SHA, SHUYING, Ph.D. Nonparametric Diagnostic Classification Analysis for Testlet
Based Tests. (2016)
Directed by Dr. Robert. A. Henson. 121pp.

Diagnostic classification Diagnostic Classification Models (DCMs) are multidimensional confirmatory latent class models that can classify individuals into different classes based on their attribute mastery profiles. While DCMs represent the more prevalent parametric approach to diagnostic classification analysis, the Hamming distance method, a newly developed nonparametric diagnostic classification method, is quite promising in that it does not require fitting a statistical model and is less demanding on sample size. However, both parametric and nonparametric approach have assumptions of local item independency, which is often violated by testlet based tests. This study proposed a conditional-correlation based nonparametric approach to assess testlet effect and a set of testlet Hamming distance methods to account for the testlet effects in classification analyses. Simulation studies were conducted to evaluate the proposed methods.

In the conditional-correlation approach, the testlet effects were computed as the average item-pair correlations within the same testlet by conditioning on attribute profiles. The inverse of the testlet effect was then used in testlet Hamming distance method to weight the Hamming distances for that particular testlet.

Simulation studies were conducted to evaluate the proposed methods in conditions with varying sample size, testlet effect size, testlet size, balance of testlet size, and balance of testlet effect size. Although the conditional-correlation based approach often underestimated true testlet effect sizes, it was still able to detect the relative size of different testlet effects. The developed testlet Hamming distance methods seem to be an improvement over the estimation methods that ignore testlet effects because they provided slightly higher classification accuracy where large testlet effects were present. In addition, Hamming distance method and maximum likelihood estimation are robust to local item dependency caused by low to moderate testlet effects. Recommendations for practitioners and study limitations were provided.

NONPARAMETRIC DIAGNOSTIC CLASSIFICATION ANALYSIS

FOR TESTLET BASED TESTS

by

Shuying Sha

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2016

Approved by

____
Committee Chair

APPROVAL PAGE

This dissertation written by Shuying Sha has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ____________________________________

Committee Members ____________________________________

____

____

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Most current large scale assessments provide a single score regarding an examinee's unidimensional ability. However, there is an increasing demand for diagnostic information about the examinee's specific skills and attributes. The test takers and stakeholders need such information to inform their learning and classroom instruction. Classical test theory and item response theory generally order people on a latent trait. These approaches are typically not useful in identifying skills and attributes that are mastered or not mastered by examinees. Diagnostic classification models (DCMs) have been developed to measure specific skills and knowledge, and thus provide information about the examinee's strengths and weaknesses in a related cognitive domain (Dibello et al, 1995; Junker & Sijtsma, 2001; Hartz, 2002; de la Torre & Douglas, 2004; Henson & Templin, 2007; Von Davier, 2008; Rupp,Templin & Henson, 2010).

A large number of diagnostic classification models have been developed in order to describe the correspondence between individuals' responses and the underlying attributes or skills that are required to correctly answer the items in a test. Most diagnostic models are constrained latent class models, in which the individuals' proficiency is described in terms of discrete attributes. Individuals are evaluated as either having mastered or not mastered each set of skills. Based on his/her mastery profile of the skills, the examinee is classified into a specific category. For example, a Number

Subtraction test measures four attributes: convert a whole number to a fraction, separate a whole number from a fraction, find a common denominator, borrow from whole number part, the individual will be classified into one of the 2^4 =16 categories based on the set of skills that have been mastered.

Different diagnostic classification models make different assumptions about how attributes are used to construct item responses. Conjunctive models assume that all measured attributes are required to correctly answer an item, whereas disjunctive models assume that only one attribute needs to be mastered in order to have a high probability of giving a correct answer to the item.

Although diagnostic classification models are gathering increased research interest and have been applied in a large number of studies such as mathematical skill diagnosis (Tatsuoka, 1983; Hartz, 2002; de la Torre & Douglas, 2004; Henson, Templin & Willse, 2009), language skill diagnosis test (Jang, 2008, 2009; Von Davier, 2008), and pathological diagnosis (Templin & Henson, 2006), they have some disadvantages. For example, diagnostic classification models heavily rely on maximum likelihood estimation (MLE) procedure with expectation maximization (EM) or Markov Chain Monte Carlo (MCMC) for model estimation. A large sample size is typically required for these estimation methods to obtain accurate parameter estimation, examinee classification and model fit testing. The necessity of a large sample size limits the application of DCM. In addition, there are always concerns that the models that are applied in diagnosis classification analysis are either incorrect or do not fit. In response to those obstacles caused by sample-size limitation and model selection in traditional diagnostic modeling, nonparametric diagnostic classification methods were developed as approximation methods to classify examinees into different attribute mastery profiles (Willse, Henson, & Templin, 2007; Ayers, Nugent, & Dean, 2008; Chiu, 2008; Chiu, Douglas, & Li, 2009; Park & Lee, 2011; Chiu & Douglas, 2013;Wang,& Douglas, 2015). Compared to the parametric methods, nonparametric methods have no requirement for large sample sizes because no parameters are estimated and they make no assumptions about population distribution.

Though parametric and nonparametric classification methods are different, the classification accuracy of both are challenged by local dependencies that exist among items because both assume or treat the items in a test as being independent from each other. Local item dependencies (LID) can come from multiple sources. In this study, the specific focus is on the LID caused by testlets or item grouping.

A testlet is a section of the test that is comprised of a group of items based on the same stimuli or shared passage (Wainer, 1977). Because it requires the examinee to have a fair amount of time and requires the mental process to read and comprehend a passage or paragraph, it will save time and cost if multiple items are created around one passage. Examples of testlets include tests in verbal proficiency, listening comprehension, analytical reading, and mathematics.

It is well known that items sharing a common stimuli yield dependence among responses of an examinee. Thus, the response of an examinee to one item could be influenced by the answer to other items in the same testlet. However, this interdependence among items is often ignored by test models that are used to score examinees. For example, both classical test theory and item response theory are based on the assumption of local item independence (LII). LII means that the examinees' scores on different items should not be related when conditioned on examinees' ability level. Nested items within the same testlet are expected to have more interdependency than the items from a different testlet.

It was shown that ignoring this dependency by using a traditional IRT model with the LII assumption will result in overestimation of measurement precision and bias in item difficulty and discrimination parameter estimates (Yen, 1993; Wainer & Lukhele, 1997; Bradlow, Wainer &Wang, 1999; Wainer, & Wang, 2000). However, the influence of testlet effects on diagnostic classification analysis is less explored. Although methods do exist in IRT and DCMs to measure local dependency (e.g., Yen's Q3, LD-X^2, conditional covariance), there is little research in measuring local dependency caused by testlets in nonparametric diagnostic classification analysis. Also, there are few existing solutions to account for local dependency in nonparametric classification analysis.

In response to the above stated obstacles in nonparametric diagnostic classification analysis, this study seeks to extend the nonparametric Hamming Distance method (NP) proposed by Chiu and Douglas (2013) to testlet-based tests with the following goals:

1) Present a nonparametric method to measure local item dependency caused by testlets in diagnostic classification analysis;

2) Present a new nonparametric method for testlet based diagnostic classification, that is, the testlet nonparametric Hamming distance (testlet NP) method;

3) Investigate the performance of nonparametric methods of local item dependency detection in different test conditions;

4) Investigate the performance of the proposed testlet NP methods in comparison to NP method and the traditional DCM methods in situations where different levels of local item dependency are present.

Findings of this study will provide some insights into the impact of testlet effect on diagnostic classification analysis and the solution to account for testlet effects. Specifically, if the proposed conditional covariance estimation method provides a heuristic approximation of the testlet effect, it can be used to refine the items and test design and increase the precision of diagnostic classification. Second, the proposed testlet NP method is an initial effort to solve the LID issue in nonparametric classification analysis. If the method is efficient, it can be applied in practical settings where only small sample sizes are available. Third, the comparison of NP methods and traditional parametric diagnostic analysis in a variety of testlet conditions will facilitate the practitioners' choice of estimation methods in specific test conditions.

CHAPTER II

LITERATURE REVIEW

The primary purpose of diagnostic classification analysis is to assign individuals to classes according to the skills or attributes they have mastered. Two major approaches exist in diagnostic classification analysis. One is the parametric method involving mathematical modeling and parameter estimation, the other is the nonparametric approach, which does not involve parameter estimation. Both approaches have the assumption of local independence. This section begins with a description of parametric and nonparametric diagnostic classification methods, then introduces an issue of local dependency in diagnostic classification analysis, followed by the attempts in solving local dependency issue in traditional diagnostic classification models.

2.1 Diagnostic Classification Modeling

Diagnostic classification models (DCMs) or cognitive diagnostic models (CDMs) are confirmatory multidimensional latent classification models (Lazarsfeld & Henry, 1968; Rupp, Templin & Henson, 2010) in that the number of classes and latent categories in DCMs are explicit. They are mathematical models that define the probability that an examinee correctly answers an item as a function of the examinee's attribute profile, i.e., the presence and absence of a set of attributes, which is typically represented by a vector $\boldsymbol{\alpha_i} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$.

$$\alpha_{ik} = \begin{cases} 1 & \text{if person i mastered attribute } k; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

An attribute profile is assumed to provide insights into the examinee's strengths and weaknesses in specific attributes. According to his/her mastery of each attribute, the examinee is classified into one of the finite number of latent classes.

Specifying the attribute mastery status of an examinee by a test requires a Q-matrix for any approach and method. The *Q matrix* represents the knowledge structures of the test and can be viewed as a loading indicator in a confirmatory factor analysis (Rupp & Templin, 2008a). The *Q matrix* is defined as a $J \times K$ matrix where J items are represented by rows and K attributes are represented by columns, the entry q_{jk} indicates whether or not attribute k is measured by item j.

$$q_{jk} = \begin{cases} 1 & \text{if item } j \text{ requires attribute } k; \\ 0 & \text{else.} \end{cases} \tag{2}$$

Thus, a test with 20 items measuring 4 attributes will also have a 20 x 4 Q matrix.

In recent decades, a large number of DCMs have been proposed (DiBello, Roussos, & Stout, 2007; Rupp &Templin, 2008a) based on the condensation rule, that is, the interaction between attributes and items. Those models in the recent literature can be categorized into the following categories: compensatory models and noncompensatory models. Under the noncompensatory models, there are conjunctive models and disjunctive models. Under the assumption of noncompensatory conjunctive models, the examinee must master all attributes required by the item in order to get the item right.

Under the disjunctive noncompensatory models, mastering a subset of required skills by the item is sufficient for having a high probability of answering the item right. Mastery of more attributes does not dramatically increase the probability. Common conjunctive models include Deterministic Input, Noisy "And" gate model (DINA; Junker & Sijtsma, 2001), Noisy Input, Deterministic "And" gate model (NIDA, Junker & Sijtsma, 2001), and the Reparametrized Unified Model (RUM; Hartz, 2002), whereas the most famous example of a disjunctive model is the Deterministic Input Noisy "Or" gate model (DINO; Templin & Henson, 2006). In contrast to noncompensatory models, compensatory models allow the probability of giving a correct answer to increase with the mastery of additional attributes. The general diagnostic model (GDM; Von Davier, 2005, 2008) and compensatory RUM (a special case of GDM; Hartz, 2002) are the two most commonly used compensatory models.

Although there are a plethora of DCM models, generalized models or frameworks have been developed to subsume many traditional DCM models, such as GDM (Von Davier, 2005, 2008), the log-linear cognitive diagnostic model (LCDM; Henson, Templin & Willse, 2009), and Generalized DINA(G-DINA; de la Torre, 2011). This study uses the LCDM as its modeling framework because it is easy to develop new models by adding or changing parameters within this framework. In the next sections, more detailed discussion of some major noncompensatory models and compensatory models, as well as the LCDM, are provided.

2.1.1 Noncompensatory Models

Henson et al. (2009) defined noncompensatory models as models where the relationship between any attribute and the item response depends on the examinee's mastery status on the remaining attributes measured by that item. Based on the dependency between item response and attribute mastery, noncompensatory models can be further divided into conjunctive and disjunctive models.

The DINA model is probably the most commonly used conjunctive model. In the DINA model, items divide the examinees into two classes, examinees who have mastered all attributes required by the item and those who have not. Let ξ_{ij} indicate whether person *i* mastered all skills required by item *j,*

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{3}$$

where *s* is the slipping parameter and represents the probability that an examinee misses item *j* when possessing all attributes required by item *j*, whereas *g*, the guessing parameter, represents the probability of an examinee giving a correct answer even if he/she hasn't mastered all attributes required by item *j*. The parameters s_j and g_j are defined as

$$s_j = P(X_{ij} = 0 \mid \xi_{ij} = 1) \tag{4}$$

$$g_j = P(X_{ij} = 0 \mid \xi_{ij} = 0) \tag{5}$$

Thus in the DINA model, each item has one slipping parameter and one guessing parameter. The probability of a person giving a correct response is defined as

$$P(X_{ij} = 1 \mid \xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})} \tag{6}$$

Although the DINA model has been widely used because of its simplicity and less demands on sample size, one concern is that the DINA model is too restrictive because it partitions examinees into only two classes per item: the examinees who have mastered all attributes required by item *j* and examinees who have not mastered all attributes. That is, the examinees lacking one attribute will have the same probability of answering the item correctly as examinees lacking more attributes. However, there are situations where the examinee has a higher probability of answering the item right when he/she only lacks one required attribute as opposed to lacking more required attributes.

Additional conjunctive models have been developed to account for this concern. One such model is the Noncompensatory Reparametrized Unified Model (NC-RUM, Dibello et al., 1995; Hartz, 2002; Dibello et al., 2007). The model has two variants, one of which is called the full NC-RUM, the other is called the reduced NC-RUM. In this section, the reduced NC-RUM is discussed.

The reduced NC-RUM accounts for different contributions of each attribute and each item. This model is based on the unified model of Dibello et al.(1995). Given an examinee's attribute profile $\boldsymbol{\alpha}_i$, the reduced RUM defines the probability that examinee *i* correctly answers item *j* as

$$P(X_{ij}=1\,|\,\alpha_i)=\pi_j^*\prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{ik})} \tag{7}$$

Where π_j^* is defined as the baseline probability of a correct answer when all the skills required by item *j* are mastered and correctly applied. When compared to the DINA model, π_j^* is equal to not slipping (i.e., 1-s_{ik}). Parameter r_{jk}^* represents the penalty to the probability of correct response to item *j* when attribute *k* is not mastered. For an examinee who has not mastered one skill, the item probability is reduced by a factor equal to r_{jk}^* for each nonmastered skill. The larger r_{jk}^* is, the smaller the penalty. The parameter r_{jk}^* is constrained to be 0 ≤ r_{jk}^* ≤1.

Both the DINA model and the reduced NC-RUM assume that the examinee should master all attributes required by the item in order to have the highest probability of giving a correct answer. However, in some situations, mastery of one attribute is enough to answer the item correctly.

Disjunctive models assume that mastery of an additional attribute does not increase the probability of a correct answer or it just increases the probability relatively little. Based on the DINA model, Templin and Henson (2006) proposed the DINO model to address this situation. Similar to the DINA model, there is only a slipping parameter s_j and a guessing parameter g_j in the DINO model. Instead of using ξ_{ij}, they used ω_{ij} to represent the latent variable and it is defined differently

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} \tag{8}$$

The value ω_{ij} indicates whether person *i* has mastered at least one skill required by item *j*, ω_{ij} =1 when the examinee mastered at least one attribute required by the item and ω_{ij} = 0 only when the examinee has not mastered any required attributes. Hence, in the DINO model, the probability that an examinee correctly answers an item is defined as

$$P(X_{ij} = 1 \mid \omega_{ij,} s_j, g_j) = (1 - s_j)^{\omega_{ij}} g_j^{\omega_{ij}} \tag{9}$$

The DINO model has similarity to the DINA model in that examinees only have two probabilities of a correct response. The class of examinees that mastered one skill have the same probability of giving a correct answer as the examinees that master all measured skills.

2.1.2 Compensatory Models

In compensatory models, the conditional association between one item and one required attribute is independent of the examinee’s mastery status of other attributes (Henson et al., 2009). Examples of compensatory DCMs include the additive GDM models (Von Davier & Yamamoto, 2004) and the compensatory version of RUM model (C- RUM; Hartz, 2002). C- RUM is a special case of GDM (Von Davier, 2005). GDM generalizes to dichotomous and polytomous responses as well dichotomous and polytomous Q matrix entry (attributes). In addition, with an interaction term added, it becomes a conjunctive model. C-RUM only considers the additive portion and dichotomous responses and its item response function is

$$P(X_{ij} = 1 | \alpha_i) = \frac{\exp(\sum_{k=1}^{k} r_{jk}^{*} \alpha_{ik} q_{jk} - \pi_j^*)}{1 + \exp(\sum_{k=1}^{k} r_{jk}^{*} \alpha_{ik} q_{jk} - \pi_j^*)} \tag{10}$$

In the C-RUM, the probability is at the lowest when no required attributes are mastered and the kernel = $-\pi_j^*$ (similar to a guessing parameter). The probability of a correct answer is increased as a function of each measured attribute that is mastered. The increase rate is defined by r_{jk}^* ($r_{jk}^* > 0$). This is different from the reduced RUM model, where the probability of a correct response decreases as a function of each required attribute not being mastered at the rate of r_{jk}^*.

2.1.3 The LCDM Framework

Henson et al. (2009) developed the LCDM by adding interaction terms to the GDM that account for the interaction between skills, and restricting the application to dichotomous item response and attribute. Therefore, as Henson et al. (2009) suggested, LCDM can also be understood a simple extension of the binary special case of the GDM.

Under LCDM, the probability that an individual with attribute profile ***ai*** giving a correct response to item *j* is defined as

$$P(X_{ij} = 1 | \alpha_i) = \frac{\exp[\lambda_{j,0} + \lambda_j^T h(q_{jk}, \alpha_{ik})]}{1 + \exp[\lambda_{j,0} + \lambda_j^T h(q_{jk}, \alpha_{ik})]} \tag{11}$$

Where the meanings of α_{ik} and q_{jk} are the same as previously described, $\lambda_{j,0}$ is the intercept and represents the log-odds when an examinee does not possess any required attributes, λ_j represents the weight for the j^{th} item, and $\lambda_j^T h(q_{jk},\alpha_{ik})$ is the sum of linear combinations of the interaction effect and all main effects of the required attributes. *h* ( ) is the mapping function which relates slope (weight), attributes, and *Q matrix* to the response function. The function $\lambda_j^T h(q_{jk},\alpha_{ik})$ is unfolded as

$$\lambda_j^T h(q_{jk},\alpha_{ik}) = \sum_{k=1}^{k} \lambda_{j,k,1}\alpha_k q_{jk} + \sum_{k=1}^{k-1}\sum_{k'>k}^{k} \lambda_{j,(k,k'),2}\alpha_k \alpha_{k'} q_{jk} q_{jk'} \ldots \tag{12}$$

Here $\lambda_{j,k,1}$ is the weight for the main effect of attribute *k* in item *j*, and $\lambda_{j,(k,k'),2}$ is the weight for the interaction effect of attribute *k* and *k'* for item *j*. There are as many main effects as the required attributes by item *j*.

By constraining slope parameters, the item response functions for many well-known DCMs such as DINA, DINO, and C-RUM can be formed. For example, when the main effects in Equation 11 are constrained to zero, and only the highest interaction remains, the probability of a correct response for the DINA model is expressed as

$$P(X_j = 1 \mid \alpha) = \frac{\exp(\lambda_{j,0} + \lambda_{j,\mathrm{C}} \prod_{k=1}^{k} \alpha_k^{q_{jk}})}{1 + \exp(\lambda_{j,0} + \lambda_{j,\mathrm{C}} \prod_{k=1}^{k} \alpha_k^{q_{jk}})} \tag{13}$$

Where *C* represents the highest interaction. If any attribute is not mastered, the whole interaction effect will be zero.

For a test that measures two attributes, when constraining $\lambda_{jk} = \lambda_{jk'} = -\lambda_{jkk'}$, Equation 11 becomes the item response function of the DINO model

$$P(X_{ij} = 1 | \alpha_i) = \frac{\exp(\lambda_0 + \lambda_{jk}\alpha_{ik} + \lambda_{k'}\alpha_{ik'} + (-\lambda_{jkk'})\alpha_{ik}\alpha_{ik})}{1 + \exp(\lambda_0 + \lambda_{jk}\alpha_{ik} + \lambda_{k'}\alpha_{ik'} + (-\lambda_{jkk'})\alpha_{ik}\alpha_{ik})} \quad (14)$$

The sign in front of λ_j can be generally determined by $(-1)^{c-1}$, where c indicates the type of effect. For example, c is equal to 1 for main effects, and equal to 2 for two-way interaction effects, and so on.

When the slope parameters for interactions are fixed at 0 and only the main effects are kept, Equation 11 becomes the item response function of the C-RUM model,

$$P(X_{ij} = 1 | \alpha_i) = \frac{\exp(\lambda_{i,0} + \sum_{k=1}^{k} \lambda_{i,1,(k)}\alpha_{ik}q_{jk})}{1 + \exp(\lambda_{i,0} + \sum_{k=1}^{k} \lambda_{i,1,(k)}\alpha_{ik}q_{jk})} \quad (15)$$

Several software packages have the capacity of estimating the LCDM, such as Mplus (Muthen & Muthen, 1998; Rupp, Templin, & Henson, 2010; Templin, 2013), R “CDM” package, and the flexMIRT computer software (Cai, 2012). In addition, the LCDM has been used in a few studies to develop new diagnostic models (Choi, 2010; Hout & Cai, 2012; Hansen, 2013).

2.2 Nonparametric Diagnostic Classification

All DCMs discussed previously have been estimated with the EM algorithm or by Markov Chain Monte Carlo (MCMC). Those estimation algorithms usually require large sample sizes and involve heavy computing procedures with special software, which limits the breadth of DCM application (Choi & Douglas, 2013; Wang & Douglas, 2015). Nonparametric classification analyses are alternatives to parametric DCMs in this aspect. Compared to DCMs, nonparametric methods do not involve any probability computation or parameter estimation, and thus typically do not require large samples size and heavy computing procedures. A few nonparametric classification methods have been proposed in recent years. In the following paragraphs, one hybrid method that includes both nonparametric computation and parametric information and two nonparametric methods are discussed.

