the random errors in the threshold estimation with lowest SE recorded in independent items models when slopes were random and highest in the testlet effects models for constant slopes. This implies that the overestimation of precision of difficulty levels is exacerbated by making use of the random slope, at the same time ignoring the random slope in item dependence effects increases the random errors. However, total errors which give a true reflection of the actual estimation errors in threshold parameters were not significantly affected by ignoring the random nature of the slope.

Embretson and Reise (2000) pointed that for data with reasonably appropriate equally discriminating items, then constrained models such as the Rasch type of models should be applied, but when varying item discrimination is unavoidable, consider more complex models with random slopes than to change the construct by deleting important items. The reason why the difference between the estimates from a constrained model and the unconstrained model are not of high magnitude is probably because the actual discriminant parameter was simulated from the log-normal distribution with mean and standard deviation of 0.2, which favours values close to 1.

The higher test reliability for the constrained model is because they overestimate the posterior variance for the ability parameter, which probably emanates for the fact that part of the omitted slope variance is incorporated into the ability variance, thus leading to overestimation. A simulation study by Hambleton, Johns and Rodgers (2006) concluded that positive errors associated with the item discrimination parameters using the 2PL from samples of different sizes could produce tests with substantially different

test information distribution which is different from the true distribution of the test information.

7.10 Conclusion

The non-parametric dual model enables the application of IRT models to variables that are unlikely to be normally distributed. A much better match between data and the analysis model has the potential to improve the measurement of constraints that form the foundation of psychometric research. The uniform distribution is less affected by local item and person dependence effects. The rank ordering of examinees according to their propensities parameters was not significantly affected by constraining the slope although the ordering of items according to their thresholds was affected. As a result, if the objective of study or program is to estimate the proficiency level of respondents, making item discrimination ability invariant to ensure model identifiability does not compromise the results. However it will be difficulty to rank the test items according to their difficulty levels.

The main effects of slope on bias and total errors were not significant. This is probably because the slope parameters were generated from the log-normal distribution with mean and variance of 0.2, favouring the simulation of values that are close to one (1), and hence using constant unit slopes had no significant effects. Further analysis using large and stochastic discrimination parameters is recommended. However, interaction between slope status and other model factors has shown ignorance of the stochastic nature of the slope to exacerbate the effects of LID and LPD on item and person parameters respectively. It is recommended that if local independence assumptions is violated, researchers employ stochastic slopes models if the items are assumed to show different discrimination abilities. Simulation research should be done to assess the effects of sample size, scale length (and category options) on the mis-specification of ability distribution (in the presence of local person dependence effects) as suggested

by Woods (2006).

Chapter 8

Application of model to operational data

8.1 Introduction

The sections above have dealt with the use of simulation studies to assess the effects of ignoring dual dependence as the sample size, number of response categories and test(let) size increase. However, in simulation studies, the true data generating process is known and this is not the case with operational data. In addition, the results in Chapter 6 have shown that the competing models can be used to estimate the latent traits and item parameters even for tests comprising of testlets with unequal number of category response options items. However, the testlets in the simulation study had testlets of 4, 3 and 2 category options.

The proposed model was applied to real life data on food insecurity measurement for urban households in Windhoek. Food security exists when all people always have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and preferences for an active health life (FAO, 1996). Based on this definition, food security involves the intersection of four dimensions: availability, access, utilization and stability (FAO, 2008; Haysom, 2017). It is important that tools and scales for food security measurement produce reliable results, particularly if they provide the evidence base for policy interventions (Deitchler et al, 2010; Jones et al, 2013). Furthermore, the ability to measure food insecurity and hunger accurately is crucial in

monitoring progress towards the attainment of SDG 2 which calls for ending hunger, achieving food security, improving nutrition, and promoting sustainable agriculture.

Household food access, is usually measured using three separate proxy measures, the Household Food Insecurity Access Score (HFIAS), the Household Dietary Diversity Score (HDDS) and the Month of Adequate Household Food Provision (MAHFP) score. The HFIAS is based on the idea that the experience of food insecurity causes predictable reactions and responses at the household level that can be quantified through a survey and summary score for each household (Coates et al., 2007; Jones et al., 2013). The HDDS, defined as the number of unique food groups consumed within the household in a given time period, is seen as a proxy for both the quality and quantity of food consumption (Hodinott & Yohannes, 2002; Jones et al., 2013; Ruel, 2003). According to Bilinsky and Swindale (2010), the MAHFP is a food security measure based on the total number of months in which the households had adequate food provision. Although the HFIAS, HDDS and MAHFP measurements are intended to measure the same dimension: food access, that are usually computed separately by taking summations of the household item category responses on the items intended to measure each of the component. This study attempts to come up with a single measure of household food access levels based on responses from the three sections. The sections are considered as testlets within a test, intended to measure the same trait level, household food access as a single dimension of food insecurity measurement.

The use of item response theory (IRT) methods to measure household food insecurity is not new. Other researchers have measured food insecurity levels using IRT methods in their studies. The IRT methods have also been used as validation tools to support the validity of summation based food security scales (see Charamba et al., 2019). Nord (2014) wrote a detailed technical paper on the use of IRT measures on the Food Insecurity Scale while the National Research Council (2006) illustrated the use of IRT on the

Household Food Security Survey Module (HFSSM). The United States Department of Agriculture (USDA) used the Rasch models to measure food insecurity levels and set a panel to review the use of the Rasch on the HFSSM scale. The panel suggested that alternatives should be made to the dichotomous Rasch so that models reflecting the nature of the data can be employed, including modelling polytomous items by polytomous items models and allowing the latent distributions to vary according to subgroups in the population of respondents. In addition, the panel suggested that threshold scores applied to estimates provided by the IRT models can be used to categorise households according to food insecurity levels.

The Windhoek survey data on household food insecurity was used to demonstrate the utility of the proposed dual model on estimation of real life data. The survey data consists of 32 items divided into 3 testlet. The first testlet is made up of 10 four-category response HFIAS items capturing household experience and perception on food insecurity. The second testlet has 12 binary responses HDDS items indicating whether or not households have consumed particular food groups in a given time frame, intended to capture household dietary diversity, while the third and last testlet has 12 binary responses MAHFP items capturing the months that the households did not have adequate food provisions. Although the testlets in the real data have 4 and 2 response category items, Lazano et al. (2008) noted that the reliability of scales of two category response items was better when compared to three category items and hence the two category response items are expected to be generally good.

The data was analysed in OpenBUGS and MultiBUGS using the 5 competing models. Priors and hyper-priors similar to those used for the simulated data were used in the real data analysis. The 10 constituencies /locations (see Nickanor et al., 2016) were used as manifest groups for the multilevel and parametric dual models which require that the group membership be known in advance. However, other group characterisations could possible explain the household food insecurity, for example the type housing categorised

as formal and informal, income categories, lived poverty index categories (cf Nickanor et al., 2016). As a result, the true ability groups in the population of households is not known, thus the Dirichlet Process priors within the stick-braking framework was used to determine the number of groups as well as the group memberships when the groups were considered to be latent (unknown) for the non-parametric model.

8.2 Results

The goodness of fit statistics for the food security data are shown in Table 8.1 and the proposed dual dependence model was selected as the best fitting model according to all fit statistics, superior to all the other models according to the large differences well above 9 and very high quantified evidences (Anderson, 2008). Although the AIC and BIC results are not shown, they all show the fitness of the proposed dual model to be significantly better than the competing parametric models.

Table 8.1: Quantified fit indices for food security survey data

Index	Model	Index value	Difference Δ_i	Likelihood L_i	Probability w_i
	GPCM	31110	2800	0.0	0.0
	Testlet	30950	2640	0.0	0.0
	Multilevel	31100	2790	0.0	0.0
	Dual	30920	2610	0.0	0.0
	DualDP	28310	0	1.0	1.0

The category characteristic curves (CCC) in Figure 8.1 show that the probability of endorsing an item response is monotonically non-decreasing for all models. The models have similar curvature functions although the GPCM and testlet models seem to estimate a narrower range for the respondent ability (food insecurity). The multilevel and dual models estimated higher traits while the non-parametric dual model has lower trait levels.

From the results shown in Table 8.2, the ability variances for the models are all close to unity (1), although slightly higher for the GPCM and testlet models. The group and testlet variances estimated by the dual and multilevel respectively, are more than 0.25, implying that there are small to moderate dependence effects within the food

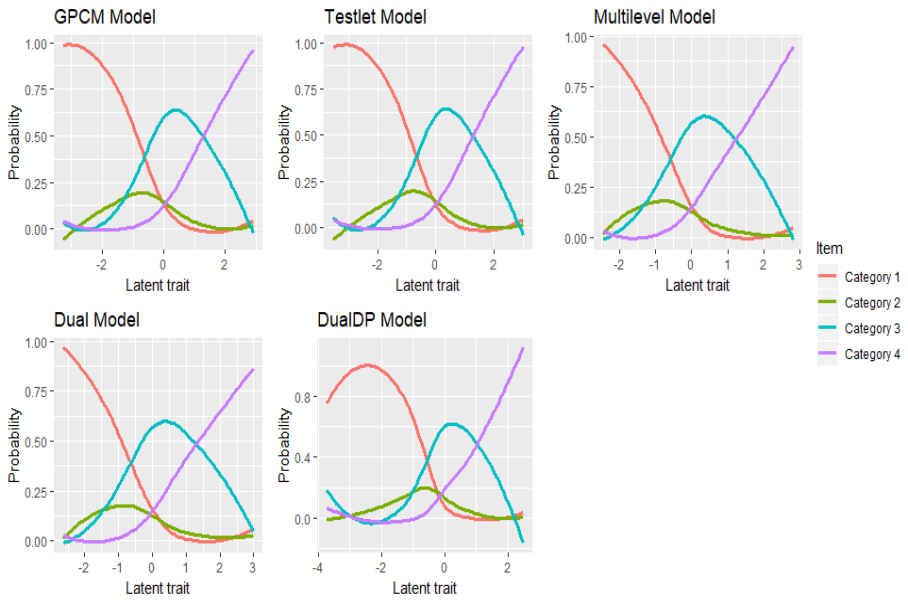


Figure 8.1: Category characteristic curves for food security survey data

Table 8.2: Ability, group and interaction variances for the food security data

Model	Ability	Group	Testlet
GPCM	1.21		
Testlet	1.27		0.38
Multilevel	0.98	0.64	
Dual	1.12	0.77	0.38
DualDP	1.16	0.81	0.41

security data for the city of Windhoek. Group variances estimated from the dual and multilevel models were close to each other. However, the interaction between testlet and group variances estimated by the testlet model was higher than the variance (effects) estimated by the dual dependence model.

Figure 8.2 shows the plots of the ability parameters for all five model. The plots are overlapping, implying that all five models managed to estimate the ability of the respondents with some degree of accuracy, in agreement with simulation results.

According to the results shown in Table 8.3, the distribution of ability (food insecurity levels) estimates for the Windhoek households are about the same for the GPCM

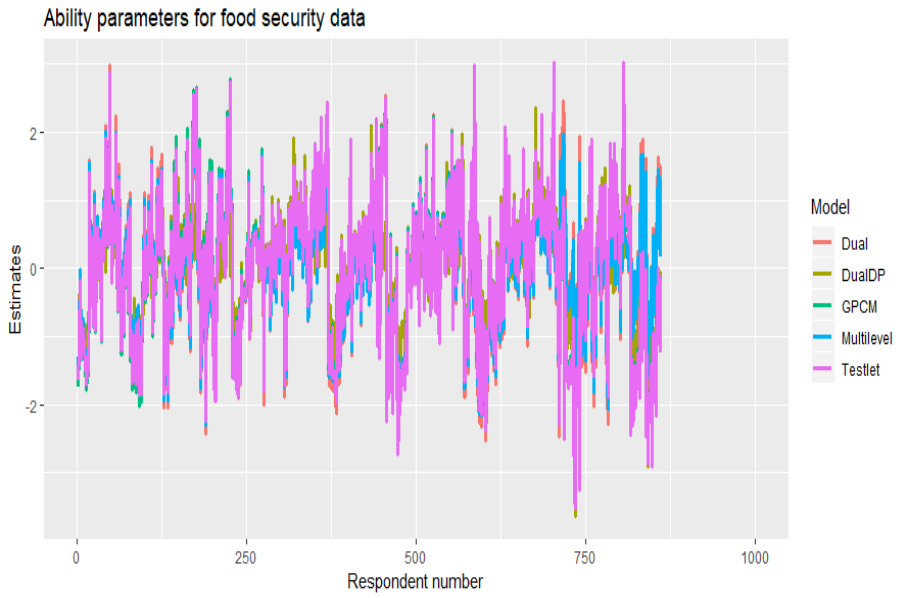


Figure 8.2: Ability parameter estimates comparison for the competing models

Table 8.3: Ability parameter estimates for food security survey data

	Minimum	Maximum	Mean	Std. Dev
GPCM	-3.28	2.92	0.00	1.06
Testlet	-3.54	3.03	-0.01	1.09
Multilevel	-2.46	2.82	0.00	0.95
Dual	-2.62	2.98	0.00	1.01
DualDP	-3.64	2.37	-0.03	0.82

and testlet models (ignoring person clustering effects), and about the same for the dual and multilevel models accounting for person clustering effects, suggesting that there are significant clustering effects within the data. The mean ability was close to zero (0) for all models although slightly negative for the proposed non-parametric dual dependence model. In addition, the range for the dual DP model are to the left of the data, implying that the food security levels generated by the dual DP model are lower than those estimated by the other parametric models. The variability for the ability parameters estimates were slightly larger for the GPCM and testlet models than the multilevel and dual DP models. Although the GPCM and testlet models do not account for person clustering effects, the parameter estimates derived from these models were closer to the estimates from the dual DP than the estimates from the dual

parametric and multilevel models.

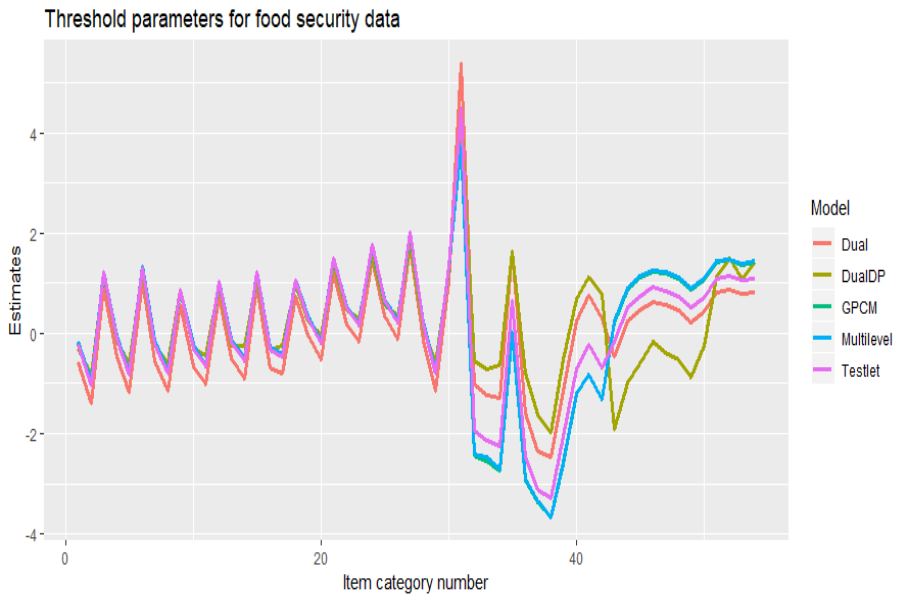


Figure 8.3: Threshold parameter estimates comparison for the five competing models

The results in Figure 8.3 show that there isn't much deviation in the difficulties as estimated by all models especially for the first testlet as all lines overlap. However, the plots differ in the last testlet although the plots for the GPCM and multilevel models overlap.

Because of the differences in number of response categories per testlet per item, the threshold / difficulty parameter was disaggregated by testlet. From the results in Table 8.4, the mean of the first threshold, transition from "never" to "rarely" was close to zero for all models although slightly lower for the dual models, implying that it was difficult to move from "never" to "rarely", thus showing monotonicity. However, the average for the second threshold was negative for all models although lowest in parametric dual and highest in the non-parametric dual, implying that it was easy to move from "rarely" to "sometimes", thereby contradicting the monotonicity assumption where higher response categories are supposed to be more difficult compared to lower response categories. However, the third threshold was higher, greater than 1 for

all models although lowest for the parametric dual model. This indicate that it was difficult to transit from "sometimes" to "often", satisfying the monotonicity assumption, that is, its only households that were higher up on the ability (food insecurity) continuum that endorsed the highest levels of food insecurity.

The mean difficulty for the second testlet measuring household dietary diversity score (HDDS) was negative for all models except for the non-parametric dual model, implying that for most of the food items were consumed by the households. For the third testlet, the difficulty parameter is positives, implying that households recorded more months of inadequate food supply when compared to months of adequate food provision. Except for the difficulty in the second testlet, the variability was lower in the dual DP model compared to all the other parametric models.

From the results in Table 8.4, the threshold parameter estimates for the polytomous items are almost the same for all models. All models recorded monotonically increasing difficulties for the three thresholds although the first thresholds are slightly higher than the second thresholds for all models. First and second thresholds lie in the medium difficulty level (-0.5 to 0.5: Barker, 2001) except for the parametric dual model that recorded the second threshold to be easy $(-2, -0.5)$, implying a higher level of disordering. However, all the models recorded the third threshold, marking the transition from sometimes to often response categories to be hard (0.5,2), implying that the often response options was only endorsed by households higher up on the food insecurity continuum. However, the difficulty parameter estimates for the last testlet (MHAFP) are lower for the non-parametric dual model.

The items in the second testlet (HDDS) were on average considered easy $(-2, -0.5)$ by the GPCM, testlet and multilevel models while both dual models reported the testlet items to be of medium difficulty level on average. Items in the third testlet (MAHFP)

Table 8.4: Threshold parameter estimates for food security survey data

	Model		N	Minimum	Maximum	Mean	Std. Deviation
Testlet 1	Threshold 1	GPCM	10	-0.30	0.64	0.05	0.34
		Testlet	10	-0.33	0.66	0.03	0.36
		Multilevel	10	-0.27	0.67	0.08	0.34
		Dual	10	-0.68	0.35	-0.31	0.37
		DualDP	10	-0.33	0.57	-0.02	0.34
	Threshold 2	GPCM	11	-1.03	1.31	-0.32	0.68
		Testlet	10	-1.03	1.32	-0.32	0.68
		Multilevel	10	-0.97	1.33	-0.28	0.67
		Dual	10	-1.41	1.03	-0.67	0.70
		DualDP	10	-0.84	1.22	-0.15	0.59
	Threshold 3	GPCM	10	0.85	1.96	1.31	0.36
		Testlet	10	0.87	2.02	1.34	0.37
		Multilevel	10	0.88	1.98	1.34	0.35
		Dual	10	0.56	1.73	1.04	0.38
		DualDP	10	0.84	1.75	1.22	0.30
Testlet2	Difficulty	GPCM	12	-3.68	3.77	-1.65	2.03
		Testlet	12	-3.27	4.49	-1.14	2.13
		Multilevel	12	-3.66	3.77	-1.63	2.02
		Dual	12	-2.45	5.39	-0.23	2.16
		DualDP	12	-1.97	4.41	0.15	1.73
Testlet 3	Difficulty	GPCM	12	0.21	1.46	1.11	0.35
		Testlet	12	-0.18	1.14	0.76	0.37
		Multilevel	12	0.24	1.48	1.14	0.34
		Dual	12	-0.47	0.87	0.49	0.37
		DualDP	12	-1.91	1.49	-0.05	1.08

were reported as hard by the testlet, GPCM and multilevel models while the parametric and non-parametric dual models reported them to be of medium difficulty levels.

From the results shown in Figure 8.4, the estimates for discriminant parameters for the first two testlet (HFIAS and HDDS) are the same, an indication that the discrimination ability for the items in these two testlets are almost the same for all model. However, the discriminant parameters for the parametric models are high and similar for the last testlet measuring the months of inadequate food supply for the Windhoek households, implying that the items housed by the testlet play the most crucial role in separating households according to their food insecurity levels. However, the discriminant parameters measured by the non-parametric dual model are lower.

