Universidad Autónoma de Madrid

Facultad de Psicología

Modelos de Diagnóstico Cognitivo: Clasificación de Atributos, Funcionamiento Diferencial del Ítem y Aplicaciones

por

Guaner Rojas

Tesis Doctoral para optar al título de Doctor en Psicología

Directores

Julio Olea

Jimmy de la Torre

Cognitive Diagnosis Models: Attribute Classification, Differential Item Functioning and Applications

by

Guaner Rojas

A thesis submitted to the

Facultad de Psicología

in conformity with the requirements for

the degree of Doctor en Psicología

Universidad Autónoma de Madrid

Madrid, Spain

October 2013

Contents

Contents i

List of Tables iii

List of Figures v

Dedication 0

Acknowledgments 1

List of Abbreviations 2

List of Symbols 4

Abstract 5

Resumen 6

Chapter 1: Introduction 7

Chapter 2: Background 12

- 2.1 Description of models 13
  - 2.1.1 The DINA model 16
  - 2.1.2 The DINO model 17
  - 2.1.3 The G-DINA model 18
  - 2.1.4 The A-CDM 18
- 2.2 Models Estimation 19
  - 2.2.1 Parameter estimation of the DINA and DINO models 20
  - 2.2.2 Parameter estimation of the G-DINA model 21
  - 2.2.3 Model fit evaluation 21
- 2.3 Previous Research 22
  - 2.3.1 Attribute Classification 22
  - 2.3.2 Differential Item Functioning 24

2.3.3 Applications 25

Chapter 3: Choosing between general and specific CDMs **27**
3.1 Study I: Examining attribute classification accuracy 28
3.1.1 Method 29
3.1.2 Results 33
3.1.3 Conclusions 48
3.2 Study II: An application of CDMs to Asperger Syndrome data 49
3.2.1 Diagnosis of asperger syndrome 49
3.2.2 Method 51
3.2.3 Results 54
3.2.4 Conclusions 61

Chapter 4: Differential Item Functioning **63**
4.1 Study III: Detecting DIF in the DINA model 64
4.1.1 New DIF statistics for the DINA model 64
4.1.2 Method 69
4.1.3 Results 73
4.1.4 Conclusions 90
4.2 An example of empirical data analysis using SDI and UDI statistics . 92

Chapter 5: A Computer Software for calibrating CDMs **95**
5.1 Availability 96
5.2 Description 96
5.2.1 Interface Characteristics 97

Chapter 6: Discussion **107**
6.1 Attribute Classification 107
6.2 Differential Item Functioning 109
6.3 Applications 111
6.4 Limitations and Future Work 111

References **113**

Appendix A: Introducción **119**

Appendix B: Discusión **125**
B.1 Clasificación de Atributos 126
B.2 Funcionamiento Diferencial del Ítem 127
B.3 Aplicaciones 129
B.4 Limitaciones y líneas futuras de investigación 130

List of Tables

2.1 *Example of Q-matrix* 14

2.2 *Example of response matrix X with $J = 7$ and $I = 10$* 15

2.3 *Example of latent classes with $K = 4$* 16

3.1 *Q-matrix for the simulated data* 30

3.2 *Example of comparison of specific CDMs* 31

3.3 *Mean of RMSE of parameters recovery with GDINA, DINA and DINO models when $N = 200$ and $K = 5$* 47

3.4 *Q-matrix for the observed data* 53

3.5 *CDMs fit indices* 55

3.6 *Percentage of classification of individuals by group and posterior probabilities* 57

3.7 *Classification of individuals by group and number of attributes* 58

3.8 *Example of attribute pattern of individuals* 59

3.9 *Example of estimated item parameters* 60

4.1 *Q-matrix for the simulated data* 70

4.2 *Summary of simulation conditions* 72

4.3 *Summary of Type I error rates by indices ($\alpha = 0.05$)* 75

4.4 *Summary of SDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.1$ 78
4.5 *Summary of SDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.2$ 79
4.6 *Summary of SDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.3$ 80
4.7 *Significant values of the empirical UDI distribution by reference item parameter value and sample size* ($\alpha = 0.05$) 81
4.8 *Summary of UDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.1$ 84
4.9 *Summary of UDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.2$ 85
4.10 *Summary of UDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.3$ 86
4.11 *Summary of Type I error rates of MHP* ($\alpha = 0.05$) 87
4.12 *Summary of MHP power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.2$ 89
4.13 *Item parameters estimates for both groups* 94

List of Figures

3.1 Proportion of correctly classified individual attribute for data generated with DINA model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 36

3.2 Proportion of correctly classified individual attribute for data generated with DINO model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 37

3.3 Proportion of correctly classified individual attribute for data generated with A-CDM model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 38

3.4 Proportion of correctly classified individual attribute for data generated with DINA model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 40

3.5 Proportion of correctly classified individual attribute for data generated with DINO model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 41

3.6 Proportion of correctly classified individual attribute for data generated with A-CDM model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models. 42

3.7 Proportion of correctly classified attribute vectors when $N = 200$ and $K = 5$. Legends correspond to fitted models and data generating models are shown on the x-axis. iq=0 high item quality, iq=1 low item quality. 44

3.8 Proportion of correctly classified attribute vectors when $N = 400$ and $K = 10$. Labels correspond to fitted models and data generating models are shown on the x-axis. iq=0 high item quality, iq=1 low item quality. 45

4.1 Example of uniform and nonuniform DIF for one item when $g_{j0} = s_{j0} = 0.2$. 67

5.1 winCDM window 97

5.2 Exit Window 98

5.3 Model Specification Window 99

5.4 Model Selection Window 100

5.5 Model Selection start message Window 101

5.6 Model Selection finish message Window 101

5.7 Relative Fit Indices Output Window 102
5.8 Item parameter estimates Output Window 103
5.9 Attribute Classification Output Window 104
5.10 Attribute Prevalences Output Window 105
5.11 Latent Classes and its Posterior Probabilities Output Window 106

Dedication

This thesis is dedicated to my family: E, E, S, N, J, M, D, D, P, I, I, and S.