2.2.1 Hamming Distance Method

Chiu and Douglas (2013) used the Hamming distance to determine the cognitive profile that generates the closest ideal response pattern to the observed response pattern. To distinguish it from other nonparametric methods, we call it Nonparametric Hamming Distance Method (NP). NP does not use any item parameters of any diagnostic models and thus can be applied with any sample size. In their simulation study, Chiu and Douglas (2013) found that NP performed perfectly when the slipping and guessing parameters are 0, and has an accurate classification rate higher than .67 when the model is the DINA or NIDA with the maximum slipping and guessing parameters no greater than .3. NP showed superiority to DINA-EM when the *Q* matrix had misspecified entries. Specifically, it deteriorated less than DINA-EM when the percentage of *Q* matrix specifications increased. However, its performance severely deteriorated when the model is misspecified.

In information theory, the Hamming distance between two equal-length strings is the number of paired symbols at the same location that are different from each other. It measures the minimum number of substitutions needed to change one string to the other string. For example, in string A = (1, 1, 0, 1), and string B = (1, 0, 1, 1), we can observe two pairs of numbers that are different. Therefore, the Hamming distance of string A and B is 2. Hamming distance is often expressed as

$$d_h(y,\eta) = \sum_{j=1}^{J} |y_j - \eta_j| \tag{16}$$

y_j is j^{th} symbol in vector $\mathbf{y}$

η_j is the j^{th} symbol in vector $\boldsymbol{\eta}$

Because the Hamming distance represents the number of paired symbols at the same location that are different from each other, it can only be applied to dichotomous DCMs, where the attribute and Q matrix entry are both dichotomous. In a test that follows the DINA rule, the combination of $\boldsymbol{\alpha_c}$ vector and Q matrix creates an ideal response (i.e., expected response) $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}$, which is the j^{th} component of the ideal response pattern $\boldsymbol{\eta_i}$. Only if the examinee has mastered all attributes that have been measured, η_{ij} =1, otherwise η_{ij}= 0. In a test that follows the DINO condensation rule, the combination $\boldsymbol{\alpha_i}$ vector and Q matrix will create an ideal response $\eta_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$. If the examinee has mastered any single attribute measured by the item, it will create a value of η_{ij}=1, if the examinee has not mastered any attribute requested by the item, η_{ij}= 1. Thus η_i is a vector filled with binary values 1's and 0's. The value of η_{ij} relies on the *Q* matrix and is a function of the underlying attribute pattern $\boldsymbol{\alpha}_{\mathrm{i}}$. For each one of the 2^K attribute patterns, an ideal response patterns η^1, η^2, η^3, η^{2k} can be constructed. Because η_{ij} is determined by α_i, the distance between the observed response pattern and the ideal response pattern under attribute $\boldsymbol{\alpha}_m$ is defined as D (y_i, α_m) for m = 1, 2,…, 2^k.

Classification is achieved through minimizing the distance between the observed response pattern and ideal response patterns under all attribute profiles, which will produce the estimator

$$\hat{\alpha}_i = \arg\min \mathrm{D}(y_i, \alpha_m) \qquad m \in (1, 2, \ldots, 2^k) \tag{17}$$

The ideal response pattern that has the minimum Hamming distance from the observed response pattern is considered the estimated response pattern, and its corresponding $\boldsymbol{a_m}$ vector will be the examinee's estimated attribute profile. Wang and Douglas (2015) mathematically proved that the consistency of the NP method does not depend on sample size.

Chiu and Douglas (2013) also proposed the weighted NP method. In this method, the distance between each pair of ideal item response and observed item response is weighted by the inverse of the observed item variance. Therefore, items with smaller variance will have larger weights.

$$d_{wn}(y,\eta)=\sum_{i=1}^{J}\frac{1}{\overline{p}_j(1-\overline{p}_j)}|y_j-\eta_j| \tag{18}$$

Chiu and Douglas's (2013) simulation study found that weighted NP resulted in fewer ties, that is, there were less ideal response patterns that have the same distance with the observed response pattern. However, this weighting scheme contradicts our knowledge in both classical test theory and item response theory, which agrees that an item with larger variance typically provides more information about the ability estimation, whereas smaller variance may indicate that this item has low discriminality.

McCoy and Willse (2014) compared the performance of NP and another diagnostic classification analysis, neural network approach (Shu, Henson, & Willse, 2013) with MLE as the baseline estimation method. Data were generated from the DINA model while manipulating several factors including item numbers, sample size, number of attributes, and correlation among attributes. Findings suggested that NP moderately outperformed neural network approach (NN) and was comparable to MLE in classifying examinees in complicated structure, and slightly outperformed MLE and NN in simple structure conditions. NN was comparable when test was short, the number of attributes was larger, had simple structure, and low correlation among each other.

2.2.2 Cluster Analysis Approach

Some researchers attempted to use cluster analysis to classify examinees (Willse, Henson, & Templin, 2007; Ayers, Nugent, & Dean, 2008; Chiu, Douglas, & Li, 2009; Park,& Lee, 2011). Take Chiu, Douglas and Li's (2009) cluster analysis approach to diagnostic classification for instance. In this method, the *ith* examinee's *kth* attribute scores $W_i = (W_{i1}, W_{i2}, ..., W_{iK})$ are first estimated by a sum-score for attribute *k* using only the items measuring each attribute

$$W_{ik} = \sum_{j=1}^{J} Y_{ij} q_{jk} \tag{19}$$

The vector ***W*** was then taken as the entry for a user-selected cluster analysis (e.g., K-mean method and hierarchical agglomerative cluster analysis) with a pre-defined number of clusters as 2^K. Chiu et al. (2009) showed that the K-means cluster analysis and hierarchical agglomerative cluster analysis (HACA) were quite comparable to DINA-EM classification when the number of test items is over 4. Although the classification results were better for DINA-MMLE and *K*-means when the sample size N=500 than when N=100, this trend was not true for HACA because it does not involve fitting either item parameters or cluster centers. This result suggests that the cluster analysis approach is influenced by test size but not by sample size.

2.2.3 Sum-Score Approach to Attribute Classification

In addition to the nonparametric and parametric methods, Henson, Templin and Douglas (2007) proposed a hybrid approach which combines attribute sum-score and mastery/nonmastery cutoffs to estimate attribute mastery. The cutoffs were estimated through the DCM model. Three different methods of computing sum-score were proposed, and they are the simple sum-score (SSS), the complex sum-score (CSS), and the weighted complex sum-score (WCSS). SSC and CSC are computed in the same way as in the cluster analysis approach (Equation 19), except that SSC is based on simple structure items, whereas CSC is based on complex structure items. The limitation of SSC and CSC is that both assume all items contribute equally in measuring the attributes. As an alternative, WSC weighs each score by terms formed by the RUM calibrated item parameters, π^* and r^* .

$$W_{ik} = \sum_{j=1}^{J} \pi_j^*(1 - r_j^*) Y_{ij} q_{jk} \tag{20}$$

Simulation studies (Henson et al.,2007) with 10,000 examinees found that the performance of the three methods are comparable to RUM classification, and WCSS is always more accurate than the other two sum-score methods in different test lengths-, attribute number- and correlation- conditions. This result suggests that the use of sum-scores combined with model-based cutoffs can be applied in settings where simple diagnostic classification is desirable. However, one limitation of WCSS is that it requires the pre-calibrated item parameters in order to find the weights, another limitation is that the cutoff scores are set by finding the cutoffs of attribute mastery in the population through model calibration, which weakens its benefit in diagnostic classification over the parametric approach.

2.3 Local Item Dependence

Conditional independence of item scores or local item independence has been assumed in classical true score theory, item response theory, latent class analysis, factor analysis, and diagnostic classification modeling (Lord & Novick, 1968; Yen, 1984, 1993;

Rupp & Templin, 2008; Rupp, Templin & Henson, 2009). Local item independence (LI) is defined such that an examinee's responses to all items are independent of each other while conditioning on his/her latent ability (or latent abilities combinations).

Because of the assumption of LI, item response theory states that given an examinee's ability θ, the probability that the examinee correctly answers K- independent dichotomous items is the product of probability of answering each item correctly.

$$P(x_1 = 1, x_2 = 1, ..., x_j = 1 \mid \theta) = P(x_1 = 1 \mid \theta) \times P(x_2 = 1 \mid \theta) \times ... \times P(x_k = 1 \mid \theta) \qquad (21)$$

For DCM, "conditional independence…means that the response on any given item is only a function of the set of measured attributes" (Rupp, Templin & Henson, 2009, p.159). Mathematically, conditional independence in DCM is expressed as

$$P(X \mid \alpha_i) = \prod_{j=1}^{J} P(x \mid \alpha_i) \qquad (22)$$

Conditional independence is also an assumption for nonparametric NP methods. Wang & Douglas (2015) explicitly specified two assumptions of NP methods: 1) for examinee *i*, his/her item responses to all *J* items are statistically independent conditional on attribute vector $\boldsymbol{\alpha_i}$; 2) for all examinees, their responses to a specific item are statistically independent. Local item independency assumption is necessary for the consistency of nonparametric classification.

Macdonald (1981, 1994) and Stout (2002) argued that the LID assumption can be weakened in a way that the item responses are mutually independent. When the weak LID holds, it is more likely that the strong LID is met (McDonald & Mok, 1995).

2.3.1 Source of Local Dependency

When there is shared variance between items conditional on the measured ability or attribute profiles, the LI assumption is violated, and the source of local item dependency (LID) should be investigated. LID can be categorized into two major categories: those caused by dimension of measurement (i.e., construct underrepresentation) and those caused by nuisance variations (i.e., construct irrelevant variance). The former should be accounted for in the modeling process. For example, a test contains items that are designed to assess distinct components belonging to a general common latent trait (Steinberg et al., 2000), or a multidimensional test that is modeled with unidimensional IRT models (Ackerman, 1992). The other causes are really considered nuisance dimensions and are hard to be accounted for by an extra dimension in the model. For example, Yen (1993) identified several potential causes of local dependencies: 1) external assistance or interference, such as instruction assistance may help students perform better on some items or disruption may influence the students' score on some items negatively; 2) item chaining, when items are organized in steps, the answer to previous items will help the answer to later items; 3) content, when items that measure the same content are often locally dependent; and 4) passage dependence, in that several items share a passage or have the same setting, LID can occur. Those items are often called testlet items. LID among testlet items could arise from the student's differential level of special interest or background knowledge about the passage or the information used to answer the items is interrelated in the passage, or the item-chaining effect.

Other sources of LID could come from speediness, fatigue, item format (construct response), and raters. Those nuisance dimensions are generally due to test design. Even though the fitted models are appropriate and number of ability dimension specified are sufficient, those nuisance dimension could still cause shared variance among items.

In diagnostic classification analysis, LID is often interpreted as the result of *under -specification* of Q matrix (Tatsuoka, 1983), where the omitted attributes might cause a dependency among items. With an incomplete Q matrix, examinees with certain attribute profiles could not be estimated (Henson, 2004). If those examinees happen to have different distribution for those unspecified attributes, differential item functioning (DIF) might occur, where examinees with the same attribute profile have different probabilities of answering an item right are from different groups (Zhang, 2006; Hou, de la Torre & Nandakumar, 2014). Similarly, when polytomous attribute spaces are modeled with dichotomous models, LID could occur because there are still unexplained variances among examinees.

In addition to the incompleteness of the *Q* matrix and differential item functioning, the previously listed sources in Yen (1993) such as item chaining and passage dependence could also cause LID in DCM. For example, many diagnostic assessments regarding English language proficiency are based on reading comprehension passages (Buck, Tatsuoka, & Kostin, 1997; Jang, 2008, 2009; Sawaki, Kim & Gentile, 2009) Though the under-specification of *Q* matrix and DIF has been widely studied in DCM literature (e.g., Zhang, 2006; Rupp & Templin, 2008b; DeCarlo, 2011; Hou, de la

Torre & Nandakumar, 2014; Macdonald, 2014), LID caused by within testlet dependency has not been frequently discussed.

Locally dependent items contribute less information about the person's assessed ability than locally independent items because the more that a pair of items are related, the more they are redundant to each other. Ignoring LID might result in biased estimation of item and person parameters, overestimation of reliability and possibly the misinterpretation of measured latent space (Yen, 1984, 1993; Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996; Chen & Thissen, 1997; Embretson & Reise, 2000). Ackerman (1987) reported that when LID exists, the item discrimination parameters of locally dependent items are over-estimated, difficulty estimates tend to homogeneous. Yen (1993) found information function is inflated when LID items were treated as independent items. When fitting a 3PL model to testlet item data, Wainer & Wang (2000) found that the estimates for the item discrimination and guessing parameters were substantially overestimated, although the item difficulties were well estimated. DeMars (2006) found that the fitted 3PL model inflated the reliability for ability estimates when the LID exists.

How examinee parameters are influenced by testlet effects or LID were not as thoroughly addressed by studies of LID. Baghaei & Aryadoust (2015) compared the multidimensional Rasch model and unidimensional model when testlet effects were present, and found that the ability estimations by the two models are close to each other. Specifically, the overall theta variance is 1.73 by the four-dimensional model and 1.70 by the unidimensional model. The study by Jiao, Kamata, Wang and Jin (2012) has similar findings where the calibration models (Testlet model, Rasch model, and Multilevel Model) do not have significant impact on person ability calibration bias. Another study by Jiao & Zhang (2014) found that ignoring item clustering effects produced higher errors on item parameter estimates but not on the accuracy of ability parameter estimates, while ignoring person clustering effects yielded higher total errors in ability parameter estimates but not in item parameter estimates.

McCoy (2015) investigated the effect of increasing systematic within-skill profile variation using DCMs caused continuous abilities variation on skill mastery classification. It was found when there was LID, the difference between nonmastery and mastery of attribute profile on a continuous ability, the classification accuracy notably dropped.

In summary, studies of LID in item response theory generally found that LID could cause inaccurate parameter estimation and overestimation of test precision but had less impact on person ability estimation. Though, the study on the influence of LID caused by testlet effects on parametric and nonparametric diagnostic classification analyses is rather scarce. While testlet effects do not have a significant impact on classification accuracy in IRT study, how testlet effects impact classification accuracy in diagnostic classification modeling is not well understood.

2.3.2 Detection of Local Item Dependency

As previously mentioned, the usefulness of latent ability estimation and the precision of item parameter estimation depends on specifying the correct form of the item response function and the assumptions of LI, monotonicity, and unidimensionality. LID and unidimensionality are usually discussed together. Because of the importance of LID, a variety of LID checking procedures have been developed, some are parametric approaches like Yen's Q3 (Yen, 1984), Chen & Thissen's (2000) G^2 and LD-X^2, the others are nonparametric procedures such as Mantel-Haenszel test and conditional covariance based approach. In parametric procedures, a unidimensional model is fit to the data, then LID is tested between each item pairs. If the LI assumption fails, a multidimensional model or a unidimensional model that allows for LID is needed. In contrast with parametric LID detection, nonparametric LID assessments do not require model specification. In the following, a few parametric LID measurement indices and a nonparametric LID detection method are discussed.

2.3.2.1 Parametric Measurement of LID

Yen's Q3 (Yen, 1984), Chen and Thissen's G^2 and LD-X^2 (Chen & Thissen, 2000) are all indices to assess item-pair LID. Among them, Yen's Q3 is most commonly used in IRT (Yen, 1984, 1993; Zenisky, Hambleton, & Sireci, 2006; Pommerich & Segall, 2008). It is defined as the correlation between a test taker's residuals on a pair of items after fitting a 3PL to the data. The computation is given by

$$d_{ij} = x_{ij} - P_{ij}(\theta_i) \tag{23}$$

$$Q_{3\,jj'} = r_{d_j d_{j'}} \tag{24}$$

Where d_{ij} is the examinee's residual of the j^{th} item, x_{ij} is the observed score of the i^{th} examine on the j^{th} item, $P_{ij}(\theta_i)$ is the probability that the i^{th} examinee gives correct response to the j^{th} item, or expected raw score for a dichotomous item. The correlation of these scores taken over examinees is Q_{3ij}. *Q3* can be transformed to a Z score which has a normal distribution. It has a mean of 0 and a variance of 1/ (N-3).

LD-X^2 reflects the discrepancy between observed and expected counts after the data is fit to a model. It is computed from the observed and expected bivariate response frequencies for a given item pair. Chen and Thissen (1997) proposed to use Pearson χ^2 and likelihood ratio G^2 to measure the discrepancy. The two statistics are computed in the following manners (Liu, 2011),

$$\chi^2 = \sum_{x_p=0}^{1} \sum_{x_q=0}^{1} \frac{(O_{x_p x_q} - E_{x_p x_q})^2}{E_{x_p x_q}} \tag{25}$$

$$G^2 = -2 \sum_{x_p=0}^{1} \sum_{x_q=0}^{1} O_{x_p x_q} \log \left(\frac{E_{x_p x_q}}{O_{x_p x_q}} \right) \tag{26}$$

Where $O_{x_p x_q}$ and $E_{x_p x_q}$ respectively are the observed and expected bivariate response frequencies for a given item pair. The observed cell counts can simply be computed by crosstabulating all the examinees' dichotomous responses, the expected (marginal) frequencies are obtained by taking the product of correct response probabilities and incorrect response probabilities of the given item pair and then integrating the products over the latent space (θ)

$$E_{x_p x_q} = N \int P_i(\theta)^p P_j(\theta)^q [1 - P_i(\theta)]^{1-p} [1 - P_j(\theta)]^{1-q} f(\theta) d\theta \tag{27}$$

It was found that Yen's Q3 often results in negative bias because the residuals are calculated by estimated θ that relies on all item responses (Yen, 1984). Most of all, if the fitted model is wrong, the resulting index might fail (Hattie, 1984). On the other hand, sufficient sample size is required for computing Chen and Thissen's (1997) LD-X^2 and G^2 that use the estimated marginal frequencies from a fitted 2PLM or 3PLM.

2.3.2.2 Item Pair Conditional Covariance

The nonparametric measurement of LID is based on the conditional covariance structure of the item scores. The conditional-covariance (CC) based approaches are widely used in nonparametric IRT (Birnbaum, 1968; Rasch, 1960) based research and application (Stout, 2001). For example, a few CC based approaches have been proposed to detect multidimensionality, such as DIMTEST (Stout, 1987; Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich & Gao, 2000), HCA/CCPROX (Roussos, Stout, & Marden, 1998), and DETECT (Kim, 1994; Zhang & Stout, 1999). In contrast with parametric methods, nonparametric procedures do not depend on any parametric form for item response functions. In addition, the previously mentioned procedures are all based on conditional covariance of the item pairs. The assumption of using item pair conditional covariance to estimate multidimensionality is that the covariance of two item response scores conditional on the target θ or θs should be zero or a small negative value. Let U_i and U_j denote all the examinees' responses to item *i* and *j*, when weak local independency holds,

$$\operatorname{cov}(U_i, U_j \mid \theta) = 0 \tag{28}$$

The number correct score $S_{(-ij)}$ is often used to represent theta when computer item-pair conditional covariance.

$$\hat{\text{cov}}(U_i, U_j \mid S_{(-ij)}) = 0 \tag{29}$$

$S_{(-ij)}$ is the sumscore with scores on items *i* and *j* excluded. Douglas, Kim, Habing and Gao (1998) further expanded the ideal of conditional covariance to detecting LID in testlet items, that is, conditional on an unidimensional θ and λ. $P(U_i = 1 \mid \theta, \lambda_i)$ and $P(U_j = 1 \mid \theta, \lambda_j)$ are increasing in each of the latent variables. The parameters θ and λ, respectively, are the target ability or the ability that a given test is assumed to measure, and the nuisance ability which is not the construct of interest but influences the examinee's response to the item. To include multidimensionality into the assumption, Douglas et al (1998) also pointed out that LID can only hold on complete space (Θ, Λ) where

$$\Lambda = (\lambda_1, ..., \lambda_n) \tag{30}$$

$$P(U_i = u_i, U_j = u_j \mid \theta, \lambda_i, \lambda_j) = P[U_i = u_i \mid \theta, \lambda_i] P[U_j = u_j \mid \theta, \lambda_j] \tag{31}$$

for $i \neq j$; λ_i and λ_j represents the nuisance dimension measured by item *i* and item *j* respectively. This situation can be found where a test consists of stand-alone items that measure a distinct nuisance dimension in addition to the target ability (Douglas et al., 1998).

In diagnostic classification modeling, unidimensionality is not assumed. Instead, the attribute profile is considered as the complete latent space. Extending CC-approach to diagnostic classification modeling, if the test meet the LI assumption, it must satisfy

$$P(U_i = u_i, U_j = u_j \mid \alpha) = P[U_i = u_i \mid \alpha]P[U_j = u_j \mid \alpha] \tag{32}$$

If LID exists among items within the same passage, the average $\text{cov}(U_i, U_j \mid \alpha) > 0$ over all item pairs within the same testlet, and large average $\text{cov}(U_i, U_j \mid \alpha)$ suggests large LID. Approximating testlet effect with LID, larger $\text{cov}(U_i, U_j \mid \alpha)$ suggests larger testlet effect size.

2.4 Strategies for Dealing with Local Item Dependency

Various approaches in IRT modeling have been proposed to account for the construct relevant and irrelevant LID. For example, the Mixture Rasch model was proposed to address LID caused by un-modeled dimensions that occurred because latent classes had been combined (Rost, 1990).

There are two existing approaches that address the issue of LID in testlet-based tests. The first approach is to fit the data with a unidimensional polytomous model where all items associated with a common stimulus are combined to create one polytomous item (Lee & Kolen, 2001; Cao, Lu &Tao, 2014). This approach is relatively easy but may lose the item response pattern information due to combining items (Sireci, Thissen & Wainer, 1991; Zenisky, Hambleton, & Sireci, 2002). The second approach retains item-level information by explicitly modeling LID, such as the bifactor model (Gibbons & Hedeker,

1992) and the testlet model (Bradshaw & Wainer, 1999; Wainer et al., 2000; Wainer et al., 2007; Wang & Wilson, 2005).

The bifactor model is a hierarchical factor model and a special case of the multidimensional model. Equation 35 is the item response function of a 2PL bifactor model (Reise, Bonifay & Haviland, 2012).

$$P_i(U_{j1} = 1) = \frac{1}{1 + \exp(-1.7(a_{j1}\theta_{j1} + a_{jk}\theta_{ik}) + d_j]} \tag{33}$$

a_{j1} = general factor discrimination parameter for item *j*,

a_{jk} = group (testlet) factor discrimination parameter for item *j*,

d_j= multidimensional intercept parameter for item *j*,

θ_{i1} = general ability score for examinee *i*, and

θ_{ik} = group (cluster specific) trait score for examinee *i*.