From the results in Table 8.5, the statistics (average, minimum, maximum and SE)

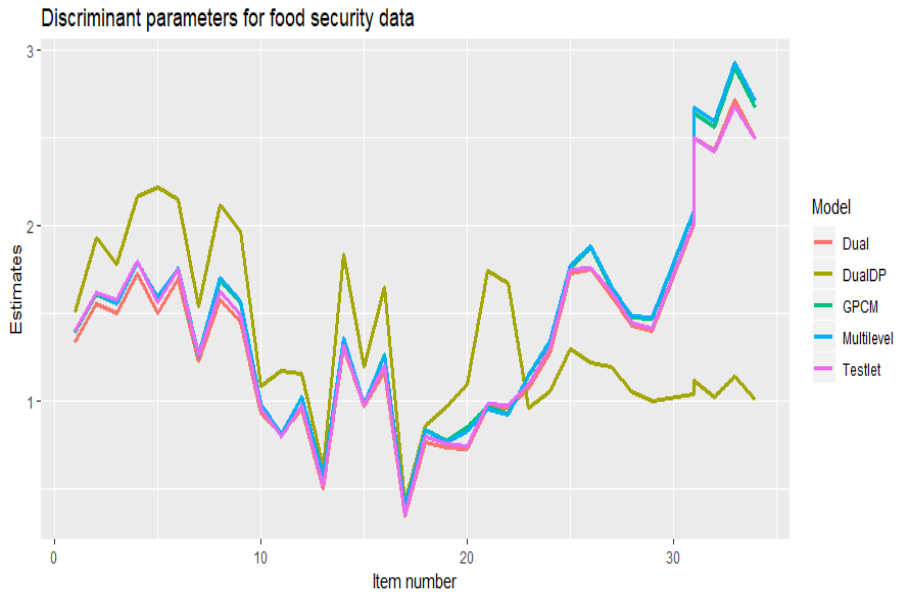


Figure 8.4: Discriminant parameter estimates comparison for the competing models

Table 8.5: Discriminant parameter estimates for food security survey data

	Minimum	Maximum	Mean	Std. Dev
GPCM	0.40	2.90	1.45	0.61
Testlet	0.35	2.67	1.41	0.57
Multilevel	0.39	2.93	1.46	0.62
Dual	0.34	2.71	1.38	0.57
DualDP	0.42	2.22	1.35	0.46

for the discriminant parameter estimates for the GPCM and multilevel (independent items models), are almost similar and the dual and testlet models controlling for testlet effects are also close. However, the non-parametric dual recorded a narrower range. The items with the lowest discrimination ability for all models was in the low (0.35 - 0.64) except for the parametric dual that recorded a slightly lower (0.35) discrimination ability (see Baker, 2001) with the maximum having very high discrimination ability ($>$ 0.70) for all models. The averages were in the high (1.35- 1.69) category for all models.

Table 8.6 gives the correlations for the ability, threshold and discriminant parameter

Table 8.6: Correlation between true-estimates for food security survey data

	Parameter	GPCM	Testlet	Multilevel	Dual	DualDP
GPCM	Ability	1	0.998	0.917	0.908	0.863
	Threshold	1	0.982	1	0.85	0.861
	Discriminant	1	0.997	1	0.998	0.187
Testlet	Ability		1	0.917	0.911	0.869
	Threshold		1	0.981	0.926	0.791
	Discriminant		1	0.996	0.999	0.236
Multilevel	Ability			1	0.998	0.773
	Threshold			1	0.848	0.679
	Discriminant			1	0.998	0.182
Dual	Ability				1	0.774
	Threshold				1	0.885
	Discriminant				1	0.207
DualDP	Ability					1
	Threshold					1
	Discriminant					1

estimates from the five competing models. There was almost perfect correlation between the ability parameter estimates from the GPCM and testlet models and almost perfect correlation between the ability parameters for the dual and multilevel models. However, the correlation were high (greater than 0.7) for all models when compared to the proposed dual dependence model although they were slightly lower for the dual and multilevel models. This implies that all the models estimated the food security levels fairly well.

8.3 Comparison with other food security measures

The food security categories obtained from the IRT models were compared to the HFIAS, HDDS and months of inadequate food supply as determined by the FANTA computations. There is no standard way of categorising the IRT categories, the HDDS and MAHFP. Suggestions by the National Research Counsil on the use of thresholds from the IRT model as cut-offs for categorising households was employed in the current study. The HDDS was categorised in a way that households that consumed 1-3 food groups were considered to be severely food insecure, those that consumed 4-6 food items moderately food insecure, 7-8 as mildly food insecure and households that consumed 10-12 food groups as food secure. The categorisation was tailor-made for the

Table 8.7: Comparison of IRT and FANTA food security measures

Model	Food secure	Mildly food insecure	Moderately food insecure	Severely food insecure
GPCM	34.0	4.8	51.6	9.6
Testlet	37.4	4.1	49.7	8.8
Multilevel	36.5	3.1	52.5	7.9
Dual	25.2	21.1	43.6	10.1
DualDP	19.3	7.0	49.8	23.9
HFIAS	16.4	3.4	13.2	67.1
HDDS	0.7	7.3	26.7	48.3
MAHFP*	67.3	19.1	3.4	10.1

current study following Oldewage-Theron and Kruder (2008) who categorised results from a test comprising of 9 food groups consumed in seven days as low for 0-3, medium for 4-6 and high for 7-9, with the food variety score categorised as 1-3, 4-6, 7-9 and 10-12. The months of inadequate household food supply were also categorised using similar categories where households that had inadequate supply for 10-12 months were considered to be severely food insecure, 7-9 months as moderately food insecure, 4-5 as mildly food insecure and households that experienced food shortages for 0-3 months as food secure. The results are shown in Table 8.7, together with the categorisation form te Household Food Insecurity Access Prevalence (HFIAP: see Nickanor et al., 2016; Charamba et al., 2019).

8.4 Discussion

Despite observing significant testlet, group and interaction effects, a great agreement was observed in the person and item parameter estimates across all models as shown by high correlations between the item and person parameter estimates derived by the models ($r > 0.7$). Considering the identifiability and convergence attained in the models, this implies that the models are measuring the same trait and it can be assumed that they estimated the food security levels for the Windhoek urban households without much loss of generality. However, the correlations with the non-parametric dual

model estimates were low, and lowest values were observed in the discriminant parameter correlations.

Correlations between models controlling for dual dependencies were very high (0.998) while correlations for independent persons and models controlling for person dependence effects were lower (0.91). Similar results were observed by Jiao and Zhang (2014) when they applied the proposed multilevel testlet model accounting for dual dependence effects on PISA data and observed similar correlation of 0.91 between models ignoring person clustering effects and models accounting for person clustering effects. The difference between correlation coefficients in the ability parameter for models controlling for person clustering effects could be because group effects estimated by multilevel and dual models were significant (> 0.25). These results are similar to results from the simulation study in group dependence effects. The simulation studies have shown that dual and testlet models accurately detected the absence of testlet $\times$ group interaction while the dual and multilevel models corrected detected the presence of group dependence effects. This implies that the significant group and testlet $\times$ group interaction effects detected in the Windhoek food insecurity survey data were reliable.

The estimated testlet effects were relatively small (0.38) but person clustering effects were large, therefore, it made sense that the non-parametric model provided better model fit as the number of groups and group membership was not absolutely known *a priori*.

The correlation between the ability estimates from the parametric dual and multilevel models are high, probably because they control for the same group memberships. The correlations were lower when the multilevel and dual models were correlated with the GPCM and testlet models not controlling for dual effects and the two were highly correlated. However, the correlations between the ability estimates from the multilevel

and parametric dual against the non-parametric dual models were lowest, an indication that they were probably controlling for different group membership.

The non-parametric dual model measure the Windhoek households to be more food insecure than the other models and recorded the lowest standard deviation, implying that the food security measurements from the models are more precise. The difference between ability parameter estimates from the dual parametric and multilevel models when compared to the non-parametric dual model may mean that the manifest groups provided for the multilevel and dual model are not the same as latent group estimated by the non-parametric model through the stick-breaking Dirichlet Process. Such differences could have adverse consequences in high stake decision making and policy formulation and should be minimised by making sure that appropriate group memberships are used for estimation.

Although the results for parametric and non-parametric dual models are almost similar for the slope parameter, the range for estimates derived from the non-parametric model is narrower. Considering the results from the simulation study where the non-parametric model recovered the group memberships, the results from the model can be assumed to be more reliable in the current study where the actual membership are not absolutely known. The discrimination parameters estimates from the non-parametric dual model were less correlated with estimates from the parametric models. This suggest that the ability clusters have an influence on the estimation of the discriminant parameter, in agreement with results from the simulation study.

The monotonicity assumption for the HFIAS section and somewhat disordered thresholds concur with the results reported by Charamba, Nickanor and Kazembe (2019) in their study to validate the use of the summation based HFIAS scale to measure the food insecurity levels of Windhoek households, implying that treating the HFIAS as

a stand alone test and as a testlet within the same test together with the HDDS and the HFIAS produced similar results. This might be due to higher levels of food insecurity in Windhoek, where households experience food insecurity "sometimes" more than "rarely". In addition, the FANTA HFIAP measurements had the lowest proportion (3.4%) of respondents in the mildly food secure category. However, according to Andrich (2006, 2011), disordered thresholds are in indication of item dependence effects, hence supporting the existence of testlets in the operational data.

The GPCM, testlet and multilevel models considered the HDDS items to be easier and less discriminating households according to their food security levels while the same treated the MAHFP items to be difficulty and highly discriminating. On the other hand, the dial models considered both testlets to have difficulty level in the medium category with the same discrimination ability although the parametric dual results are closer to the results observed from the other parametric models. Considering the number of respondents that reported to have accessed adequate food for 10 to 12 months (67.3%), it is only reasonable that the dual models reported the testlet items to be of medium difficulty ability compared to the testlet, GPCM and multilevel models that rendered the testlet items to be hard, making the results from the dual models more reliable.

8.5 Conclusion

The study findings have shown that group and item clustering effects do exist in operational data. In general, the real data analysis revealed results similar to findings from the simulation study. Ability parameters were affected by person clustering effects as correlations between models controlling for group effects were high. The discrimination parameters were affected by the presence of local person dependence effects. The evidence from simulation studies and the small fit statistic and standard errors for the non-parametric outperforms other models in ability and items parameter estimation

due to its ability to detect unknown group membership. This may mean that the group membership data supplied for the multilevel and parametric dual model are not the actual groups that define the respondent abilities (food insecurity). However, when the modal group estimated by the non-parametric model were used for the multilevel and dual models, similar results were obtained. As a result, it is recommended that when groups and group membership are not absolutely known *a priori*, researchers should detect them using the non-parametric model despite its computational challenges.

Although this is probably the first time to come up with a single food insecurity measure from the three sections that are usually treated as separate measures of the same food insecurity aspect "food access".The comparison between individual FANTA food security measures and the combined IRT food access measures suggest that aggregation of food access measures by treating them as testlets attempting to measure the same dimension "food access", can be achievable and the current study can be set as a basis for further research, deliberations, debates and suggestions.

The use of thresholds need to be examined further. In addition, the interchanging of the first and second thresholds that were disordered is a subjects for suggestions, debates and further research. Other ways of setting cut-offs can be examined. For example, Johnson (2004) suggested mathematical ways of determining cut-off points for classifying respondents into one of the three food security classes and / or estimate the proportion of individuals in the population that fall into each of these classes.

Chapter 9

Conclusions and Recommendations

9.1 Conclusion

Ability estimation is one of the main components in IRT framework. However, the simultaneous impact of examinee characteristic differences and dependence of the test items has been ignored. Regardless, in educational assessment, clinical trials and biological assays, it is common to observe respondents from different groups characterised by gender, social levels, ethnicity, schools grades. In addition, test items can be categorised into item bundles measuring a common, sub-stimuli of the main stimuli being measured by the whole test. Item and person group heterogeneity can reflect different behaviours and respondent groups may be manifest, where they are known and can be supplied as model data, or may not be known *a priori*. It is important to take cognisance of item and person groups into account when modelling person proficiency and item parameters.

To address the item and person dependence and inadequacy of the normal ability parameters in the presence of clustering effects, this study presented a non-parametric multilevel polytomous dual dependence level which involves the modification of standard IRT models to account for both item and person clustering effects simultaneous, complementing early researches (Zhang, 2010; Jiao & Zhang, 2014; Jiao et al., 2010; Jiao, Kamata & Binici, 2010) to get more accurate estimation of ability and item parameters. However, the model differs from Jiao and Zhang (2014) model in that it assumes the number of groups and group membership to be latent, and have to be

inferred from the data. The model also assumes interaction between group and testlet effects.

The model presented, which assume latent group membership estimated using the stick-breaking Dirichlet Process prior, was compared with a similar dual model which assumes group membership to be known *a priori*, and models ignoring either testlet effects, group effects or both, as the dependence levels, sample size, test(let) length, response category options and the ability distribution varied. The current research compliments earlier research on person and item dependency in that it allows for detection, explicit modelling and evaluation of the extent of group and item effects when the number of groups and group membership are not known in advance. The model explicitly separates the group specific ability parameter from the person specific ability parameter, assuming interaction between person and item clusters where respondents in one group can respond similarly to question in one testlet.

The study was conducted using simulated data to allow for manipulation of item and person data to give different examination conditions and to assess goodness of the models in retaining simulated true values. In addition, the models were compared on real life data with unknown parameters and dependence conditions. The simulation study, the DIC correctly identified the true data generation model probably because the DIC does not always favour the model with the highest number of parameters as the complexity is compensated for in computation of the index. To that effect, the GPCM was selected as the best fitting model in local independence conditions despite the models having the fewest number of parameters.

The results from the non-parametric dual model were close to those from the dual parametric model assuming group membership known *a priori*, with the differences affected by the ability to effectively detect group membership. The proposed model recovered

the group membership fairly well, with the recovery of group membership aided by increasing sample size and testlet length and compromised by the increase in person dependence levels. Item parameters were not compromised as testlets were known in advance. In addition, the proposed model performed well for operational data where the actual group membership was not definitely known in advance, resulting in more reliable results. The independent persons models overestimated the ability parameter variance in the presence of LPD while the independent items models underestimated the ability parameter variance, increasingly so as the dependence levels increase. Interaction and group variances were accurately estimated, enabling the proposed model to be able to not just detect the group membership and number of latent classes, and group membership, but to be able to detect the absence and presence of person and item dependence effects.

The ability parameter was generally well recovered by the dual and multilevel models in person and dual dependence effects, implying LPD has more negative effects on ability parameters than LID. The average bias in item and person parameters were not much affected by LID, LPD and other factors as the means were constrained to zero (0) for identifiability purposes. In addition, regardless of the magnitude of deviation from the mean for data set observations, they usually cancel out to give a zero average. However, the absolute bias differed significantly and bias in the person and item parameter estimates were more for models ignoring LPD and LID respectively. SE and RMSE were affected by LID and LPD, leading to the conclusion that ignoring dual dependence does not affect the average bias but will affect the model efficiency. In general, in dual dependence effects, the dual models retained true ability parameter variances, item and person parameter estimates. The estimation accuracy measured by SE, RMSE and bias decrease as group dependency increase.

In general, simulation results show that estimation bias was not very much affected

by the model factor although the absolute bias, standard errors and root mean square errors were in some circumstances affected. The results also show that sample size has a significant impact on estimation in the presence of dual dependence effects as the overestimation of ability parameters increased with sample size. Contrary to expectation, the bias in ability parameters estimates (and hence total errors) for ignored dual dependence effects increases with sample size, implying that the discrepancy between true and estimated values increases with sample size. In addition, the degree of overestimation of test reliability increases with sample size.

The results have shown that increasing the testlet length when group effects were not accounted for negatively impact on the ability parameter estimation when person dependence effects were not accounted for (that is, higher bias, RMSE, overestimation of ability variance and overestimation of test reliability). This is probably because increasing testlet length in the presence of group effects and interaction between group and testlet variance increases the number of correlated responses, thus aggravating the negative consequences of ignoring group effects on ability parameter measurement. On the other hand, elongating the testlet in the presence of local item dependence resulted in negative consequences on item parameter estimation. This is because increasing the testlet length increased the number of relating items, hence the effects of dependence are strengthened, and these have been seen to impact more on item than group recovery in addition to underestimation of standard errors of ability measurement and overestimation of test reliability and test information both of which have been noted to increase with increase in test length when dependence conditions were not accounted for.

Estimation of ability parameters generally improves when response options were increased, except when item dependence conditions were ignored. This is probably because increasing the number of response options did not increase the number of

responses but increases the variation in responses. When responses are more variant,it becomes easier to disaggregate respondents according to their trait levels/groups. However, variant response options might increase the dependence of items within a testlet, thus escalating the negative consequences of ignoring testlet effects in item response theory measurement. In the same arena, ignorance of testlet effects resulted in erroneous estimation of item parameters increasingly with the number of response options. Moreover, the overestimation of test reliability was exacerbated by increasing response category options. According to Lee and Paek (2014), a scale with a smaller number of response categories can be as effective as a scale with greater number of response points as long as it uses more. However, although this can be inferred from the current results where increasing both test length and number of response options led to improved psychometric properties, further studies are recommended where both factors are manipulated and the interaction effects evaluated.

Ignoring the random slope did not compromise the ranking of respondents according to traits, implying that the constrained model can be effectively used if the program objective is to rank respondents according to traits. However, ability bias was escalated by constraining the slope when person clustering effects were not accounted for. The rank ordering of items according to their difficulty levels was compromised by ignoring the random slope. In addition, constraining the slope worsened overestimation of test reliability due to the increase in ability variances, which is probably a result of part of the ignored slope variance being incorporated into the ability variance.

In summary, if items are reasonably equally discriminating, the Rasch type of models that assume equal discrimination ability, not only because constraining the slope aids identifiability but also because the models with fewer parameters are parsimonious. However, when varying item discrimination is unavoidable, then complex models with

random slope parameters should be considered rather than dropping some items especially when the local independence assumptions has been violated.

The calibration models under comparison retained the rank ordering of respondents and items according to their traits and difficulty levels respectively (correlations > 0.7) although the correlations in the discrimination parameters were low. However, the estimates from the testlet and GPCM models were biased even for local independence conditions. The total errors were higher than their counterparts for testlets of items with equal number of response category options and were greater than 0.4 across all models. However, although further research needs to be undertaken on ways to reduce the estimation errors, they provide a plausible way of estimation when items are mixed, that is they provide a way of coming up with a single food security measurement from dietary diversity scores, months of adequate food security and household food insecurity assessment score items which have binary and polytomous items.

In accordance with the simulation study and the operational data results, several findings can be inferred about ignoring dual dependence effects in IRT modelling. The results suggest that failure to account for dual dependence effects bias ability and item parameter estimates, overestimates precision of ability parameter measurement, test information and test reliability. The recovery of item and person parameters for local independence conditions and when local dependence is accounted for, improve with testlet length and sample size and the number of category options. The recovery of class membership was best when sample sizes were larger and testlets were longer. However, when local dependence effects are ignored, the psychometric properties deteriorate increasingly with increase in sample size, test length and number of category options. This implies that when faced with dual dependence in testing, test developers can manipulate the 3 conditions, (sample size, testlet length and number of response options per item) in order to optimize their psychometric scales.

Clustering effects cannot be ignored when groups such as schools or classrooms are randomly assigned to research status as ignoring both item and person clustering effects can lead to overestimation of the precision levels. However, the results have also shown that introducing spurious sources of clustering can lead to serious underestimates of precision levels and inflation of measurement errors in ability and item parameter estimation. As a result, the test makers must clearly specify the sources of clustering under their designs or the underlying assumptions and ascertain the presence of item or person dependence effects before applying the best model for estimation. This makes the proposed non-parametric model handy as it has the ability not only to detect the presence of groups in a sample but also to cluster respondents based on the observed data.

An important concern for implementation of the proposed non-parametric dual dependence model is that the MCMC estimation requires substantial computing to complete a single replication and the increase is a direct function of the model complexity. Since multiple starting points are necessary for detecting convergence and label switching, the amount of computing time required may increase abruptly. The current model required 24 hours on a 3.0GHz with a 16GB of ram and 4 cores to run 10 000 iterations of a single replication. In order to implement such a model in operational situations, much faster software and algorithms would be required such as might be obtained using some partial means rather than full information. However, the advent of advanced versions of BUGS software such as MultiBUGS and High performance computing panels can make the operational implementation of the model doable.