Acknowledgments

I would like to express my gratitude to the Universidad de Costa Rica and Fundación Carolina for funding this project. I am also grateful for the support from the Cátedra de Modelos y Aplicaciones Psicométicos at Universidad Autónoma de Madrid. I especially wish to thank my advisors Dr. Julio Olea and Dr. Jimmy de la Torre at the Rutgers University for their guidance throughout the research.

List of Abbreviations

AIC	Akaike information criterion
ACA	Attribute classification accuracy
A-CDM	Additive cognitive diagnosis model
BIC	Bayesian information criterion
CDMs	Cognitive diagnosis models
CTT	Classical test theory
DIF	Differential item functioning
DINA	Deterministic input noisy and gate model
DINO	Deterministic input noisy or gate model
EAP	Expected a posteriori method
G-DINA	Generalized DINA model
IRF	Item response function
IRT	Item response theory
$-2LL$	Deviance
MMLE	Marginal maximum likelihood estimation
PCA	Proportion of classified attribute
PCV	Proportion of classified vector
SDI	Signed difference statistic
UDI	Unsigned difference statistic

List of Symbols

$\boldsymbol{\alpha}$	Attribute pattern
g_j	Guessing parameter of item j
J	Number of items
K	Number of attributes
L	Number of latent classes
I	Number of individuals
s_j	Slipping parameter of item j

Abstract

At present, a variety of cognitive diagnosis models (CDMs) that vary in generality (i.e., complexity) have been proposed. As with most psychometric models, parameters of more general models require larger sample size to be calibrated accurately. In the current work, commonly used general and specific cognitive diagnosis models are systematically explored in terms of attribute classification accuracy (ACA) and differential item functioning (DIF). It is also provided a detailed investigation to help researchers and practitioners evaluate conditions where a general or specific model can be more appropiate. Conditions such as item quality, sample size, test length, true model, and number of attributes are considered in a ACA simulation study, whereas factors such as sample size, item quality, DIF size, DIF type, and number of attributes per item are investigated in two DIF simulation studies, in which it is proposed two new indices for DIF detection. In addition to ACA and DIF studies, the present project provides two examples using real data. One of the data sets comes from an application of a scale designed to detect individuals with Asperger Syndrome and the other comes from TIMSS 2007 fourth grade mathematics assessment. Finally, a special purpose software was designed and develop to perform CDMs estimation.

Resumen

En el contexto de los modelos de diagnóstico cognitivo (MDC) se ha propuesto modelos que varían en complejidad. Como es de esperar, los MDC más generales requieren tamaños de muestra más grandes para obtener estimaciones más precisas. En este trabajo se investiga sistematicamente un MDC general y varios específicos en términos de la precisión de la clasificación de atributos (ACA) y el funcionamiento diferencial del ítem (DIF). También, se expone una investigación detallada para ayudar a investigadores y profesionales a evaluar las condiciones donde un modelo general o específico podría ser más apropiado. Las condiciones de calidad de ítems, tamaño de la muestra, longitud de test, modelo verdadero y número de atributos se han considerado en un estudio de simulación de ACA, mientras que los factores de tamaño de la muestra, calidad de ítems, tipo de DIF, tamaño de DIF, y número de atributos se ha analizado en dos estudios de simulación de DIF, en los cuales se han propuesto dos índices para la detección de DIF. Además, el presente proyecto proporciona dos ejemplos con datos reales. Uno de los datos provienen de una aplicación de una escala para detectar personas con Síndrome de Asperger y el otro conjunto de datos pertenece a la aplicación del 2007 de TIMMS. Finalmente, se presenta un programa diseñado y desarrollado para realizar estimación de los MDC.

Chapter 1

Introduction

Psychological testing plays an important role in settings such as educational, clinical and organizational psychology. For example, educators are using test scores to determine who will be admitted to university, clinician psychologist are using tests to help diagnose psychological disorders, and organizational psychologist have test to select people for jobs. Within the psychological testing, a concern which is closely to psychological measurement is the assignment of candidate score according to a specific measurement theory.

Two commonly used measurement models are the classical test theory (CTT) and the latent variable models. The central concept in the first model is the expected value of the observed score, while the second conceptualizes theoretical attributes as latent variables. An important idea associated with latent variable theory is the use of statistical models fitted to the observed data to estimate respondent's scores.

Common statistical models, as in confirmatory factor analysis (CFA) and uni- and multi-dimensional item response theory (IRT), the respondent's score on latent variables are assumed to be continuous. Based on that continuous score that has been assigned to examinees, a classification into different levels on the assessment can be made by researchers identifying cut-scores on the continuous latent scale.

In spite of the popularity of both CTT and IRT approaches, models known as cognitive diagnosis models (CDMs) have seen an increase in the recent measurement literature (de la Torre, 2011; 2009; de la Torre & Lee, 2010; Junker & Sijtsma, 2001; Henson, Templin & Willse, 2009; Huebner, 2010; Rupp, Templin & Henson, 2010; von Davier, 2005) and the foremost international conferences such as the meeting of the Psychometric Society and the National Council on Measurement in Education. Most of the CDMs developments have focused on the formulations and estimation of new models.