In a bifactor model, an item *j* loads on two dimensions: a cluster specific factor *k* and a general factor "1". The cluster-specific dimensions are independent of each other conditioning on the general factor. The specification of a general factor is to account for the association of items that is not explained by cluster-specific factor. If the item discrimination parameters within a testlet are constrained to be equal, that is, remove the subscript *j* in a_{jk}, the bifactor model becomes a 2PL testlet model.

As in IRT, LID in DCM is related either to measured attributes or nuisance dimensions. The former may be caused by an underspecified *Q* matrix or when multiple strategies are used by examinees. There are models developed to account for the multiple strategies that examinees may use to approach an item, such as the multiple-choice model multiple-strategy deterministic, inputs, noisy ''and'' gate model (MS- DINA; de la Torre & Douglas, 2008; Huo & de la Torre, 2014). The incompleteness of *Q* matrix specification can be solved by specifying additional attributes in the *Q* matrix. Instead of specifying another *Q* matrix or additional attributes, the full NC-RUM model includes a continuous residual ability η_c to capture the influence of the attributes that are not captured by the *Q* matrix. The full NC-RUM is defined as

$$\pi_{ic} = P(X_{ic} = 1 \mid \alpha_c, \eta_c) = [\pi_i^* \Pi r_{i\alpha}^{*(1-\alpha_{ca})q_{ia}}] P_{ci}(\eta_c) \tag{34}$$

where $\boldsymbol{\alpha}_c$ is the vector with all attribute mastery indicators for latent class *c,* π_i^* and r_{ia}^* has the same meaning as in reduced NC-RUM, and respectively is, the baseline probability of a correct answer when all the skills required by item *j* are mastered and correctly applied, and the penalty to the probability of correctly answering item *j* when attribute *k* is not mastered. $P_{ic}(\eta_c)$ is the probability for item *j* with difficult parameter c, and it is defined as

$$P_{cj}(\eta_c) = \frac{\exp(\eta_c + c_j)}{1 + \exp(\eta_c + c_j)} \tag{35}$$

Equation 35 implies that $P_{ci}(\eta_c)$ gets smaller when the value of c_i gets smaller, a large value of c_i indicates that the item is not influenced much by the ability beyond the attributes specified in the *Q* matrix. On the other hand, a low c_i, (e.g. $c_i < 1$) indicates that the item requires more on unspecified attributes in the *Q* matrix.

In the full NC-RUM model, η_c is a measure of the undifferentiated "ability" of the respondent in class *c* that is associated with all the unspecified attributes (Rupp, Templin & Henson, 2010). As discussed in the previous paragraph, this residual ability η_c is only relevant when c_i is small. When c_i is large, $P_{ci}(\eta_c)$ is very small for lower η_c values, and essentially 1 for medium to high η_c values. In another words, a large c_i indicates that only respondents with lower ability draw on η_c, the unspecified attributes (Rupp, Templin & Henson, 2010). Since η_c absorbs all unspecified attributes or unaccounted shared variance, it can also be used to explain the testlet-specific abilities.

In the testlet model, a testlet effect only accounts for the shared variance of the items within the same testlet, the number of testlet effects corresponding to the number of testlets in a test. Both η_c and the testlet effect are considered as random effects. That is, regardless of their mastery profile, all examinees are equally likely to be at a certain level of the residual ability. Examinees that mastered all *Q* matrix specified attributes might have lower residual ability, or have high residual ability but need not apply it, whereas examinees who have not mastered the specified attributes might be high in that residual ability.

A similar approach to the full NC-RUM in accounting for LID in diagnostic classification modeling is the DCM Mixture Rasch Model (DCMixRM; Choi, 2010). DCMixRM combines the LCDM and Rasch models in order to model both discrete attributes and the continuous latent ability. Specifically, the LCDM portion of the model provides detailed profile information, and the Rasch portion captures the quantitative difference between persons within a latent class.

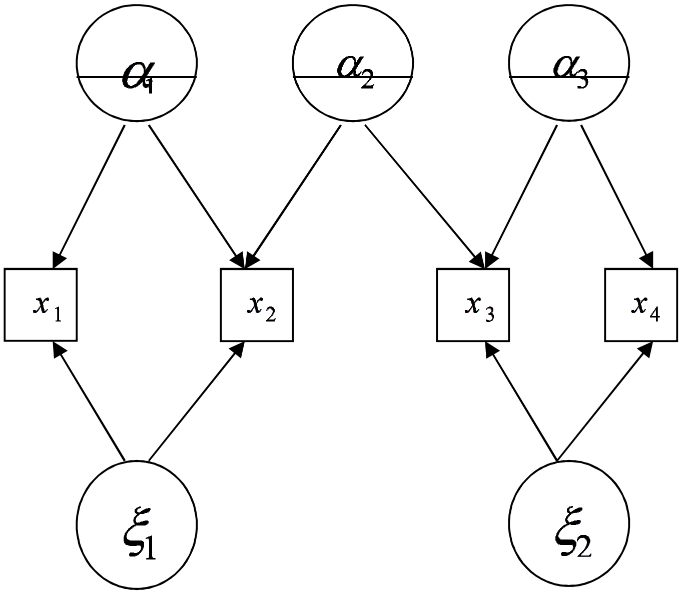
The second approach is to model the dependency but consider it as a nuisance dimension without estimating it. For example, Hansen (2013) extended the development in hierarchical item factor analysis to diagnostic classification modeling and proposed a hierarchical item response model (i.e., testlet DCM) to account for LID caused by nuisance dimension. A random effect (error effect) was added to the LCDM framework to account for dependency among items within the same item cluster (i.e., testlet). For a polytomous item response, the cumulative response probability of the hierarchical item response model for two attributes is given by

$$P(X_{ij} \geq k \mid \alpha_i, \xi_s) = \frac{\exp(\lambda_0 + \lambda_1\alpha_{i1} + \lambda_2\alpha_{i2} + \lambda_{j12}\alpha_{i1}\alpha_{i2} + \beta_{j,s}\xi_s)}{1 + \exp(\lambda_0 + \lambda_1\alpha_{i1} + \lambda_2\alpha_{i2} + \lambda_{j12}\alpha_{i1}\alpha_{i2} + \beta_{j,s}\xi_s)} \tag{36}$$

where $\beta_{j,s}$ is the slope of item j on the cluster-specific factor ξ_s. ξ_s is the random error and assumed to be normally distributed among the examinees. Each item is only allowed to load on one cluster-specific dimension. When constraining the random intercept to be the same across items within a testlet, that is, removing the subscript of j in $\beta_{j,s}$, all $\beta_{j,s} = \beta_s$, this model becomes the testlet DCM model. When further constraining the number of score categories to two, the model is a testlet LCDM model. Figure 1 presents a path diagram for a special case of the resulting model-testlet C-RUM model.

By constraining the intercept and slope parameter as previously described in the LCDM section, a testlet DINA model, a testlet DINO model, and a testlet C-RUM model can be developed from Equation 36.

Figure 1. Diagram for Testlet LCDM Model



Simulation studies (Hansen, 2013) showed that in all conditions, the testlet DCMs provided higher classification accuracy and better calibrated EAP scores than the traditional DCM models when LID was present.

Despite that the testlet DCM models have been proposed to account for LID in parametric DCMs, no effort has been made to account for the LID in nonparametric diagnostic classification analysis (i.e., the NP method). When using both parametric and nonparametric methods, Chiu and Douglas (2013) found that the larger the guessing /slipping parameter is, the lower the classification rate. Testlet effects add randomness to the classification result, possibly deteriorating the performance of nonparametric methods. This result creates some uncertainty concerning the examinee’s attribute mastery status. This research intends to propose a variation of the nonparametric Hamming distance method in order to account for the LID that exists in testlet-based tests.

In summary, there is a need to develop a nonparametric testlet effect detection method and a new nonparametric method that could account for the testlet effect. The next chapter is devoted to describing the nonparametric LID detection method and several variations of the testlet Hamming distance method, and the design of a simulation study for evaluating the new methods.

CHAPTER III

METHOD

3.1 Testlet Hamming Distance Method

The development of the testlet Hamming distance method (testlet NP) intends to improve the performance of NP methods in situations where testlets might cause LID between items. To account for the LID among items within the same testlet, we propose weighting the distance between observed item response and ideal item response by the inverse of a parameter corresponding to the testlet effect. Therefore, the Hamming distance between two item response patterns is computed as

$$d_{th}(y,\eta)=\sum_{s=1}^{S}\sum_{j=1}^{J_s}\frac{1}{\gamma_s}|y_{sj}-\eta_{sj}| \tag{37}$$

where *J* is the number of items within a testlet, *S* is the number of testlets in a test, γ_S is a parameter based on the testlet effect for a particular testlet in which item *j* is located. When there are no testlet effects, $\gamma_S = 1$ and therefore there is no additional association among items after conditioning on the attribute profile, thus the weight $1/\gamma_S = 1$ for all items. In contrast, when all items in a testlet are perfectly correlated and altogether contribute as much information as one single item, the information contributed by each item is one over the number of items, that is, $1/\gamma_S = 1/J_s$, hence the weight is constrained by $1/J_s \leq 1/\gamma_S \leq 1$.

The value of γ_s is computed based on a heuristic used to approximate the testlet effect size of the S_{th} testlet. By weighting the Hamming distance with the inverse of the γ_s, items with larger testlet effects will be penalized more than items with smaller testlet effects.

Hansen (2013) applied LD-X^2 to detect local dependency caused by testlets However, LD-X^2 requires fitting the item response data to a testlet DCM and therefore demands large sample sizes. A method that does not require the fitting of a mathematical model and has less demand on sample size, that is, a nonparametric approach to testlet effect detection remains to be developed. In this study, a method was proposed to approximate the parameter γ_s using the average conditional correlation like the CC approach to LID detection in IRT. The CC approach to LID detection in IRT, the conditional variable is often the observed test total score or true score. However, the conditional variable in diagnostic classification analysis is the examinee's attribute profile. If the test items are independent of each other, the correlation between item pairs should be close to zero conditional on attribute profile; if LID exists within a testlet, when conditioned on attribute profile, the correlations between item pairs within the same testlet should be larger than the conditional correlation between items from different testlets.

The question then arises: how is the attribute profile estimated prior to completing the conditional correlations? In this study, two methods are proposed to approximate the attribute profile. The first is simply to estimate the attribute profile with the NP method, the second is to approximate the attribute profile with the attribute sum-scores. However, if the raw sumscores are used as conditional variables, there will be a large number of attribute profiles. For example, if four attributes are measured in a test, and each attribute is measured by 10 items, there would be 10^4 possible sum-score combinations and result in 10,000 possible conditional attribute profiles. If there are 50 examinees, it is possible that no correlation matrices could be computed because of the scarceness of examinees in each sum-score profile. However, cutoff scores can be set for attribute sumscores and classify the examinee into the mastery or nonmastery group based on his/her attribute sumscores. If the examinee's sumscore of one attribute is higher than the corresponding cutoff, the examinee will be classified as the master of that specific attribute, otherwise as the nonmaster. For convenience, the average attribute sumscore across examinees will be used as the cutoff score. An α_k will represent the individual's k_{th} attribute sum-score, if α_k is equal to or above the mean, α_k is set to be 1, and otherwise is set to be 0.

3.1.1 Testlet NP Penalized by Conditional Correlation

Conditional covariance is influenced by item difficulty, and inflated by large item variance. Therefore, average item item-pair correlation was proposed to estimate LID- γ_S. For testlet *S*, the average correlation $\overline{r_s}$ will be computed and 1 / γ_S can be defined as

$$\frac{1}{\gamma_s} = 1 - \frac{Js - 1}{Js}\overline{r_s} \tag{38}$$

The weight or penalty parameter 1/ γ_S is equal to "1" when all items are completely independent of each other ($\overline{r}$ =0), and equal to 1/ J_s when the testlet items are perfectly correlated ($\overline{r_s}$ =1). To accommodate situations where the standard deviation of the response scores is zero, define N_s as the number of items where the variances are not equal to zero. When the weight of Hamming distance is computed as in Equation 38, the testlet NP method is called testlet NP penalized by conditional correlation.

NP penalized by conditional correlation minimizes the penalized distance between the ideal response pattern and the observed response pattern. With this approach, examinees are first classified using one of the nonparametric classification methods (such as the NP method and attribute sum-score method), then the conditional correlation is computed conditional on the examinees' attribute profile (latent class). For each estimated attribute profile and testlet, a conditional correlation matrix is computed with item-pair correlations as the entries, and Fisher's Z transformation is conducted for each of the entries, then the average item-pair correlation for latent class C, $\overline{r}_{sc}$ is computed across the matrix entries (Eq. 39).

$$\overline{r}_{sc} = \frac{\sum_{i \neq j}^{J_s} \sum^{J_s} Zcor(U_{s,\mathrm{i}}, U_{s,j})}{J_s(J_s - 1)/2} \tag{39}$$

where *c* is the *C^th* latent class. The testlet-specific average conditional correlation $\overline{r}_s$ is computed by weighing the $\overline{r}_{sc}$ with sample size in the latent class.

$$\overline{r}_s = \frac{\sum_{c=1}^{C} Nc \times \overline{r}_{sc}}{\sum_{c=1}^{C} Nc} \tag{40}$$

where *Nc* represents the number of examinees in the C^{th} latent class. $\overline{r_s}$ is then transformed back to Pearson's *r*. For a test that measures four attributes, the maximum number of item-pair correlation matrix for each testlet is 16.

If the LI assumption is met, correlations between items within the same testlet should be equal to the correlations between items from different testlets. Though not necessarily, for the convenience of computation, it is assumed that testlet effects are the equal across items. Therefore, all items within the same testlet are given the same weight in Hamming distance calculation.

As the conditional correlation is computed by conditioning on the attribute profile, and the examinee's attribute profile can only be estimated through other methods, the value of conditional correlation depends on how the examinees' attribute profiles are initially estimated. To show the dependency of conditional correlations on attribute profile classification, a simulation study was run. In this study, data were generated with the testlet DINA model (Hansen, 2013) for 1000 examinees to take three tests measuring four attributes (K = 4). Each test contains five items belonging to one single testlet. The three tests varied in testlet effect size (i.e., 0, 1, and 2). The correlation matrix of each testlet was computed conditional on the true attribute profiles, NP estimated attribute profiles, and attribute-sumscore estimated attribute profiles.

Table 1 presents the correlation matrices for examinees with true attribute profile α = (0, 0, 0, 1). As can be seen, the correlation values are larger and positive when the testlet effect is large, whereas the correlation values are small and tend to be negative when testlet effect is small. In general, conditional correlations increase when testlet effects increase.

Table 2 displays the average item-pair correlation for each of the 16 latent classes. There is no obvious relationship between latent class and the average item-pair correlation. The value of testlet effect when using the testlet model is the value of the β parameter in Equation 36. When β= 0, there is no testlet effect, β =1 or 2 indicates lower and higher testlet effect, respectively. Notice that the three βs are in three different tests. If the three βs are for three different testlets in the same test, results will be different.

Table 1. Item Correlation Matrix of the Three Testlets for α = (0, 0, 0, 1)

Testlet effect=0					
Item1	-.28				
Item2	-.32	.2			
Item3	.08	.19	.05		
Item4	-.18	.03	.08	-.07	
Item5	.09	-.15	-.2	-.12	-.1
Testlet effect=1					
Item1	.23				
Item2	.2	.04			
Item3	.19	.09	.36		
Item4	.17	.11	.28	.24	
Item5	.21	.24	.18	.1	.15
Testlet effect=2					
Item1	.27				
Item2	.29	.44			
Item3	.23	.42	.44		
Item4	.26	.32	.54	.37	
Item5	.22	.45	.33	.25	.25

Table 2. Average Item-Pair Correlation Conditioning on the True Attribute Profile

		Testlet Effect Size (β)		
	Profile	0	1	2
1	0000	-.01	.12	.39
2	0001	-.01	.17	.41
3	0010	.01	.14	.29
4	0011	-.03	.14	.34
5	0100	-.01	.15	.30
6	0101	-.03	.16	.30
7	0110	-.02	.17	.31
8	0111	.01	.14	.35
9	1000	.04	.14	.35
10	1001	.01	.21	.30
11	1010	-.03	.07	.46
12	1011	-.01	.06	.37
13	1100	.01	.09	.16
14	1101	.01	.09	.25
15	1110	-.04	.07	.30
16	1111	.00	.13	.31

Table 3 and 4, respectively, represent the average item-pair correlations conditioned on the NP estimated attribute profiles and the sum-score estimated attribute profiles. Comparing Table 2, Table 3, and Table 4, the relationship between the average conditional correlation and testlet effect are the same across the three tables. In other words, the average correlations are larger in situations where the testlet effect size is larger. However, there are some exceptions for Table 3 and Table 4, where the relationship between testlet effect size and average conditional correlation are not truly reflected, such as profiles 1 (0000), 5(0100), 9(1000), and 13 (1100) in Table 3 and profile 12(1011) and 13 in Table 14. Under close inspection, it can be seen that the average correlations conditioned on the NP estimated attribute profiles do not completely reflect the true conditional item correlation in Table 2. Specifically, they underestimate the local dependency in many occasions.

Same as the average correlations when conditioning on NP estimated profiles, when the testlet effect is present, the average correlations conditioned on attribute sum-score estimated profiles are also smaller than those conditioned on true attribute profiles.

Table 3. Average Item-pair Correlations Conditioned on NP-Estimated Attribute Profiles

		Testlet Effect Size (β)		
	Profile	0	1	2
1	0000	-.02	.08	.06
2	0001	-.02	.04	.08
3	0010	-.01	.04	.06
4	0011	-.03	.02	.08
5	0100	.01	-.01	.14
6	0101	-.02	.10	.16
7	0110	-.01	.06	.18
8	0111	.00	.04	.12
9	1000	.01	.18	.09
10	1001	.00	.03	.09
11	1010	.01	.05	.09
12	1011	-.01	.06	.09
13	1100	-.02	.03	.10
14	1101	-.02	.06	.17
15	1110	.00	.09	.15
16	1111	.02	.11	.14

Table 4. Average Item-pair Correlations Conditioned on Sum-score Estimated Attribute Profile

		Testlet Effect Size (β)		
	Profile	0	1	2
1	0000	-.01	.08	.13
2	0001	-.02	.06	.05
3	0010	-.04	.03	.08
4	0011	-.02	.03	.09
5	0100	-.05	.05	.11
6	0101	-.12	.09	.18
7	0110	-.04	.03	.20
8	0111	-.02	.05	.14
9	1000	-.02	.11	.07
10	1001	-.05	.05	.13
11	1010	.00	.08	.11
12	1011	-.02	.07	.07
13	1100	-.05	.00	.11
14	1101	-.02	.07	.17
15	1110	-.01	.06	.16
16	1111	.02	.11	.16

Although the idea of estimating testlet effect from the conditional correlation perspective may be possible, there are several problems that limit its application. With small sample size, it is likely that only a few or no examinees belong to certain attribute profiles. Although the computation of the Hamming distance does not depend on sample size, the accuracy of the correlation estimates is related to sample size. Small sample sizes might result in less accurate estimation of conditional item-pair correlations. In addition, when all examinees of the same attribute profile give the same response to an item, the variance of that item will be zero, thus the item-pair correlation cannot be estimated. However, when better methods to estimate the testlet effect are developed or when a reasonable approximation is known, applying the testlet NP method in diagnostic classification can still be plausible.

3.1.2 Testlet NP Penalized by Known Testlet Effect

Because the correlation estimation can be inaccurate with small sample size and examinees' homogeneous responses to the item, the testlet NP method might not work well in a situation where sample size is small. However, assume that the testlet effect size or the conditional correlation is known, γ_s can be used to represent the relative testlet effect size within a testlet. In the testlet NP method, it is not the exact value of testlet effect, but the relative weight for each item that is important. For example, if $\overline{r} = 0$ represents no testlet effect, $\overline{r} = .1$ represents low testlet effect, $\overline{r} =.4$ represents a higher testlet effect, and larger numbers indicate larger testlet effects. For the five-item testlets, we can define the penalty parameter $1/\gamma_S$ by Equation 38. Correspondingly, $1/\gamma_S$ for each of the testlets has the value of 1, .92, and .68. When $\gamma_S = 1$, the testlet NP method in Eq. 37 is the NP method.

$$d_{wh}(y,\eta) = \sum_{s=1}^{S}\sum_{j=1}^{J}\frac{1}{1}|y_{sj} - \eta_{sj}| = \sum_{s=1}^{S}\sum_{j=1}^{J}|y_{sj} - \eta_{sj}| \qquad (41)$$

In this section, a few variants of the NP methods that account for testlet effects were presented and discussed. They are the NP approach penalized by the testlet conditional correlation and the NP penalized by known testlet effects. Among the NP penalized by conditional correlation, two ways to compute the conditional correlation are presented. One is to compute the conditional correlations conditioning on the NP estimated attribute profile, the other is to compute the conditional correlations conditional on the sumscore estimated attribute profiles. In the next section, a simulation study is proposed to evaluate these NP methods.

3.2 Simulation Study

In Chapter 2, a literature review for the parametric and nonparametric classification analysis as well as the methods and strategies used to deal with local dependency was provided. In the first section of Chapter Three, the development of the testlet Hamming distance nonparametric (testlet NP) method was presented. The purpose of developing a new method was to account for LID caused by testlets in nonparametric classification analyses. Though conceptually these methods can be explained, the performance of the new methods also depends on how LID is computed. The simulation studies described in this section were proposed to evaluate the performance of the newly developed methods in various practical conditions in comparison with the NP method and traditional DCM. In both the NP method and traditional DCM classification analyses, testlet effects are ignored.

3.2.1 Research Design

To be informative and realistic, simulation studies should be representative of the real world. However, some real world situations are too complicated to be represented in a single study. Therefore this simulation study will only include factors that are considered to be most important based on the literature (Table 5) and pertaining to the research questions.

The first step of the testlet NP procedure is to estimate the testlet effect. In section 3.1, it was proposed that the average conditional correlation could be used to approximate the testlet effect. Therefore, it is important to evaluate how well the conditional correlation can be estimated by the two proposed methods: 1) the correlation when conditioning on the NP estimated attribute profiles, so called NP method (CC-NP); 2) the correlation when conditioning on the sumscore attribute profile, so called sumscore method (CC-Sumscore).