The benefits of the proposed model should be weighted against model complexity in data analysis as the aim of model selection is not just maximum fit to the data set but also the model that best captures the characteristics or trends underlying the cognitive

process of interest. In sum, the best model matches the purpose of study and can explain the important features of the actual data without adding unnecessary complexity and loss of generality. The model need to be used only for complicated situations where both items and persons are clustered, and if the number of groups and group membership is not known prior to analysis, otherwise simpler models can be employed, or the parametric counterpart can be used if the group memberships are already known. If there is no such structure in the data, the model does not need to be considered.

In addition, spuriously accounting for absent clustering effects can inflate the measurement errors in parameter estimation. When the data was simulated with independent items, independent person or both, adding extra parameters where they were not needed over-capitalised on chance and increased the error variance. To avoid the errors introduced by using an inappropriate, over-fitting model when items and persons are independent, the dual model should be used first as a method of detecting the presence and absence of these dependence effects before the appropriate model can be selected for estimation. The dependence effects should be deemed significant enough to warrant modelling if they are more than 0.25.

9.2 Recommendations for further studies

It has been discovered that in the presence of LID and LPD, models accounting for dual dependency effects provide a better model-data harmony while the models accounting for group effects when present provide better ability parameter estimates than models ignoring these effects. On the other hand, models accounting for LID provide better item parameter estimates when compared to models ignoring these effects. It is suggested that one has to ascertain the presence of item or person dependence effects before applying the best model for estimation. Researchers believe that testlet effects of 0.25 and above are considered significant and hence are worthy accounting for in

measurement. However, the current study only investigated a small range of testlet effects (0, 0.25 & 1). Larger values of LID are worthy further investigations.

There are psychometric tests that incorporate a combination of independent items and items that are nested within testlets. Such tests were not considered in the current study and further studies may be conducted to evaluate the performance of the proposed dual dependence model in parameter estimation and detection of ability groups in the presence of person clustering effects.

It would be nice to see some future work examining parameter recovery for models under a varying number of response categories for varying sample sizes and varying testlet sizes. This will also be in line with Sahin and Amil (2016) who argued that rather than sample size and test length, the combination of these two variables is important. In addition,the effects of increasing quadrature points should be studied and the effects of priors analysed. Priors might be estimated from factor analysis as suggested by Reise and Yu (1990). Further more, simulation studies could be done to assess the effects of changing sample size, number of response category options, testlet length on the effects of mis-specifying the true ability distribution and ignoring the random slope in the presence of local dependence effects. Different distributions could be suggested for the discriminant parameter and the effects of ignoring the random slope can then e assessed for such distributions.

Since the effectiveness of the proposed model is inherent in its ability to detect the latent classes which worsened when the number of groups increased especially for smaller samples, the development and utilisation of techniques to improve group membership detection is recommended. The current study assumes the class membership to be independent and non-overlapping, that is, each examinee belongs to one latent class. However, there might be situations where one class membership overlap and this might

be subject for further research.

The WinBUGS code for this study was not designed to prevent label switching. Usually restrictions are set on parameters of likelihood functions to minimise label switching, based on empirical justification. This study did not consider such constraints either in the simulation study or operational data. In addition, the current study examined the use of Markov Chain Monte Carlo (MCMC) simulation methods for the non-parametric model, other estimation methods such as the maximum likelihood methods can be explored.

The current study was based on the assumption of no differential item functioning (DIF), that is, the items are interpreted by respondents similarly with the same difficulty level but it is rather the ability levels that varies across latent groups. Exploration could be extended to model DIF and differential testlet functioning.

9.3 Limitations of study

The proposed model requires high computing power, with a cluster of parallel linked computers as it can take several hours before the estimated models can converge. The research is less likely to have such cluster of linked computers and hence it might take time before the proposed models can converge and produce results. Because of the computational performance required for the models to converge, the number of replicates simulated for each condition was limited to 10. In addition, researcher ended up combining some simulation conditions aimed at addressing different objectives in order to minimise the number of data sets and hence the number of models to be estimated.

One of the objectives of the current study was to come up with a software to estimate the proposed models. However, this objective was not achieved in the current study as only WinBUGS codes were prepared.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, *22(1)*, 47-76.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547-573.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215-222). Springer, New York, NY.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of educational statistics*, *17(3)*, 251-269.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*(6A), 3099-3132.

Anderson, D. R. (2008). Model based inference in the life sciences: A primer on evidence. New York, NY: Springer.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics*, 403-415.

Andrich, D. (2006). Item discrimination and Rasch-Andrich thresholds revisited. *Rasch Measurement Transactions*, *20*(2), 1055-1057.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert review of pharmacoeconomics & outcomes research*, *11*(5), 571-585.

Ansari, A., & Iyengar, R. (2006). Semiparametric Thurstonian models for recurrent choices: A Bayesian analysis. *Psychometrika*, *71*(4), 631.

Azevedo, C. L., Andrade, D. F., & Fox, J. P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational statistics & data analysis*, *56*(2), 4399-4412.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling*, *52*(3), 313.

Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing*, *15*(1), 71-87.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica*, *37*(1), 85-104.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Binici, S. (2007). *Random-effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods*. The Florida State University.

Birnbaum, A., Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability. Addison-Wesley, Reading, MA*.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, *1*(2), 353-355.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, *1*(2), 121-143.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(2), 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Mislevy, R. J. (1989). A hierarchical item response model for educational testing. In *Multilevel analysis of educational data* (pp. 57-74). Academic

Press.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In *Handbook of modern item response theory* (pp. 433-448). Springer, New York, NY.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153-168.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434-455.

Brooks, S. P., & Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, *8*(4), 319-335.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical models: Applications and data analysis methods.

Bush, C. A., & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, *83*(2), 275-285.

Casabianca, J. M., & Lewis, C. (2015). IRT Item Parameter Recovery With Marginal Maximum Likelihood Estimation Using Loglinear Smoothing Models. *Journal of Educational and Behavioral Statistics*, *40*(6), 547-578.

Cassella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Chang, Y., & Wang, J. (2010). Examining testlet effects on the PIRLS 2006 assessment. In *4th IEA International Research Conference, Gothenburg, Sweden.*

Charamba, V., Nickanor, N. & Kazembe, L. N. (2019). HCP Discussion Paper No. 37: Validation of the HCP Survey Tool for Measuring Urban Food Insecurity: An Item Response Theory Approach

Cho, S. J., Cohen, A. S., & Kim, S. H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*(2), 278-306.

Cho, S. J., Cohen, A. S., & Kim, S. H. (2014). A mixture group bifactor model for

binary responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 375-395.

Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, *1*(2), 114-142.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Congdon, P. (2003). A multivariate model for spatio-temporal health outcomes with an application to suicide mortality. *Geographical Analysis*, *36*(3), 234-258.

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational statistics & data analysis*, *50*(2), 346-357.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied psychological measurement*, *23*(1), 3-19.

De Boeck, P., & Wilson, M. (2004). Explanatory item response models: a generalized linear and nonlinear approach. *Psychometrika*, *71*(4), 787-787.

De Boeck, P., & Wilson, M. (Eds.). (2013). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of marketing research*, *45*(1), 104-115.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied psychological measurement*, *27*(4), 275-288.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of educational measurement*, *43*(2), 145-168.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Chapman and Hall/CRC.

Draper, D. (2008). Bayesian multilevel analysis and MCMC. In *Handbook of multilevel analysis* (pp. 77-139). Springer, New York, NY.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41-66.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*(1), 39-61.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175-186.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: L.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, *90*(430), 577-588.

Escobar, M. D., & West, M. (1998). Computing nonparametric hierarchical models. In *Practical nonparametric and semiparametric Bayesian statistics* (pp. 1-22). Springer, New York, NY.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209-230.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271-288.

Fujimoto, K. A., & Karabatsos, G. (2014). Dependent Dirichlet process rating model. *Applied Psychological Measurement*, *38*(3), 217-228.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). *Model determination using predictive distributions with implementation via sampling-based methods* (No. TR-462). STANFORD UNIV CA DEPT OF STATISTICS.

Gelman, A., & Hill, J. (2007). Data analysis using regression and hierarchical/multilevel models. *New York, NY: Cambridge.*

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Goldman, S. H., & Raju, N. S. (1986). Recovery of one-and two-parameter logistic item parameters: An empirical study. *Educational and psychological measurement*, *46*(1), 11-21.

Goudie, R. J. B., Turner, R. M., De Angelis, D. & Thomas, A. (2020) MultiBUGS: A parallel implementation of the BUGS modelling framework for faster Bayesian inference. Journal of Statistical Software, *95(7).* doi:10.18637/jss.v095.i07

Grilli, L., & Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, *28*(1), 31-44

Grilli, L., & Rampichini, C. (2009). *Measurement error in multilevel models with sample cluster means.* Electronic Working Papers 6/2009, Department of Statistics-University of Florence.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2-3), 57-63.

He, Q., & Wheadon, C. (2013). The effect of sample size on item parameter estimation for the partial credit model. *International Journal of Quantitative Research in Education*, *1*(3), 297-315.

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*(4), 973-985.

Ip, E. H., Smits, D. J., & De Boeck, P. (2009). Locally dependent linear logistic test

model with person covariates. *Applied Psychological Measurement*, *33*(7), 555-569.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161-173.

Ishwaran, H., & James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics*, *11*(3), 508-532.

Ishwaran, H., & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, *87*(2), 371-390.

Ishwaran, H., & Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, *12*(3), 941-963.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 50-67.

Jiao, H., & Zhang, Y. (2014). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 65-83.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82-100.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement*, *50*(2), 186-203.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kinderman, R. P., & Snell, L. (1980). Introduction to Markov Random Fields.

Kogar, E. Y., & Kelecioglu, H. (2017). Examination of Different Item Response Theory Models on Tests Composed of Testlets. *Journal of Education and Learning*, *6*(4), 113-126.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational*

and Psychological Measurement, *61*(**??**), 958-975.

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, *32*(7), 663-673.

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325-337.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, *28*(25), 3049-3067.

Luo, Y. (2018). Parameter Recovery with Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation for the Generalized Partial Credit Model. , (), https://arxiv.org/p

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(3), 149-174.

McLachlan, G., & Peel, D. (2000). Mixtures of factor analyzers. In *In Proceedings of the Seventeenth International Conference on Machine Learning.*

Mislevy, R. J. (1987). Chapter 6: Recent Developments in Item Response Theory with Implications for Teacher Certification. *Review of research in education*, *14*(1), 239-275.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*(4), 661-679.

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*(3), 375-393.

Molenaar, I. W. (1995). Estimation of item parameters. In *Rasch models* (pp. 39-51). Springer, New York, NY.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30.

Muthukumarana, P. S. (2010). *Bayesian methods and applications using WinBUGS* (Doctoral dissertation, Science: Department of Statistics and Actuarial Science).

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture

models. *Journal of computational and graphical statistics*, *9*(2), 249-265.

Nickanor, L., Kazembe, L., Crush, J. & Wagner, J. (2016). The Supermarket Revolution and Food Security in Namibia. AFSUN Urban Food Security Series No. 26, Cape Town.

Ohlssen, D. I., Sharples, L. D., & Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in medicine*, *26*(9), 2088-2112.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral Statistics*, *24*(2), 146-178.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of educational and behavioral statistics*, *24*(4), 342-366.

Prenovost KM, Fihn SD, Maciejewski ML, Nelson K, Vijan S, et al. (2018) Using item response theory with health system data to identify latent groups of patients with multiple health conditions. PLOS ONE *13*(11). https://doi.org/10.1371/journal.pone.0206915

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and computing*, *10*(4), 339-348.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self–reported criminal behavior. *Sociological methodology*, *33*(1), 169-211.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation*, *20*.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). Springer, New York, NY.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of educational Measurement*, *27*(2), 133-144.

Rijmen, F. (2008). Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, *48*(2), 659-666.

Rijmen, F., De Boeck, P., & Leuven, K. U. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*(3), 271-285.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological methods*, *8*(2), 185.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349-359.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75-92.

Sahin, A., & Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory and Practice*, *17*(1n), 321-335.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Santos, J. R., Azevedo, C. L., & Bolfarine, H. (2013). A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework. Journal of Applied Statistics, 40(10), 2129-2149.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica*, 639-650.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Smyth, R. (2015). Item Response Theory for Polytomous Items.
https://www.uwo.ca/fhs/tc/labs/12.PolytomousIRT.pdf

Sperrin, M., Jaki, T., & Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, *20*(3), 357-366.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583-639.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS Version 1.4. *MRC Biostatistics Unit, Cambridge, UK*.

Stanke, L., & Bulut, O. (2019). Explanatory Item Response Models for Polytomous Item Responses. *International Journal of Assessment Tools in Education*, *6*(2), 259-278.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795-809.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of applied measurement*, *5*(1), 48-61.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27-51.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247-260.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 51-74.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217-221.

Verbeke, G., & Lesaffre, E. (1996). *Large sample properties of the maximum likelihood*

estimators in linear mixed models with misspecified random-effects distributions. Technical Report 1996.1, Catholic University of Leuven, Biostatistical Centre for Clinical Trials, Leuven.

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383-401.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157-86.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement*, *24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*(1), 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning definitions and detection. *ETS Research Report Series*, *1991*(1), i-42.

Wang, S., Jiao, H., & He, W. (2011). Effect of Person Cluster on Accuracy of Ability Estimation of Computerized Adaptive Testing in K-12 Education Assessment. *Online Submission.*

Wang, S., Jiao, H., Jin, Y., & Thum, Y. M. (2010). Investigating Effect of Ignoring Hierarchical Data Structures on Accuracy of Vertical Scaling Using Mixed-Effects Rasch Model. *Online Submission.*

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*(1), 109-128.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. Journal of Machine Learning Research, 14(Mar), 867-897.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(Dec), 3571-3594.

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories?. *Assessment*, *21*(6), 765-774.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, *22*(2), 144-152.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied psychological measurement*, *26*(3), 339-352.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281.

Woods, C. & Lin, N. (2009). Item Response Theory With Estimation of the Latent Density Using Davidian Curves. Applied Psychological Measurement - APPL PSYCHOL MEAS. 33. 102-117. 10.1177/0146621608319512.

Wright, R. E. (1995). Logistic regression.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.

Yoes, M. (1995). An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model. *Saint Paul, MN: Assessment Systems Corporation.*

Zhang, O., Shen, L., & Cannady, M. (2010). Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations (Doctoral dissertation, University of Florida).

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589-600.

Appendix A: Additional Tables of Results

Table A1: True-estimated ability correlations for changing group and sample sizes

Sample size	Groups	LID	LPD	GPCM	Testlet1	Multilevel	Dual	DualDP
400	1	None	None	0.97	0.97	0.96	0.97	0.96
		Large	None	0.91	0.87	0.94	0.95	0.90
	5	None	Large	0.73	0.77	0.97	0.96	0.94
		Large	Large	0.61	0.78	0.92	0.95	0.87
	20	None	Large	0.69	0.71	0.95	0.95	0.87
		Large	Large	0.75	0.78	0.93	0.95	0.85
	40	None	Large	0.72	0.73	0.92	0.82	0.77
		Large	Large	0.69	0.76	0.90	0.82	0.79
1000	1	None	None	0.97	0.97	0.97	0.97	0.98
		Large	None	0.86	0.76	0.94	0.95	0.96
	5	None	Large	0.70	0.75	0.95	0.95	0.90
		Large	Large	0.77	0.78	0.96	0.96	0.93
	20	None	None	0.80	0.82	0.96	0.96	0.97
		Large	Large	0.60	0.63	0.94	0.94	0.86
	40	None	Large	0.71	0.72	0.95	0.75	0.75
		Large	Large	0.66	0.70	0.94	0.75	0.74
2000	1	None	None	0.97	0.97	0.97	0.97	0.97
		Large	None	0.89	0.88	0.96	0.97	0.97
	5	None	Large	0.66	0.71	0.96	0.96	0.88
		Large	Large	0.57	0.60	0.94	0.96	0.87
	20	None	Large	0.72	0.78	0.96	0.96	0.83
		Large	Large	0.72	0.76	0.95	0.96	0.84
	40	None	Large	0.68	0.70	0.95	0.95	0.74
		Large	Large	0.69	0.77	0.98	0.84	0.74

Table A2: Quantified evidence in model comparison for first replication for effects of ignoring dual dependence

	AIC				BIC				DIC			
Model	Value	KL Δ_i	Likelihood L_i	Probability w_i	Value	KL Δ_i	Likelihood L_i	Probability w_i	Value	KL Δ_i	Likelihood L_i	Probability w_i
NoneNone												
GPCM	48720	0	1.0	0.5	49140	0	1.0	0.5	49710	100	0.0	0.0
Testlet GPCM	48920	200	0.0	0.0	49310	200	0.0	0.0	50000 3	90	0.0	0.0
Multilevel GPCM	48720	0	1.0	0.5	49110	0	1.0	0.5	49610	0	1.0	1.0
Dual Parametric	48750	30	0.0	0.0	49150	40	0.0	0.0	49820	210	0.0	0.0
Dual Dirichlet												
NoneSmall												
GPCM	50880	4500	0.0	0.0	51280	4510	0	0.0	51770	4330	0.0	0.0
Testlet GPCM	46450	70	0.0	0.0	46840	70	0.0	0.0	47520	80	0.0	0.0
Multilevel GPCM	50880	4500	0.0	0.0	51270	4500	0	0.0	51770	4330	0.0	0.0
Dual Parametric	46380	0	1.0	0.999999	46770	0	1	0.9999997	47440	0	1.0	1.0
Dual Dirichlet												
NoneLarge												
GPCM	43380	490	0.0	0.0	43770	490	0.0	0.0	44350	370	0.0	0.0
Testlet GPCM	43500	610	0.0	0.0	43890	610	0.0	0.0	44650	670	0.0	0.0
Multilevel GPCM	43110	220	0.0	0.0	43500	220	0.0	0.0	43980	0	1.0	1.0
Dual Parametric	43080	190	0.0	0.0	43470	190	0.0	0.0	44140	160	0.0	0.0
Dual Dirichlet	42890	0	1.0	1.0	43280	0	1.0	1.0				
SmallNone												
GPCM	49710	40	0.0	0.0	50140	40	0.0	0.0000	50650	10	0.0067	0.0067
Testlet GPCM	49710	0	1.0	0.4983	50100	0	1.0	0.3333	50710	70	0.0	0.0
Multilevel GPCM	49760	50	0.0	0.0	50150	50	0.0	0.0000	50640	0	1.0	1.0
Dual Parametric	49710	0	1.0	0.4983	50100	0	1.0	0.3333	50710	70	0.0	0.0
Dual Dirichlet	49720	10	0.0067	0.0034	50100	0	1.0	0.3333	50720	80	0.0	0.0
SmallSmall												
GPCM	45420	3210	0.0	0.0	45810	3200	0.0	0.0	46480	3200	0.0	0.0
Testlet GPCM	42310	380	0.0	0.0	42700	370	0.0	0.0	43420	500	0.0	0.0
Multilevel GPCM	45390	3180	0.0	0.0	45790	3180	0.0	0.0	46270	2990	0.0	0.0
Dual Parametric	42210	0	1.0	1.0	42610	0	1.0	1.0	43280	0	1.0	1.0
Dual Dirichlet	42800	590	0.0	0.0	43200	590	0.0	1.0				

Table A3: Quantified evidence in model comparison for first replication for effects of ignoring dual dependence