CDMs are multidimensional and confirmatory models developed specifically for diagnosing the presence or absence of multiple attributes required for solving test items. Attribute is a term referred to latent variable which is assumed to be discrete. Multidimensional nature of CDMs sets multiple attributes to be measured by a test, and its confirmatory aspect associates a prior structure based on substantive theory. Thus, a CDMs conceptual key focuses on a matrix of attribute specifications, called the Q-matrix. The Q-matrix is crucial for model parameters estimation because it describes which item loads on each attribute.

In addition to the multidimensional and confirmatory characteristic, the item responses are modeled by the item parameters and attribute patterns. The number of items parameters depends on the generality of the model used to describe the observe data. For example, the DINA model (Junker & Sijtsma, 2001), one of the parsimonious models has two item parameters to be estimated per item. However, models such as G-DINA model (de la Torre, 2011) has parameters depending on the number of attributes involving an item. Regardless the implemented CDM, a vector or pattern containing the attribute probabilities are estimated. The attribute vector of probabilities is usually expressed as zeros and ones. The probabilities closest to one are transformed to one, and this value in a attribute pattern indicates that a person has the attribute of interest.

The main goal of CDMs is to classified individuals into a set of predefined categories or latent classes. The categories come from the number of attributes measured by a test. By implementing CDMs as assessment tool, each person receives a profile with information that can be used by researchers, teachers or psychologist to develop action plans in educational and psychological settings. For example, in clinical psychology, the attribute patterns can help clinicians with information that may be useful in the treatment of a disorder (Templin & Henson, 2006). Moreover, in educational context, the profiles would provide clues in designing instructional or learning activities for a given classification outcome (DiBello, Roussos & Stout, 2006).

One of the most important purposes of an assessment is to obtain valid and accurate estimates of examinees in the latent variable of interest. The examinee scores are expressed by the attribute classification in the attribute pattern. Estimation of the attribute classification is affected by conditions such as number of attributes, sample size, item quality and test length (e.g., de la Torre, Hong & Deng, 2010; Rupp & Templin, 2008a; Rupp & Templin, 2008b). Simulation studies of von Davier (2004) and de la Torre and Douglas (2004) showed that CDMs such as the general diagnostic model and the DINA model can offer attribute classification accuracy at individual atrribute level greater than 90% when the model underlying to the data is correct. However, there are no definitive answers regarding sample size requirements when researchers choose a model for attribute classification purposes.

Another statistical and methodological issue which arise in the CDMs paradigm is the item bias, in which little research has been done (Rupp & Templin, 2008a; Li, 2008; Zhang, 2006). Because each item should contribute to the discrimination between latent classes and the attribute probabilities are estimated by assuming known item parameters, the question of item non-invariance is most relevant one in attribute classification across subgroups of respondents. The item non-invariance can be explored through differential item functioning (DIF). The presence of DIF could influence the item parameter estimates, and this may have an effect on attribute classification.

One of the reasons to arrive at the benefits of the CDMs implementation concerns to software for CDMs estimation. Programs such as R (R Core Team, 2013) and Ox (Doornik, 2003) use a programing code to fit the CDMs. Researchers, practitioners or test developers may have less experience with environments that require programming language to do analyses rather than commonly used point and click software. Programs based on a graphical user interface may help researchers to perform CDMs estimation without programming skill requirements.

A major concern associated with this thesis is attribute classification assessed through two methodological issues: attribute classification accuracy and differential item functioning assessment in the CDMs framework. Hence, there were three main goals for this thesis. The first goal was to systematically compare the impact of small sample size on the attribute classification accuracy of general and specific CDMs. The second goal of this dissertation was to introduce a new procedure for identifying item differential functioning in the CDMs context. The final goal of the thesis was to develop a computer program for calibrating item and person parameters for CDMs.

This dissertation will be divides into seven chapters. In chapter two is provided an introduction to CDMs framework. Chapter three describes a simulation study, which was implemented to compare a general and three specific CDMs in terms of attribute classification accuracy. Chapter three also describes the implementation details and the analysis of a clinical tool in the context of CDMs. Chapter four proposes and systematically analyzes a new method for differential item functioning detection in the DINA model. Chapter six describes a point and click computer software developed specifically for calibrating CDMs. Chapter seven concludes and outlines future work.

Chapter 2

Background

Several CDMs make specific assumptions about how attributes combine or interact to produce an item response. An important distinction in commonly used is that of the model being either conjunctive or disjunctive (Rupp, Templin & Henson, 2010). Models are conjunctive if all the required attributes are necessary for successful completion of the item. In contrast, models are disjunctive if the absence of one attribute can be made up for the presence of other attributes. Other CDMs assume that mastery of attributes has an additive effect. Examples of specific CDMs are the DINA (*deterministic input, noisy "and" gate*; Junker & Sijtsma, 2001; de la Torre, 2009) model, DINO (*deterministic input, noisy "or" gate*; Templin & Henson, 2006) model, and the A-CDM (*additive* CDM; de la Torre, 2011). The DINA model is said to be conjunctive, and the DINO model is disjunctive.

According to Rupp, Templin and Henson, (2010), other well known CDMs are the NIDA model (*noisy input deterministic and*; Junker & Sijtsma, 2001, Maris, 1999), the NIDO (*noisy input deterministic or*, Templin, Henson, and Douglas, 2006) model, and the R-RUM (*reduced reparametrized unified model*; Hartz, 2002). Moreover, researchers have proposed general CDMs which reflect the assumptions of specifics models (see, e.g., Henson, Templin & Willse, 2009; von Davier, 2005). Examples of general CDMs are the G-DINA (*generalized* DINA; de la Torre, 2011) model, the log-linear cognitive dianosis model (LCDM; Henson, Templin & Willse, 2009), and the general diagnostic model (GDM; von Davier, 2005). These models describes the probability of success in terms of the sum of the effects due the presence of specific attributes and their interactions. In the next section, four of the most common CDMs as described.