In section 3.1, the three variants of testlet NP methods were discussed: testlet NP penalized by correlation conditioning on the NP estimated attribute profile, testlet NP penalized by the correlation conditioning on sum-score estimated attribute profile, and testlet NP penalized by preknown testlet effect . The former two are based on the proposed conditional-correlation estimation methods, the third one is based on known testlet effects. In fact, the third method is not a completely different method but is used to determine whether or not the idea of penalizing the Hamming distance for testlet effects is effective while avoiding the statistical estimation of the testlet effect. To evaluate the performance of the testlet NP methods, the DINA model is chosen as the baseline model, that is, all data are estimated as though they follow the DINA condensation rule. MLE of DINA and the NP method were chosen to compare with the proposed testlet NP methods to determine if the testlet NP methods show improvement in classification accuracy. Throughout the study, the following questions were considered in evaluating the estimation efficiency of conditional correlation methods, the classification performance of testlet NP methods, and the impact of testlet effect on different classification methods.

1. How well can the testlet effect be represented through average item-pair conditional correlations?

    1.1 What is the relative performance of the NP method and the sum-score method in item-pair conditional correlation estimation with the correlation estimation conditional on the true attribute profile as the baseline?

    1.2 How does the sample size influence the performance of conditional correlation estimation by the NP method and the sum-score method?

2. How do the testlet NP methods perform compared to DINA-MLE estimation and NP in different test situations?

    2.1. How does the testlet effect size affect the performance of the NP method, testlet NP methods, and DINA-MLE in diagnostic classification analysis?

    2.2. How does sample size affect the performance of the NP method, testlet NP methods, and DINA-MLE in diagnostic classification analysis?

To answer the above questions, simulation studies were conducted. The simulation design is presented in the next section.

3.2.2 Simulation Design

In reviewing the literature pertaining to nonparametric classification analysis, it was found that several factors are commonly manipulated in previous studies (a summary as seen in Table 5). Those factors include sample size, test length, the values for slipping and guessing parameters, number of attributes, attribute correlation, the correct specification of the *Q* matrix, and the matching between data generation models and estimation models. In general, the number of attributes influences both the performance of the parametric and nonparametric methods but not the relative performance when compared to each other. Slipping and guessing parameters have a significant impact on the performances of both the parametric and nonparametric methods, specifically, the larger the two parameters are, the worse the classification accuracy (e.g., Chiu, Douglas & Li, 2009; Henson, Templin & Douglas, 2007), although larger sample sizes and longer tests increased CCRs when using both and NP classification (Chiu & Douglas, 2009; Wang & Douglas, 2015; McCoy& Willse, 2015). Misspecification of the *Q* matrix and the misspecification of model affected the classification accuracy of both parametric and nonparametric methods (Chiu & Douglas, 2013; Wang & Douglas, 2015).

A portion of the factors that seemed most important were manipulated and they are presented in Table 6 with their levels that are proposed for the study. To facilitate understanding, Table 7 explicates the conditions related to testlets.

Table 5. Summary of Relevant Simulation Studies in Diagnostic Classification Analyses

Study	Estimation model	Attribute Number	Number of items	Sample size	Profile simulation distribution	Generation models	Item parameters
Henson (2007)	Sum score	3, 5, 8	20,40	10000	MVN* R= .3; .5	Reduced RUM	
Chiu (2009)	DINA-EM K-mean HACA	3, 4	20,40,80	100,500	*MVN* R =*.25 .5;* *And Uniform*	DINA, NIDO RUM, comprensatory GDM	*s*, *g*, U(0,.15) and U (0,.3)
Chiu (2015)	Cluster analysis	Same as above	Same as above	Same as above	Same as above	DINO,DINA	Same as above
Chiu (2013)	NP, NPW DINA DINO	3, 4	20,40	10000	MVN R=0, .3, .5 Uniform	DINA, NIDA	uniform distribution 0-.1, .3, or .5

McCoy (2014)	NP, NPW and NN	4, 8	20,50	20,50,100, 500	0,.333,.5, .7	DINA	U(0,.10)
							U(.15, .25)
							U(.35, .45)

Hansen (2013)	Testlet DCMs (DINA, DINO, C-RUM) And traditional DCM	4	20,120 Clusters=1, 2,20	20000	Higher order model	Testlet DINA, DINO, C-RUM	*s* beta(.02, .05) *g* beta(.01, .05)

Note, MVN: multivariate normal distribution; R: attributes correlation; s: slipping parameter; g: guessing parameter; NPW: Weighted Nonparametric analysis; beta: beta distribution; U: uniform distribution

Factors including test length, testlet size (number of items within the same testlet), and the number of testlets contained in a test were not manipulated in this study because the three factors are confounded. One factor cannot be changed while keeping the level of other two factors constant. In realistic test situations, it is not likely that all testlets in a test have the same number of items and/or LID, and the testlet NP shows no necessity in addition to NP method when all the testlets have equal LID because all items receive the same weight (i.e., results will be identical to NP method). Therefore, the equality of testlet size and testlet effect size is manipulated. In some simulated situations, the testlet size and testlet effect size are held constant across testlets; in the other simulated conditions, they vary among testlets.

Table 6. Simulation Design

Factor	N of Levels	Level Values
Attribute	1	4
Attribute correlation	1	.5
N of items	1	24
Model generating type	1	Testlet DINA
Model application	1	DINA
Estimation	3	NP, Testlet NP, MLE
N of clusters	3	2, 4
Testlet effect Size	5	β = 0, .5, 1, 2, 3
Equality of testlet effect size	2	Equal, Unequal
Equality of testlet size	2	Equal, Unequal
Sample size (N)	5	50, 100, 500, 1000, 10000

Four factors were manipulated in this study, they are the number of testlets, the equality of testlet size across testlets, testlet effect size, and the equality of testlet effect size across testlets. There were two levels for the number of testlets factor: the two-testlets condition and four-testlets condition. Within each of the two conditions, the equality of testlet size (number of items) was manipulated. In the “equal” condition, all testlets in a test contain the same number of items (either 6 or 12 items depending on the number of testlets). In the “unequal” condition, the number of items was different across testlets. Specifically, in the two-testlets test, one testlet contains 2 items and the other contains 18. In the four-testlet test, the four testlets contain 2, 4, 8, and 10 items respectively (for specific information, see Table 7).

In testlet IRT, the magnitude of testlet effect is indicated by the variance of the random testlet effect (Wainer & Wang, 2000; Wang, Bradlow & Wainer, 2002; Wang, Chen, & Willson, 2005). The testlet effect variance indicates the degree of LID among the items within a given testlet. For example, in Jiao et al. (2013), a variance of .25, .56 and 1 represented small, moderate and large testlet effect, respectively; in Wang et al (2005), a variance of 0.25, 0.5, 0.75, and 1.00 represented small to large effects. In the present study, data were simulated using testlet DINA model (Hout & Cai, 2012; Hansen, 2013) because the equality of testlet effect size across testlets in a test can be manipulated by varying the testlet specific parameter β_s as in Equation 36. In the “unequal” condition, the testlet effect size varied across different testlets. For example, in the 2-testlet condition, data for one testlet was simulated with β =1, the other was simulated with β = 2; in the “equal” condition, testlet effects were the same across all testlets. Within the condition of equal testlet effect size, the factor-testlet effect magnitude was controlled by manipulating β between 0-3 (β = 0, .5, 1, 2, 3) with the mean and variance of the random error ξ_s fixed at 0 and 1 respectively. The square of β corresponds to the testlet variance (i.e., testlet effect). A β value "1" corresponds to testlet variance of 1, a β value of "2" corresponds to testlet variance of 4, and so on. Therefore, in this study, while the β value of 0 and .5 represents no testlet effects and small testlet effect respectively, β =1, 2, and 3 all represent a large testlet effect. The reason that large testlet effects were used is because this study intends to examine 1) what degree that the classification methods ignoring testlet effects can tolerate LID in terms of classification accuracy and 2) at what conditions, the proposed CC methods and testlet NP methods show advantages.

The testlet effect size was not fully crossed with the factor-number of testlets. For example, in two-testlet tests with different testlet effect size, the parameter β was constrained to be 1 and 2. Table 7 provides detailed information about the testlet structures described in the above simulation design.

Table 7. Testlet Design

Number of testlets	Equality of testlet size	Testlet size	Equality of test effect size	Testlet effect size
2	Different	8+16	Same	0, .5, 1, 2, 3
			Different	1 + 2
	Same	12+12	Same	0, .5, 1, 2, 3
			Different	1+2
4	Different	2+4+8+10	Same	0, .5, 1, 2, 3
			Different	.5 + 1 + 2 + 3
	Same	6+6+6+6	Same	0, .5, 1 , 2, 3
			Different	.5 + 1 + 2 + 3

For each test condition, item response data of five sample sizes was simulated (*N*=50, 100, 500, 1000, 10000). The relatively small sample sizes were chosen to determine to what extent the nonparametric classification analyses demonstrate advantages in small sample size conditions.

In summary, there were a total of 2 (Number of testlets) x 2 (Equality of Testlet Size) x 5 (Testlet Effect Size of Equal Condition) +2 (Number of testlets) x 2 (Equality of Testlet Size) x1 (Testlet Effect Size of Unequal Condition) = 24 test generation conditions. As five sample sizes were simulated for each test condition, there were a total of 24 x 5=120 data generation conditions.

3.2.3 Data Generation

3.2.3.1 Q Matrix Generation

The number of attributes measured by one test was fixed at K=4 in all simulation conditions. For different models, items were constrained to load on no more than two attributes. In generating the Q matrix, a balanced design was first used, that is, there was an equal number of items under each loading pattern. However, the model can become unidentified if all items measure more than one attribute (Chiu, Douglas & Li 2009; Madison & Bradshaw, 2014). A possible limitation of Hansen's study (Hansen, 2013) is that all items were designed to measure two attributes. Therefore, to ensure that the model is identified, eight of the 24 items were constrained to have simple structure, those items only measured one attribute. The resulting *Q* matrix is presented in Table 8, where all attributes were measured by the same number of items.

Table 8. Q Matrix for a Test of 24 Items

K=4			
0	0	1	1
1	1	0	0
0	1	0	1
1	0	0	1
0	1	0	0
0	1	1	0
1	0	1	0
0	1	0	1
1	1	0	0
1	1	0	0
1	0	0	0
0	1	1	0
1	0	1	0
1	0	0	0
1	0	0	1
0	1	1	0
0	0	0	1
0	0	1	1
0	0	0	1
0	0	1	0
0	0	1	1
1	0	1	0
0	0	0	1
0	1	0	0

3.2.3.2 Attribute Generation

Examinee attribute profiles were generated from a multivariate normal distribution so the attribute correlations could be controlled. In this model, discrete attribute profile $\boldsymbol{\alpha}$ was linked to multidimensional abilities with an underlying multivariate normal distribution, $MVN\,(0, \Sigma)$, where the covariance matrix is expressed as

$$\Sigma = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix} \tag{42}$$

In this study, ρ =.5 for all conditions was used as in Henson et al. (2007), Chiu et al. (2009) and McCoy & Willse (2015). After the four sets of θs were generated from the MVN distribution, they were further converted into 1's and 0's based on the following transformation

$$\alpha_{ik} = \begin{cases} 1, \text{ if } \theta_{ik} > 0; \\ \\ 0, \text{ otherwise.} \end{cases} \tag{43}$$

3.2.3.3 Item Parameter and Response Data Generation

Item response data was generated using 50 replications with a special case of Hansen's (2013) test DCM (Equation 36). That is, the responses were constrained to have only two categories and the cluster-specific parameter to be equal across items within the same testlet. For example, when measuring two attributes, the item response function of the testlet DINA can be formed through additional constraints of the item intercept and slope parameter,

$$P(X_{ij} = 1 \mid \alpha_i, \xi_s) = \frac{\exp(\lambda_{0,j} + \lambda_j \prod_1^K \alpha_{ik}^{q_{jk}} + \beta_s \xi_s)}{1 + \exp(\lambda_{0,j} + \lambda_j \prod_1^K \alpha_{ik}^{q_{jk}} + \beta_s \xi_s)} \tag{44}$$

The guessing parameters g_j and slipping parameters s_j were both simulated from a uniform distribution U (0, .2), and then transformed into LCDM intercept parameters λ_0 and slope parameter λ as described in Henson et al (2009). For example, when define LCDM as function of DINA parameters,

$$\lambda_{0,j} = \ln(\frac{g_j}{1-g_j}) \tag{45}$$

$$\lambda_{C,j} = -\lambda_{0,j} + \ln(\frac{1-s_j}{s_j}) \tag{46}$$

3.2.4 Examinee Classification

First, the performances of CC-NP (attribute pattern estimated by NP method) and CC-Sumscore (the attribute classification based on attribute sum-score) were investigated to see which of the two methods provided average conditional correlation estimations that were more reflective of the true testlet effect size. The average correlation conditional on attribute pattern estimated by attribute sumscores and the average correlation conditional on attribute pattern estimated through NP method were compared with the correlations conditioned on generated attribute profiles. All conditional correlations were used in computing the penalty parameter in testlet NP penalized by conditional correlation.

For each of the generated data sets, both parametric and nonparametric classification methods were used for examinee classification. For parametric estimation, all data were fitted using the DINA model and estimated using the MLE with an EM algorithm (Bock & Aitkin, 1981). For nonparametric classification, the NP method, testlet NP penalized by conditional correlations, and testlet NP penalized by preknown testlet effect were applied. The "CDM" package (Robitzsch, 2015) in R was used to perform DINA-MLE estimation, AlphaNP function from "NPCD" package (Zheng, Chiu & Douglas, 2015) in R was used to perform NP estimation, and testlet NP methods was programmed in R by the author.

3.2.5 Evaluation of Examinee Classification

The performance of the traditional classification modeling and nonparametric method was evaluated through correct classification rates (CCRs), which is the agreement between the estimated and the known true classification. Like Chiu et al. (2013) two indices were employed to summarize the results. One is the pattern-wise agreement rate (PAR)-the proportion of attribute patterns accurately estimated, the other is the attribute-wise agreement rate (AAR)-the proportion of individual attributes that were classified correctly. The two indices were defined as:

$$PAR = \sum_{i=1}^{N} \frac{I \mid \hat{\alpha}_i = \alpha_i \mid}{N} \tag{47}$$

$$AAR = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{I \mid \hat{\alpha}_{ik} = \alpha_{ik} \mid}{NK} \tag{48}$$

Mean and standard deviation of the CCRs for the 50 replications for each condition and estimation were calculated. In Monte Carlo study, standard deviation is the standard estimation error that provides the precision information of each estimation method in different test conditions.

CHAPTER IV

RESULTS

The purpose of the simulation study was to investigate the performance of the item-pair conditional correlation in estimating testlet effects and the classification accuracy of the proposed testlet Hamming distance methods in conditions with varying testlet effect, sample size, equality of testlet effect, and equality of testlet size. Results are presented to address the following two major research questions:

1. How well can testlet effects be identified using average item-pair conditional correlations?
2. How do testlet NP methods (weighted Hamming distance methods) perform compared to the NP (unweighted Hamming distance) method and MLE method?

Because the proposed testlet NP methods are essentially weighted Hamming distance methods, and the weights are determined by the testlet-specific average item-pair conditional correlation, answers to the first question are expected to provide some information for selecting the appropriate method used to estimate the weighting coefficients, and some explanation for differential performances of the testlet NP methods.

4.1 Item-pair Conditional Correlation Estimation

The testlet-specific average item-pair conditional correlation was proposed in Chapter Three to measure the testlet effect. The conditional variables, attribute profiles, were estimated via the NP method or the attribute-sumscore method. Correspondingly, the two conditional correlation estimation methods are represented using CC-NP and CC-Sumscore, respectively.

To examine to what a degree that CC-NP and CC-Sumscore are able to detect the true testlet effect, the conditional correlations estimated by the two methods were compared to that when the conditional variable is the true attribute profile (CC-True). In this section, the average item-pair conditional correlation by CC-NP, CC-Sumscore, and CC-True are presented separately for the three major test conditions: the condition with equal testlet sizes and equal testlet effect sizes, the condition with equal testlet effect sizes and unequal testlet sizes, and the condition with unequal testlet effect sizes and equal testlet sizes. Within each test condition, the impact of sample size and testlet effect size on the estimation of conditional correlation was studied.

4.1.1 Equal Testlet Size Equal Testlet Effect

Average item-pair conditional correlations for equal-testlet-size and equal-testlet effect tests were summarized across the 50 replications by estimation method, sample size and testlet effect size. Standard deviations and means were presented in Table 9. As the testlet effects were the same for all the testlets in the same test, only the results for the first testlet were presented.

In simulation studies, a small standard deviation across replications represents a small estimation error and indicates a more stable estimation, whereas a large standard deviation indicates a less stable estimation. Across the three estimation methods, estimation errors in small sample size or larger testlet effect conditions were larger than those in large sample size or smaller testlet effect conditions. In the same test condition, estimation errors of the three methods were close to each other.

It was suggested in Rosenbaum (1985) and Douglas et al. (1998) that zero or a small negative value for conditional correlation should be found when independency exists between an item pair. When N = 500, the average item-pair conditional correlations from the three methods were all close to zero. Based on the standard error of estimation (SD in Table 8), their upper limits of 90% confidence intervals at β = .5 were still smaller than the average conditional correlations at β = 1. For example, when $\beta = .5,$ and $N = 500,$ the upper limit of 90% confidence interval of the conditional correlation estimated by CC-NP is .022 + 1.97 x .09 = .059, which is smaller than .066, the conditional correlation estimated by CC-NP at $\beta = 1.$ Although the estimation error decreased with the increase of the sample size, the conditional correlation values across sample size (N = 500, 1000, and 10000) were close to each other regardless of the estimation method applied.

Table 9. Summary of Item-pair Conditional Correlations for Equal Testlet Size Equal Testlet Effect Condition

		2-Testlet						4-Testlet					
		True		NP		Sumscore		True		NP		Sumscore	
Sample	β	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
50	0	.016	.091	.001	.258	-.018	.128	.040	.105	.054	.186	.023	.113
	.5	.073	.096	.094	.261	.031	.153	.064	.113	.076	.173	.071	.103
	1	.168	.092	.129	.161	.092	.126	.150	.074	.180	.205	.115	.117
	2	.350	.077	.327	.272	.261	.192	.340	.064	.333	.174	.301	.143
	3	.456	.066	.361	.288	.337	.187	.435	.061	.442	.159	.433	.148
100	0	.009	.085	.011	.116	.001	.121	.029	.076	.030	.097	.010	.080
	.5	.048	.083	.030	.108	.001	.122	.045	.074	.028	.086	.012	.084
	1	.186	.076	.147	.123	.096	.125	.161	.076	.138	.084	.123	.092
	2	.381	.061	.264	.160	.215	.134	.357	.063	.329	.118	.285	.096
	3	.498	.052	.311	.150	.268	.113	.490	.060	.430	.117	.371	.085
500	0	-.001	.012	.003	.011	-.002	.013	.002	.018	.007	.018	-.001	.018
	.5	.034	.008	.022	.019	.017	.012	.037	.017	.032	.018	.027	.019
	1	.118	.015	.066	.014	.069	.013	.118	.016	.082	.022	.086	.015
	2	.318	.020	.131	.018	.156	.015	.325	.021	.202	.027	.230	.019
	3	.449	.020	.172	.026	.187	.028	.449	.018	.290	.035	.313	.022
1000	0	.000	.005	-.001	.005	-.002	.006	-.001	.007	.001	.009	-.004	.008
	.5	.028	.005	.015	.005	.018	.005	.029	.007	.018	.008	.021	.007
	1	.108	.009	.052	.008	.066	.007	.105	.010	.068	.009	.078	.010
	2	.299	.016	.119	.009	.148	.009	.299	.011	.188	.014	.214	.012
	3	.428	.016	.147	.015	.183	.014	.429	.011	.281	.017	.305	.014

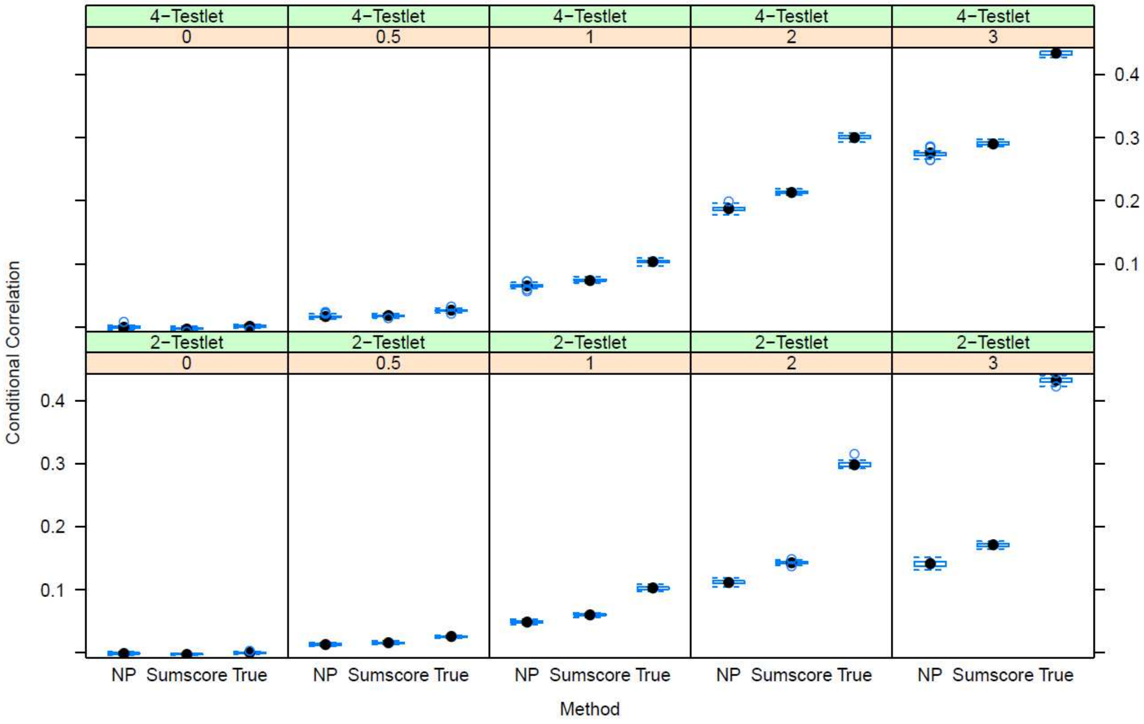
Table 9. Continued

		2-Testlet						4-Testlet					
		True		NP		Sumscore		True		NP		Sumscore	
Sample	β	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
10000	0	.000	.001	-.001	.001	-.002	.001	.000	.002	.000	.002	-.003	.002
	.5	.026	.001	.014	.001	.016	.001	.026	.002	.016	.002	.018	.002
	1	.103	.003	.049	.002	.060	.002	.103	.003	.065	.003	.074	.002
	2	.299	.004	.112	.003	.143	.002	.301	.003	.187	.005	.213	.002
	3	.432	.004	.142	.005	.171	.003	.434	.003	.275	.005	.291	.003

Notes, True: Conditional correlation estimated by CC-True; NP: Conditional correlation estimated by CC-NP; Sumscore: Conditional correlation estimated by CC-Sumscore.