	AIC				BIC				DIC			
Model	Index	KL Δ_i	Likelihood L_i	Probability w_i	Index	KLΔ_i	Likelihood L_i	Probability w_i	Index	KL Δ_i	Likelihood L_i	Probability w_i
SmallLarge												
GPCM	45330	3790	0.0	0.0	45730	3800	0.0	0.0	46340	3750	0.0	0.0
Testlet GPCM	41650	550	0.0	0.0	42040	550	0.0	0.0	42720	640	0.0	0.0
Multilevel GPCM	45120	3580	0.0	0.0	45510	3580	0.0	0.0	45990	3400	0.0	0.0
Dual Parametric	41540	0	1.0	1.0	41930	0	1.0	1.0	42590	0	1.0	1.0
Dual Dirichlet	43150	1610	0.0	0.0	43540	1610	0.0	0.0				
LargeNone												
GPCM	56760	10570	0.0	0.0	57160	10580	0.0	0.0	57690	10420	0.0	0.0
Testlet GPCM	46780	590	0.0	0.0	47180	600	0.0	0.0	47890	620	0.0	0.0
Multilevel GPCM	56700	10510	0.0	0.0	57090	10510	0.0	0.0	57580	10310	0.0	0.0
Dual Parametric	46190	0	1.0	1.0	46580	0	1.0	1.0	47270	0	1.0	1.0
Dual Dirichlet	46420	230	0.0	0.0	48810	2230	0.0	0.0				
LargeSmall												
GPCM	45750	9380	0.0	0.0	46140	9380	0.0	0.0	46720	9300	0.0	0.0
Testlet GPCM	36510	910	0.0	0.0	36900	920	0.0	0.0	37650	1140	0.0	0.0
Multilevel GPCM	45620	9250	0.0	0.0	46010	9250	0.0	0.0	46480	9060	0.0	0.0
Dual Parametric	36370	0	1.0	1.0	36760	0	1.0	1.0	37420	0	1.0	1.0
Dual Dirichlet												
LargeLarge												
GPCM	49960	9070	0.0	0.0	50350	9060	0.0	0.0	51040	9080	0.0	0.0
Testlet GPCM	41020	1000	0.0	0.0	41120	990	0.0	0.0	42120	1200	0.0	0.0
Multilevel GPCM	49910	9020	0.0	0.0	50310	9020	0.0	0.0	50780	8820	0.0	0.0
Dual Parametric	40890	0	1.0	1.0	41290	0	1.0	1.0	41960	0	1.0	1.0
Dual Dirichlet												

Table A4: Mean estimates of ability, group and interaction variances for 10 replicates

Model	Ability	Group	Interaction	Model	Ability	Group	Interaction
SmallNone				NoneNone			
GPCM	0.81	-	-	GPCM	0.90	-	-
Testlet GPCM	1.06	-	0.52	Testlet GPCM	1.04	-	0.03
Multilevel GPCM	0.82	0.11	-	Multilevel GPCM	0.94	0.09	-
Dual Parametric	1.04	0.12	0.66	Dual Parametric	1.20	0.10	0.03
Dual Dirichlet	1.06	0.12	0.56	Dual Dirichlet	1.14	0.12	0.03
SmallSmall				NoneSmall			
GPCM	0.84	-	-	GPCM	0.99	-	-
Testlet GPCM	1.16	-	0.42	Testlet GPCM	1.21	-	0.03
Multilevel GPCM	0.75	0.28	-	Multilevel GPCM	1.01	0.19	-
Dual Parametric	1.01	0.30	0.51	Dual Parametric	1.20	0.19	0.03
Dual Dirichlet	1.07	0.53	0.47	Dual Dirichlet	1.11	0.32	0.04
SmallLarge				NoneLarge			
GPCM	1.16	-	-	GPCM	1.60	-	-
Testlet GPCM	1.43	-	0.53	Testlet GPCM	1.93	-	0.04
Multilevel GPCM	0.62	0.79	-	Multilevel GPCM	0.97	0.90	-
Dual Parametric	0.89	1.16	0.56	Dual Parametric	1.17	1.12	0.03
Dual Dirichlet	0.99	0.96	0.61	Dual Dirichlet	0.99	0.96	0.61
LargeSmall				LargeNone			
GPCM	0.43	-	-	GPCM	0.50	-	-
Testlet GPCM	1.42	-	1.60	Testlet GPCM	1.44	-	1.31
Multilevel GPCM	0.40	0.24	-	Multilevel GPCM	0.45	0.14	-
Dual Parametric	1.02	0.32	2.06	Dual Parametric	0.89	0.14	1.67
Dual Dirichlet	1.04	1.33	1.98	Dual Dirichlet	0.89	0.13	1.57
LargeLarge							
GPCM	0.74	-	-				
Testlet GPCM	1.94	-	1.50				
Multilevel GPCM	0.58	0.95	-				
Dual Parametric	0.89	1.76	1.75				
Dual Dirichlet	1.05	1.76	1.64				

Table A5: Quantified evidence in model selection for the first replication for 400, 1000 and 2000 respondents with no dependency

	400 respondents				1000 respondents				2000 respondents			
Model	Index value	KL Δ_i	Likelihood L_i	Probability w_i	Index value	KL Δ_i	Likelihood L_i	Probability w_i	Index value	KL Δ_i	Likelihood L_i	Probability w_i
NoneNone												
GPCM	21160	10	0.0	0.0	49710	100	0.0	0.0	105300	0	1.0	1.0
Testlet GPCM	21230	80	0.0	0.0	49700	90	0.0	0.0	106500	1200	0.0	0.0
Multilevel GPCM	21150	0	1.0	1.0	49610	0	1.0	1.0	106400	1100	0.0	0.0
Dual	21230	80	0.0	0.0	49700	90	0.0	0.0	106400	1100	0.0	0.0
DualDP												
LargeLarge												
GPCM	19540	3700	0.0	0.0	50820	8860	0.0	0.0	90720	17520	0.0	0.0
Testlet GPCM	15970	130	0.0	0.0	42120	160	0.0	0.0	73640	440	0.0	0.0
Multilevel GPCM	19460	3620	0.0	0.0	50780	8820	0.0	0.0	90570	17370	0.0	0.0
Dual Parametric	15840	0	1.0	1.0	41960	0	1.0	1.0	73200	0	1.0	1.0
DualDP												

Table A6: Ability, Group and interaction variances for 400, 1000 and 2000 respondents

	400 respondents			1000 respondents			2000 respondents		
Model	Ability	Group	Interaction	Ability	Group	Interaction	Ability	Group	Interaction
NoneNone									
GPCM	1.17	-	0.90	-	-	0.85	-	-	
Testlet GPCM	1.25	-	0.04	1.10	-	0.03	1.05	-	0.02
Multilevel GPCM	1.11	0.15	-	0.94	0.09	-	0.95	0.07	-
Dual Parametric	1.26	0.12	0.04	1.20	0.10	0.03	1.14	0.08	0.02
Dual Dirichlet	1.16	0.14	0.05	1.14	0.12	0.03	1.14	0.09	0.02
NoneSmall									
GPCM	1.26	-	-	1.10	-	-	1.05	-	-
Testlet GPCM	1.26	-	0.04	1.21	-	0.03	1.20	-	0.02
Multilevel GPCM	0.99	0.38	-	1.01	0.29	-	0.99	0.26	-
Dual Parametric	1.11	0.38	0.04	1.12	0.29	0.03	1.12	0.30	0.02
Dual Dirichlet	1.09	0.41	0.04	1.11	0.32	0.04	1.09	0.28	0.03
NoneLarge									
GPCM	1.40	-	-	1.60	-	-	2.23	-	-
Testlet GPCM	1.53	-	0.05	1.93	-	0.04	1.35	-	0.05
Multilevel GPCM	0.93	0.71	-	0.97	0.90	-	1.11	1.29	-
Dual Parametric	1.06	0.65	0.05	1.17	1.12	0.03	1.33	1.50	0.02
Dual Dirichlet	1.04	0.74	0.05	1.06	1.08	0.04	1.23	1.54	0.03
SmallNone									
GPCM	0.83	-	-	0.81	-	-	0.78	-	-
Testlet GPCM	1.09	-	0.03	1.06	-	0.53	1.01	-	0.38
Multilevel GPCM	0.82	0.14	-	0.82	0.11	-	0.83	0.09	-
Dual Parametric	1.05	0.23	0.34	1.04	0.12	0.66	1.02	0.07	0.47
Dual Dirichlet	1.10	0.64	0.39	1.06	0.12	0.56	1.01	0.09	0.51
SmallSmall									
GPCM	1.17	-	-	0.84	-	-	1.01	-	-
Testlet GPCM	1.36	-	0.03	1.16	-	0.41	1.38	-	0.36
Multilevel GPCM	0.85	0.42	-	0.75	0.43	-	0.81	0.44	-
Dual Parametric	1.16	0.62	0.34	1.03	0.50	0.51	1.01	0.70	0.45
Dual Dirichlet	1.14	0.64	0.39	1.07	0.53	0.47	1.11	0.77	0.51

Table A7: Ability, Group, Testlet and interaction variances for 400, 1000 and 2000 respondents

	400 respondents			1000 respondents			2000 respondents		
Model	Ability	Group	Interaction	Ability	Group	Interaction	Ability	Group	Interaction
SmallLarge									
GPCM	1.29	-	-	1.16	-	-	2.14	-	-
Testlet GPCM	1.57	-	0.34	1.43	-	0.53	2.26	-	0.54
Multilevel GPCM	0.70	0.67	-	0.62	0.79	-	0.73	1.47	-
Dual Parametric	0.97	1.04	0.38	0.89	1.16	0.56	0.90	1.97	0.47
Dual Dirichlet	1.02	1.02	0.39	0.99	0.96	0.61	0.99	1.82	0.61
LargeNone GPCM	0.45	-	-	0.50	-	-	0.65	-	-
Testlet GPCM	1.19	-	1.23	1.44	-	1.31	1.54	-	1.67
Multilevel GPCM	0.42	0.11	-	0.45	0.13	-	0.46	0.14	-
Dual Parametric	0.88	0.63	1.49	0.89	0.11	1.67	0.87	0.13	2.14
Dual Dirichlet	088	0.64	1.39	0.89	0.13	1.57	0.90	0.13	1.99
LargeSmall									
GPCM	0.65	-	-	0.43	-	-	0.54	-	-
Testlet GPCM	1.62	-	1.42	1.42	-	1.60	1.39	-	1.45
Multilevel GPCM	0.50	0.26	-	0.39	0.24	-	0.30	1.31	-
Dual Parametric	1.17	0.95	1.62	1.02	1.32	2.06	0.44	1.06	0.92
Dual Dirichlet	1.11	1.11	1.57	1.04	1.33	1.98	0.45	1.04	1.15
LargeLarge									
GPCM	0.77	-	-	0.74	-	-	0.40	-	-
Testlet GPCM	1.42	-	1.56	1.94	-	1.50	2.67	-	1.86
Multilevel GPCM	0.45	0.41	-	0.43	0.95	-	0.44	0.73	-
Dual Parametric	0.98	1.05	1.74	1.05	1.76	1.75	1.02	1.43	2.22
Dual Dirichlet	1.01	1.06	1.61	1.05	1.76	1.64	0.96	1.47	1.87

Table A8: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the Ability parameters for different sample and group sizes

			SE					Bias					RMSE				
Sample	Groups	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
	1	NoneNone	0.23	0.25	0.30	0.32	0.38	-0.03	-0.05	-0.01	0.00	0.00	0.32	0.34	0.40	0.42	0.44
400		LargeNone	0.20	0.27	0.24	0.31	0.31	0.00	0.02	0.02	0.01	0.01	0.43	0.36	0.43	0.42	0.33
	5	NoneLarge	0.28	0.28	0.28	0.29	0.30	-0.20	-0.16	-0.01	0.00	-0.09	0.87	0.78	0.37	0.38	0.45
		LargeLarge	0.25	0.29	0.25	0.30	0.32	-0.18	-0.23	-0.07	-0.04	-0.08	0.82	0.67	0.45	0.41	0.47
	20	NoneLarge	0.29	0.32	0.35	0.37	0.39	-0.05	-0.01	0.04	0.05	0.07	0.90	0.82	0.45	0.47	0.47
		LargeLarge	0.22	0.29	0.26	0.34	0.36	0.01	0.08	-0.02	-0.03	-0.06	0.63	0.61	0.49	0.46	0.52
	40	NoneLarge	0.25	0.28	0.37	0.39	0.40	0.07	0.07	-0.01	-0.04	0.06	0.73	0.71	0.51	0.53	0.54
		LargeLarge	0.22	0.30	0.28	0.39	0.40	0.03	0.06	0.01	0.00	0.06	0.66	0.61	0.54	0.53	0.53
1000	1	NoneNone	0.23	0.25	0.26	0.27	0.30	0.04	0.04	0.05	0.04	0.04	0.31	0.33	0.35	0.37	0.39
		LargeNone	0.20	0.27	0.22	0.29	0.29	0.01	-0.02	0.01	0.00	0.00	0.49	0.38	0.45	0.40	0.45
	5	NoneLarge	0.30	0.36	0.30	0.30	0.32	-0.11	0.01	-0.01	-0.01	-0.03	0.99	0.95	0.40	0.40	0.42
		LargeLarge	0.20	0.33	0.20	0.27	0.34	0.02	0.04	-0.01	0.01	-0.01	0.59	0.59	0.42	0.36	0.40
	20	NoneLarge	0.25	0.29	0.27	0.29	0.30	-0.04	-0.04	-0.03	-0.03	-0.04	0.68	0.69	0.36	0.39	0.45
		LargeLarge	0.23	0.31	0.24	0.30	0.33	-0.05	-0.18	0.01	0.01	0.09	0.77	0.63	0.47	0.43	0.45
	40	NoneLarge	0.29	0.30	0.34	0.36	0.39	0.02	0.01	0.01	0.00	0.03	0.88	0.85	0.45	0.48	0.51
		LargeLarge	0.24	0.30	0.28	0.35	0.39	0.01	0.01	0.05	0.05	0.04	0.79	0.76	0.79	0.47	0.50
2000	1	NoneNone	0.23	0.25	0.25	0.26	0.27	-0.02	-0.02	-0.02	-0.02	-0.02	0.31	0.33	0.34	0.34	0.34
		LargeNone	0.21	0.30	0.22	0.27	0.27	0.00	0.01	0.01	0.01	0.01	0.47	0.59	0.42	0.36	0.37
	5	NoneLarge	0.25	0.34	0.27	0.26	0.28	-0.05	-0.04	-0.01	-0.04	-0.03	0.53	0.53	0.36	0.36	0.37
		LargeLarge	0.25	0.33	0.25	0.28	0.37	0.01	-0.10	0.03	0.03	0.04	0.93	0.43	0.46	0.39	0.43
	20	NoneLarge	0.25	0.33	0.28	0.29	0.30	0.02	0.02	0.02	0.02	0.02	0.74	0.74	0.38	0.39	0.42
		LargeLarge	0.22	0.38	0.23	0.28	0.31	-0.02	0.09	-0.03	-0.03	-0.03	0.62	0.59	0.42	0.38	0.41
	40	NoneLarge	0.28	0.30	0.31	0.34	0.36	0.01	0.02	0.02	0.02	0.02	0.91	0.92	0.42	0.45	0.51
		LargeLarge	0.17	0.19	0.19	0.20	0.24	-0.04	-0.04	-0.02	-0.04	-0.06	0.69	0.67	0.39	0.65	0.66

Table A9: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the threshold parameters for different sample and group sizes

			SD					Bias					RMSE				
Samples	Groups	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
400	1	NoneNone	0.17	0.20	0.18	0.27	0.31	-0.04	-0.03	-0.21	-0.28	-0.24	0.21	0.23	0.22	0.39	0.33
	1	LargeNone	0.19	0.24	0.20	0.36	0.35	0.40	0.10	-0.15	0.02	-0.08	0.75	0.73	0.71	0.48	0.51
	5	NoneLarge	0.19	0.22	0.20	0.27	0.28	0.29	0.32	0.00	0.03	-0.05	0.37	0.37	0.25	0.42	0.44
	5	LargeLarge	0.19	0.27	0.20	0.29	0.28	0.41	0.55	0.06	-0.13	-0.16	0.70	0.70	0.65	0.45	0.80
	20	NoneLarge	0.17	0.21	0.20	0.27	0.28	0.10	0.14	-0.10	-0.11	-0.09	0.24	0.30	0.26	0.42	0.45
	20	LargeLarge	0.17	0.23	0.19	0.30	0.33	-0.27	-0.25	-0.13	-0.07	0.03	0.61	0.55	0.59	0.58	0.58
	40	NoneLarge	0.13	0.12	0.12	0.12	0.12	-0.04	-0.08	0.08	0.19	0.06	0.22	0.22	0.22	0.36	0.29
	40	LargeLarge	0.13	0.14	0.13	0.13	0.14	-0.06	-0.05	-0.01	0.13	0.02	0.59	0.54	0.59	0.40	0.44
1000	1	NoneNone	0.11	0.15	0.11	0.22	0.25	0.05	0.22	0.12	0.14	0.14	0.16	0.16	0.16	0.42	0.27
	1	LargeNone	0.11	0.16	0.18	0.30	0.30	-0.18	-0.29	-0.09	0.15	-0.08	1.36	1.73	1.41	0.77	0.83
	5	NoneLarge	0.11	0.28	0.14	0.30	0.31	0.44	0.40	-0.41	0.12	-0.30	0.45	0.65	0.44	0.54	0.50
	5	LargeLarge	0.13	0.25	0.14	0.31	0.20	-0.43	-0.23	-0.05	-0.03	0.03	0.94	1.09	0.97	0.79	0.50
	20	NoneLarge	0.12	0.17	0.12	0.23	0.24	0.03	0.00	-0.03	0.05	0.00	0.16	0.23	0.16	0.56	0.58
	20	LargeLarge	0.12	0.19	0.20	0.27	0.27	0.40	0.44	-0.16	0.10	-0.07	0.82	0.92	0.79	0.47	0.37
	40	NoneLarge	0.09	0.08	0.09	0.08	0.08	0.03	0.02	0.01	0.03	0.05	0.16	0.16	0.16	0.35	0.24
	40	LargeLarge	0.11	0.10	0.09	0.09	0.09	0.12	0.12	-0.09	-0.15	-0.13	0.50	0.40	0.52	0.44	0.39
2000	1	NoneNone	0.08	0.08	0.09	0.16	0.12	0.01	0.07	0.10	-0.04	0.08	0.15	0.13	0.13	0.59	0.17
	1	LargeNone	0.09	0.17	0.11	0.28	0.18	0.40	0.32	-0.69	0.44	-0.19	0.60	0.61	0.67	0.71	0.71
	5	NoneLarge	0.09	0.28	0.12	0.33	0.28	0.58	0.80	-0.78	0.28	-0.21	0.58	0.60	0.49	0.61	0.62
	5	LargeLarge	0.09	0.22	0.24	0.30	0.24	0.69	0.68	1.10	0.96	0.40	0.92	0.53	0.90	0.53	0.53
	20	NoneLarge	0.09	0.18	0.13	0.21	0.18	0.22	0.18	-0.43	0.30	0.36	0.27	0.43	0.45	0.72	0.67
	20	LargeLarge	0.08	0.23	0.23	0.22	0.19	0.12	0.12	-0.09	-0.15	-0.18	0.60	0.44	0.70	0.42	0.39
	40	NoneLarge	0.08	0.07	0.07	0.06	0.06	0.22	0.24	0.10	0.34	0.31	0.24	0.27	0.15	0.41	0.38
	40	LargeLarge	0.19	0.11	0.11	0.10	0.11	-0.49	-0.51	-0.30	-0.44	-0.48	0.59	0.59	0.78	0.57	0.54

Table A10: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the discriminant parameters for different sample and group sizes