2.1 Description of models

Most of the CDMs utilize a Q-matrix (Tatsuoka, 1983) to organize the attributes that are believed to be involved to solving the test items. An attribute is a task, subtask, cognitive process, or skill involve in answering an item. The Q-matrix is binary and of order J items by K attributes, as in, $Q = \{q_{jk}\}$, where $j = 1, \ldots, J$ and $k = 1, \ldots, K$; if item j involves attribute k then $q_{jk} = 1$, and $q_{jk} = 0$ otherwise. An example of a Q-matrix is displayed in Table 2.1, in which $J = 7$ and $K = 4$. For instance, item five measures attributes first, third and fourth, but no the second.

Table 2.1. *Example of Q-matrix*

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	1	0	1
5	1	0	1	1
6	0	1	1	0
7	1	1	1	0

In addition to the Q-matrix, CDMs generally requires a binary response matrix X of order I examinees by J items. The response vector of examinee i will be denoted by $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{iJ})$, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Table 2.2 depicts the responses of 10 individuals to seven items. The correct response is represented by the number one in the response matrix of the Table 2.3.

Table 2.2. *Example of response matrix X with $J = 7$ and $I = 10$*

Person	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
1	0	1	0	1	1	0	1
2	1	1	1	1	1	1	0
3	0	0	1	1	0	0	0
4	1	1	1	1	1	1	1
5	1	0	0	1	1	1	1
6	0	1	1	1	1	0	0
7	1	0	1	1	1	0	1
8	1	1	1	1	1	1	1
9	1	0	1	1	1	1	1
10	0	1	1	1	0	0	0

The primary objective of CDMs is to classify examinees into 2^K latent classes for an assessment diagnosing K attributes. Each latent class is denoted by α_l, where $l = 1, \ldots, 2^K$. CDMs assign to each examinee i an attribute vector of length K denoted by $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik}, \ldots, \alpha_{iK})$. Specifically, $\alpha_{ik} = 1$ if the k^{th} has been mastered by the i^{th} examinee, and $\alpha_{ik} = 0$ if the k^{th} attribute has not been mastered. Each attribute vector or pattern defines a unique latent class, thus, K attributes define 2^K latent classes. Moreover, all the CDMs express by $P(X_j = 1 \mid \alpha_l)$ the conditional probability of success on item j given the attribute vector of latent class l, where $l = 1, \ldots, 2^K$. Based on the Table 2.1, the $K = 4$ attributes define 16 latent classes expressed by Table 2.3. For instance, an attribute pattern of $\alpha_i = (1, 0, 1, 0)$

indicates that person i possesses the first and third attribute, but not the second and fourth, and then person i is classified into latent class $l = 7$ of the Table 2.3.

Table 2.3. *Example of latent classes with* $K = 4$

Latent class l	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	1	1	0	0
7	1	0	1	0
8	1	0	0	1
9	0	1	1	0
10	0	1	0	1
11	0	0	1	1
12	1	1	1	0
13	1	1	0	1
14	0	1	1	1
15	1	0	1	1
16	1	1	1	1

2.1.1 The DINA model

The DINA model partitions the latent classes into two groups for each item j. The DINA model has one s_j slip parameter and one g_j guessing parameter per item j. The model specifies that, for item j, only examinees who have mastered all the required attributes will have probability of success equal to $1 - s_j$, whereas all other examinees will have a chance of success equal to g_j. Given the slip and guessing parameters s_j and g_j, the item response function (IRF) is written as

$$P(X_j = 1 \mid \alpha_l) = P(X_j = 1 \mid \eta_{jl}) = g_j^{(1-\eta_{jl})}(1 - s_j)^{\eta_{jl}} \tag{2.1}$$

where $\eta_{jl} = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}}$ is the deterministic component of the model. Note that the η_{jl} is a binary indicator signifying whether or not the i^{th} examinee possesses all the required skills for item j.

The slip parameter s_j is the probability that the examinees in latent class l whose $\eta_{jl} = 1$ will slip and incorrectly answer item j (i.e., an incorrect response despite the examinee having mastered all the required skills for that item), and the guessing parameter g_j is the probability that the examinees in latent class l whose $\eta_{jl} = 0$ will guess and correctly answer the item (i.e., a correct response despite the examinee not having mastered all the required skills for that item). Formally, s_j and g_j are defined as $s_j = P(X_j = 0 \mid \eta_{jl} = 1)$ and $g_j = P(X_j = 1 \mid \eta_{jl} = 0)$.

2.1.2 The DINO model

The DINO model also partitions the latent classes into two groups for each item j. It is assumed that an item can be answered correctly if at least one of the required attributes involved in the item has been mastered. Given the slip and guessing parameters s'_j and g'_j, its IRF is written as

$$P(X_j = 1 \mid \alpha_l) = P(X_j = 1 \mid \zeta_{jl}) = {g'_j}^{(1-\zeta_{jl})}(1 - s'_j)^{\zeta_{jl}}, \tag{2.2}$$

where $\zeta_{jl} = 1 - \prod_{k=1}^{K}(1 - \alpha_{lk})^{q_{jk}}$ is the deterministic component of the model.