The accuracy of conditional correlation estimation was found to be related to the sample size. For CC-NP, CC-Sumscore and CC-True, the standard error of estimation became smaller when sample sizes increased. Because large sample size was related to more accurate estimation of conditional correlation, comparison of different estimation methods is more valid when the sample size is larger. Therefore, in this study, the discussion is mainly based on sample size *N*=10,000.

Figure 2. Distribution of Average Item-pair Conditional Correlations for Testlets in Equal Testlet Size Equal Testlet Effect Condition (N =10,000)



Box-plots in Figure 2 display the distributions of the estimated average item-pair conditional correlations via the three methods in both 2-testlet and 4-testlet conditions. The graph should be read left-to-right and bottom-to-top. From left to right, the β value increases from 0 to 3. From bottom to top, the number of testlets that a test contains increases from two to four.

When there was no testlet effect (β= 0), the average item-pair conditional correlations estimated by CC-NP and CC-Sumscore were close to that estimated by CC-True in both 2-testlet and 4-testlet conditions. When testlet effects were present, both methods underestimated the true conditional correlation, and the size of underestimation increased as the testlet effect increased. It should also be noted that the item-pair conditional correlations estimated by CC-NP and CC-Sumscore were larger in 4-testlet conditions than in the 2-testlet condition. That is, the two methods provided larger underestimation in the 2-testlet condition than in the 4-testlet condition. There are two possible explanations for this difference between the 2-testlet condition (12 items in each testlet) and 4-testlet condition (6 items in each testlet). First, compared to the smaller testlet with the same true testlet effect, the large testlet might exert more influence on the attribute profile classification, and the estimated attribute profiles might account for more variance in the item response patterns of the large testlet. Therefore, there is less shared variance left unexplained for the large testlet after conditioning on the estimated attribute profile, resulting in smaller average item-pair correlation. Second, it is expected that in large testlet conditions, the average conditional correlations are more accurately estimated based on the central limit theorem.

Comparing CC-NP and CC-Sumcore, it can be observed that CC-NP provided slightly larger underestimation than CC-Sumscore. A close examination of the classification accuracy showed that the NP method provided higher classification accuracy than the attribute-sumscore method. Another interpretation of this phenomenon is that the conditional correlations estimated by CC-NP and CC-Sumscore reflect not only testlet effects but the unexplained shared variances caused by inaccurate profile classifications. Because the attribute sum-score method of classification provided lower classification accuracy rates, it most likely left a larger unexplained shared variance between items.

4.1.2 Unequal Testlet Size Equal Testlet Effect

In this section, the estimated item-pair conditional correlations for all testlets are presented to demonstrate whether or not unequal testlet size influences the performance of the two conditional correlation estimation methods. Table 10 presents the means and standard deviations of the average conditional correlations for each testlet in the 2-testlet tests. Because the relationship between standard deviations and average conditional correlations were similar in 2-testlet and 4-testlet conditions, that is, large testlet effects were related to large standard deviations, only means were presented for the 4-testlet tests in Table 11.

Similar to the condition with equal testlet size and equal testlet effect, the estimation errors were larger in conditions with smaller sample sizes and large testlet effects. Item-pair conditional correlations were underestimated when testlet effects were presented. As in the equal testlet size condition, the average item-pair conditional correlations and magnitude of underestimation were related to the size of the testlets. When the testlet size increased, the estimated conditional correlations became smaller for both CC-NP and CC-Sumscore, though CC-True stayed the same.

Table 10. Summary of Item-pair Conditional Correlations in 2-Testlet Unequal Testlet Size Condition

		Testlet 1 (8 item)						Testlet 2 (16 item)					
		True		NP		Sum		True		NP		Sum	
N	β	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
50	0	.059	.760	.039	.756	.054	.690	.121	.664	.064	.612	-.129	.532
	.5	.378	.661	.164	.705	.123	.602	.204	.672	.117	.677	.082	.609
	1	.593	.527	.443	.584	.319	.511	.723	.361	.427	.566	.225	.521
	2	.916	.230	.503	.561	.604	.372	.908	.189	.532	.474	.307	.461
	3	.912	.305	.741	.367	.657	.462	.970	.076	.400	.575	.301	.507
100	0	.148	.402	.143	.381	.128	.354	.123	.352	.130	.349	.039	.330
	.5	.141	.381	.120	.287	.085	.388	-.026	.364	.062	.323	-.043	.335
	1	.295	.334	.182	.305	.254	.309	.351	.355	.284	.274	.206	.285
	2	.593	.240	.520	.280	.426	.308	.749	.135	.430	.245	.217	.245
	3	.811	.123	.596	.221	.650	.226	.827	.103	.415	.269	.391	.254
500	0	.007	.044	-.003	.032	.008	.043	.008	.018	.007	.015	.017	.028
	.5	.031	.039	.034	.034	.029	.043	.034	.021	.028	.016	.026	.022
	1	.118	.042	.084	.027	.092	.035	.117	.022	.068	.016	.052	.031
	2	.327	.034	.22	.029	.237	.034	.324	.028	.106	.014	.079	.013
	3	.482	.034	.334	.045	.356	.041	.473	.026	.130	.015	.106	.022
1000	0	.001	.011	-.002	.010	.008	.02	.001	.005	.002	.008	.005	.009
	.5	.023	.009	.016	.010	.023	.014	.023	.006	.02	.008	.017	.006
	1	.091	.014	.068	.013	.081	.017	.093	.013	.052	.007	.034	.006
	2	.295	.017	.202	.016	.220	.019	.288	.013	.092	.007	.070	.008
	3	.453	.016	.296	.019	.327	.020	.441	.015	.112	.007	.093	.008

Table 10. Continued

		Testlet 1 (8 item)						Testlet 2 (16 item)					
		True		NP		Sum		True		NP		Sum	
N	*β*	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
10000	0	.000	.002	-.002	.002	.003	.002	.000	.001	.000	.001	.001	.001
	.5	.019	.002	.014	.002	.019	.002	.019	.001	.015	.001	.011	.001
	1	.084	.004	.065	.003	.069	.004	.081	.003	.047	.002	.033	.002
	2	.285	.005	.196	.004	.208	.004	.274	.004	.089	.003	.067	.002
	3	.447	.006	.289	.005	.312	.006	.435	.004	.108	.002	.092	.002

Table 11. Summary of Average Item-pair Conditional Correlations in 4-Testlet Unequal Testlet Size Condition

		Testlet 1 (2 items)			Testlet 2 (4 items)			Testlet 3 (8 items)			Testlet 4 (10 items)		
N	*β*	True	NP	Sum	True	NP	Sum	True	NP	Sum	True	NP	Sum
50	0	.075	.080	.143	.017	.005	.023	.204	.186	.000	-.079	-.099	-.139
	.5	.191	.165	.062	.001	.009	.032	.089	-.041	-.052	.328	.221	.078
	1	.199	.144	.205	.185	.199	.110	.552	.313	.166	.689	.493	.288
	2	.721	.531	.471	.709	.621	.519	.888	.646	.632	.881	.615	.378
	3	.805	.540	.559	.871	.670	.652	.919	.773	.537	.925	.654	.452
100	0	.033	.059	.031	.112	.103	.032	.025	.054	-.041	.038	.037	-.005
	.5	.032	.001	.033	.151	.181	.118	.161	.149	.118	.066	.134	-.036
	1	.139	.160	.092	.229	.253	.096	.333	.242	.164	.323	.260	.178
	2	.422	.405	.383	.580	.516	.472	.582	.478	.399	.699	.511	.410
	3	.636	.510	.524	.748	.603	.593	.764	.552	.496	.809	.562	.353

Table 11. Continued

		Testlet 1 (2 items)			Testlet 2 (4 items)			Testlet 3 (8 items)			Testlet 4 (10 items)		
N	β	True	NP	Sum	True	NP	Sum	True	NP	Sum	True	NP	Sum
500	0	.004	-.010	-.003	.012	-.005	-.001	.012	.008	-.011	.008	.010	.004
	.5	.025	.029	.023	.058	.036	.029	.027	.026	.009	.035	.025	.030
	1	.111	.093	.082	.165	.099	.087	.117	.088	.060	.111	.081	.058
	2	.340	.288	.288	.318	.274	.259	.320	.210	.171	.325	.147	.116
	3	.474	.422	.431	.472	.414	.382	.478	.296	.254	.470	.190	.160
1000	0	-.003	-.004	.000	.002	.004	.013	.002	.001	-.010	.000	.007	.000
	.5	.028	.025	.027	.021	.025	.018	.023	.018	.004	.022	.020	.017
	1	.108	.098	.101	.081	.080	.073	.097	.076	.042	.087	.058	.038
	2	.317	.287	.283	.266	.247	.226	.303	.196	.154	.285	.127	.099
	3	.477	.431	.436	.430	.382	.360	.454	.274	.226	.438	.164	.137
10000	0	-.002	-.002	-.007	.000	.000	.000	.000	.000	-.011	.000	.001	.001
	.5	.027	.025	.018	.015	.014	.012	.021	.018	.004	.017	.015	.011
	1	.109	.098	.085	.067	.064	.057	.090	.071	.045	.076	.054	.039
	2	.317	.285	.281	.262	.237	.212	.291	.192	.148	.266	.126	.094
	3	.468	.424	.425	.426	.377	.347	.449	.270	.220	.428	.162	.130

To obtain a better understanding of the results that larger testlets produced smaller average item-pair conditional correlations, conditional correlation matrices from CC-Sumscore were closely examined for the item response data of a two-testlet test. Testlet 1 and 2 consists of 8 and 16 items, respectively. The data were simulated with the testlet parameter β =3 for 10,000 examinees. Table 12 lists the range and mean of the conditional correlation matrix of each testlet for five randomly selected attribute profiles.

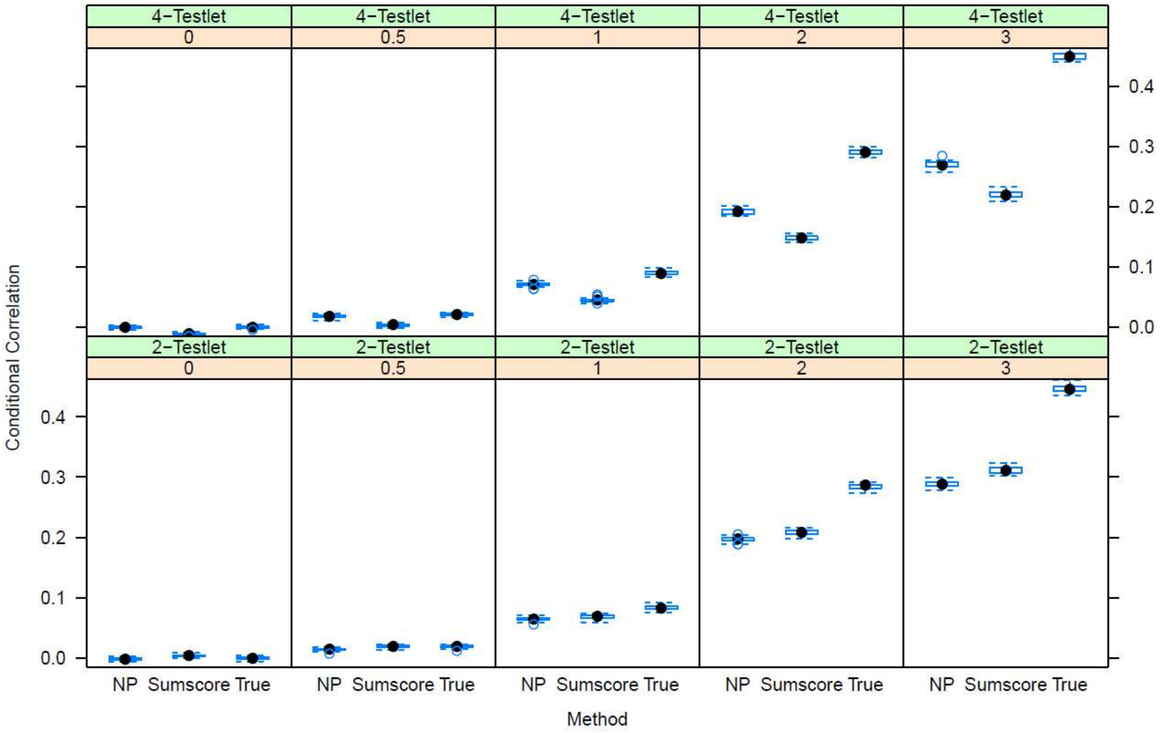
Table 12. Ranges and Means of Correlation Matrix for Each Testlet and Selected Attribute Profiles

Class	Attribute	Testlet 1 (8 items)		Testlet 2 (16 items)	
	Profile (N)	Range	Mean	Range	Mean
1	0 0 0 0 (3366)	.073, .480	.239	-.012, .626	.099
2	1 0 0 0 (369)	-.028, .425	.260	-.313, .555	.058
3	1 1 1 0 (433)	.130, .470	.250	-.255, .345	.047
4	1 1 0 1 (467)	.093, .611	.265	-.301, .354	.044
5	1 1 1 1 (2893)	.420, .547	.481	-.014 , .353	.146

Across attribute profiles, the ranges of item-pair correlations were larger for testlet 2 (the large testlet) than for testlet 1(the smaller testlet). For example, about 1/3 of the 10,000 examinees were classified in class 1 that has the attribute profile α = c (0, 0, 0, 0). For examinees in that class, the range of item-pair conditional correlation is .480 -.073 = .41 for the smaller testlet, and .626 - (-.012) =.64 for the large testlet. Furthermore, there were also more negative values of item-pair correlations for the 16-item testlet than for the 8-item testlet.

Figure 3 visually displays the distribution of estimated average conditional correlations for sample size =10,000. Item-pair conditional correlations estimated by CC-NP were close to CC-Sumcore in 2-testlet conditions but consistently smaller than CC-NP in the 4-testlet conditions. Therefore, it is difficult to determine which estimation method is better as inconsistent results were discovered in 2-testlet conditions and 4-testlet conditions.

Figure 3. Distribution of Estimated Item-pair Conditional Conditions for Unequal Testlet Size Tests 8-items Condition (N=10,000)



4.1.3 Unequal Effects Equal Testlet Sizes

Table 13 summarizes testlet-specific average conditional correlations for tests with unequal testlet effects by sample size and number of testlets. In the 2-testlet conditions, item response data were simulated with β=1 for one testlet and β= 2 for the other. In the four-testlet conditions, data were simulated with β = .5, 1, 2, and 3 respectively. To be comparable with the 2-testlet conditions, only the conditional correlations for testlets with generating β =1 and 2 in the 4-testlets conditions are presented in Table 13. The complete results for 4-testlet conditions are displayed in Appendix 1.

The standard error of the conditional correlation estimates in the unequal testlet size are larger in small-sample-size conditions and in the large-testlet-effect conditions. In this study, a sample size N = 500 was sufficient to produce stable estimation (small estimation error), the 90 percent confidence interval of the mean conditional correlation in any testlet effect condition did not overlay with each other. However, to be consistent with the previous two sections, discussions regarding the relative performance of the estimation methods were based on sample size N=10,000.

It can be observed that, across estimation methods, the conditional correlation for testlets with generating effect size β = 2 is approximately twice as large as that for testlets with generating effect size β =1. This result indicates that the proposed conditional correlation methods can be used to indicate the relative testlet effect difference among testlets.

The CC-Sumscore method overestimated true conditional correlations when the testlet effect was small (β =1) and underestimated true conditional correlations when the testlet effect was large (β =2). However, because the deviance between CC-Sumscore and CC-True at β =1 is so small, it can be considered random error instead of overestimation or positive bias. In contrast, CC-NP underestimated the true conditional correlation across all conditions.

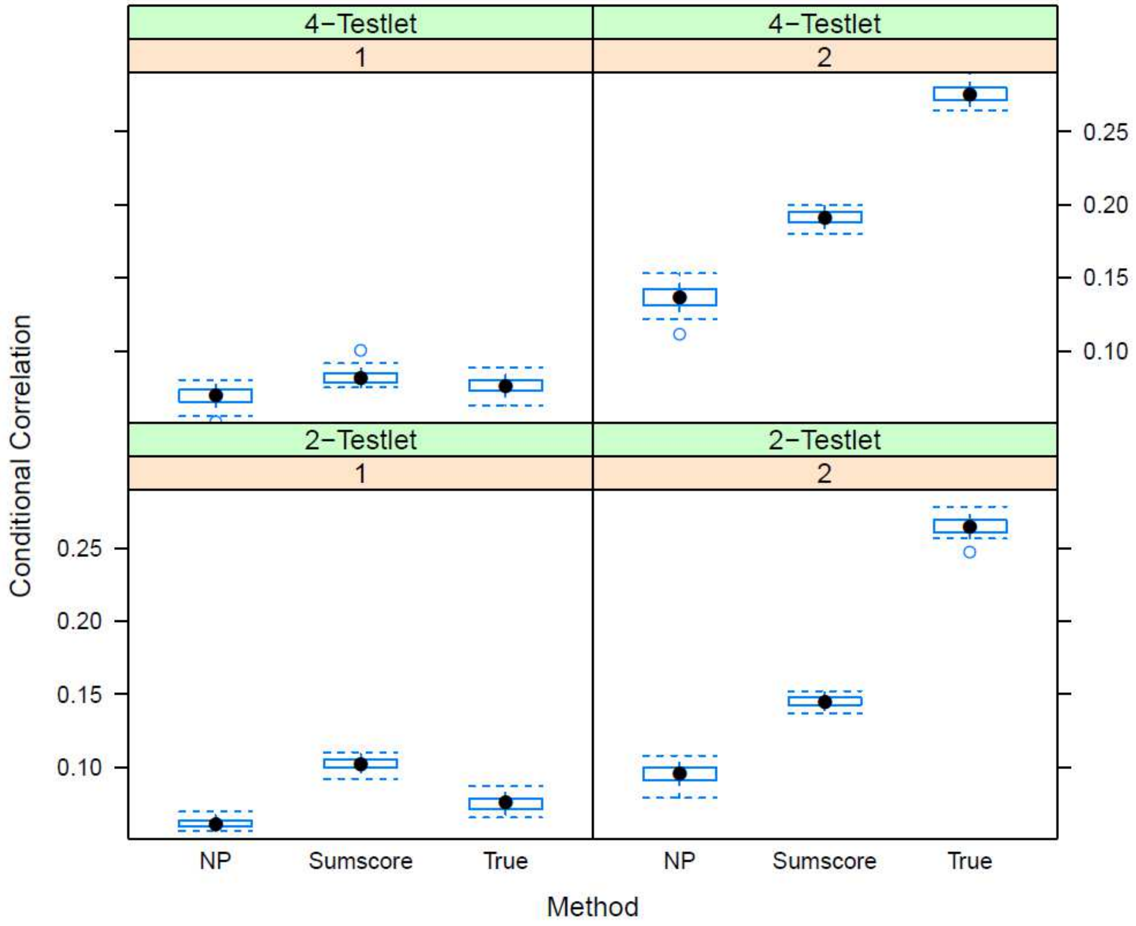
Similar to the equal testlet effect conditions, the estimated conditional correlation for the large testlet size (2-testlet tests) condition was smaller than that of the small testlet size (4-testlet tests) condition. That is, CC-NP and CC-Sumscore underestimated the conditional correlations more for the large-testlet-size conditions than for the small-testlet-size conditions. One possible explanation for this phenomenon is that it is more difficult for the estimated attribute profile to account for the variation of response patterns in four testlets than for that in two testlets. Therefore, the shared variance among items might be captured more in the 2-testlet condition (larger testlet condition) than in the 4-testlet condition (smaller testlet condition).

The information described above can also be found in Figure 4. The graph is read the same way as Figure 2. From left to right, when testlet effect β increased from 1 to 2, the estimated conditional correlation and standard error of estimation both increased. From bottom to top, when the number of testlets increased (size of testlet decreased), both CC-NP and CC-Sumscore increased in magnitude.

Table 13. Summary of Item-pair Conditional Correlation for Equal Testlet Size & Unequal Testlet Effect Conditions

		2-Testlet						4-Testlet					
		True		NP		Sumscore		True		NP		Sumscore	
Sample	*β*	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
50	1	.568	.129	.602	.129	.493	.061	.438	.08	.541	.144	.520	.141
	2	.514	.056	.612	.143	.587	.122	.506	.059	.636	.161	.537	.135
100	1	.421	.130	.442	.127	.344	.069	.327	.08	.441	.169	.371	.149
	2	.434	.053	.470	.188	.426	.096	.453	.065	.461	.133	.437	.089
500	1	.179	.029	.138	.026	.136	.030	.132	.028	.143	.039	.156	.045
	2	.294	.028	.155	.032	.206	.020	.299	.031	.186	.049	.247	.026
1000	1	.134	.012	.088	.012	.103	.014	.102	.019	.097	.022	.114	.016
	2	.272	.021	.113	.017	.170	.012	.277	.025	.151	.029	.206	.019
10000	1	.102	.004	.062	.003	.075	.005	.076	.005	.069	.006	.082	.005
	2	.265	.007	.096	.006	.145	.004	.276	.006	.136	.009	.191	.005

Figure 4. Distribution of Estimated Item-pair Conditional Correlations for Unequal-Testlet-Effect-Size Tests (N=10,000)



4.1.4 Summary of the Main Findings

In summary, this section found: 1) Small sample size and large testlet effects contributed to large estimation errors. 2) Both CC-Sumscore and CC-NP underestimated the true conditional correlations. 3) In equal-testlet-effect conditions, CC-Sumscore demonstrated more underestimation than CC-NP, although the relationship was reversed in unequal-testlet-effect conditions. 4) The magnitude of underestimation for both methods increased when testlet effects increased.