			SD					Bias					RMSE				
Samples	Groups	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
400	1	NoneNone	0.13	0.13	0.13	0.12	0.13	-0.01	-0.05	-0.03	-0.07	-0.09	0.17	0.18	0.17	0.18	0.20
	1	LargeNone	0.14	0.13	0.14	0.13	0.13	0.01	0.03	-0.01	0.00	0.03	0.26	0.18	0.27	0.18	0.91
	5	NoneLarge	0.13	0.13	0.13	0.13	0.13	0.01	-0.02	0.00	-0.02	-0.02	0.17	0.17	0.17	0.17	0.17
	5	LargeLarge	0.14	0.15	0.14	0.13	0.13	0.02	0.02	-0.01	-0.06	-0.06	0.37	0.19	0.37	0.18	0.19
	20	NoneLarge	0.12	0.12	0.12	0.11	0.13	-0.04	-0.10	-0.04	-0.09	-0.09	0.16	0.19	0.16	0.18	0.19
	20	LargeLarge	0.13	0.13	0.13	0.13	0.14	0.06	0.08	0.05	0.06	0.03	0.29	0.19	0.30	0.18	0.19
	40	NoneLarge	0.13	0.12	0.12	0.12	0.12	0.08	0.02	0.06	0.00	-0.08	0.17	0.16	0.16	0.15	0.17
	40	LargeLarge	0.13	0.14	0.13	0.13	0.14	0.06	0.11	0.05	0.06	-0.04	0.22	0.21	0.22	0.19	0.19
1000	1	NoneNone	0.10	0.09	0.09	0.09	0.09	0.04	-0.04	0.03	-0.04	-0.42	0.13	0.12	0.12	0.13	0.48
	1	LargeNone	0.10	0.10	0.11	0.10	0.09	0.03	0.05	0.07	0.01	-0.82	0.26	0.16	0.33	0.13	0.86
	5	NoneLarge	0.09	0.13	0.09	0.09	0.09	-0.06	-0.21	-0.12	-0.12	-0.12	0.12	0.26	0.16	0.16	0.16
	5	LargeLarge	0.10	0.16	0.10	0.09	0.09	0.11	-0.13	0.10	-0.02	-0.06	0.35	0.22	0.37	0.12	0.12
	20	NoneLarge	0.09	0.11	0.09	0.08	0.09	-0.05	-0.14	-0.02	-0.10	-0.09	0.12	0.18	0.12	0.14	0.13
	20	LargeLarge	0.09	0.14	0.10	0.10	0.10	0.05	0.03	0.04	0.03	0.04	0.29	0.18	0.31	0.14	0.14
	40	NoneLarge	0.09	0.08	0.09	0.08	0.08	-0.05	-0.06	0.00	-0.09	-0.08	0.13	0.13	0.12	0.14	0.13
	40	LargeLarge	0.11	0.10	0.09	0.09	0.09	0.01	0.03	-0.05	-0.02	-0.04	0.22	0.15	0.19	0.13	0.16
2000	1	NoneNone	0.09	0.07	0.07	0.07	0.07	0.05	-0.05	0.00	-0.05	-0.41	0.12	0.11	0.10	0.10	0.49
	1	LargeNone	0.08	0.12	0.08	0.07	0.07	-0.06	-0.08	-0.04	-0.02	-0.91	0.24	0.16	0.24	0.10	0.10
	5	NoneLarge	0.08	0.17	0.07	0.07	0.07	0.09	-0.04	-0.01	0.02	0.01	0.13	0.19	0.09	0.09	0.09
	5	LargeLarge	0.08	0.13	0.09	0.07	0.07	0.13	-0.04	0.08	0.02	-0.04	0.45	0.18	0.46	0.09	0.10
	20	NoneLarge	0.08	0.12	0.07	0.07	0.07	0.11	-0.07	0.05	0.01	0.01	0.14	0.16	0.10	0.09	0.09
	20	LargeLarge	0.07	0.15	0.07	0.07	0.08	-0.12	-0.29	-0.12	-0.07	-0.08	0.21	0.33	0.21	0.10	0.12
	40	NoneLarge	0.08	0.07	0.07	0.06	0.06	0.01	-0.09	-0.02	-0.12	-0.20	0.09	0.12	0.08	0.14	0.21
	40	LargeLarge	0.19	0.11	0.11	0.10	0.11	0.36	0.17	0.07	0.09	-0.03	0.42	0.20	0.16	0.14	0.11

Table A11: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the ability parameters for different sample sizes

	SE						Bias						RMSE					
	400		1000		2000		400		1000		2000		400		1000		2000	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
SmallNone																		
GPCM	0.22	0.03	0.22	0.03	0.22	0.03	0.10	0.32	0.02	0.33	0.04	0.31	0.37	0.16	0.36	0.17	0.35	0.15
Testlet	0.26	0.03	0.25	0.03	0.24	0.04	0.10	0.36	0.02	0.49	0.05	0.45	0.42	0.18	0.49	0.24	0.46	0.23
Multilevel	0.31	0.02	0.26	0.03	0.25	0.03	0.09	0.32	0.02	0.33	0.05	0.27	0.42	0.16	0.40	0.16	0.36	0.13
Dual	0.35	0.03	0.30	0.03	0.27	0.03	0.08	0.30	0.01	0.32	0.05	0.27	0.45	0.12	0.42	0.13	0.37	0.12
DualDP	0.31	0.03	0.28	0.03	0.25	0.04	0.08	0.28	0.01	0.30	0.05	0.26	0.41	0.13	0.39	0.13	0.35	0.12
SmallSmall																		
GPCM	0.24	0.04	0.22	0.03	0.24	0.06	-0.03	0.57	0.03	0.48	-0.02	0.54	0.53	0.32	0.46	0.25	0.51	0.30
Testlet	0.28	0.04	0.25	0.03	0.26	0.06	-0.03	0.58	0.03	0.52	0.00	0.63	0.57	0.31	0.52	0.28	0.59	0.35
Multilevel	0.38	0.03	0.30	0.04	0.29	0.05	-0.02	0.40	0.03	0.34	-0.01	0.28	0.53	0.17	0.43	0.15	0.38	0.13
Dual	0.40	0.03	0.31	0.03	0.28	0.05	0.00	0.40	0.02	0.32	-0.03	0.28	0.54	0.17	0.43	0.14	0.38	0.13
Dual DP	0.34	0.03	0.28	0.03	0.27	0.05	0.00	0.20	0.01	0.16	-0.01	0.14	0.39	0.07	0.32	0.06	0.30	0.07
SmallLarge																		
GPCM	0.24	0.05	0.23	0.06	0.27	0.10	-0.11	0.70	-0.02	0.83	0.02	1.18	0.65	0.38	0.71	0.49	1.00	0.70
Testlet	0.29	0.05	0.27	0.06	0.54	0.18	-0.13	0.74	0.00	0.82	0.03	1.14	0.71	0.38	0.74	0.44	1.13	0.59
Multilevel	0.33	0.03	0.27	0.05	0.28	0.08	-0.07	0.38	-0.03	0.39	0.02	0.35	0.48	0.18	0.44	0.20	0.41	0.19
Dual	0.39	0.04	0.31	0.05	0.30	0.08	-0.06	0.34	-0.03	0.35	0.03	0.32	0.50	0.14	0.44	0.17	0.41	0.17
Dual DP	0.34	0.04	0.29	0.06	0.42	0.13	-0.03	0.17	-0.01	0.17	0.01	0.16	0.37	0.07	0.33	0.08	0.44	0.14
LargeSmall																		
GPCM	0.21	0.04	0.18	0.02	0.21	0.02	0.01	0.61	0.03	0.59	0.00	0.53	0.56	0.32	0.53	0.33	0.50	0.28
Testlet	0.30	0.06	0.29	0.05	0.28	0.10	0.07	0.80	0.00	0.75	0.00	1.73	0.77	0.38	0.70	0.39	1.45	0.99
Multilevel	0.29	0.03	0.23	0.02	0.22	0.02	-0.01	0.51	0.03	0.51	0.00	0.43	0.52	0.26	0.49	0.28	0.42	0.23
Dual	0.41	0.05	0.34	0.04	0.29	0.03	-0.02	0.41	0.04	0.35	0.00	0.28	0.55	0.17	0.47	0.15	0.39	0.12
Dual DP	0.35	0.05	0.31	0.04	0.29	0.05	-0.01	0.20	0.02	0.18	0.00	0.14	0.40	0.07	0.36	0.07	0.32	0.06

Table A12: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the threshold parameters for changing sample sizes

	SE						Bias						RMSE					
	400		1000		2000		400		1000		2000		400		1000		2000	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
NoneNone																		
GPCM	0.12	0.02	0.11	0.02	0.11	0.03	0.00	0.10	0.01	0.08	0.12	0.09	0.15	0.04	0.14	0.04	0.06	0.03
Testlet	0.12	0.02	0.08	0.02	0.07	0.02	-0.06	0.10	-0.05	0.08	-0.02	0.08	0.16	0.05	0.12	0.05	0.09	0.06
Multilevel	0.12	0.02	0.09	0.02	0.07	0.02	-0.01	0.10	-0.01	0.08	0.03	0.09	0.15	0.04	0.11	0.04	0.03	0.02
Dual	0.12	0.02	0.08	0.02	0.07	0.02	-0.08	0.10	-0.09	0.08	-0.07	0.09	0.16	0.06	0.13	0.06	0.20	0.19
Dual DP	0.12	0.02	0.09	0.02	0.07	0.02	-0.17	0.10	-0.07	0.08	0.01	0.08	0.22	0.08	0.13	0.06	0.19	0.20
NoneSmall																		
GPCM	0.17	0.04	0.11	0.04	0.08	0.02	-0.04	0.15	0.06	0.11	0.08	0.10	0.22	0.08	0.16	0.07	0.14	0.06
Testlet	0.18	0.04	0.12	0.03	0.09	0.02	-0.03	0.16	0.05	0.10	0.08	0.10	0.23	0.08	0.16	0.06	0.15	0.06
Multilevel	0.18	0.04	0.11	0.03	0.10	0.02	-0.11	0.15	0.15	0.10	0.05	0.09	0.25	0.08	0.20	0.06	0.14	0.04
Dual	0.31	0.04	0.20	0.04	0.19	0.06	-0.20	0.37	0.35	0.57	-0.01	0.77	0.48	0.21	0.62	0.31	0.63	0.49
DualDP	0.24	0.04	0.15	0.03	0.14	0.03	-0.14	0.27	0.25	0.39	0.02	0.52	0.36	0.15	0.44	0.22	0.42	0.34
NoneLarge																		
GPCM	0.13	0.02	0.09	0.02	0.08	0.01	0.08	0.11	-0.05	0.09	0.01	0.04	0.17	0.06	0.13	0.04	0.09	0.02
Testlet	0.12	0.02	0.08	0.02	0.07	0.01	0.02	0.11	-0.06	0.10	-0.09	0.06	0.16	0.04	0.13	0.06	0.12	0.05
Multilevel	0.12	0.02	0.09	0.02	0.07	0.01	0.06	0.11	0.00	0.09	-0.02	0.05	0.16	0.05	0.12	0.04	0.08	0.02
Dual	0.12	0.02	0.08	0.02	0.06	0.01	0.00	0.10	-0.09	0.09	-0.12	0.06	0.15	0.04	0.14	0.06	0.14	0.05
Dual DP	0.12	0.02	0.08	0.02	0.06	0.01	-0.08	0.10	-0.08	0.09	-0.20	0.07	0.17	0.05	0.13	0.06	0.21	0.06
LargeNone																		
GPCM	0.13	0.02	0.12	0.03	0.08	0.01	-0.03	0.22	0.07	0.35	-0.17	0.27	0.24	0.10	0.33	0.17	0.26	0.20
Testlet	0.13	0.02	0.09	0.02	0.07	0.02	0.02	0.15	0.00	0.10	-0.01	0.06	0.19	0.05	0.12	0.05	0.09	0.03
Multilevel	0.13	0.02	0.10	0.02	0.07	0.01	-0.05	0.22	0.00	0.32	-0.08	0.28	0.24	0.11	0.28	0.17	0.24	0.17
Dual	0.12	0.02	0.09	0.02	0.06	0.02	0.00	0.14	-0.01	0.07	-0.07	0.06	0.18	0.06	0.11	0.04	0.10	0.04
Dual DP	0.13	0.02	0.09	0.02	0.07	0.02	-0.14	0.15	-1.25	0.28	-0.21	0.11	0.22	0.11	1.25	0.28	0.22	0.11
LargeLarge																		
GPCM	0.13	0.03	0.11	0.02	0.19	0.13	0.06	0.20	0.01	0.27	0.36	0.39	0.22	0.10	0.22	0.19	0.42	0.40
Testlet	0.14	0.03	0.10	0.03	0.11	0.07	0.11	0.17	0.03	0.14	0.17	0.11	0.21	0.12	0.15	0.08	0.20	0.13
Multilevel	0.13	0.03	0.09	0.02	0.11	0.06	0.05	0.20	-0.05	0.27	0.07	0.13	0.22	0.10	0.19	0.21	0.16	0.10
Dual	0.13	0.03	0.09	0.03	0.10	0.06	0.06	0.15	-0.02	0.11	0.09	0.05	0.19	0.08	0.13	0.07	0.14	0.07
Dual DP	0.14	0.03	0.09	0.02	0.11	0.07	-0.04	0.14	-0.04	0.18	-0.03	0.02	0.19	0.07	0.16	0.14	0.11	0.07

Table A13: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the threshold parameters for changing sample sizes

	SE						Bias						RMSE					
	400		1000		2000		400		1000		2000		400		1000		2000	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
SmallNone																		
GPCM	0.17	0.05	0.12	0.04	0.08	0.02	0.11	0.44	-0.23	0.52	-0.12	0.39	0.42	0.24	0.49	0.32	0.36	0.20
Testlet	0.19	0.05	0.14	0.03	0.11	0.03	0.12	0.41	-0.23	0.47	-0.12	0.35	0.42	0.21	0.48	0.24	0.36	0.12
Multilevel	0.18	0.05	0.12	0.03	0.09	0.02	0.10	0.45	-0.17	0.52	-0.08	0.39	0.43	0.24	0.47	0.29	0.36	0.18
Dual	0.28	0.04	0.22	0.03	0.26	0.09	0.07	0.36	-0.23	0.45	0.18	0.49	0.43	0.15	0.53	0.16	0.55	0.20
Dual DP	0.23	0.04	0.17	0.03	0.18	0.05	0.08	0.28	-0.23	0.41	0.08	0.37	0.35	0.14	0.46	0.21	0.39	0.18
SmallSmall																		
GPCM	0.14	0.03	0.12	0.03	0.10	0.01	-0.03	0.17	0.16	0.16	0.02	0.12	0.19	0.11	0.22	0.13	0.15	0.06
Testlet	0.14	0.04	0.10	0.03	0.07	0.01	-0.03	0.15	0.10	0.10	0.01	0.09	0.18	0.10	0.16	0.06	0.10	0.04
Multilevel	0.13	0.03	0.10	0.03	0.06	0.01	-0.09	0.17	0.07	0.15	-0.12	0.13	0.20	0.13	0.17	0.09	0.16	0.09
Dual	0.12	0.04	0.10	0.03	0.06	0.01	-0.13	0.15	0.04	0.10	-0.03	0.07	0.20	0.12	0.13	0.06	0.09	0.03
Dual DP	0.13	0.04	0.10	0.03	0.07	0.01	-0.13	0.15	-0.04	0.10	-0.12	0.09	0.21	0.12	0.13	0.07	0.15	0.06
SmallLarge																		
GPCM	0.14	0.03	0.12	0.03	0.08	0.02	0.07	0.17	0.13	0.15	0.01	0.10	0.21	0.09	0.21	0.10	0.11	0.06
Testlet	0.14	0.03	0.10	0.02	0.04	0.01	0.06	0.13	0.06	0.09	-0.57	0.15	0.19	0.07	0.14	0.05	0.57	0.15
Multilevel	0.14	0.03	0.09	0.02	0.07	0.01	0.06	0.17	0.04	0.15	0.02	0.11	0.21	0.09	0.16	0.10	0.11	0.06
Dual	0.13	0.03	0.09	0.03	0.07	0.02	0.00	0.13	0.01	0.08	0.05	0.05	0.18	0.05	0.12	0.04	0.10	0.03
Dual DP	0.13	0.03	0.09	0.03	0.05	0.01	-0.06	0.13	-0.08	0.10	-0.06	0.07	0.19	0.06	0.14	0.07	0.09	0.05
LargeSmall																		
GPCM	0.14	0.03	0.16	0.12	0.09	0.02	0.04	0.20	0.25	0.47	0.00	0.21	0.22	0.10	0.45	0.34	0.18	0.14
Testlet	0.14	0.03	0.11	0.04	0.07	0.02	0.05	0.13	0.04	0.12	0.08	0.07	0.19	0.06	0.16	0.07	0.11	0.05
Multilevel	0.14	0.03	0.11	0.04	0.07	0.01	0.03	0.20	0.02	0.35	-0.01	0.21	0.22	0.10	0.31	0.19	0.17	0.14
Dual	0.13	0.03	0.10	0.04	0.07	0.02	-0.01	0.11	-0.03	0.10	0.00	0.06	0.17	0.05	0.14	0.05	0.09	0.03
Dual DP	0.13	0.03	0.11	0.04	0.07	0.02	-0.18	0.18	-0.09	0.14	-0.11	0.10	0.25	0.15	0.18	0.09	0.14	0.07

Table A14: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the discriminant parameters for changing sample sizes

	SE						Bias						RMSE					
	400		1000		2000		400		1000		2000		400		1000		2000	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
NoneNone																		
GPCM	0.12	0.02	0.11	0.02	0.11	0.03	0.00	0.10	0.01	0.08	0.12	0.09	0.15	0.04	0.14	0.04	0.17	0.07
Testlet	0.12	0.02	0.08	0.02	0.07	0.02	-0.06	0.10	-0.05	0.08	-0.02	0.08	0.16	0.05	0.12	0.05	0.10	0.06
Multilevel	0.12	0.02	0.09	0.02	0.07	0.02	-0.01	0.10	-0.01	0.08	0.03	0.09	0.15	0.04	0.11	0.04	0.10	0.06
Dual	0.12	0.02	0.08	0.02	0.07	0.02	-0.08	0.10	-0.09	0.08	-0.07	0.09	0.16	0.06	0.13	0.06	0.11	0.07
Dual DP	0.12	0.02	0.08	0.02	0.07	0.02	-0.17	0.10	-0.07	0.08	0.01	0.08	0.22	0.08	0.13	0.06	0.09	0.06
NoneSmall																		
GPCM	0.13	0.02	0.12	0.02	0.07	0.01	0.00	0.10	0.07	0.09	0.07	0.06	0.16	0.04	0.15	0.05	0.11	0.04
Testlet	0.12	0.02	0.09	0.02	0.06	0.01	-0.05	0.11	-0.06	0.08	-0.02	0.06	0.16	0.05	0.13	0.04	0.08	0.03
Multilevel	0.13	0.02	0.09	0.02	0.07	0.01	0.00	0.10	-0.01	0.08	0.05	0.06	0.16	0.04	0.12	0.03	0.10	0.03
Dual	0.12	0.02	0.08	0.01	0.06	0.01	-0.06	0.11	-0.11	0.08	-0.08	0.06	0.17	0.05	0.15	0.06	0.11	0.05
Dual DP	0.12	0.02	0.09	0.02	0.06	0.01	-0.04	0.10	-0.05	0.08	-0.03	0.06	0.16	0.04	0.12	0.04	0.09	0.03
NoneLarge																		
GPCM	0.13	0.02	0.09	0.02	0.08	0.01	0.08	0.11	-0.05	0.09	0.01	0.04	0.17	0.06	0.13	0.04	0.09	0.02
Testlet	0.12	0.02	0.08	0.02	0.07	0.01	0.02	0.11	-0.06	0.10	-0.09	0.06	0.16	0.04	0.13	0.06	0.12	0.05
Multilevel	0.12	0.02	0.09	0.02	0.07	0.01	0.06	0.11	0.00	0.09	-0.02	0.05	0.16	0.05	0.12	0.04	0.08	0.02
Dual	0.12	0.02	0.08	0.02	0.06	0.01	0.00	0.10	-0.09	0.09	-0.12	0.06	0.15	0.04	0.14	0.06	0.14	0.05
Dual DP	0.12	0.02	0.08	0.02	0.06	0.01	-0.08	0.10	-0.08	0.09	-0.20	0.07	0.17	0.05	0.13	0.06	0.21	0.06
LargeNone																		
GPCM	0.13	0.02	0.12	0.03	0.08	0.01	-0.03	0.22	0.07	0.35	-0.17	0.27	0.24	0.10	0.33	0.17	0.26	0.20
Testlet	0.13	0.02	0.09	0.02	0.07	0.02	0.02	0.15	0.00	0.10	-0.01	0.06	0.19	0.05	0.12	0.05	0.09	0.03
Multilevel	0.13	0.02	0.10	0.02	0.07	0.01	-0.05	0.22	0.00	0.32	-0.08	0.28	0.24	0.11	0.28	0.17	0.24	0.17
Dual	0.12	0.02	0.09	0.02	0.06	0.02	0.00	0.14	-0.01	0.07	-0.07	0.06	0.18	0.06	0.11	0.04	0.10	0.04
Dual DP	0.13	0.02	0.09	0.02	0.07	0.02	-0.14	0.15	-0.01	0.08	-0.21	0.11	0.22	0.11	0.12	0.05	0.22	0.11
LargeLarge																		
GPCM	0.13	0.03	0.11	0.02	0.19	0.13	0.06	0.20	0.01	0.27	0.36	0.39	0.22	0.10	0.22	0.19	0.42	0.40
Testlet	0.14	0.03	0.10	0.03	0.11	0.07	0.11	0.17	0.03	0.14	0.17	0.11	0.21	0.12	0.15	0.08	0.20	0.13
Multilevel	0.13	0.03	0.09	0.02	0.11	0.06	0.05	0.20	-0.05	0.27	0.07	0.13	0.22	0.10	0.19	0.21	0.16	0.10
Dual	0.13	0.03	0.09	0.03	0.10	0.06	0.06	0.15	-0.02	0.11	0.09	0.05	0.19	0.08	0.13	0.07	0.14	0.07
Dual DP	0.14	0.03	0.09	0.03	0.11	0.07	-0.04	0.14	-0.04	0.18	-0.03	0.02	0.19	0.07	0.16	0.14	0.11	0.07