The slip parameter s'_j is the probability that the examinees in latent class l whose $\zeta_{jl} = 1$ will slip and incorrectly answer the item j, and the guessing parameter g'_j is the probability that the examinees in latent class l whose $\zeta_{jl} = 0$ guess and correctly answer the item. Formally, s'_j and g'_j are defined as $s'_j = P(X_j = 0 \mid \zeta_{jl} = 1)$ and $g'_j = P(X_j = 1 \mid \zeta_{jl} = 0)$.

2.1.3 The G-DINA model

The G-DINA model partitions the latent classes into 2^{K_j} groups for each item j, where $K_j = \sum_{k=1}^{K} q_{jk}$ represent the required attributes for item j, and q_{jk} is the k^{th} element of the j^{th} row of the Q-matrix. The G-DINA model describes the probability of success on item j in terms of the sum of the effects of involved attributes, and their interactions. Specifically, the probability that examinees with attribute pattern α_l will answer item j correctly is expressed by

$$P(X_j = 1|\alpha_l) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (2.3)$$

where δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j}$.

2.1.4 The *A*-CDM

The *A*-CDM model has $K_j + 1$ parameters for item j. This model indicates that mastering attribute α_k increases the probability of success on item j, and its contribution is independent of the contributions of the other attributes. By constraining the parameters of the G-DINA model, de la Torre (2011) has shown that the general formulation above reduce to some commonly used CDMs. These models include the A-CDM, the DINA model, and the DINO model. For instance, the A-CDM model can be obtained from the G-DINA model by setting all interaction effect to zero. The IRF is written as

$$P(X_j = 1|\alpha_l) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk}\alpha_{lk} \tag{2.4}$$

2.2 Models Estimation

A commonly used technique to estimate the CDMs parameters is marginal maximun likelihood estimation (MMLE; de la Torre, 2009). Under this approach, the item parameters are assumed to be known, and then attribute patterns are obtained through expected a posteriori method. In this section, the MMLE procedure is presented.

The marginal probability can be written as

$$L(X_i) = \sum_{l=1}^{L} L(X_i \mid \alpha_l)p(\alpha_l) \tag{2.5}$$

where $L(X_i \mid \alpha_l) = \prod_{j=1}^{J} P_j(\alpha_i)^{X_{ij}}(1 - P_j(\alpha_i))^{1-X_{ij}}$ is the likelihood of the response vector of examinee i conditional on attribute profile α_l, and $p(\alpha_l)$ is the prior probability of α_l in the population.

The marginal likelihood of the response data is written as

$$L(X) = \prod_{i=1}^{I} L(X_i) = \prod_{i=1}^{I}\sum_{l=1}^{L} L(X_i \mid \alpha_l)p(\alpha_l) \tag{2.6}$$

The log-marginalized likelihood of the response data is written as

$$l(X) = \ln L(X) = \ln \prod_{i=1}^{I} L(X_i) = \ln \prod_{i=1}^{I} \sum_{l=1}^{L} L(X_i \mid \alpha_l) p(\alpha_l) \tag{2.7}$$

Based on the equation (2.7), the parameter estimates of the DINA, DINO, and G-DINA can be obtained.

2.2.1 Parameter estimation of the DINA and DINO models

To obtain the maximum likelihood estimate of the structural parameters $= g_j$ and s_j of the DINA model, the equation (2.7) is maximized by taking the derivate of $l(X)$ (i.e., $\partial l(X)$) with respect to $= g_j$ and s_j respectively. According to de la Torre (2009), maximization of $\partial l(X)$ gives the estimator $\hat{g} = \frac{R_{jl}^{(0)}}{I_{jl}^{(0)}}$, where $I_{jl}^{(0)}$ is the expected number of examinees lacking at least one of the required attributes for item j and where $R_{jl}^{(0)}$ is the expected number of examinees among $I_{jl}^{(0)}$ correctly answering item j. Similarly, the estimator $\hat{s}$ can be expressed as $\hat{s} = \frac{I_{jl}^{(1)} - R_{jl}^{(1)}}{I_{jl}^{(1)}}$, where $I_{jl}^{(1)}$ and $R_{jl}^{(1)}$ represent the examinees with all the required attributes for item j.

Finally, if $\beta = (g_j, s_j)$, the root of the diagonal elements of $I^{-1}(\hat{\beta})$ represents the $SE(\hat{\beta})$ and the information matrix $I(\beta) = -E(\frac{\partial^2 l(X)}{\partial \beta})$ is the expectation of the second partial derivate of the equation (2.7) with respect to β.

In the DINO model the estimator $\hat{g'} = \frac{R_{jl}^{(0)}}{I_{jl}^{(0)}}$, where $I_{jl}^{(0)}$ is the expected number of examinees lacking all of the required attributes for item j and where $R_{jl}^{(0)}$ is the expected number of examinees among $I_{jl}^{(0)}$ correctly answering item j. Similarly, the estimator $\hat{s'}$ can be expressed as $\hat{s} = \frac{I_{jl}^{(1)} - R_{jl}^{(1)}}{I_{jl}^{(1)}}$, where $I_{jl}^{(1)}$ and $R_{jl}^{(1)}$ represent the examinees with at least one of the required attributes for item j.

2.2.2 Parameter estimation of the G-DINA model

By considering equations 2.3 and 2.7, computation of the derivate of equation 2.7 with respect to $P(\alpha_{lj})$, and solving for $P(\alpha_{lj})$, it is obtained an approximation, $\hat{P}(\alpha_{lj})$, expressed by the number of examinees in the latent group α_{lj} expected to answer item j correctly, over the number of examinees expected to be in the latent group α_{lj}. Similarly, as in DINA model the second derivate of the equation 2.7 offers the standard error of $\hat{P}(\alpha_{lj})$.