The CC-Sumscore method produced less underestimation than CC-NP except in the 4-testlet unequal testlet size conditions. Therefore, conclusions cannot be made about which method is better based only on the results of this study.

4.2 The Performance of Testlet NP Methods

In this section, the classification accuracy for the proposed testlet NP methods (weighted Hamming distance methods) are reported and compared to the NP method and the MLE method. In Chapter Three, the weighting coefficient in testlet NP methods is defined as a function of the average conditional correlation and testlet size. In addition, it was proposed that the conditional correlation should be estimated by the method that approximates the true conditional correlation most accurately. However, the results of conditional correlation estimation did not provide an optimal method and therefore, both CC-NP and CC-Sumscore methods were used to estimate the weights. Weights were also estimated by CC-True. The respective testlet NP methods are named Testlet NP based on NP estimated profiles (NPT), Testlet NP based on attribute sum-score estimated profiles (Sumscore), and Testlet NP based on true attribute profile (True). Throughout the remainder of the document, “testlet NP methods” was used interchangeably with “weighted methods” depending on the circumstances. Similarly, the “unweighted methods” were also used to represent the NP method and MLE.

The correct classification rates (CCRs) including AARs and PARS were summarized by sample size, testlet effect size in Tables 12-14 and Figures 4-7. Results for each test condition were presented in the following order: the condition with equal testlet sizes and equal testlet effects, the condition with unequal testlet sizes and equal testlet effects, and the condition with unequal testlet effects and equal testlet sizes. As testlet NP methods are mainly weighted by the testlet effect, it is anticipated that the results are more influenced by testlet effect size rather than sample size. Therefore, information in Tables 14-16 was organized differently from that in the previous sections about conditional correlation estimation. Specifically, the CCRs were organized first by testlet effect and then by sample size.

4.2.1 Equal Testlet Size Equal Testlet Effect

Table 14 summarizes the classification accuracy rate of the three weighted methods and two unweighted methods for the condition with equal testlet size and equal test effect condition. It should be noticed that AARs are always higher than PARs and decreased in a low-rate than PARs when testlets effect increased.

Table 14. Average CCRs for Equal Testlet Size Equal Testlet Effect Condition

		AAR					PAR				
β	N	True	NPT	Sum-score	NP	MLE	True	NPT	Sum-score	NP	MLE
2-Testlet											
0	50	.972	.973	.973	.971	.973	.900	.903	.904	.897	.905
	100	.971	.970	.970	.973	.971	.895	.891	.892	.901	.897
	500	.971	.971	.970	.971	.972	.896	.897	.894	.895	.899
	1000	.969	.969	.969	.969	.970	.890	.889	.890	.890	.891
	10000	.970	.970	.970	.970	.969	.893	.892	.893	.893	.887
.5	50	.964	.965	.964	.967	.966	.878	.881	.876	.886	.885
	100	.964	.965	.965	.965	.964	.872	.874	.875	.874	.870
	500	.966	.966	.966	.966	.968	.881	.880	.880	.878	.887
	1000	.963	.964	.963	.963	.966	.871	.872	.870	.871	.879
	10000	.964	.964	.964	.964	.964	.872	.872	.872	.872	.871
1	50	.946	.948	.946	.945	.948	.822	.830	.824	.827	.831
	100	.941	.939	.939	.940	.944	.803	.799	.797	.799	.813
	500	.942	.943	.942	.942	.948	.808	.811	.808	.808	.825
	1000	.941	.941	.940	.940	.945	.803	.804	.803	.802	.815
	10000	.942	.942	.942	.942	.945	.808	.808	.809	.808	.812
2	50	.863	.860	.863	.863	.858	.622	.619	.623	.623	.611
	100	.866	.864	.867	.867	.867	.621	.617	.623	.622	.621
	500	.862	.861	.860	.862	.866	.622	.618	.619	.620	.626
	1000	.860	.860	.860	.859	.863	.617	.619	.620	.616	.616
	10000	.860	.860	.860	.860	.863	.618	.619	.620	.618	.615
3	50	.773	.769	.770	.772	.750	.448	.443	.446	.446	.402
	100	.784	.782	.784	.783	.771	.479	.478	.478	.472	.451
	500	.783	.784	.784	.783	.776	.475	.479	.477	.475	.453
	1000	.777	.777	.777	.777	.768	.468	.470	.469	.469	.443
	10000	.781	.781	.780	.781	.773	.474	.474	.474	.473	.447

Table 14. Continued

		AAR					PAR				
β	N	True	NPT	Sum-score	NP	MLE	True	NPT	Sum-score	NP	MLE
					4-Testlet						
0	50	.973	.973	.973	.974	.973	.905	.908	.904	.908	.908
	100	.973	.972	.973	.971	.972	.901	.898	.900	.896	.899
	500	.970	.970	.970	.971	.972	.894	.894	.894	.898	.899
	1000	.970	.970	.970	.970	.970	.891	.891	.891	.892	.893
	10000	.970	.969	.970	.970	.969	.893	.891	.892	.893	.888
.5	50	.967	.967	.967	.968	.969	.889	.888	.887	.892	.893
	100	.965	.965	.964	.966	.966	.876	.875	.872	.875	.877
	500	.965	.964	.965	.965	.968	.878	.875	.876	.877	.889
	1000	.964	.964	.963	.963	.965	.873	.872	.871	.870	.877
	10000	.965	.964	.965	.965	.965	.878	.875	.877	.877	.874
1	50	.952	.951	.952	.951	.957	.837	.833	.836	.833	.851
	100	.947	.947	.946	.945	.949	.819	.818	.815	.813	.826
	500	.951	.951	.951	.950	.955	.835	.834	.833	.830	.848
	1000	.948	.948	.947	.947	.952	.825	.822	.821	.821	.833
	10000	.950	.949	.950	.949	.951	.829	.826	.827	.825	.831
2	50	.884	.881	.886	.887	.880	.656	.650	.662	.658	.649
	100	.884	.882	.884	.884	.885	.644	.639	.645	.646	.648
	500	.884	.883	.883	.884	.890	.653	.647	.649	.654	.667
	1000	.883	.882	.883	.882	.888	.649	.645	.647	.648	.660
	10000	.885	.882	.883	.883	.889	.654	.642	.646	.647	.663
3	50	.816	.807	.811	.815	.792	.505	.488	.496	.500	.449
	100	.817	.815	.819	.818	.801	.515	.510	.514	.507	.470
	500	.812	.811	.810	.811	.803	.501	.497	.495	.498	.473
	1000	.807	.805	.806	.807	.796	.490	.483	.486	.489	.461
	10000	.810	.806	.806	.808	.799	.496	.484	.483	.491	.460

Note, True: Testlet NP weighted by conditional correlation based on true attribute profile; NPT: Testlet NP weighted by conditional correlation based on NP estimated attribute profile; Sumscore: Testlet NP weighted by conditional correlation based on Sumscore estimated attribute profile. NP: Original Hamming distance method.

If the testlet effect size for all of the testlets in a test are equal, it is the same as no weighting. Therefore, it is expected when the conditional correlations are accurately estimated, there should be no difference between the weighted methods and unweighted methods. Compared to the NP method that ignored the testlet effect, testlet NP methods did not show dramatic improvement regarding classification accuracy, though, testlet NP-True did demonstrate higher classification accuracy in 2-testlet conditions when testlet effect was large ($\beta = 3$) and in 4-testlet conditions when $\beta \geq 1$. The differences between the three testlet NP methods were minor.

Figures 5 and 6 visually display the distribution of PAR by test condition. The graph is read left-to-right and top-to-bottom. From left-to-right, when the testlet effect sizes increase, the classification accuracy decreases and the standard error of estimations increase. From top to bottom, when sample sizes increase, the standard error of estimates decrease. However, the change was not dramatic in terms of classification accuracy, which was true for both the unweighted and unweighted methods, and both the parametric method (MLE) and nonparametric methods.

Figure 5. Distribution of PARs for 2-testlet Equal Testlet Size Equal Testlet Effect Condition

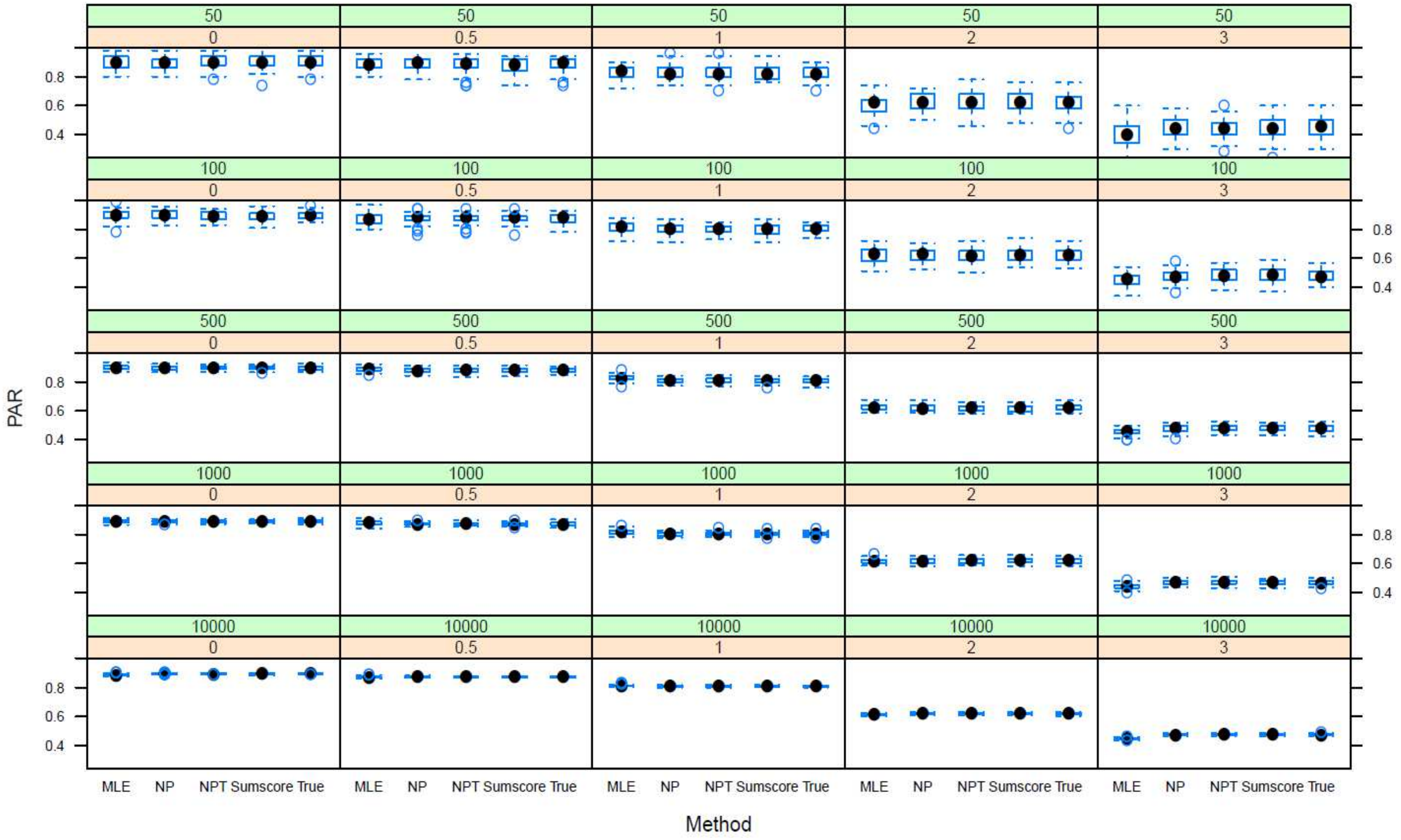
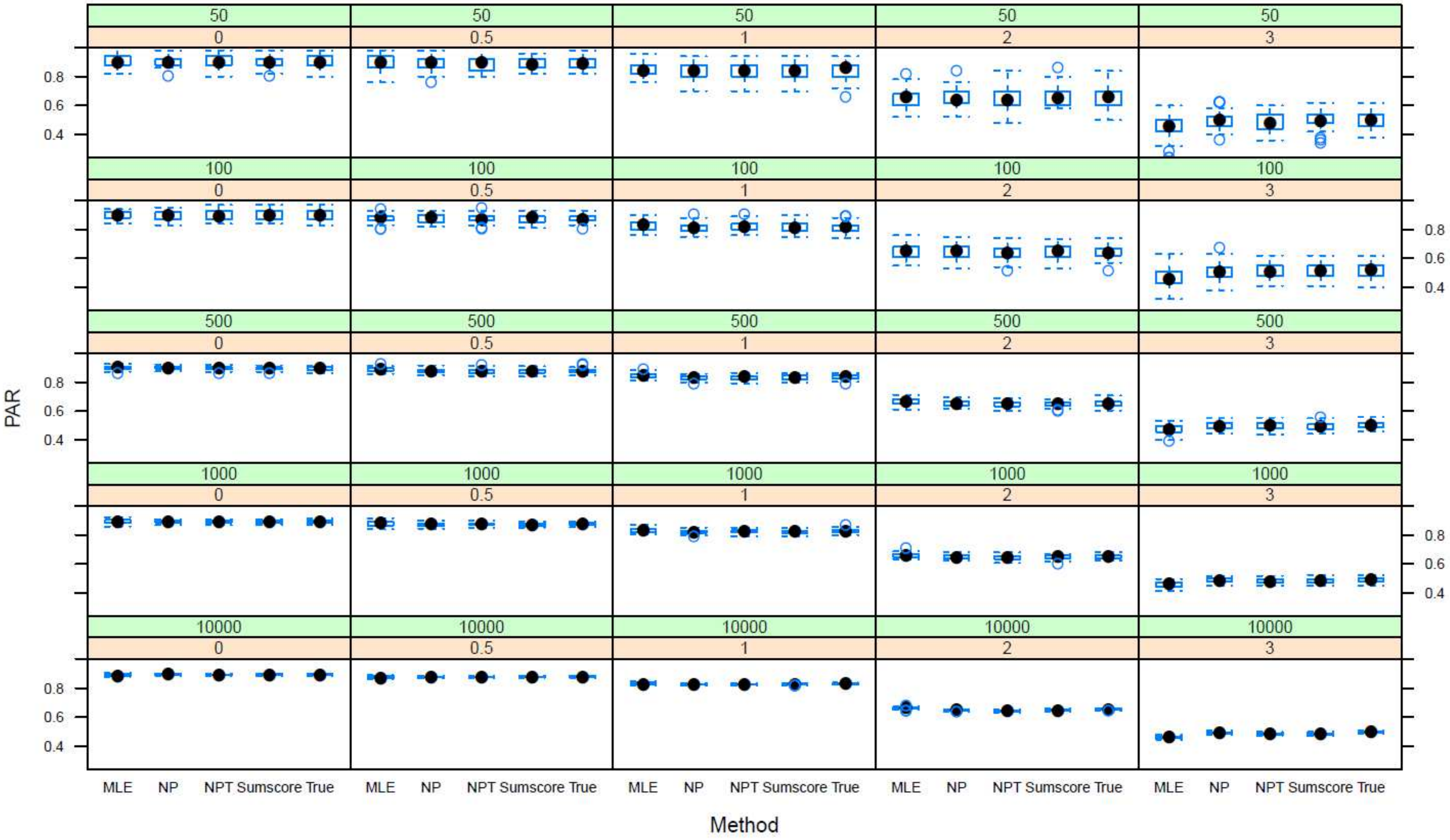


Figure 6. Distribution of PARs for 4-testlet Equal Testlet Size Equal Testlet Effect Condition



4.2.2 Unequal Testlet Effects Equal Testlet Sizes

Table 15 summarizes AARs and PARs of the testlet NP methods in comparison to the unweighted methods in unequal-testlet-effect conditions. As described in the simulation design of Chapter Three, conditions regarding testlet size and testlet effect size for the unequal-testlet-effects condition were predetermined. That is, in the 2-testlet conditions, parameter β in the data simulation model is “1” for testlet 1 and “2” for testlet 2; in the four-testlet condition, β is equal to .5, 1, 2, and 3 for each testlet, respectively. As such, the average testlet effect for the 2-testlet test is smaller than that of the 4-testlet test. Therefore, higher CCRs were produced in 2-testlet tests.

The standard deviations of AARs and PARs were similar, thus only the distribution of the PAR are summarized in Figure 7. As it can be observed, smaller sample sizes are related to larger standard deviations.

Table 15. Average CCRs for Unequal Testlet Effect Equal Testlet Size Condition

	AAR					PAR				
			Sum-					Sum-		
N	True	NPT	score	NP	MLE	True	NPT	score	NP	MLE
						2-Testlet				
50	.910	.914	.911	.905	.912	.728	.737	.733	.717	.738
100	.912	.911	.910	.909	.916	.731	.729	.726	.721	.743
500	.913	.915	.912	.910	.923	.739	.744	.736	.731	.764
1000	.909	.911	.907	.905	.916	.725	.729	.721	.715	.741
10000	.911	.913	.909	.907	.919	.733	.737	.727	.722	.747
						4-Testlet				
50	.905	.906	.906	.898	.906	.710	.695	.710	.683	.714
100	.911	.912	.910	.901	.918	.721	.712	.719	.700	.735
500	.912	.911	.911	.902	.920	.731	.737	.728	.703	.748
1000	.905	.905	.905	.896	.913	.712	.714	.710	.687	.726
10000	.909	.909	.909	.899	.917	.724	.728	.723	.697	.738

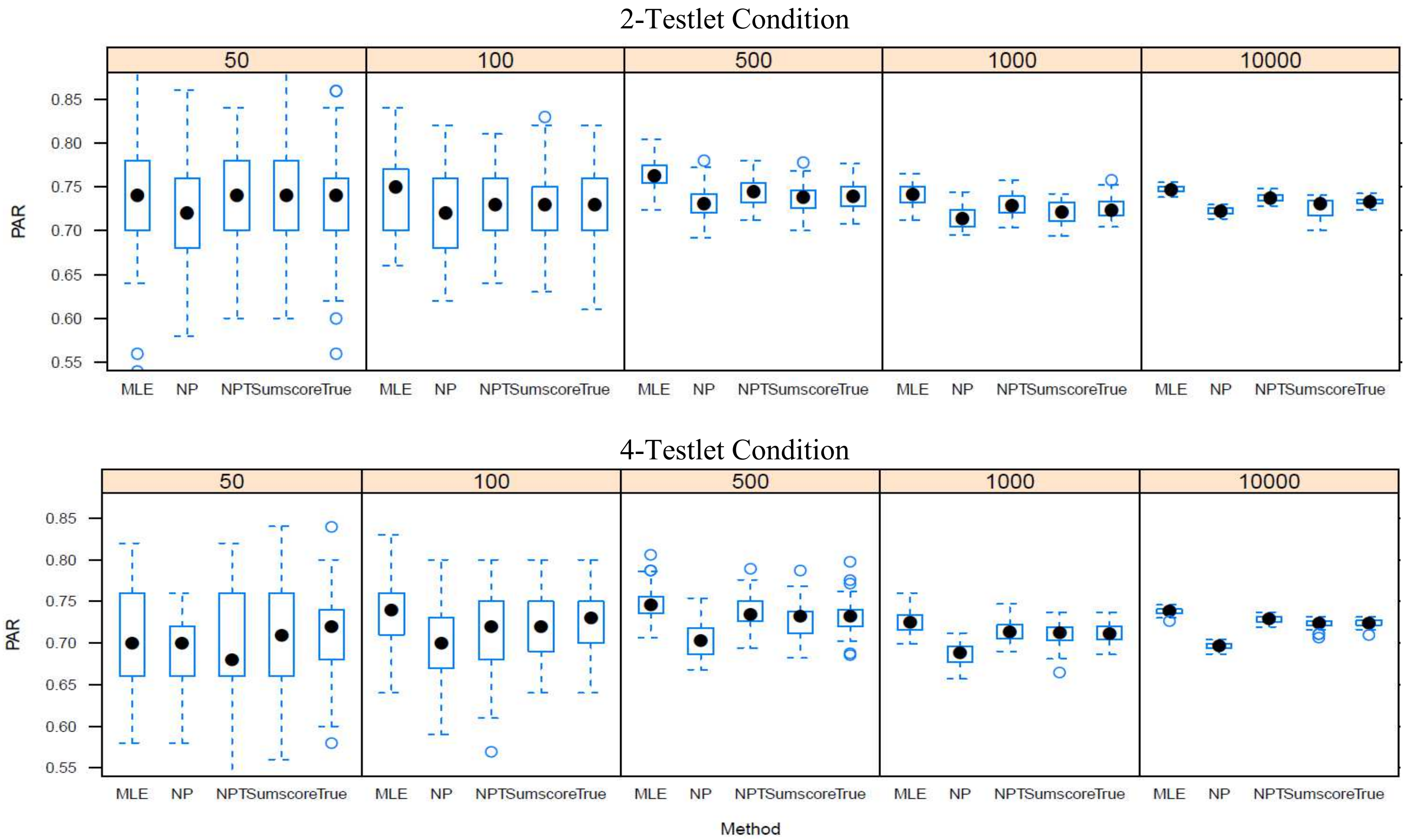
Notes, In two-testlet tests, β=1, 2; in four-testlet tests, β=.5, 1, 2, 3

MLE produced the highest AAR and PAR across all sample sizes and testlet effects. However, weighted Hamming distance methods provided higher classification accuracies than those of the unweighted NP method across all conditions. The difference between classification accuracies of weighted and unweighted NP methods is as high as .03 in terms of PAR. Among the three testlet NP methods, NPT (testlet NP weighted by the conditional correlation based on NP estimated attribute profile) showed a slight advantage over the other two testlet NP methods.

The influence of sample size on classification accuracies for all nonparametric methods in both 2-testlet conditions and 4-testlet conditions is small. However, it should be noted that the AARs and PARs for N= 500 are consistently better than that in other sample size conditions (N= 50, 100, 1000, 10,000). This result is contrary to our expectation, as in general, the larger the sample size, the better the classification accuracy.