Table A15: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the discriminant parameters for changing sample sizes

	SE						Bias						RMSE					
	400		1000		2000		400		1000		2000		400		1000		2000	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
SmallNone																		
GPCM	0.14	0.03	0.12	0.02	0.07	0.01	0.03	0.17	0.03	0.14	-0.01	0.11	0.20	0.09	0.17	0.08	0.12	0.07
Testlet	0.14	0.03	0.09	0.01	0.07	0.01	0.04	0.15	0.04	0.10	-0.01	0.06	0.19	0.07	0.13	0.05	0.09	0.03
Multilevel	0.13	0.03	0.09	0.01	0.07	0.01	0.01	0.17	-0.01	0.14	-0.07	0.11	0.20	0.10	0.15	0.08	0.13	0.08
Dual	0.13	0.03	0.09	0.01	0.07	0.01	0.00	0.15	0.00	0.10	-0.04	0.05	0.19	0.07	0.12	0.05	0.09	0.03
Dual DP	0.13	0.03	0.09	0.01	0.07	0.01	0.01	0.15	0.01	0.10	-0.03	0.07	0.19	0.08	0.13	0.06	0.10	0.04
SmallSmall																		
GPCM	0.14	0.03	0.12	0.03	0.10	0.01	-0.03	0.17	0.16	0.16	0.02	0.12	0.19	0.11	0.22	0.13	0.15	0.06
Testlet	0.14	0.04	0.10	0.03	0.07	0.01	-0.03	0.15	0.10	0.10	0.01	0.09	0.18	0.10	0.16	0.06	0.10	0.04
Multilevel	0.13	0.03	0.10	0.03	0.06	0.01	-0.09	0.17	0.07	0.15	-0.12	0.13	0.20	0.13	0.17	0.09	0.16	0.09
Dual	0.12	0.04	0.10	0.03	0.06	0.01	-0.13	0.15	0.04	0.10	-0.03	0.07	0.20	0.12	0.13	0.06	0.09	0.03
Dual DP	0.13	0.04	0.10	0.03	0.07	0.01	-0.13	0.15	-0.04	0.10	-0.12	0.09	0.21	0.12	0.13	0.07	0.15	0.06
SmallLarge																		
GPCM	0.14	0.03	0.12	0.03	0.08	0.02	0.07	0.17	0.13	0.15	0.01	0.10	0.21	0.09	0.21	0.10	0.11	0.06
Testlet	0.14	0.03	0.10	0.02	0.04	0.01	0.06	0.13	0.06	0.09	-0.57	0.15	0.19	0.07	0.14	0.05	0.57	0.15
Multilevel	0.14	0.03	0.09	0.02	0.07	0.01	0.06	0.17	0.04	0.15	0.02	0.11	0.21	0.09	0.16	0.10	0.11	0.06
Dual	0.13	0.03	0.09	0.03	0.07	0.02	0.00	0.13	0.01	0.08	0.05	0.05	0.18	0.05	0.12	0.04	0.10	0.03
Dual DP	0.13	0.03	0.09	0.03	0.05	0.01	-0.06	0.13	-0.08	0.10	-0.06	0.07	0.19	0.06	0.14	0.07	0.09	0.05
LargeSmall																		
GPCM	0.14	0.03	0.16	0.12	0.09	0.02	0.04	0.20	0.25	0.47	0.00	0.21	0.22	0.10	0.45	0.34	0.18	0.14
Testlet	0.14	0.03	0.11	0.04	0.07	0.02	0.05	0.13	0.04	0.12	0.08	0.07	0.19	0.06	0.16	0.07	0.11	0.05
Multilevel	0.14	0.03	0.11	0.04	0.07	0.01	0.03	0.20	0.02	0.35	-0.01	0.21	0.22	0.10	0.31	0.19	0.17	0.14
Dual	0.13	0.03	0.10	0.04	0.07	0.02	-0.01	0.11	-0.03	0.10	0.00	0.06	0.17	0.05	0.14	0.05	0.09	0.03
Dual DP	0.13	0.03	0.11	0.04	0.07	0.02	-0.18	0.18	-0.09	0.14	-0.11	0.10	0.25	0.15	0.18	0.09	0.14	0.07

Table A16: Ability, Group, Testlet and interaction variances for changing testlet items and category options

Items	Categories	Condition	GPCM			Testlet			Multilevel			Dual			DualDP		
			σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ
3	3	NoneNone	0.91	-	-	1.27	-	0.04	0.95	0.11	-	1.22	0.12	0.04	0.87	0.14	0.10
		NoneLarge	1.7	-	-	1.91	-	0.07	1	0.73	-	1.24	0.83	0.04	0.95	0.79	0.07
		LargeNone	0.5	-	-	1.61	-	2.08	0.38	0.15	-	0.97	1.02	2.13	1.02	1.11	0.07
		LargeLarge	0.66	-	-	1.57	-	1.97	0.39	0.4	-	0.95	1.46	1.74	0.91	1.49	1.89
	4	NoneNone	0.93	-	-	1.15	-	0.03	1.02	0.1	-	1.24	0.11	0.04	0.96	0.10	0.06
		NoneLarge	1.51	-	-	1.18	-	1.28	0.8	0.98	-	0.94	0.64	0.09	0.92	0.64	0.11
		LargeNone	0.4	-	-	1.1	-	1.24	0.28	0.12	-	0.82	0.1	1.44	1.07	0.11	0.67
		LargeLarge	0.71	-	-	1.46	-	1.3	0.35	0.45	-	1.19	1.46	1.59	1.41	0.78	1.59
	5	NoneNone	0.78	-	-	1.05	-	0.03	0.83	0.09	-	1.04	0.1	0.03	1.07	0.11	o.07
		NoneLarge	1.54	-	-	2.26	-	0.03	0.74	1.2	-	0.81	1.31	0.03	0.91	1.21	0.09
		LargeNone	0.24	-	-	1.31	-	1.29	0.21	0.1	-	1.01	0.81	1.69	0.91	0.62	1.86
		LargeLarge	0.5	-	-	2	-	2.67	0.21	0.37	-	0.82	1.56	1.96		1.43	2.1
6	3	NoneNone	0.95	-	-	1.04	-	0.03	0.94	0.09	-	1.2	0.09	0.03	1.14	0.12	0.03
		NoneLarge	1.6	-	-	1.93	-	0.04	0.97	1.07	-	1.17	1.12	0.03	1.06	1.08	0.04
		LargeNone	0.5	-	-	1.44	-	1.31	0.45	0.13	-	0.88	0.14	1.67	0.89	0.13	1.54
		LargeLarge	0.74	-	-	1.94	-	1.5	0.43	0.95	-	1.05	1.76	1.75	1.05	1.76	1.64
	4	NoneNone	0.94	-	-	1.11	-	0.02	0.98	0.13	-	1.16	0.1	0.02	1.16	0.12	0.03
		NoneLarge	1.79	-	-	2.1	-	0.02	0.98	1.12	-	1.03	1.3	0.03	1.04	1.27	0.03
		LargeNone	0.41	-	-	1.19	-	1.23	0.46	0.13	-	0.8	0.16	1.74	0.9	0.17	1.81
		LargeLarge	0.72	-	-	1.88	-	1.96	0.41	0.61	-	1.07	1.64	2.11	1.08	1.71	1.98
	5	NoneNone	0.97	-	-	1.13	-	0.02	0.98	0.1	-	1.27	0.11	0.02	1.21	0.12	0.03
		NoneLarge	1.82	-	-	1.77	-	0.02	0.88	0.79	-	1.08	1.05	0.02	1.11	1.1	0.03
		LargeNone	0.41	-	-	1.18	-	1.28	0.23	0.19	-	0.88	0.17	1.6	0.89	0.21	1.57
		LargeLarge	0.74	-	-	1.63	-	1.27	0.44	1.01	-	1.08	1.29	1.44	1.09	1.26	1.39
10	3	NoneNone	1.13	-	-	1.41	-	0.04	0.95	0.11	-	1.19	0.12	0.02	0.98	0.13	0.04
		NoneLarge	2.14	-	-	2.33	-	0.09	1.07	1.13	-	1.22	1.34	0.03	0.95	1.41	0.05
		LargeNone	0.5	-	-	1.39	-	1.43	0.44	0.24	-	0.86	0.81	1.64	0.93	0.21	1.86
		LargeLarge	1.02	-	-	2.47	-	1.71	0.5	0.62	-	0.99	2	1.78	0.82	1.82	1.58
	4	NoneNone	0.82	-	-	0.92	-	0.02	0.84	0.11	-	0.94	0.10	0.02	0.89	0.13	0.04
		NoneLarge	1.4	-	-	1.48	-	1.38	0.85	0.91	-	0.95	0.67	0.05	1.01	0.72	0.05
		LargeNone	0.37	-	-	1.43	-	1.61	0.31	0.16	-	0.79	0.76	1.33	0.76	0.29	0.68
		LargeLarge	0.57	-	-	1.86	-	1.86	0.3	0.34	-	0.93	1.05	1.78	1.09	1.19	2.26
	5	NoneNone	0.8	-	-	1.14	-	0.03	0.82	0.09	-	0.92	0.08	0.02	0.96	0.06	0.04
		NoneLarge	1.52	-	-	2.18	-	0.09	0.81	0.97	-	0.88	1.18	0.02	0.92	1.36	0.03
		LargeNone	0.2	-	-	1.23	-	1.41	0.18	0.1	-	0.75	0.39	1.1	0.67	0.46	0.96
		LargeLarge	0.47	-	-	1.81	-	2.24	0.39	1.46	-	0.74	1.25	1.59	0.82	1.33	0.06

Table A17: Standard errors (SE), Bias and Root Mean Square Errors (RMSE) in the ability parameters for changing items and category options

Items	Categories	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	NoneNone	0.31	0.37	0.35	0.39	0.42	-0.01	-0.01	-0.01	-0.01	-0.01	0.42	0.48	0.48	0.51	0.53
		NoneLarge	0.35	0.40	0.38	0.42	0.44	0.01	0.01	0.00	0.01	0.01	0.77	0.80	0.51	0.55	0.60
		LargeNone	0.26	0.40	0.28	0.38	0.40	-0.02	0.00	-0.02	-0.01	-0.01	0.53	0.77	0.53	0.51	0.54
		LargeLarge	0.28	0.40	0.29	0.39	0.43	-0.05	-0.06	-0.04	-0.04	-0.04	0.69	0.80	0.54	0.51	0.54
	4	NoneNone	0.26	0.29	0.32	0.34	0.36	0.05	0.06	0.05	0.03	0.03	0.35	0.39	0.42	0.42	0.44
		NoneLarge	0.29	0.41	0.32	0.37	0.40	-0.08	-0.06	-0.04	-0.01	-0.01	0.76	0.98	0.46	0.49	0.52
		LargeNone	0.22	0.32	0.23	0.34	0.37	0.04	0.02	0.04	0.04	0.04	0.53	0.59	0.52	0.46	0.49
		LargeLarge	0.24	0.34	0.25	0.40	0.42	0.02	-0.02	0.03	0.02	0.03	0.71	0.70	0.54	0.52	0.54
	5	NoneNone	0.21	0.25	0.26	0.28	0.31	-0.02	-0.09	-0.03	-0.04	-0.04	0.31	0.33	0.36	0.37	0.40
		NoneLarge	0.25	0.34	0.29	0.30	0.33	-0.03	0.02	0.05	0.05	0.04	0.87	0.98	0.44	0.44	0.46
		LargeNone	0.15	0.27	0.17	0.31	0.33	0.02	0.02	0.01	0.01	0.01	0.57	0.63	0.56	0.43	0.45
		LargeLarge	0.18	0.31	0.19	0.31	0.34	-0.09	0.25	-0.04	-0.04	0.05	0.69	1.10	0.52	0.44	0.45
6	3	NoneNone	0.23	0.24	0.28	0.30	0.32	0.01	0.02	0.02	0.00	0.01	0.31	0.32	0.36	0.38	0.35
		NoneLarge	0.29	0.30	0.34	0.36	0.38	0.02	0.01	0.01	0.00	0.01	0.88	0.85	0.45	0.48	0.51
		LargeNone	0.18	0.26	0.21	0.31	0.33	-0.02	-0.02	-0.02	-0.02	-0.01	0.49	0.45	0.47	0.42	0.38
		LargeLarge	0.24	0.30	0.28	0.35	0.36	0.01	0.01	0.05	0.05	0.01	0.79	0.76	0.49	0.47	0.64
	4	NoneNone	0.19	0.21	0.25	0.27	0.29	-0.07	-0.08	-0.05	-0.06	-0.08	0.26	0.29	0.34	0.35	0.38
		NoneLarge	0.22	0.24	0.28	0.30	0.32	-0.04	-0.03	-0.05	-0.06	-0.04	0.89	0.91	0.41	0.41	0.44
		LargeNone	0.15	0.21	0.17	0.26	0.29	-0.05	-0.05	-0.06	-0.07	-0.06	0.49	0.64	0.50	0.38	0.41
		LargeLarge	0.18	0.25	0.21	0.30	0.32	0.05	0.07	0.03	0.02	0.03	0.69	0.82	0.47	0.40	0.42
	5	NoneNone	0.16	0.18	0.22	0.25	0.28	-0.02	-0.01	-0.03	-0.03	-0.03	0.22	0.23	0.30	0.33	0.35
		NoneLarge	0.18	0.20	0.25	0.27	0.29	-0.01	-0.05	0.03	0.03	0.00	0.75	0.77	0.37	0.38	0.42
		LargeNone	0.12	0.18	0.15	0.25	0.26	0.04	-0.03	0.05	0.05	0.04	0.54	0.57	0.53	0.36	0.41
		LargeLarge	0.13	0.19	0.15	0.24	0.26	-0.02	-0.07	-0.02	-0.02	-0.03	0.61	0.85	0.51	0.36	0.40
10	3	NoneNone	0.18	0.22	0.27	0.26	0.29	-0.03	0.00	-0.05	-0.06	-0.06	0.24	0.30	1.21	0.36	0.39
		NoneLarge	0.21	0.26	0.27	0.30	0.32	0.03	0.02	-0.01	-0.01	-0.03	0.93	0.96	0.38	0.41	0.43
		LargeNone	0.16	0.23	0.20	0.25	0.28	-0.06	-0.11	-0.05	-0.04	-0.03	0.43	0.64	0.41	0.37	0.39
		LargeLarge	0.18	0.28	0.22	0.29	0.33	0.01	0.09	0.00	-0.01	-0.02	0.70	1.06	0.43	0.39	0.41
	4	NoneNone	0.14	0.15	0.21	0.22	0.25	-0.01	-0.01	-0.01	0.01	0.02	0.20	0.21	0.29	0.29	0.32
		NoneLarge	0.17	0.22	0.23	0.24	0.26	0.04	0.04	0.05	0.00	0.02	0.71	0.79	0.33	0.33	0.35
		LargeNone	0.12	0.20	0.15	0.20	0.21	-0.02	-0.09	0.01	0.01	0.01	0.44	0.61	0.45	0.34	0.36
		LargeLarge	0.13	0.24	0.16	0.23	0.25	0.04	0.23	0.01	0.01	0.00	0.59	0.80	0.48	0.37	0.39
	5	NoneNone	0.12	0.17	0.17	0.19	0.22	-0.05	0.00	-0.03	-0.04	-0.06	0.20	0.22	0.26	0.26	0.30
		NoneLarge	0.16	0.23	0.21	0.22	0.24	0.19	0.17	0.09	0.08	0.06	0.77	0.90	0.32	0.32	0.34
		LargeNone	0.09	0.24	0.11	0.19	0.22	0.02	-0.01	0.01	0.01	0.02	0.53	0.46	0.52	0.31	0.34
		LargeLarge	0.10	0.19	0.10	0.20	0.22	0.08	0.05	0.08	-0.02	-0.02	0.62	0.82	0.62	0.32	0.34

Table A18: Standard errors , Bias and Root Mean Square Errors in the threshold parameters for changing items and category options

Items	Categories	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	NoneNone	0.11	0.14	0.12	0.30	0.24	-0.01	0.01	-0.03	-0.06	-0.05	0.15	0.22	0.16	0.37	0.35
		NoneLarge	0.12	0.14	0.12	0.33	0.27	-0.03	-0.01	-0.20	0.12	-0.02	0.16	0.21	0.25	0.53	0.54
		LargeNone	0.11	0.22	0.14	0.30	0.32	0.13	0.24	0.09	0.30	0.14	0.60	0.56	0.62	0.53	0.54
		LargeLarge	0.12	0.20	0.14	0.24	0.25	0.04	0.12	0.05	0.00	0.00	0.68	0.64	0.69	0.43	0.42
	4	NoneNone	0.12	0.13	0.14	0.30	0.22	0.06	0.04	0.08	0.06	0.06	0.16	0.18	0.18	0.34	0.32
		NoneLarge	0.16	0.26	0.15	0.21	0.24	-0.06	-0.08	-0.11	-0.07	-0.08	0.31	0.37	0.63	0.56	0.54
		LargeNone	0.16	0.21	0.15	0.31	0.25	-0.09	-0.05	-0.06	-0.02	0.03	0.62	0.64	0.64	0.57	0.54
		LargeLarge	0.15	0.19	0.17	0.31	0.30	0.04	0.01	0.03	0.01	0.00	0.80	0.57	0.75	0.52	0.53
	5	NoneNone	0.17	0.21	0.18	0.23	0.24	0.01	-0.01	0.02	-0.01	0.04	0.32	0.30	0.30	0.46	0.50
		NoneLarge	0.21	0.30	0.19	0.24	0.23	-0.21	-0.18	-0.06	-0.06	-0.07	0.45	0.44	0.34	0.50	0.51
		LargeNone	0.17	0.25	0.20	0.25	0.26	-0.12	-0.02	-0.07	-0.04	-0.03	0.69	0.62	0.73	0.61	0.60
		LargeLarge	0.16	0.27	0.18	0.21	0.25	0.09	0.05	0.08	0.02	0.02	0.72	0.43	0.77	0.35	0.35
6	3	NoneNone	0.12	0.12	0.11	0.21	0.18	0.02	0.01	0.04	0.17	0.09	0.16	0.16	0.16	0.42	0.44
		NoneLarge	0.12	0.12	0.14	0.23	0.18	0.07	0.06	0.01	0.09	0.05	0.17	0.17	0.18	0.45	0.37
		LargeNone	0.16	0.19	0.12	0.21	0.20	0.04	0.03	0.01	0.03	0.35	0.66	0.53	0.60	0.38	0.33
		LargeLarge	0.12	0.18	0.17	0.25	0.21	0.08	0.06	-0.05	-0.04	-0.03	0.64	0.45	0.69	0.43	0.44
	4	NoneNone	0.13	0.14	0.16	0.21	0.17	-0.08	-0.06	-0.16	-0.17	-0.15	0.18	0.20	0.25	0.32	0.31
		NoneLarge	0.13	0.14	0.15	0.20	0.17	-0.33	-0.33	-0.19	-0.24	-0.25	0.36	0.37	0.26	0.34	0.34
		LargeNone	0.14	0.19	0.15	0.22	0.21	-0.18	-0.22	-0.17	-0.11	-0.15	0.76	0.38	0.77	0.43	0.40
		LargeLarge	0.13	0.20	0.14	0.21	0.21	-0.32	-0.33	-0.21	-0.10	-0.16	0.74	0.39	0.72	0.37	0.39
	5	NoneNone	0.15	0.17	0.16	0.21	0.19	-0.02	-0.01	-0.05	0.01	0.01	0.22	0.24	0.23	0.39	0.37
		NoneLarge	0.15	0.16	0.15	0.20	0.18	0.12	0.10	0.06	0.02	0.04	0.27	0.24	0.24	0.30	0.28
		LargeNone	0.16	0.19	0.20	0.23	0.21	0.14	0.24	-0.10	-0.10	0.05	0.79	0.47	0.81	0.40	0.44
		LargeLarge	0.17	0.18	0.18	0.23	0.23	-0.01	0.08	-0.01	-0.07	-0.03	0.84	0.43	0.84	0.42	0.41
10	3	NoneNone	0.12	0.15	0.13	0.20	0.16	-0.03	-0.01	-0.07	0.07	0.17	0.16	0.21	0.34	0.35	0.27
		NoneLarge	0.11	0.17	0.13	0.18	0.19	-0.16	-0.20	-0.20	-0.09	0.04	0.21	0.29	0.25	0.34	0.30
		LargeNone	0.12	0.15	0.19	0.19	0.19	0.39	0.54	0.16	0.23	0.01	0.74	0.35	0.71	0.28	0.30
		LargeLarge	0.12	0.20	0.12	0.20	0.19	-0.24	-0.25	-0.03	0.02	0.10	0.73	0.48	0.72	0.42	0.43
	4	NoneNone	0.14	0.15	0.14	0.22	0.18	-0.02	-0.01	-0.08	-0.28	-0.24	0.21	0.21	0.22	0.47	0.36
		NoneLarge	0.14	0.21	0.14	0.17	0.16	0.05	0.06	0.13	-0.01	0.03	0.20	0.28	0.23	0.28	0.30
		LargeNone	0.14	0.17	0.14	0.17	0.18	0.06	0.14	-0.05	-0.24	-0.06	0.73	0.49	0.73	0.37	0.39
		LargeLarge	0.16	0.21	0.17	0.20	0.19	-0.24	-0.21	-0.08	0.00	0.06	0.86	0.36	0.85	0.35	0.37
	5	NoneNone	0.17	0.22	0.14	0.22	0.20	-0.03	0.01	-0.01	-0.08	0.05	0.26	0.30	0.25	0.41	0.32
		NoneLarge	0.17	0.25	0.17	0.16	0.17	0.00	-0.03	0.00	0.29	0.21	0.30	0.37	0.24	0.38	0.35
		LargeNone	0.18	0.28	0.18	0.18	0.18	0.17	0.19	0.09	-0.09	0.10	0.85	0.48	0.85	0.41	0.42
		LargeLarge	0.17	0.23	0.22	0.17	0.17	0.01	0.06	0.16	0.06	0.16	0.80	0.47	0.57	0.47	0.37