2.2.3 Model fit evaluation

For inferences from CDMs to be valid, it is necessary to evaluate the fit of the model to the observed data. To do so, it should be used statistics to compare different CDMs and Q matrices. Fit statistics for evaluating model-data fit such as deviance ($-2LL$; minus twice the maximum log-likelihood), Akaike information criterion (AIC; Akaike, 1974; the deviance plus twice the number of parameters) and Bayesian information criterion (BIC; Schwarz, 1978; the deviance plus the logarithm of the sample size times the number of parameters) can be also used to compare different CDMs.

Based on the the maximun likelihood of expression 2.7, the deviance, AIC and BIC are written as

$$deviance = -2l(X), \tag{2.8}$$

$$AIC = deviance + 2P \tag{2.9}$$

and

$$BIC = deviance + P\ln(I), \tag{2.10}$$

respectively, where P is the number of model parameters and I is the sample size. According to Chen, de la Torre and Zhang (2012), if J represents the test length, then $P = 2J + 2^K - 1$ for the DINA model. For the G-DINA model $P = \sum_{j=1}^{J} 2^{K_j^*} + 2^K - 1$ where K_j^* denotes requiered attributes for item j and K is the number of attributes measured by a test.

In addition to the AIC and BIC criteria, absolute fit indices such as the proportion correct, transformed correlation, and log-odds ratio have been studied by Chen, de la Torre and Zhang (2012). These indices were proposed with their corresponding standard error, and then the z-scores of the three statistics can be derived to test whether the residuals differ significantly from zero.

2.3 Previous Research

2.3.1 Attribute Classification

In CDMs the responses to test items provide the estimated item and person parameters. The item parameters estimates give a probability of correctly answering an item, and the person parameters estimates are expressed as attribute mastery pattern.

Despite the factors affecting item and person (i.e, attribute pattern) parameters estimation when the reduced DINA and DINO models, the A-CDM and the G-DINA model alluded above have been investigated, there is no consensus on how factors (e.g., number of attributes, sample size, item quality, test length) interact to affect attribute classification accuracy.

A review of the literature on both simulation studies and application examples with CDMs shows that most works use about four to ten attributes (see, e.g., de la Torre, 2011; de la Torre, Hong & Deng, 2010; Rupp & Templin, 2008a; Rupp & Templin, 2008b; Templin & Henson, 2006). The simulation studies indicate that level of guessing and slip parameter can have a profound impact in minimizing the attribute misclassification rate (de la Torre, Hong & Deng, 2010). Rupp and Templin (2008b) has shown that the proportion of items measuring an attribute and the number of attributes measured by an item can affect estimation accuracy.

Regarding to sample size, there is no consensus on minimum sample size, for instance, de la Torre, Hong and Deng (2010) suggested that a sample of 1000 examinees would be sufficient accurate estimation of the DINA model parameters, whereas Rupp and Templin (2008a) recommended that for the DINA model and DINO model a sample size consisting of a few hundred respondents is sufficient for convergence, especially when the number of attributes measured by test is not too large, such as four to six, and the tests are of moderate length (e.g., 20 or 40 items).

De la Torre, Hong and Deng (2010) also used a Q-matrix and test length, two factors that have direct impact on the attribute classification rate, but the number of attributes and test length were fixed to $K = 6$ and $J = 15$, repectively. Similarly, fixed conditions such as sample size, number of items, and number of attributes (i.e., $N = 2000$, $J = 30$ and $K = 5$) which were examined in the de la Torre (2011) simulation study, provided accurate estimation of the G-DINA model parameters. Despite the detail description of the conducted simulation studies, factors such a sample size and test length were not manipulated.

Choosing between a general or specific model is an important issue in applying

CDMs because a general model implies more item parameters than a specific model, which in turn require a larger sample size to obtain stable item parameter estimates. For example, when the number of required attributes for item j is $K_j = 3$, the DINA and DINO models have two parameters, the A-CDM have $K_j + 1 = 4$ parameters, whereas the G-DINA model has $2^{K_j} = 8$ parameters. Because of this, although the G-DINA model subsumes the many specific CDMs, including the three CDMs alluded above, it is not clear whether it is the model that should be used when the sample size is small.

2.3.2 Differential Item Functioning

In the CDMs, it can be said that differential item functioning (DIF) is present when the probability of correctly responding to a particular item differs across manifest groups of individual with the same attribute mastery pattern. Thus far only a few DIF detection studies have been reported within the CDMs. Zhang (2006) examined DIF by matching the examinees on their attribute profile scores from the DINA model (de la Torre, 2009; Junker & Sijtsma, 2001) to investigate the efficiency of Mantel-Haenszel (Holland & Thayer, 1988) and SIBTEST (Shealy & Stout, 1993) procedures with attribute profile score as the matching criterion for DIF detection.

Li (2008) used a modified higher-order DINA (de la Torre & Douglas, 2004) model for separating the source of construct relevant (i.e., benign) DIF from construct irrelevant DIF (i.e., adverse). The higher-order DINA model was calibrated with different sets of item parameters for the reference and focal groups, and then the DIF was studied by using the marginalized differences in probabilities of success of an item.

Potential limitations in the studies described above need to be addressed. First, in Zhang (2006) study the attribute patterns for the reference and focal group were not separately estimated. This means that the estimates of attribute vectors are biased, and the matching variable is contaminated. Besides, methods of Mantel-Haenszel and SIBTEST showed lower power for detecting nonuniform DIF. Second, according to Li (2008), Type I error rates in some simulation conditions appeared to be out of control. Third, both the Zhang (2006) and Li (2008) studies implemented a relatively small numbers of replications for each simulated condition.