The results described above were similar for 2-testlet and 4-testlet conditions. However, the advantage of the weighted Hamming distance methods over unweighted Hamming distance method was slightly larger in 4-testlet conditions. As the average testlet effect in a 4-testlet test is larger than that in a 2-testlet test. This result suggests that the weighted methods have more advantage in larger testlet effect conditions. The same was found in the previous section (the condition with equal testlet size and equal testlet effect), where the weighted methods exceeded the unweighted methods the most when the testlet effect size $\beta = 3$.

Figure 7. Distribution of the PARs by Sample Size for Unequal Testlet Effect and Equal Testlet Size Condition


2-Testlet Condition

4-Testlet Condition

Notes, True: Hamming distance weighted by conditional correlation based on true attributes profile; Sumscore: Hamming distance weighted by conditional correlation based on attribute-sumscore estimated attribute profiles; NPT: Hamming distance weighted by conditional correlation based on NP-estimated attribute profiles.

4.2.3 Unequal Testlet Size and Equal Testlet Effect

It was expected that the weighted methods in the unequal testlet size condition would not perform as well as they did in the equal testlet size condition because the conditional correlations were not accurately estimated. Because the AAR and PAR have the same pattern across all simulation conditions, only the PARs are summarized in Table 16. In addition, the distribution of PARs across all conditions are presented using boxplots in Figures 8 and 9.

Similar to what was found in the equal-testlet-size condition, the estimation error decreased with the increase of sample size and the decrease of testlet effect, the PAR of MLE increased more than the other methods when sample size increased, and decreased more than the other methods when testlet effect increased. This result indicated that MLE was more influenced by sample size and teslet effect than the other methods.

Overall, MLE slightly outperformed the other methods in most test conditions. PARs for weighted Hamming distance methods were close to those of the NP method in most conditions except when sample size was as small as 50 and 100. When N= 500 or 1000, weighted Hamming distance methods provided lower PARs than the unweighted Hamming distance method. This result is as expected for the accurate estimation of the weight coefficient-function of a conditional correlation- relies on large sample size.

Table 16. Averaged PARs from the Weighted Methods versus Unweighted Methods for Unequal Testlet Size Equal Testlet Effect Tests

		NP-	NP-				
β	N	True	Sumscore	Sum	NPT	NP	MLE
			2-Testlet				
0	50	.897	.914	.782	.905	.928	.936
	100	.896	.899	.770	.900	.905	.923
	500	.920	.921	.713	.919	.922	.941
	1000	.911	.910	.756	.911	.911	.934
	10000	.914	.914	.742	.914	.914	.936
.5	50	.882	.899	.770	.887	.920	.930
	100	.888	.891	.761	.889	.891	.913
	500	.909	.908	.703	.908	.907	.932
	1000	.896	.896	.744	.897	.896	.923
	10000	.902	.902	.732	.902	.902	.927
1	50	.832	.852	.695	.838	.867	.876
	100	.834	.841	.709	.842	.848	.870
	500	.867	.863	.655	.865	.866	.896
	1000	.852	.851	.691	.850	.852	.886
	10000	.858	.855	.666	.856	.857	.891
2	50	.667	.674	.516	.660	.699	.681
	100	.678	.688	.567	.695	.699	.714
	500	.700	.699	.534	.700	.698	.720
	1000	.695	.693	.541	.692	.695	.710
	10000	.702	.698	.545	.698	.700	.727
3	50	.506	.524	.371	.519	.526	.496
	100	.546	.550	.463	.549	.552	.546
	500	.560	.565	.437	.563	.561	.560
	1000	.553	.556	.434	.556	.554	.548
	10000	.555	.558	.443	.558	.557	.554

Table 16. Continued

β	N	NP-True	NP-Sumscore	Sum	NPT	NP	MLE
			4-Testlet				
0	50	.902	.910	.782	.894	.930	.940
	100	.895	.894	.779	.896	.905	.927
	500	.919	.917	.715	.920	.918	.942
	1000	.914	.909	.756	.913	.909	.934
	10000	.919	.910	.743	.919	.914	.936
.5	50	.883	.899	.775	.892	.927	.936
	100	.881	.885	.762	.881	.891	.908
	500	.907	.911	.706	.910	.910	.931
	1000	.897	.902	.746	.898	.898	.924
	10000	.899	.907	.735	.899	.903	.928
1	50	.822	.840	.725	.842	.880	.887
	100	.844	.845	.724	.844	.857	.883
	500	.867	.870	.671	.868	.873	.904
	1000	.854	.866	.707	.855	.862	.895
	10000	.860	.868	.682	.858	.866	.899
2	50	.628	.668	.538	.658	.716	.710
	100	.687	.706	.606	.703	.721	.744
	500	.730	.723	.570	.723	.731	.758
	1000	.712	.714	.570	.705	.717	.747
	10000	.721	.727	.577	.712	.724	.755
3	50	.502	.539	.421	.523	.575	.526
	100	.543	.558	.492	.560	.574	.557
	500	.574	.571	.460	.566	.574	.563
	1000	.571	.575	.464	.563	.574	.558
	10000	.575	.583	.472	.566	.578	.565

Notes, NP-True: Hamming distance weighted by conditional correlation based on true attributes profile; NP-Sumscore: Hamming distance weighted by conditional correlation based on Sumscore-estimated attribute profiles; Sum: Attribute Sumscore method; NPT: Hamming distance weighted by conditional correlation based on NP-estimated attribute profiles; NP: Hamming distance method.

The performance of the weighted methods in the unequal-testlet-size condition deteriorated compared to the equal-testlet-size condition. In equal-testlet-size condition, the weighted methods provided slightly higher CCRs than unweighted methods (i.e., the NP method) when testlet effects were large (e.g., β=3); in unequal-testlet-size condition, their CCRs were lower than the NP method. Recall that in conditional correlation estimations, the magnitude of conditional correlations was related to testlet sizes, specifically, the CC-NP and CC-Sumscore estimated average conditional correlations were larger for the small testlet than for the large testlet although the two testlets had the same true testlet effects (i.e., simulated with the same β value). The inaccurate estimation of conditional correlations led to the wrong weighting coefficients. That is, the items in smaller testlets received a larger penalty than those in larger testlets. It can be observed that in zero to small testlet effect conditions ($\beta \leq .5$), the weighted methods provided lower PARs than the unweighted methods. When testlet effects increased, the difference between weighted methods and the unweighted methods decreased.

The CCRs of the three weighted methods were almost identical. Testlet NP-True did not provide higher CCRS than any of the other weighted methods. This similarity between the weighted methods was unexpected because more accurate estimations of conditional correlations were anticipated to lead to higher classification accuracy.

When there were not testlet effects, PARs from the weighted methods were close to each other for the 2-testlet conditions and 4-testlet conditions. When the testlet effect increased, PARs for the 4-testlet test conditions became increasingly higher than the PARs of 2-testlet conditions. This difference may be due to the fact that the testlet size variation in 2- and 4-testlet tests were different from each other. The testlet size variation in the 4-testlet test (consists of 2-, 4-, 8-, and 10- items testlets) are smaller than that in the 2-testlet test (consists of 8-, 16- items testlets). When the true testlet effects within a test are equal, the test with the larger testlet size variation will result in large variation among estimated weights. Therefore, Hamming distance was weighed incorrectly because all testlets should be penalized equally if they have the same testlet effect sizes.

Figure 8. Distribution of PARs for Unequal Testlet Size 2-Testlet Tests

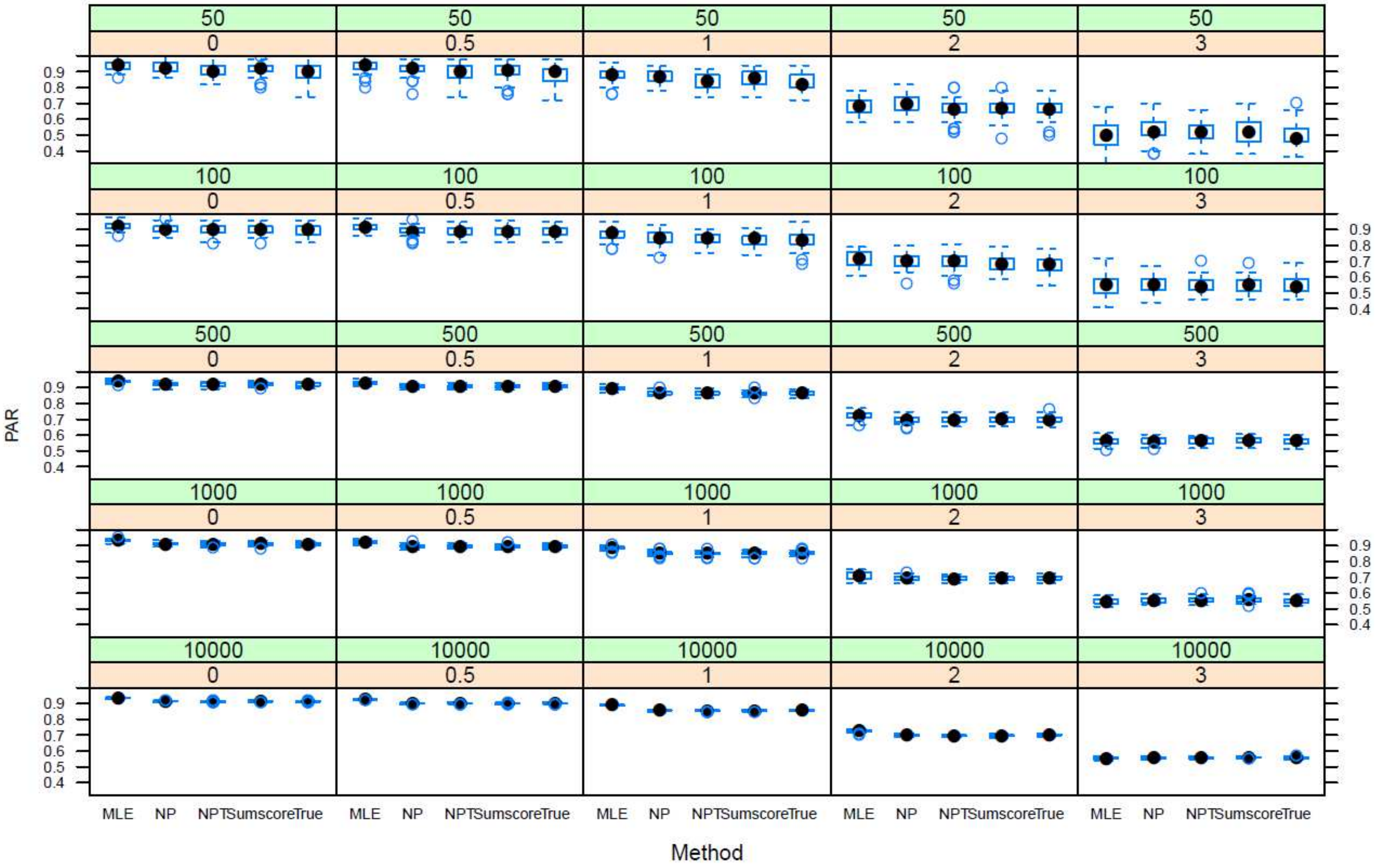
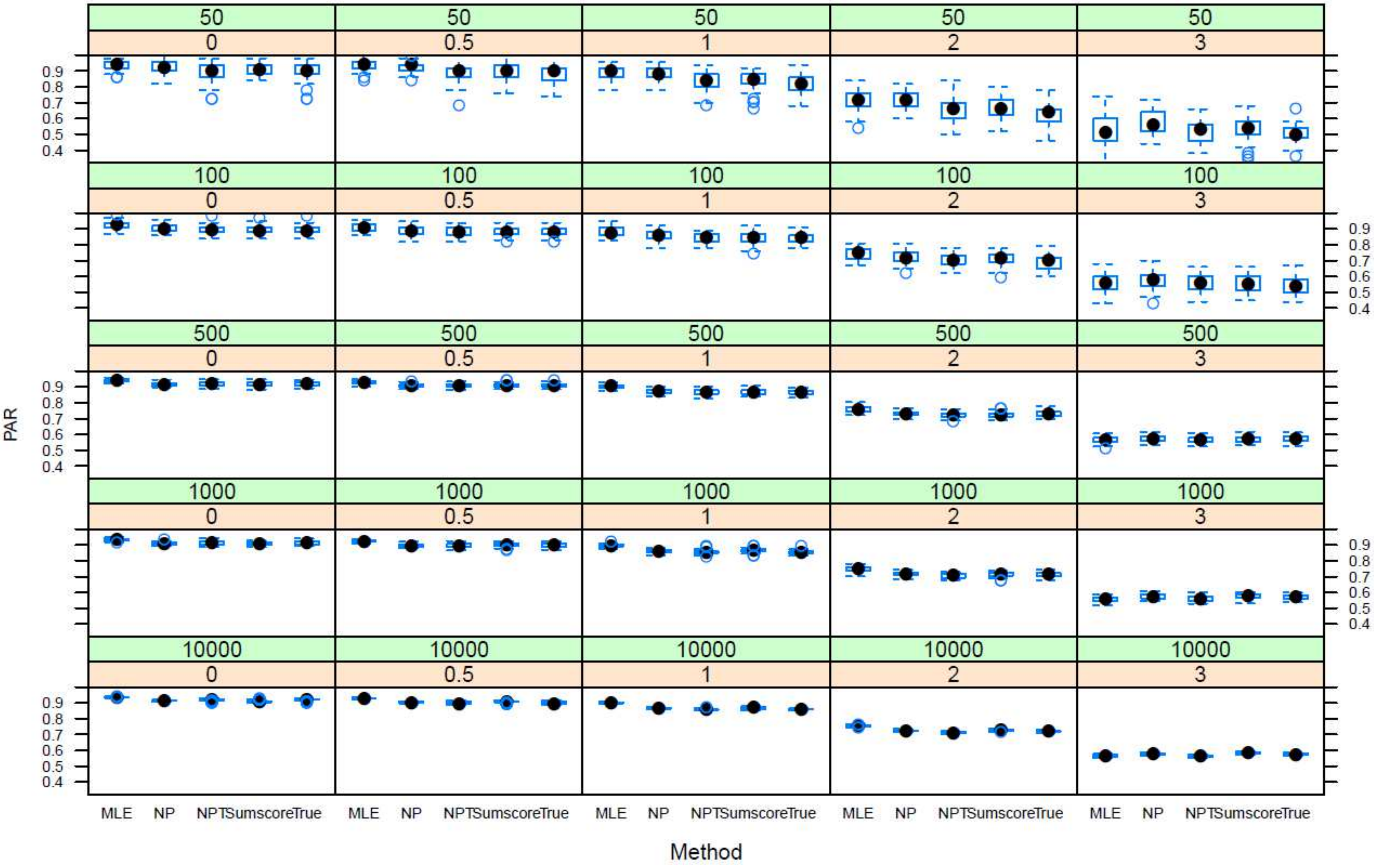


Figure 9. Distribution of PARs for Unequal Testlet Size 4-Testlet Tests



4.2.4 Summary of the Testlet NP Results

Testlet NP methods provided higher classification accuracy than NP in conditions where the testlet effects were large. However, across all conditions, MLE produced the highest classification accuracy except where the testlet effects were extremely large. Next, the results are summarized by the factors in testlet design.

4.2.4.1 Equality of Testlet Effect

The weighted Hamming distance methods provided higher classification accuracy than unweighted Hamming distance method when the testlet effects were unequal across testlets. It can be concluded that weighting the Hamming distance with a function of the average item-pair conditional correlation (Equation 38) improved classification accuracy. The purpose of testlet-NP methods is to penalize the items with larger testlet effect smore than items having smaller testlet effects. However, it should be noted that the advantage of weighted Hamming distance methods was found in tests that consisted of equal-size testlet.

4.2.4.2 Equality of Testlet Size

A comparison of Table 14 and Table 16 revealed that the classification accuracy of the proposed testlet NP methods in conditions with unequal testlet size were lower than those in conditions with equal testlet sizes. In addition, in conditions with equal testlet size, the weighted methods provided classification accuracies that were either similar or slightly higher than the unweighted method, whereas in conditions with unequal testlet size, the weighted methods produced lower classification accuracies than the unweighted methods. The deteriorated performance of weighted methods in unequal testlet-size condition suggests that there may be better alternatives defining a weight parameter. As was found in the section of conditional correlation estimation, the size of CC-NP and CC-Sumscore estimated item-pair conditional correlations were dependent on testlet size.

4.2.4.3 Testlet Size

The difference between classification accuracy for the 2-testlet condition and the 4-testlet condition was negligible in conditions with equal testlet size. However, in conditions with unequal testlet size, the classification accuracies for the 4-testlet tests were slightly higher than those of the 2-testlet tests, which might be due to a smaller difference in weights among items in 4-testlet tests when compared to 2-testlet tests. Because an interdependency was found between estimation of conditional correlations and testlet size in conditions with unequal testlet size, a conclusion cannot be arrived at whether or not testlet size influences the performance of weighted Hamming distance methods.

4.2.4.4 Testlet Effect

When the other factors were held constant, the weighted Hamming distance methods provided higher classification accuracies than the unweighted Hamming distance method (NP) in the large testlet effect conditions. Although not the focus of this study, it should be noticed that NP methods were comparable to the MLE when there was no testlet effects or small testlet effects, and had higher classification accuracies than MLE when testlet effect is large ($\beta = 3$).

4.2.4.5 Sample Size

When sample size increased and the testlet effects were fixed, the difference of classification accuracy between the weighted methods and the unweighted methods became smaller, and the weighted methods provided higher classification accuracy. The influence of sample size on weighted methods is due to the fact that the calculation of the weights (function of conditional correlation) is not independent of sample size. Large sample sizes provided more accurate estimation of conditional correlations.

Another thing about the impact of sample size on classification accuracy is that the N=500 in most conditions provide slightly higher CCRs than that in other sample size conditions. This result is contrary to our expectation, as in general, the larger the sample size, the better the classification accuracy. Future research might replicate the simulation study to investigate the impact of sample size on classification accuracy of different classification analysis methods.

CHAPTER V

DISCUSSION AND CONCLUSION

Local item dependency (LID) is an assumption for many psychometric models, such as item response models and diagnostic classification models. When the assumption of LID is met, there should be no significant covariance between items after conditioning on the respondents' ability (abilities, attribute profiles). As with other statistical models, inferences drawn from diagnostic classification analyses are valid if this assumption is reasonable

Oftentimes associations between item responses still exist even after conditioning on the attribute profile. This association indicates that the assumption of LID is violated and the validity of the inferences drawn from the analysis is challenged. LID can be caused by multiple sources as described in Chapter one. Item bundle or passage dependency is one of the causes that has been studied in IRT. Because of the popularity of testlets in today's assessment (Rosenbaum, 1984; Wainer, Bradlow & Wang, 2007; Lu, 2010; Zhang, 2010), it is necessary to investigate the issues related to testlet effects in diagnostic classification analysis.

Psychometric models have been developed to account for testlet effects, such as the testlet IRT models (Wainer & Wang, 2000; Wainer, Bradlow & Wang, 2007) and testlet diagnostic classification models (Hout & Cai, 2012; Hansen, 2013). In application, practitioners often must choose between the accuracy and efficiency (the ease and/or speed of administration). More accurate estimation often requires large sample size and more computation time. If the LID does not pose a serious threat to classification accuracy, practitioners often choose the model that is more parsimonious. In addition, modeling testlet effects that are negligible results in a more complicated model than necessary and potentially increases the error of parameter estimation (Demars, 2012). Therefore, it is necessary to understand the size of LID or testlet effects that exist among the item bundles and to what an extent classification accuracy can be impacted.

As discussed in Chapter One, traditional methods of LID detection are not practical in situations where nonparametric classification methods are applied. The conditional-correlation (CC) approach to measure testlet effects was then developed to provide a general estimation of testlet effect. Similar to the conditional-covariance approach of detecting item dependency in IRT (Stout et al., 1996; Douglas et al., 1998), this study assumes that the association between item pairs within a testlet should be close to zero or a small negative value if the LID assumption is met.

If the testlet effects seriously threaten test validity, it should be accounted for in classification analyses. It is assumed that penalizing the Hamming distance with a coefficient related to the testlet effect, 1-(nitem-1)**r*/nitem, and hence assigning more weight to the items that are less interdependent might increase the classification accuracy. Based on how initial attribute profiles are estimated, three weighted Hamming distance (testlet NP) methods for diagnostic classification analysis were proposed: the Hamming distance method weighted by CC-NP, the Hamming distance method weighted by CC-Sumscore, and the Hamming distance method weighted by CC-True. A simulation study was conducted to investigate whether or not the newly proposed testlet NP methods provide better estimations than the methods that ignore testlet effects. In the following discussion, a summary of the findings is first provided with respect to each of the two general research questions, followed by the implications and recommendations.

5.1 Can Item-pair Conditional Correlation be Used to Estimate Testlet Effect

Findings of the current study with respect to conditional correlation suggest several implications for practitioners. First, it was found that when the generated testlet-effect increased, the estimated average item-pair conditional correlation increased. The mean values presented in Tables 9-12 in Chapter Four provide some insights in the size of conditional correlation that suggests a violation of LID in diagnostic classification analysis. Because conditional correlations accessed in this study can be computed when performing diagnostic classification, practitioners can calculate this statistic first and inspect its magnitude before interpreting the results or applying more complicated classification methods. However, because the CC approach also requires large sample sizes to achieve stable estimates, the results listed in Chapter IV should be considered specific to particular sample sizes and number of measured attributes.

Second, sample size had a noticeable impact on the estimation of conditional correlation. In general, the larger the sample size, the larger the standard error of estimation, and the smaller the magnitude of the estimated conditional correlation. However, when the sample size reached 1000, the decrease of estimated conditional correlation was barely noticeable. On one hand, as sample size goes up, correlation coefficients fluctuate less around the "true" magnitude for the population *r*; therefore, the estimation error decreased. On the other hand, it is more likely to calculate a larger correlation value with a smaller sample size than with a larger sample size because it is easier to fit a linear relationship for less data points. An extreme case is the linear correlation between two data points A and B on a two-dimensional space; you can always fit a line through these two points. In addition, although the testlet component ξ_s was always generated with N (0, 1), the resulted variance from smaller sample size was always larger than that with larger sample size. For example, the resulted variance was 1.12 for *N*=500, but 1.00 for *N*=1000.