Table A19: Standard errors, Bias and Root Mean Square Errors in the discriminant parameters for changing items and category options

Items	Categories	Condition	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP	GPCM	Testlet	Multilevel	Dual	DualDP
3	3	NoneNone	0.10	0.09	0.10	0.09	0.10	0.02	-0.15	-0.01	-0.14	-0.08	0.13	0.18	0.12	0.17	0.14
		NoneLarge	0.10	0.10	0.10	0.10	0.10	-0.01	-0.01	0.01	-0.01	0.01	0.12	0.17	0.12	0.16	0.14
		LargeNone	0.11	0.12	0.11	0.10	0.11	0.01	-0.06	-0.02	0.00	0.02	0.24	0.17	0.25	0.13	0.17
		LargeLarge	0.11	0.11	0.09	0.10	0.11	0.02	-0.06	-0.04	0.02	0.00	0.23	0.17	0.22	0.15	0.17
	4	NoneNone	0.09	0.08	0.09	0.08	0.10	-0.02	-0.03	-0.08	-0.03	-0.03	0.12	0.17	0.13	0.20	0.15
		NoneLarge	0.12	0.13	0.10	0.10	0.10	0.13	-0.19	0.10	-0.06	-0.04	0.18	0.24	0.15	0.13	0.13
		LargeNone	0.12	0.14	0.11	0.11	0.11	0.11	0.01	0.12	0.01	-0.01	0.32	0.19	0.32	0.17	0.14
		LargeLarge	0.10	0.11	0.10	0.10	0.11	0.03	0.08	0.02	-0.02	0.03	0.24	0.16	0.24	0.14	0.16
	5	NoneNone	0.15	0.10	0.14	0.10	0.11	0.10	-0.03	0.12	-0.02	0.02	0.24	0.13	0.20	0.13	0.14
		NoneLarge	0.20	0.10	0.13	0.17	0.10	0.04	-0.06	0.08	0.03	-0.03	0.27	0.22	0.26	0.25	0.17
		LargeNone	0.12	0.11	0.13	0.12	0.11	0.10	0.03	0.05	0.04	0.05	0.39	0.19	0.38	0.16	0.16
		LargeLarge	0.12	0.11	0.13	0.13	0.11	-0.02	-0.15	-0.04	0.01	0.02	0.35	0.22	0.34	0.16	0.16
6	3	NoneNone	0.11	0.08	0.09	0.08	0.08	0.01	-0.02	-0.01	-0.02	-0.02	0.14	0.12	0.11	0.13	0.14
		NoneLarge	0.09	0.08	0.09	0.08	0.08	-0.05	-0.06	0.00	-0.03	-0.03	0.13	0.13	0.12	0.14	0.13
		LargeNone	0.12	0.09	0.10	0.09	0.09	0.07	0.00	0.00	-0.01	-0.01	0.33	0.12	0.28	0.11	0.12
		LargeLarge	0.11	0.10	0.09	0.09	0.09	0.01	0.03	-0.05	-0.02	-0.04	0.22	0.15	0.19	0.13	0.16
	4	NoneNone	0.09	0.09	0.09	0.08	0.09	0.01	-0.09	-0.02	-0.11	-0.10	0.14	0.15	0.13	0.16	0.16
		NoneLarge	0.08	0.08	0.08	0.08	0.08	0.07	0.00	0.07	-0.02	-0.02	0.13	0.11	0.13	0.11	0.14
		LargeNone	0.10	0.10	0.10	0.09	0.10	0.06	0.03	0.05	0.01	0.06	0.23	0.17	0.23	0.15	0.15
		LargeLarge	0.09	0.09	0.09	0.09	0.09	0.00	0.03	-0.02	0.00	0.01	0.27	0.13	0.27	0.12	0.174
	5	NoneNone	0.09	0.09	0.09	0.08	0.08	0.07	-0.03	0.05	-0.09	-0.11	0.15	0.14	0.14	0.16	0.15
		NoneLarge	0.09	0.09	0.09	0.08	0.09	0.13	0.06	0.12	0.01	-0.02	0.16	0.12	0.16	0.11	0.11
		LargeNone	0.10	0.10	0.10	0.09	0.09	0.10	0.18	0.06	0.10	-0.02	0.30	0.21	0.29	0.15	0.16
		LargeLarge	0.10	0.11	0.09	0.10	0.10	0.09	0.07	0.08	0.05	0.02	0.28	0.21	0.28	0.19	0.18
10	3	NoneNone	0.09	0.10	0.08	0.09	0.09	0.00	-0.01	-0.01	-0.01	0.01	0.12	0.17	0.29	0.13	0.12
		NoneLarge	0.08	0.09	0.08	0.08	0.09	0.03	-0.06	0.02	0.05	0.05	0.11	0.15	0.11	0.11	0.14
		LargeNone	0.10	0.10	0.10	0.09	0.09	0.04	0.05	0.01	0.07	0.04	0.17	0.14	0.16	0.13	0.12
		LargeLarge	0.09	0.12	0.09	0.10	0.09	0.03	-0.03	0.02	0.02	0.00	0.21	0.16	0.21	0.13	0.12
	4	NoneNone	0.10	0.09	0.10	0.09	0.09	0.11	0.04	0.09	0.02	0.03	0.16	0.13	0.15	0.13	0.13
		NoneLarge	0.10	0.11	0.09	0.10	0.09	0.06	-0.09	0.07	0.03	0.03	0.14	0.16	0.13	0.13	0.12
		LargeNone	0.10	0.10	0.09	0.10	0.10	0.08	0.05	0.04	0.03	0.03	0.31	0.16	0.29	0.14	0.12
		LargeLarge	0.10	0.09	0.10	0.09	0.09	0.05	-0.03	0.06	0.05	0.04	0.27	0.19	0.28	0.14	0.14
	5	NoneNone	0.10	0.10	0.10	0.10	0.10	0.17	-0.04	0.14	0.04	0.05	0.22	0.16	0.19	0.17	0.16
		NoneLarge	0.14	0.15	0.11	0.09	0.09	0.23	-0.04	0.14	0.04	0.04	0.28	0.17	0.18	0.13	0.13
		LargeNone	0.13	0.11	0.12	0.11	0.10	0.22	0.00	0.17	0.20	0.11	0.16	0.20	0.33	0.14	0.15
		LargeLarge	0.12	0.16	0.13	0.12	0.10	0.15	0.10	0.15	0.09	0.08	0.36	0.21	0.25	0.14	0.15

Table A20: Ability, group and interaction variances for models mis-specifying ability parameter distribution

			Bimodal			Skewed			Uniform			Normal		
LID	LPD	Model	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ	σ^2_θ	σ^2_δ	σ^2_γ
None	None	GPCM	1.01			0.74			1.22			0.90		
		Testlet	1.19		0.03	1.02		0.03	1.31		0.03	1.04		0.03
		Multilevel	1.07	0.10		0.91	0.10		1.19	0.11		0.94	0.09	
		Dual	1.27	0.11	0.03	1.06	0.11	0.03	1.42	0.12	0.03	1.20	0.10	0.03
		DualDP	1.21	0.11	0.03	1.05	0.10	0.03	1.31	0.11	0.03	1.14	0.12	0.03
None	Large	GPCM	1.75			1.73			1.72			1.66		
		Testlet	2.27		0.04	2.07		0.03	2.03		0.04	1.93		0.04
		Multilevel	1.02	1.20		0.88	1.08		0.96	1.08		0.97	0.90	
		Dual	1.11	1.28	0.03	1.03	1.23	0.03	1.10	1.17	0.03	1.17	1.12	0.03
		DualDP	1.07	1.23	0.03	1.05	1.11	0.03	1.13	1.21	0.03	1.06	1.08	0.04
Large	None	GPCM	0.51			0.35			0.48			0.50		
		Testlet	1.25		1.15	1.28		1.86	1.09		1.39	1.44		1.31
		Multilevel	0.48	0.16		0.33	0.16		0.46	0.14		0.45	0.14	
		Dual	1.11	0.18	1.76	0.85	0.19	2.00	0.97	0.15	1.56	0.89	0.11	1.67
		DualDP	1.11	0.11	1.75	0.89	0.08	1.98	0.91	0.13	1.64	0.89	0.13	1.57
Large	Large	GPCM	1.01			0.89			1.11			0.74		
		Testlet	1.74		1.43	1.78		1.41	1.73		1.3	1.94		1.5
		Multilevel	0.44	0.79		0.43	0.70		0.50	0.74		0.43	0.5	
		Dual	0.88	1.88	1.58	0.95	1.80	1.60	0.96	1.49	1.39	1.11	1.76	1.75
		DualDP	0.93	1.91	1.58	1.01	1.93	1.67	0.94	1.53	1.41	1.05	1.76	1.64

Table A21: SE, Bias and RMSE in the ability for different ability distributions

		Standard error				**Bias**				**RMSE**			
Condition	Model	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform
NoneNone	GPCM	0.23	0.23	0.21	0.25	0.03	0.01	-0.01	-0.02	0.31	0.31	0.32	0.33
	Testlet	0.25	0.24	0.25	0.26	0.03	0.02	0.00	-0.01	0.33	0.32	0.33	0.35
	Multilevel	0.28	0.28	0.28	0.30	0.03	0.02	-0.01	-0.01	0.37	0.36	0.37	0.40
	Dual	0.31	0.30	0.30	0.32	0.04	0.00	-0.03	-0.03	0.41	0.38	0.39	0.43
	Dual DP	0.28	0.27	0.27	0.29	0.03	0.00	-0.02	-0.02	0.36	0.35	0.36	0.38
NoneLarge	GPCM	0.25	0.29	0.25	0.26	0.01	0.02	0.02	0.01	0.83	0.88	0.85	0.86
	Testlet	0.29	0.30	0.28	0.29	-0.01	0.01	0.02	0.01	0.93	0.85	0.91	0.91
	Multilevel	0.33	0.34	0.31	0.32	-0.02	0.01	0.00	-0.01	0.43	0.45	0.44	0.45
	Dual	0.34	0.36	0.33	0.35	-0.01	0.00	0.00	-0.01	0.44	0.48	0.46	0.47
	Dual DP	0.31	0.33	0.31	0.32	0.00	0.01	0.01	0.00	0.49	0.53	0.51	0.52
LargeNone	GPCM	0.20	0.18	0.18	0.20	0.00	-0.02	0.01	-0.01	0.48	0.49	0.49	0.50
	Testlet	0.27	0.26	0.26	0.26	0.01	-0.02	0.03	-0.01	0.56	0.45	0.66	0.56
	Multilevel	0.24	0.21	0.22	0.24	0.00	-0.02	0.01	-0.01	0.44	0.47	0.47	0.47
	Dual	0.32	0.31	0.31	0.32	0.00	-0.02	0.00	-0.02	0.42	0.42	0.42	0.43
	Dual DP	0.29	0.29	0.29	0.29	0.00	-0.02	0.00	-0.01	0.40	0.41	0.42	0.43
LargeLarge	GPCM	0.23	0.24	0.21	0.23	0.03	0.01	0.05	-0.02	0.80	0.79	0.77	0.74
	Testlet	0.28	0.30	0.29	0.28	0.04	0.01	0.07	-0.02	0.96	0.76	0.67	0.81
	Multilevel	0.26	0.28	0.25	0.26	0.03	0.05	0.03	-0.01	0.50	0.49	0.49	0.48
	Dual	0.33	0.35	0.34	0.32	0.02	0.05	0.03	-0.01	0.47	0.47	0.46	0.43
	Dual DP	0.30	0.33	0.32	0.30	0.03	0.03	0.04	-0.01	0.52	0.50	0.50	0.48

Table A22: SE, Bias and RMSE in the threshold parameter for different ability distributions

		Standard error				**Bias**				**RMSE**			
Condition	Model	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform
NoneNone	GPCM	0.11	0.12	0.14	0.11	0.02	0.02	0.15	-0.03	0.14	0.16	0.23	0.15
	Testlet	0.12	0.12	0.10	0.11	0.01	0.01	-0.07	-0.03	0.16	0.16	0.15	0.15
	Multilevel	0.12	0.11	0.10	0.12	0.02	0.04	-0.01	0.02	0.15	0.16	0.14	0.14
	Dual	0.22	0.21	0.09	0.22	-0.02	0.17	-0.09	0.15	0.44	0.42	0.16	0.52
	Dual DP	0.17	0.17	0.09	0.17	-0.01	0.12	-0.01	0.09	0.31	0.31	0.14	0.35
NoneLarge	GPCM	0.11	0.12	0.11	0.11	0.05	0.07	0.11	-0.09	0.18	0.17	0.17	0.17
	Testlet	0.13	0.12	0.09	0.12	0.03	0.06	-0.04	-0.12	0.17	0.17	0.11	0.18
	Multilevel	0.16	0.14	0.09	0.16	0.02	0.01	0.02	0.03	0.20	0.18	0.11	0.19
	Dual	0.20	0.23	0.09	0.24	-0.12	0.09	-0.06	-0.07	0.35	0.45	0.12	0.37
	Dual DP	0.16	0.17	0.09	0.18	-0.06	0.08	-0.01	-0.08	0.26	0.33	0.10	0.27
LargeNone	GPCM	0.13	0.16	0.15	0.13	0.11	0.40	0.14	0.19	0.64	0.66	0.41	0.65
	Testlet	0.17	0.19	0.10	0.18	0.10	0.43	0.07	0.22	0.67	0.53	0.16	0.66
	Multilevel	0.14	0.12	0.11	0.13	0.07	0.41	0.02	0.21	0.66	0.60	0.32	0.68
	Dual	0.22	0.21	0.09	0.22	0.20	0.33	0.02	0.19	0.53	0.58	0.14	0.54
	Dual DP	0.20	0.20	0.10	0.20	0.17	0.35	0.06	0.19	0.47	0.57	0.20	0.50
LargeLarge	GPCM	0.13	0.12	0.13	0.12	0.19	0.18	0.19	0.07	0.68	0.64	0.27	0.61
	Testlet	0.18	0.18	0.10	0.17	0.29	0.16	0.09	0.11	0.56	0.52	0.18	0.56
	Multilevel	0.15	0.17	0.10	0.15	0.22	-0.10	0.04	-0.02	0.70	0.69	0.19	0.62
	Dual	0.26	0.25	0.10	0.28	0.23	-0.21	-0.01	0.02	0.59	0.59	0.13	0.40
	Dual DP	0.22	0.22	0.10	0.22	0.21	-0.08	0.06	0.04	0.51	0.53	0.14	0.39

Table A23: SE, Bias and RMSE for discriminant parameter estimates for changing ability distributions

		Standard error				**Bias**				**RMSE**			
Condition	Model	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform	Bimodal	Normal	Skewed	Uniform
NoneNone	GPCM	0.10	0.11	0.14	0.11	-0.01	0.08	0.15	-0.05	0.13	0.15	0.23	0.13
	Testlet	0.10	0.09	0.10	0.08	-0.11	-0.04	-0.07	-0.09	0.16	0.11	0.15	0.13
	Multilevel	0.09	0.09	0.10	0.08	-0.06	0.03	-0.01	-0.05	0.12	0.12	0.14	0.11
	Dual	0.08	0.05	0.09	0.08	-0.14	-0.07	-0.09	-0.14	0.17	0.10	0.16	0.17
	Dual DP	0.09	0.07	0.09	0.08	-0.10	-0.02	-0.01	-0.11	0.15	0.09	0.14	0.14
NoneLarge	GPCM	0.11	0.12	0.11	0.10	0.14	0.09	0.11	0.12	0.20	0.17	0.17	0.18
	Testlet	0.09	0.09	0.09	0.08	-0.06	-0.07	-0.04	-0.01	0.13	0.13	0.11	0.12
	Multilevel	0.09	0.09	0.09	0.09	0.00	-0.01	0.02	0.05	0.12	0.12	0.11	0.12
	Dual	0.08	0.08	0.09	0.08	-0.05	-0.12	-0.06	-0.02	0.12	0.16	0.12	0.11
	Dual DP	0.09	0.09	0.09	0.08	0.02	-0.05	-0.01	0.03	0.12	0.12	0.10	0.12
LargeNone	GPCM	0.14	0.11	0.15	0.13	0.06	-0.01	0.14	0.09	0.27	0.19	0.41	0.22
	Testlet	0.10	0.09	0.10	0.10	0.05	0.04	0.07	0.06	0.15	0.12	0.16	0.15
	Multilevel	0.10	0.10	0.11	0.10	-0.03	-0.03	0.02	0.00	0.23	0.17	0.32	0.18
	Dual	0.09	0.09	0.09	0.10	-0.05	0.02	0.02	-0.01	0.13	0.11	0.14	0.13
	Dual DP	0.10	0.09	0.10	0.10	-0.01	0.01	0.06	0.03	0.15	0.12	0.20	0.15
LargeLarge	GPCM	0.15	0.13	0.13	0.11	0.20	0.15	0.19	0.07	0.32	0.33	0.27	0.24
	Testlet	0.10	0.10	0.10	0.09	0.13	0.11	0.09	0.07	0.18	0.17	0.18	0.14
	Multilevel	0.10	0.09	0.10	0.09	0.08	0.06	0.04	0.03	0.23	0.27	0.19	0.21
	Dual	0.09	0.09	0.10	0.09	0.04	0.07	-0.01	0.03	0.13	0.13	0.13	0.12
	Dual DP	0.10	0.09	0.10	0.09	0.10	0.10	0.06	0.04	0.17	0.18	0.14	0.15