According to these limitations, new effective methods for DIF detection need to be implemented based on the CDMs framework. Particularly, the new methods need to include separate item and attributes pattern parameters calibration for comparison groups, and the method should have higher power of detecting both uniform and nonuniform DIF.

2.3.3 Applications

At present, there is a growing interest among researchers and practitioners to use CDMs in applied situations (e.g., Leighton & Gierl, 2003; Roussos, Templin, & Henson, 2007, 2009; Embretson, 2010). Primarily, applications of CDMs such as DINA model have been used in educational measurement, in which the information is used for diagnosing students' strengths and weaknesses, giving researchers and teachers information that can be used to design treatments and supports. Although educational applications have dominated most of the CDM developments, these models are general diagnostic tools that can be applied outside educational contexts.

Recent works in psychological measurement have focused on providing detailed diagnostic information to patients. Results of previous studies using real data sets in psychological measurement have shown how CDMs can be used to diagnose and study the psychological disorders. For example, Templin and Henson (2006) used the DINO model to evaluate and diagnose pathological gamblers using a set of dichotomous Statistical Manual of Mental Disorders (DSM-IV-TR, American Psychiatric Association, 2000) criteria. de la Torre (2011) also reported an example of the G-DINA model using the Millon Clinical Multiaxial Inventory-III (MCMI-III) to diagnose personality and other clinical disorders. However, examination of the literature on psychological assessment reveals a dearth of applications under CDM framework thus far.

Chapter 3

Choosing between general and specific CDMs

This chapter begins with a study about the characteristics that can affect the attribute classification (i.e., person parameters estimates) using generated data. It was studied in detail five conditions that can be framed in CDMs: item discrimination, sample size, number of test items, true model, and number of attributes. By taking as a reference the results obtained in the simulation study, the second study is intended to give researchers a description of the methodology conducted in real data with CDMs.

The particular data used in the empirical study has been previously analyzed with CTT, and evidences in support which model underlie the data have not been explored before implementing CDMs. Because in the simulation study, it was demonstrated that the G-DINA model provided ACA as good as the specific model when the true model is not known, the real data application put particular attention to the attribute classification, but other aspects such as model fit to observed data and item level information is interpreted.

3.1 Study I: Examining attribute classification accuracy

In the absence of an a priori reason to believe that a cognitive diagnosis model assumes a particular form, cognitive diagnosis models (CDMs) with general formulations are preferable over specific CDMs in that former subsume the latter, and thus, will provide a better fit to the data. However, its is also true the general CDMs are more complex (i.e., they have more parameters), and require a larger sample size to be estimated reliably. As such, it is not clear whether general CDMs are to be preferred over specific CDMs when the sample size is small. In particular, it is not clear to what extent instability in the item parameter estimates will affect the attribute classification accuracy (ACA) of the general models. In this study, we systematically compare the impact of small sample size on the ACAs of general and specific CDMs, with the goal of providing practical guidance to researchers and practitioners in selecting the appropriate CDMs when the sample size is relatively small.

The primary objective of this study is to use a simulation study to examine how the ACA of a general CDM (i.e., G-DINA model) at the attribute and vector levels compares with those of specific CDMs, specifically, the DINA model, DINO model and A-CDM when the sample size is small. The impact of other factors such as test length, number of attributes, and assumption about the underlying process are also considered.

3.1.1 Method

Design

In the simulation study, we consider five factors: item quality (high or low), sample size (N=100, 200, 400, 800 or 1600), test length (J=15, 30 or 60), true model (DINA, DINO, A-CDM), and number of attributes ($K = 5$ or 10). All attribute patterns were generated with equal probability. The high item quality refers to items with lowest and highest probabilities of success of .10 and .90, respectively; the low item quality refers to items with lowest and highest probabilities of .20 and .80, respectively. One of the Q-matrix ($K = 5$) used in this simulation study, which represents a subset of the 32 possible attribute patterns, can be found in Table 4.1. This Q-matrix was constructed such that each attribute appears alone, in a pair, or in a triple the same number of times as other attributes. For $J = 15$, we used 1 to 5, 11, 14, 15, 18, 20, 21, 23, 26, 27, 30; for $J = 30$, all the items were used; and for $J = 60$, each item was used twice.

Table 3.1. *Q-matrix for the simulated data*

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	11	1	1	0	0	0	21	1	1	1	0	0
2	0	1	0	0	0	12	1	0	1	0	0	22	1	1	0	1	0
3	0	0	1	0	0	13	1	0	0	1	0	23	1	1	0	0	1
4	0	0	0	1	0	14	1	0	0	0	1	24	1	0	1	1	0
5	0	0	0	0	1	15	0	1	1	0	0	25	1	0	1	0	1
6	1	0	0	0	0	16	0	1	0	1	0	26	1	0	0	1	1
7	0	1	0	0	0	17	0	1	0	0	1	27	0	1	1	1	0
8	0	0	1	0	0	18	0	0	1	1	0	28	0	1	1	0	1
9	0	0	0	1	0	19	0	0	1	0	1	29	0	1	0	1	1
10	0	0	0	0	1	20	0	0	0	1	1	30	0	0	1	1	1

A computer program was implemented in Ox (Doornik, 2008) for data generation and performed 100 replications under each condition. All the data sets were analyzed using the DINA, DINO, and the G-DINA models. When the underlying process corresponds to the fitted model, we assumed that the model is known; otherwise the model is considered unknown. The item parameters were estimated via marginal maximum likelihood estimation (MMLE), and the vectors of attribute classification were obtained based on expected a posteriori estimation. These procedures were implemented by de la Torre (2009) using the computer program Ox (Doornik, 2003).