Third, the estimated conditional correlation by both CC-Sumscore and CC-NP was negatively related with testlet size. That is, when the other factors were fixed, the larger the testlet, the smaller the conditional correlations estimated by CC-NP and CC-Sumscore. However, there was not such a relationship for CC-True. It is expected that the larger testlets exert a larger influence on the attribute profile estimation, which makes the estimated attribute profiles explain more variance in the larger testlet and leave less shared variance unexplained. In CC-True, the attribute profiles were not estimated but the true generated profiles, the variation among the influences exerted by different testlets did not exist, therefore, the magnitude of conditional correlation was not dramatically different across testlets of different sizes. Because of the above stated problem, it is not recommended that the proposed methods are used to compare testlet effect sizes of multiple testlets when they vary in sizes. Future studies should investigate the relationship between testlet size and estimated conditional correlation in different conditions other than those in this study. It is also helpful to see if LD-X^2 and Yen's Q3 discover similar relationships. If such a negative relationship is not found by other LID measurement methods, it may suggest a problem with the proposed conditional correlation method.

Fourth, the magnitude of CC underestimation was not related to the accuracy of the initial attribute profile estimation. For example, although the Hamming distance method provided a higher classification accuracy than the attribute-sumscore method, CC-NP always underestimated CC-true more than CC-Sumscore did, except in the unequal testlet-size and 4-testlet condition. The unexplained shared variance estimated by the more precise classification method and less precise classification method are different. The conditional correlation based on NP estimated attribute profiles is more likely to be related to the testlet effect, whereas the conditional correlation based on attribute-sumscore estimated profiles is probably due to unexplained variances caused by inaccurate attribute classification. Considering there is no distinct difference between testlet NP based on CC-Sumscore and testlet NP based on CC-NP, the practitioners may choose either method to detect LID caused by the testlet.

Lastly, although CC-NP and CC-Sumscore methods both underestimated the true conditional correlation, the ordinal relationship between testlets with differing testlet effects was still preserved. Based on the results from the simulation studies, the following conclusions may be drawn: if the attribute profile is estimated through the Hamming distance method or attribute-Sumscore method with sufficient sample size, an average conditional correlation larger than .01 indicates the presence of a small LID. An average conditional correlation larger than .05 indicates the presence of a moderate LID, and an average conditional correlation larger than .1 signals a large LID. When the average conditional correlation is larger than .1, and the pattern-wise classification accuracy (PAR) is below .8, it is reasonable to consider using the testlet DCM to model the local item dependency.

CC-NP and CC-Sumscore both underestimated the true LID when the testlet effect was large. The bias in estimations could be a result of the method used to compute the conditional correlation. The initial attribute pattern was estimated from all item responses including the testlet items. This method of obtaining attribute pattern estimates may lead to a poor estimation of attribute profiles, as a result, the average item-pair correlation is computed based on an incorrect conditional variable. The above stated problem also exists in Yen's Q3. Practitioners may consider estimating the conditional correlation for each testlet by conditioning on attribute profile estimates based on all other items not included in that testlet.

Although correlations conditioned on attribute profiles in this study was developed to detect testlet effect, like Yen's *Q3*, it has the potential to be used to detect LID caused by other sources, such as incomplete/underspecified *Q matrix*, test speediness, etc. For example, in detecting LID caused by an incomplete or underspecified *Q matrix*, the conditional correlation can be calculated for all possible item pairs conditional on attribute profiles.

5.2 Can Testlet Hamming Distance Method Improve Classification Accuracy

The proposed testlet NP methods weight the original Hamming distance with a function of the testlet-specific average item-pair conditional correlation. Based on the method used to estimate the conditional correlation, three testlet NP methods were examined in the simulation study. The results demonstrated that there were no distinct differences in terms of classification accuracy between testlet NP method based on CC-true and testlet NP method based on either CC-NP or CC-Sumscore. This result suggests that the estimation method used to obtaining conditional correlations does not influence the performance of weighted Hamming distance methods. Therefore, in the following discussion, the three different weighted Hamming distance methods are not differentiated.

The weighted Hamming distance methods provided higher classification accuracy than the unweighted Hamming distance method (i.e., the NP method) when testlet effects were large (β =1) regardless of sample size. However, in small sample size conditions, this advantage of unweighted Hamming distance methods decreased when the testlet effect increased. In extremely large testlet effect conditions ($\beta = 3$), the weighted and unweighted Hamming distance methods all provided higher classification accuracy than MLE. In other testlet effect conditions, MLE provided the highest classification accuracy.

The influence of sample sizes on the classification accuracy of all methods was limited. Though the classification accuracy increased when the sample sizes became larger, the magnitude of this improvement was less than .01. In practice, if the number of measured attributes is small, sample size should not be a big concern in diagnostic classification analysis, especially when using the NP method and DINA-MLE. However, it should be noted that this conclusion is drawn from simulation conditions where the number of measured attributes was four. It is expected that the influence of sample size will be larger when the number of attributes increases. Future research might investigate the rate at which the classification accuracy deteriorates.

Consistent with what was found in Hansen (2013), where estimation bias for item parameters only occurred at β = 2, this study found that the impact of the testlet effect on DINA-MLE and the NP method was small when the testlet effect was within a reasonable range ($\beta < 2$). These findings suggests that both methods are quite robust to testlet effects. Therefore, when the average item-pair conditional correlation is less than .1, the impact of the testlet effect might not be a big concern for classification analysis. Based on the results from this study, it could be concluded that model techniques that account for the inter-item dependency should be implemented only when the average item-pair conditional correlation is greater than .1. This finding is also in line with what was found in testlet IRT studies (DeMars, 2012; Jiao & Zhang, 2014; Baghaei & Aryadoust, 2015), in which testlet effects had no noticeable impact on ability parameter estimation.

This study found that the NP method was comparable to MLE regarding classification accuracy, which is consistent with findings in Chiu and Douglas (2013). In fact, the NP method in this study even provided slightly higher classification accuracy when large LID was present. The finding described above indicates that the NP method is more robust to the violation of LID in terms of classification accuracy. It is probably because the MLE method needs to estimate both item and person parameters. As shown in studies of testlet IRT (Jiao et al., 2012; Jiao & Zhang 2014; Baghaei & Aryadoust, 2015), LID had more impact on item parameter estimations. In turn, item parameter estimations exert influence on person parameter estimations. Although the impact on person parameters is small, it should not be ignored when the testlet effects are large. In contrast, there is no parameter estimation in the NP method, which may reduce the impact of testlet effects on the examinee classification. Therefore, in testlet-based tests, the NP method can be used as an alternative to MLE when diagnostic assessment follows either a conjunctive condensation rule or a disjunctive condensation rule.

The criticism of parametric classification analysis is mainly due to its high demands of large sample sizes for model fitting (Chiu & Douglas, 2013; Wang & Douglas, 2015; Chiu & Köhn, 2015). Surprisingly, few studies have investigated how sample size influences the classification accuracy of parametric methods in comparison to the NP methods. Most simulation studies of diagnostic classification approaches used extremely large sample sizes to obtain stable estimations. However, diagnostic classification analyses are often based on small to medium scale assessments such as in classroom settings (Wang & Douglas, 2015) and psychiatric domains (Henson & Templin, 2006). Unexpectedly, the MLE method in this study provided classification accuracy as high as the NP method with a small sample size (N=50). Because this study only included tests that measure a small number of attributes as in Chiu & Douglas (2013), it can be concluded that sample size should not be a major concern for the MLE method if the number of measured attributes is no larger than four. Therefore, for practitioners, it is recommended that if the sample size is ≥ 50, the MLE method is still a reasonable option.

In summary, the performance of the testlet NP methods depends on the accuracy of estimation of the testlet effect. It is expected that when the weighting formula is improved, the testlet NP methods should provide higher classification accuracy.

5.3 Limitations and Future Research

With the increasing interest in diagnostic classification modeling (Huff & Goodman, 2007), there are still questions and problems left for its application and interpretation, such as differential item functioning, testlet effects, and item bias (Rupp & Templin, 2008). The current study investigates the problem related to testlet effects. Nonparametric methods were developed to detect testlet effects and then incorporate the testlet effects into the classification analyses. A simulation study was conducted to evaluate the proposed methods. Results of the simulation study should be cautiously interpreted because of the following limitations:

First, this study only included tests where all items belong to a testlet, and did not consider tests with both independent items and testlet items. If the classification is based purely on responses to interdependent items, the accuracy could be lower than when based on responses that include independent items. The estimated testlet effects could not reflect the true LID when classification was not accurate, hence, the interpretation of testlet effect becomes problematic. When the attribute profile is conditioned on more accurately estimated profiles, the item-pair conditional correlations will be more accurately estimated. Future research should consider including both independent items and testlet items in designing diagnostic assessments or conducting simulation studies.

Second, the choice of testlet effect conditions in this study was based on a previous simulation study (Hansen, 2013) rather than real test conditions. In IRT, testlet effects are measured by testlet variance and the variance rarely exceeds “3” (Wainer et al, 2007; Zhang, 2010; Jiao et al., 2012; Eckes, 2014). For example, Wainer et al. (2007) found that in the four testlets of the 1994-1995 administration of the North Carolina Test of Computer Skills exam, testlet variance ranged widely between .03 and 2.8. In other studies such as Jiao et al. (2012), the estimated testlet variance of a reading comprehension test could be very small (< .27). Papp, Glas and Veldkamp (2012) stated that a value 1.00 or larger is often found in real data-sets. Compared to studies in testlet IRT (unidimensional models), the testlet effect was rarely estimated or measured through fitting a DCM model with Hansen (2013) as an exception. In Hansen (2013), the LD-X^2 was used to measure LID caused by testlet effect when fitting the testlet DCM. The detected values of LD-X^2 for the two tests, PISA and TIMSS, were quite small. The reason that large testlet effects were used is because this study intended to examine 1) to what a degree that the classification methods ignoring testlet effect are robust to LID in terms of classification accuracy and 2) at what condition, the proposed CC methods and testlet NP methods work well. More studies need to be conducted with realistic β parameter to reflect testlet effect in reality.

Third, the number of attributes measured per test was fixed at a small value “4” in this study. In reality, the number of measured attributes could dramatically vary. For example, in Von Davier (2009), TOEFL iBT was retrofitted to measure three skills. The PISA 2000 reading comprehension test measured four attributes (Hansen, 2013). In other situations, there may be a need to learn about an examinee's attribute mastery at a finer-grain size or for specific curriculum standards. In that case, the number of measured attributes may be large. For example, TIMSS 2007 was retrofitted to measure 15 attributes in Lee et al (2011). Future studies should investigate how the testlet NP methods perform in assessments that measure a larger number of attributes.

Fourth, one potential reason that the testlet NP methods did not provide significantly higher classification accuracy than the NP method is that the accurate estimation of conditional correlation still depends on a large sample size. Oftentimes there are specific latent classes with only a few examinees. The conditional correlation estimation with examinees in that class was far from stable and accurate. For example, if two people have that particular latent class, the correlation between their responses to two items will be either "1" or "-1". To reduce the influence of inaccurately estimated conditional correlation caused by small sample size in a particular latent class, in obtaining the testlet-specific average conditional correlation, this study weighted the conditional correlation for each latent class by the corresponding number of respondents in that latent class. However, the estimated conditional correlation can be still inaccurate if the total number of examinees that take the assessments is small. For future research, the approximate estimation of testlet effects can be achieved through content experts' rating a testlet with respect to the inter-item dependency. Because the ratings are based on non-statistical item properties, the estimation can be done before the item response data is collected. For example, in Baldonado, Svetin, and Gorin's (2015) study, the linguistic experts were asked to rate the testlet items with respect to the common necessary information required to correctly answer those items. The expert used “0” to represent item pairs that were not “connected” by necessary information for correct responses, and “1” to represent those they were “connected”. If the testlet effect of each testlet is preknown from the interdependency rating by content experts, the rating can be incorporated in testlet NP methods. When rating “1” represents no testlet effect, rating “2” represents low testlet effect, rating “3” represents medium testlet effect, and so on, the penalty parameter in teslet NP methods (Equation 37) can be defined as γ_s = 1, 2, 3, … When there is no testlet effect, that is, γ_s = 1, the testlet NP method is the NP method.

Fifth, using R^2, the square of the correlation coefficient to show how much of the variation in two variables are associated, is probably more intuitive for approximating the shared variance among item pairs. In this study, condition correlation-Pearson’s r was used as a heuristic to approximate LID. The assumption behind using r to approximate LID is that the relationship between responses to the item pair is linear. If the relationship between items is not linear, Pearson’s r might underestimate the correlation between items. Therefore, future studies might consider approximating LID with R^2 and using testlet-specific average item-pair R^2 to approximate the testlet effect. This approach might increase the precision of weighting coefficients in testlet NP methods, and in turn, increase the classification accuracy rate.

Lastly, this study found that sample size 500 consistently produced the highest CCR among all sample size conditions and across classification methods. This is against our general knowledge about the impact of sample size on estimation. Future research might replicate the simulation study to investigate the impact of sample size on classification accuracy of different classification analysis methods. In addition, future research may also investigate the performance of CC approach and testlet NP methods in other test conditions such as test length, number of attributes, and especially item parameters. In this study, clean item parameters were chosen so that the noise caused by testlet effect in the NP classification is not confounded with that caused by slipping and guessing. However, clean item parameters are generally not realistic in real practice. In addition, although a previous study (Chiu & Douglas, 2013) found that both the NP method and DINA-MLE were impacted by slipping and guessing parameters, it will be interesting to understand how the sizes of slipping and guessing parameters are reflected in LID estimation.

Given the possible limitations, the main contributions of this paper are as follows: First, it contributes to a research gap in diagnostic classification analysis by presenting the nonparametric testlet effect detection methods: CC-NP and CC-Sumscore. Though those method underestimated the true conditional correlation in most cases, it did differentiate the testlets that vary in testlet effect size (magnitude of LID). Second, the proposed testlet NP methods represent an initial effort to account for testlet effects in nonparametric classification diagnostic analysis. The testlet NP methods generated higher classification accuracy than methods that ignore testlet effects in various test conditions. Though small, the improvement of the testlet NP methods from the NP is still encouraging. In high-stakes assessments, such as assessments that assign people into different remediation groups, a slight increase of .01 in terms of a classification accuracy rate can still create serious impact or consequences.

The proposed testlet effect detection method can be used in educational assessment settings where teachers or schools need to diagnose students' mastery status of a set of learning objectives, standards, or problem solving skills. As many reading tests include large proportion testlet items (e.g., North Carolina End of Grade Reading Test, NCDPI), it is important to measure the LID magnitude in these tests first and then give cautious explanation of the latent space. If the LID is moderate to large, more complicated models that account for testlet effect (e.g., testlet DCM models) or well-developed testlet NP methods should be applied.

REFERENCES

Ackerman, T. A. (1987). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. *ACT research report series*, 87-14. Iowa City, IA: American College Testing.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification*, 30, 195-224.

Bock, R. D., & Atkin, M. (1981). Marginal maximum likelihood/EM approach to item parameter estimation. *Psychometrika,* 46, 443-459.

Boughton, K.A., & Yamamoto, K. (2007). A HYBIRD model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp147-156). New York: Springer.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning,* 47, 423-466.

Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 589–624). Jyväskylä, Finland: University of Jyväskylä.

Cai, L. (2012). FlexMIRT: Flexmible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cao, Y., Lu, R.,& Tao, W. (2014). Effect of item response theory model selection on testlet-based test equating. *ETS Research Report Series*, 2, 1 -13.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37, 419-437.

Chen, J., & de la Torre, J. (2014). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50, 123-14.

Chiu, C., & Köhn, H. (2015). Consistency of cluster analysis for cognitive diagnosis: the DINO model and the DINA model revisited. *Applied Psychological Measurement*, 1-15.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnostic theory and applications. *Psychometrika*, 74, 633-665.

Chiu, C., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response pattern. *Journal of Classification*, 3, 225-25.

Choi, H. J. (2010). A model that combines diagnostic classification assessment with mixture item response theory models (*Unpublished doctoral dissertation*). University of Georgia, Athens, GA.

DeCarlo, L.T. (2011). On the analysis of fraction Subtraction data: The DINA model, classification, latent class sizes, and the Q matrix. *Applied Psychological Measurement,* 35, 8-26.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.

de la Torre, J. (2008), An empirically based method of Q matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple choice options. *Applied Psychological Measurement*, 33, 163-183.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43, 145-168.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, 36, 104-121.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327–359). Erlbaum: Hillsdale.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R Rao & S. Sinharay (Eds.) *Handbook of Statistics*, 26, (pp. 979-1030). Amsterdam: Elsevier.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, 23, 129-151.

Eckes, T. (2013). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, 31, 39-61.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Multivariate Applications Books Series. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, 57, 423-436.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 333-352.

Hansen, M. P. (2013). Hierarchical Item Response Models for Cognitive Diagnosis. (Unpublished doctoral dissertation), University of California at Los Angeles.

Hartz, S. (2002). A Bayesian framework for the united model for assessing cognitive abilities: Blending theory with practicality (*Unpublished doctoral dissertation*).Champaign, IL: University of Illinois.

Hartz, S. (2008). The Fusion Model for Skill diagnosis Blending theory with practicality (ETS Research Report RR-08-71).

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement*, 44, 361-376.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika,* 74, 191-21.

Hou, L, de la Torre, J. & Nandakumar, R. (2014). Diagnostic Modeling: applying of Wald test to investigate DIF in the DINA model, *Journal of Educational Measurement*, 51, 98-125.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge: Cambridge University Press.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chappelle, Y.R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, 26, 1-73.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49, 82-100.

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, 65-83.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.

Karelitz, T. M. (2004). Ordered category attribute coding framework for cognitive assessments (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods,* 8, 305-321.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin.

Lee, G, Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of tetslets. *Applied Psychological Measurement*, 25, 357-372.

Liu, Y. (2011). Identifying Local Dependence with a Score Test Statistic Based on the Bifactor 2-Parameter Logistic Model. (Unpublished Master Thesis), University of North Carolina at Chapel Hill.

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

Lu, R (2010). Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions. (Unpublished Master Thesis), University Maryland at College Park.

Macdonald, G. T. (2014). The Performance of the Linear Logistic Test Model When the Q matrix is Misspecified: A Simulation Study. (Unpublished Doctoral dissertation). University of South Florida.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537-563.

McCoy, T. & Willse, J. (2014). Accuracy of neural network versus nonparametric approaches in diagnostic classification. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.

McCoy, T. P. (2015). The Effects of Mixture-Induced Local Dependence on Diagnostic Classification. (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern Theories in*

Measurement: Problems and Issues (pp. 31-61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 54, 483-495.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, 18, 41-68.

Paap, M. C. S., & Veldkamp, B. P. (2012). Minimizing the testlet effect: Identifying critical testlet features by means of tree-based regression. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), Psychometrics in Practice at RCEC. Enschede, The Netherlands: RCEC, Cito/University of Twente.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2012).Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95, 129–14.

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2014). CDM: Cognitive Diagnosis Modeling. R Package Version 4.1. http://CRAN.Rproject.org/package=CDM

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425-435.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

Rupp, A. A. & Templin, J. L. (2008a). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219-262.

Rupp, A. A., & Templin, J. (2008b). The effects of Q matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68, 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. New York: Guilford Press.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q matrix construction: defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6, 190-209.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, 31, 1-33.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer et al. (Eds.), Computerized adaptive testing: A primer (2nd ed.). Hillsdale, NJ: Erlbaum.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively Diagnostic Assessment (pp. 327–359). Hillsdale NJ: Erlbaum.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12, 55-73.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & Safto, M. (Eds.). *Monitoring skills and knowledge acquisition* (pp.453-488). Hillsdale, NJ; Erlbaum.

Templin, J. (2004).*Generalized linear mixed proficiency models for cognitive diagnosis.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods,* 11, 287-305.

Templin, J*., &* Hoffman, L. *(*2013*).* Obtaining diagnostic classification model estimates using M*plus. Educational Measurement: Issues and Practice,* 32**,** 37-5.

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287-307.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37, 203-22.

Wang, S. & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, 80, 85-10.

Wang, W. C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29, 126-149.

Willse, J., Henson, R., & Templin, J. (2007). Using sum-scores or IRT in place of cognitive diagnosis models: can existing or more familiar models do the job?

Paper presented at the *Annual Meeting of the National Council on Measurement in Education*. Chicago, Illinois.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Effects of local item dependencies on the validity of item, test, and ability statistics. *Journal of Educational Measurement*, 39, 1-16.

Zhang, W. (2006). Detecting Differential Item Functioning Using the DINA Model (Unpublished doctoral dissertation). University of North Carolina at Greensboro.

Zheng, Y., Chiu, C.-Y., & Douglas, J. A. (2013). The NPCD Package: Nonparametric Methods for Cognitive Diagnosis, R package, http://cran.r-project.org/web/packages/NPCD/index.html.

APPENDIX A

SUMMARY OF TESTLET-SPECIFIC AVERAGE ITEM-PAIR CONDITIONAL CORRELATIONS FOR 4-TESTLET TESTS WITH UNEQUAL EFFECT SIZE

		TRUE		NP		Sumscore	
N	β	Mean	SD	Mean	SD	Mean	SD
50	.5	.394	.080	.536	.175	.494	.169
	1	.438	.080	.541	.144	.520	.141
	2	.506	.059	.636	.161	.537	.135
	3	.530	.049	.521	.121	.554	.147
100	.5	.234	.092	.379	.119	.320	.168
	1	.327	.080	.441	.169	.371	.149
	2	.453	.065	.461	.133	.437	.089
	3	.492	.059	.484	.154	.479	.121
500	.5	.063	.025	.124	.026	.128	.030
	1	.132	.028	.143	.039	.156	.045
	2	.299	.031	.186	.049	.247	.026
	3	.423	.041	.270	.062	.329	.032
1000	.5	.042	.013	.062	.015	.110	.023
	1	.102	.019	.097	.022	.114	.016
	2	.276	.025	.151	.029	.206	.019
	3	.401	.022	.225	.040	.284	.019
10,000	.5	.016	.004	.034	.005	.065	.005
	1	.076	.005	.069	.006	.082	.005
	2	.276	.006	.136	.009	.191	.005
	3	.399	.007	.208	.012	.272	.006

Note, the β value is corresponding to the four testlets in the same test.