Appendix B: Model codes

R data generating code

```
personL <- rnorm(1000,0,1);np<-length(personL)
thresholds <- t(apply(matrix(rnorm(36*2, 0, 1), 36), 1, cumsum))
thresholds <- -(thresholds - rowMeans(thresholds));
thresholds <- thresholds + rnorm(36);thresholds <- thresholds*-1
difficulty = rowMeans(thresholds);step1 = thresholds[,1] - difficulty
step2 = thresholds[,2] - difficulty; step3 = thresholds[,3] - difficulty
steps <- cbind(difficulty,step1, step2);beta<-as.vector(t(cbind(step1,step2)))
beta4<-as.vector(t(cbind(thresholds[,1],thresholds[,2])))
itema <-rlnorm(36,0.2,0.2);ni<-nrow(steps);na<-length(itema);
gamm<- rnorm(240,0,1);gammam<- matrix(gamm,ncol=6,byrow=T);
#tdif<-rnorm(6,0,1);lpd<-rnorm(40,0,1);group<-40;K<-3;totmatrix<-np*ni;
lpdL<-rep(lpd,each=25);personL4<- personL + lpdL;
personL5<-matrix(personL4,nrow=40,byrow=T)
responseL <- matrix(rep(NA, totmatrix), np, ni);for (gg in 1:40){
thetag<-lpd[gg];for(pp in 1:25){theta<-personL5[gg,pp];#theta<-personL1[gg,pp]
 + lpd[gg];
print(pp);np1<-25*(gg-1)+pp;for (tt in 1:6){bet<-tdif[tt];gamma<-gammam[gg,tt];
for (ii in 1:6){nitem<-6*(tt-1)+ii;alpha<-itema[nitem];bb <-difficulty[nitem];
measure=0;p <- vector();p[1] <- 1;for(k in 2:3){dd<-steps[nitem,k];
measure<-measure+ alpha*(theta+gamma-bb-dd);p[k] <- p[(k-1)] + exp(measure)};
U <- runif(1, 0, 1);U = U * p[3];for(k in 1:3) {if(U <= p[k])
{responseL[np1,nitem] <- (k);break}}}}}} responseL#transposing data
 for use in WinBUGS format;responseL1<-t(responseL)
```

Generalised Partial Credit Model

```
model{for (j in 1:J) {for (i in 1:I) {r[j,i]<-response[j,i];
r[j, i] ~ dcat(prob[j, i, 1:K[i]])};theta[j]~dnorm(0.00E+00,tau)}
for (j in 1:J) {for (i in 1:I) {for (k in 1:K[i]) {
eta[j, i, k] <- alpha[i] * (theta[j]-beta[i, k])
psum[j, i, k] <- sum(eta[j, i, 1:k])
exp.psum[j, i, k] <- exp(psum[j, i, k])
prob[j, i, k] <- exp.psum[j, i, k]/sum(exp.psum[j,i,1:K[i]])}}}
for (i in 1:I) {alpha[i] ~ dlnorm(m.alpha, pr.alpha)
beta[i, 1] <- 0.00000E+00;for (k in 2:K[i]) {
beta[i, k] ~ dnorm(m.beta, pr.beta)}};tau ~ dgamma(1, 1)
var <- 1/tau ;pr.alpha <- pow(s.alpha, -2);pr.beta <- pow(s.beta,-2)
m.alpha<-0.2;s.alpha<-0.2;m.beta<-0;s.beta<-1 ;for (j in 1:J) {
```

```
for (i in 1:I) {lik[j,i]<-   log(prob[j,i,r[j,i]])}};
loglik <-  sum(lik[1:J,1:I])AIC <-  -2*(loglik - np);
BIC <-  -2*loglik + np*log(J)
#Data
list(I=36,J=1000, np=, K=c(3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3,3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3),
response=structure(.Data=c(), .Dim = c(1000,36)))
```

Testlet GPCM model

```
model{
for (j in 1:J) {theta[j] ~ dnorm(0.00000E+00, tau)
for (i in 1:I) {r[j, i] <- response[j, i]
for (k in 1:K[i]) {r[j, i] ~ dcat(prob[j, i, 1:K[i]])}}}
for (j in 1:J) {for (i in 1:I) {for (k in 1:K[i]) {
eta[j, i, k] <- alpha[i] * (theta[j] + gamma[group[j],test[i]]-beta[i,k])
psum[j, i, k] <- sum(eta[j, i, 1:k]);exp.psum[j, i, k] <- exp(psum[j,i,k])
prob[j, i, k] <- exp.psum[j, i, k]/sum(exp.psum[j, i, 1:K[i]])}}}
for (g in 1:G){ gamma[g, 1] <- 0.00000E+00;for( t in 2:T){
gamma[g, t] ~ dnorm(0.00000E+00, taug)}}for (i in 1:I) {
alpha[i] ~ dlnorm(m.alpha, pr.alpha);beta[i,1]<-0.00000E+00
for (k in 2:K[i]) {beta[i, k] ~ dnorm(m.beta, pr.beta)}} ;tau ~ dgamma(1,1)
var <- 1/tau;taug ~dgamma(1, 1);varg <- 1/taug;pr.alpha <- pow(s.alpha,-2)
pr.beta <- pow(s.beta, -2);m.alpha<-0.2;s.alpha<-0.2;m.beta<-0;s.beta<-1
for (t in 1:T) {sigma.gamma[t] ~ dunif(0.00000E+00, 10)
pr.gamma[t] <- pow(sigma.gamma[t], -2)}
for (j in 1:J) {for (i in 1:I) {   lik[j,i]<- log(prob[j,i,r[j,i]])}}
loglik <-  sum(lik[1:J,1:I]);AIC <-  -2*(loglik - np)
BIC <-  -2*loglik + np*log(J)}
list(I=36,J=1000, np=80,T=6,test=c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,
4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,6,6,6),K=c(3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3,3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3),group=c(), response=structure(.Data=c(), .Dim = c(1000,36)))
```

Multilevel model

```
model{for(j in1:J){theta[j]~dnorm(0.00000E+00, tau);for(i in 1:I){
r[j,i]<-response[j,i]}}
## Specify Mixture Model;for(j in 1:J){for(i in 1:I){for(k in 1:K[i]){
eta[j, i, k] <- alpha[i] * (theta[j] + delta[group[j]]-beta[i,k])
psum[j, i, k] <- sum(eta[j,i,1:k]);exp.psum[j,i,k] <- exp(psum[j,i,k])
p[j, i, k] <- exp.psum[j, i, k]/sum(exp.psum[j,  i, 1:K[i]])}
r[j, i] ~ dcat(p[j, i, 1:K[i]])}}
#### Specify Priors for Item and Class Parameters
for (i in 1:I) {alpha[i]~dlnorm(m.alpha,pr.alpha)
beta[i,1]<- 0.00E+00;for (k in 2:K[i]){beta[i,k]~dnorm(0,1)}
for (g in 1:G) {delta[g] ~ dnorm(0, taug)}tau ~ dgamma(1,1)
var <- 1/tau ;taug ~ dgamma(1,1);varg<-1 / taug
pr.alpha <- pow(s.alpha, -2);m.alpha<-0.2;s.alpha<-0.2;for(j in 1:J){
```

```
for (i in 1:I) {lik[j,i]<-log(p[j,i,r[j,i]])}}loglik<-sum(lik[1:J,1:I])
AIC <-  -2*(loglik - np)BIC <-  -2*loglik + np*log(J)}
list(I=36,J=1000, np=80,K=c(3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3),
group=c(),response=structure(.Data=c(), .Dim = c(1000,36)))
```

Dual GPCM model

```
model{for (j in 1:J) {theta[j] ~ dnorm(0.00000E+00, tau)
for (i in 1:I) {r[j, i] <- response[j, i]}}for (j in 1:J) {
for (i in 1:I) { for (k in 1:K[i]) {
eta[j, i, k] <- alpha[i] * (theta[j] + delta[group[j]] +
gamma[group[j], test[i]] - bet[test[i]] - beta[i,k])
psum[j, i, k] <- sum(eta[j, i, 1:k])
exp.psum[j, i, k] <- exp(psum[j, i, k])
p[j, i, k] <- exp.psum[j, i, k]/sum(exp.psum[j,  i, 1:K[i]])}
r[j, i] ~ dcat(p[j, i, 1:K[i]])}},#delta[1]<-0.00000E+00
for (g in 1:G){ delta[g] ~ dnorm(0.00000E+00, taud)  }
for( g in 1:G){gamma[g, 1] <- 0.00000E+00for( t in 2:T){
gamma[g, t] ~ dnorm(0.00000E+00, taug)}}
#bet[1]<- 0.00000E+00for(t in 1:T){
bet[t] ~ dnorm(0.00000E+00, taub)}mu ~ dnorm(0.00000E+00, 1)
tau ~ dgamma(1, 1);var <- 1/tau ;taub ~ dgamma(1, 1)
varb <- 1/taub ;taud ~dgamma(1, 1);vard <- 1/taud
taug ~dgamma(1, 1);varg <- 1/taug;for (i in 1:I) {
alpha[i]~dlnorm(m.alpha,pr.alpha)};pr.alpha<-pow(s.alpha,-2)
m.alpha<-0.2;s.alpha<- 0.2;vara<-1/pr.alpha;for (i in 1:I ) {
beta[i, 1] <- 0.00E+00;for (k in 2:K[i]) {
beta[i,k]~dnorm(m.beta, pr.beta) }}pr.beta <- pow(s.beta, -2)
m.beta<-0;s.beta<-1;varbeta<-1/pr.betafor (j in 1:J) {
for (i in 1:I) {lik[j, i] <- log(p[j, i, r[j, i]])}}
loglik <- sum(lik[1:J, 1:I]);AIC <- -2 * (loglik - np)
BIC <- -2 * loglik + np * log(J)}
list(I=36,J=1000, np=80,T=6,test=c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,
4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,6,6,6),
K=c(3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,3,
 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3),
group=c(),response=structure(.Data=c(), .Dim = c(1000,36)))
```

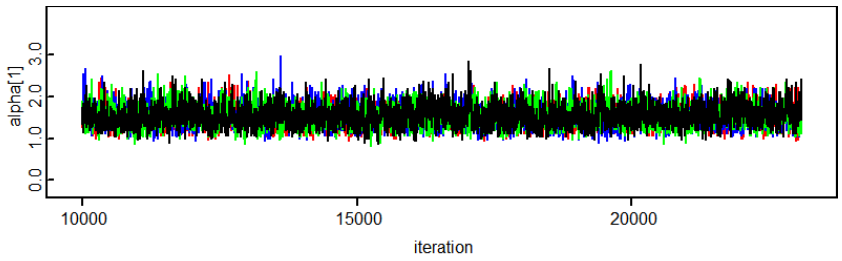
Dual Dirichlet Process Mixture model

```
model{for (j in 1:J) {theta[j] ~dnorm(mut[j],taut[j])group[j] ~ dcat(pj[])
for (i in 1:I) {r[j, i] <- response[j, i]}}for (j in 1:J) {for (i in 1:I){
for (k in 1:K[i]) {eta[j, i, k] <- alpha[i] * (theta[j] + delta[group[j]]
+gamma[group[j],test[i]] - beta[i,k])psum[j, i, k] <- sum(eta[j, i, 1:k])
exp.psum[j,i,k]<-exp(psum[j,i,k])p[j,i,k]<-exp.psum[j,i,k]
/sum(exp.psum[j,i,1:K[i]])}r[j, i] ~ dcat(p[j, i, 1:K[i]])}for(g in 1:G){
SC[j, g] <- equals(g, group[j])}}for(j in 1:J){mut[j] <- mu[group[j]]
taut[j]<-tau[group[j]]}for(gin 1:G){mu[g]<-0.00000E+00tau[g]~dgamma(1,1)
```

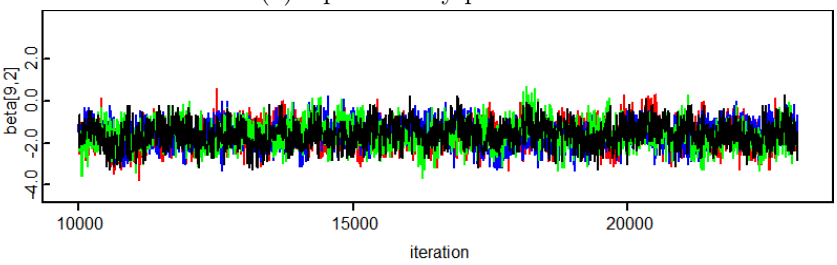
```
var[g]<-1/tau[g]delta[g]~dnorm(0.00000E+00,taud)gamma[g, 1]<-0.00000E+00
for(t in 2:T){gamma[g, t] ~ dnorm(0.00000E+00, taug)}}# Precision Parameter
alphadp <- 1# Constructive DPP for (g in 1:G){q[g] ~ dbeta(1, alphadp)
pj[g] <- pd[g] / p.sum} pd[1] <- q[1]for (g in 2 : G) { pd[g] <- q[g] *
(1 - q[g - 1]) * pd[g -1 ] / q[g - 1]}# scaling to ensure sum to 1
p.sum <- sum(pd[])# total clusters K.star <- sum(cl[])for (g in 1 : G) {
sumSC[g] <- sum(SC[ , g])cl[g] <- step(sumSC[g] -1)}for (i in 1:I) {
alpha[i] ~ dlnorm(m.alpha, pr.alpha)};pr.alpha <- pow(s.alpha, -2)
m.alpha<-0.2;s.alpha<-0.2; for (i in 1:I ) {beta[i, 1] <- 0.00000E+00
for (k in 2:K[i]) {beta[i,k] ~ dnorm(0.00000E+00, 1) }}taud ~ dgamma(1, 1);
vard <- 1/taud; taug ~ dgamma(1, 1);varg <- 1/taug;taub ~ dgamma(1, 1)
varb <- 1/taug; for (j in 1:J) {for (i in 1:I) {
lik[j, i] <- log(p[j, i, r[j, i]])}loglik <- sum(lik[1:J, 1:I])
AIC <- -2 * (loglik - np);BIC <- -2 * loglik + np * log(J)}
```

Appendix C: Convergence checks graphs

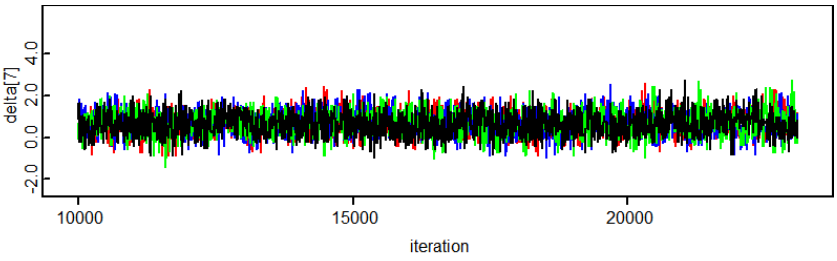
Figure C.1: History plots for convergence checks



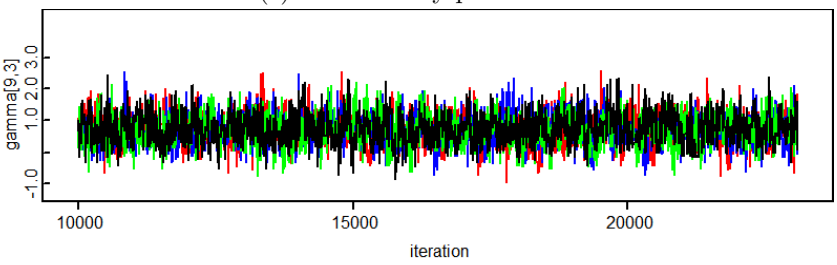
(a) alpha history plot



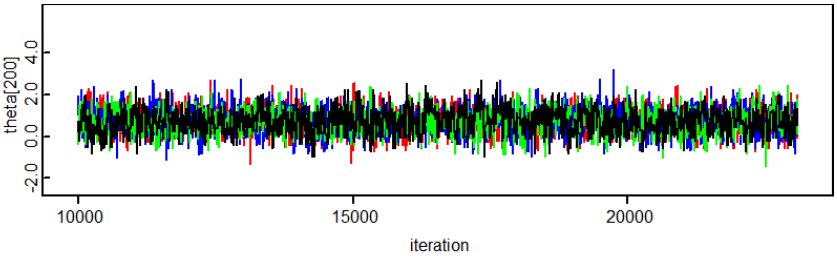
(b) beta history plot



(c) delta history plot

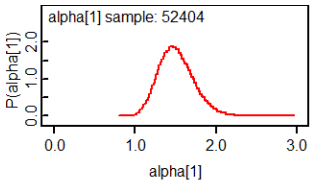


(d) gamma history plot

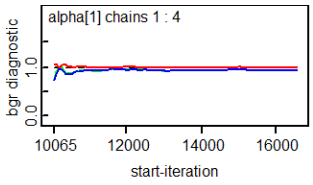


(e) theta history plot

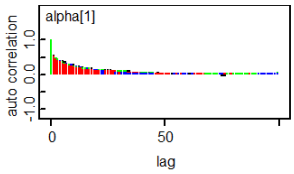
Figure C.2: History plots for convergence checks



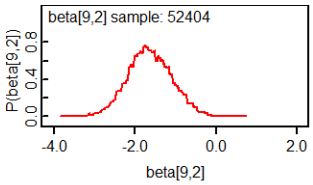
(a) alpha density



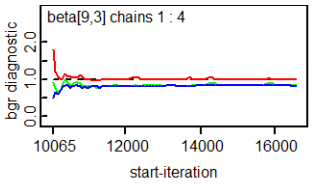
(b) alpha BGR



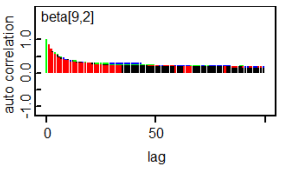
(c) alpha autocorrelation



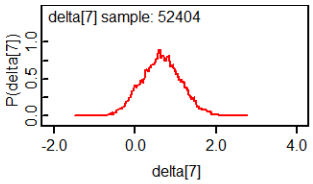
(d) beta density



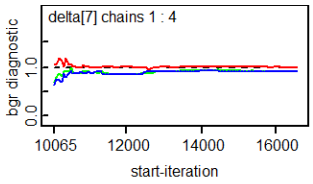
(e) beta BGR



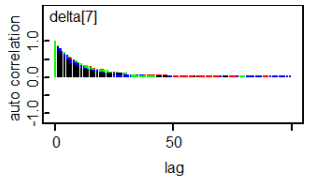
(f) beta autocorrelation



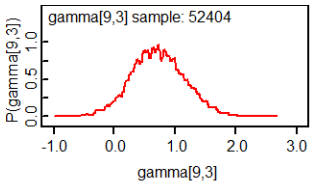
(g) delta density



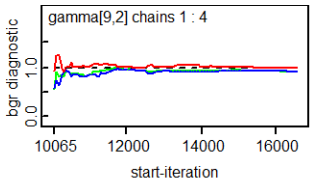
(h) delta BGR



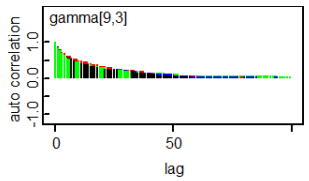
(i) deltaautocorrelation



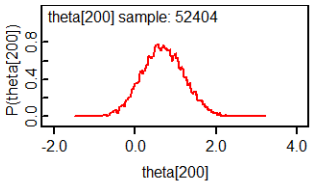
(j) gamma density



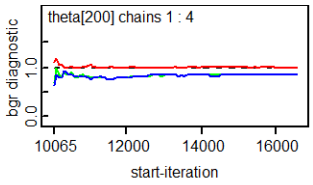
(k) gamma BGR



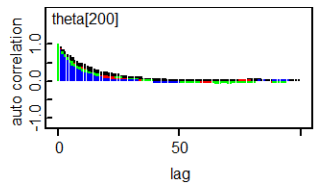
(l) gamma autocorrelation



(m) theta density



(n) theta BGR



(o) theta autocorrelation

Figure C.3: Density, BGR and autocorrelation plots for convergence checks

Appendix D: HCP Permission letter

BALSILLIE SCHOOL
OF INTERNATIONAL AFFAIRS
67 Erb Street West, Waterloo, Ontario Canada N2L 6C2

8 June 2021

To Whom It May Concern:

Ms Vonai Chiramba has permission to utilize the data from a household food security survey conducted by the Hungry Cities Partnership in partnership with University of Namibia in 2015 for her doctoral dissertation at the University. Please ensure that the funders of the survey are fully acknowledged in the dissertation as follows:

> The 2015 Hungry Cities Partnership household food security survey in Windhoek was funded by the International Partnerships in Sustainable Societies (IPaSS) Program of the International Development Research Centre (IDRC) and the Social Sciences and Humanities Research Council of Canada (SSHRC), as well as the Open Society Foundation (South Africa).

In addition, please send an electronic copy of the finalized dissertation to me at jcrush@balsillieschool.ca

Sincerely,

Professor Jonathan Crush