Despite the differences in model formulations, it is possible to compare the item parameter estimates with the parameters of the true model (i.e, known model). Item parameter comparison can be done by taking into account the probability of a correct response for each of the latent group implied by the required attributes for an item. For example, let the number of required attributes be equal to $K_j = 2$, the probability of success for each of the latent group under the DINA model, A-CDM, and the DINO model are given in Table 3.2.

Table 3.2. *Example of comparison of specific CDMs*

	Probability of success			
Model	$P(\{00\})$	$P(\{10\})$	$P(\{01\})$	$P(\{11\})$
DINA	.10	.10	.10	.90
DINO	.10	.90	.90	.90
A-CDM	.10	.50	.50	.90

Thus, although the DINA and DINO models are typically specified with only two parameters, and the A-CDM with $K_j + 1$ parameters, they can be expanded to provide probabilities for all the 2^{K_j} latent groups of interest. In this way, they can be compared to each other and to G-DINA model which has 2^{K_j} parameters.

The ACA under each condition were determined. The ACA were computed at the individual attribute and attribute vector levels. In addition to the ACA, root mean square error (RMSE) of the item parameter estimates across the replications was computed.

Let α_{ik} be the true classification of attribute k for examinee i where $i = 1, \ldots, I$ and let be $\widehat{\alpha_{ik}}$ the estimated classification of attribute k. The proportion of correctly classified individual attribute (PCA) k and the proportion of correctly classified attribute vectors (PCV) are given by

$$PCA_k = \frac{1}{I} \sum_{i=1}^{I} \mathcal{I}(\alpha_{ik} = \widehat{\alpha_{ik}}) \tag{3.1}$$

$$PCV = \frac{1}{I} \sum_{i=1}^{I} \prod_{k=1}^{K} \mathcal{I}(\alpha_{ik} = \widehat{\alpha_{ik}}) \tag{3.2}$$

where $\mathcal{I}(\cdot)$ is the indicator.

In general, the accuracy of the item parameter estimates for item j with K_j attributes can be computed as:

$$RMSE(P_j(\alpha_l)) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \sum_{l=1}^{2^{K_j}} w_j(\alpha_l)(P_j(\alpha_l) - \widehat{P_{jr}(\alpha_l)})^2} \tag{3.3}$$

where r, $w_j(\alpha_l)$, $P_j(\alpha_l)$ y $\widehat{P_{jr}(\alpha_l)}$ are the number of replication, weight, true probability of success, and estimated probability of success, respectively, of the latent group with the attribute pattern α_l. The weight was computed as $w_j(\alpha_l) = (2^{K_j})^{-1}$ and R was equal to 100.

The results of the simulation study to evaluate the ACA are reported into two sections: accuracy of attributes classification and item parameters estimates. Each simulated data set is based on the DINA model, the DINO and A-CDM. Each model condition were fitted using the G-DINA model, the DINA model and the DINO model.

3.1.2 Results

Accuracy of attributes classification

The simulation study results to examine the ACA at the attribute and vector levels are presented in Figures 3.1 through 3.8. Results are presented for each condition of number of attributes. By using the equation (3.1), the Figures 3.1 through 3.6 contains the mean of proportion of correctly classified individual attribute as a function of sample size for the three test length and item quality, whereas the Figures 3.7 and 3.8 presents the estimated mean of proportion of correctly classified attribute vectors using the equation (3.1). Each x-axis corresponds to data generating model and fitted models are shown on the legends.

Figures 3.1 to 3.3 summarize the results for each fitted model measuring 5 attributes, whereas Figures 3.4 through 3.6 show the results for each fitted model when 10 attributes were assessed. As expected, in general for each model, the data generated using high item quality provided higher ACA for all attributes. With low level of item quality, ACA was low for all attributes.

Figure 3.1 contains the mean of proportion of correctly classified individual attribute estimated with G-DINA, DINO and DINA models for the data generated with the DINA model. It should be noted that under high item quality, at least sample size of 100 is required to have similar degree of proportion between DINA model and G-DINA model for each test length. When the item quality was low, using a test length of 60 items and sample size of 200, the degree of proportion between DINA model and G-DINA was similar. Further, if the test length was decreased to 30 items, then a sample size of 400 was required to obtained similar proportions. For the high item quality and sample size of 100, the mean of proportions was low as 0.908 for the G-DINA model, 0.919 for the DINA model and test length of 15 items; 0.971 for the G-DINA model, 0.975 for the DINA model and test length of 30 items; and 0.995 for both G-DINA and DINA models and test length of 60 items. For the low item quality, the mean of proportions was low as 0.817 for the G-DINA model, 0.829 for the DINA model with sample size of 800 and test length of 15 items; 0.901 for the G-DINA model, 0.911 for the DINA model with sample size of 400 and test length of 30 items; and 0.960 for the G-DINA model, and 0.967 for the DINA model with sample size of 200 and test length of 60 items.

Figure 3.2 contains the mean of proportion of correctly classified individual attribute estimated with G-DINA, DINO and DINA models for the data generated with the DINO model. When the data were simulated with DINO model, sample size of 200 provided similar degree of proportion between DINO model and G-DINA model under high item quality and considering a test length of 30 or 60 items. Also, if the test length was decreased to 15 items, then a sample size of 400 was required to obtained similar proportions. In addition, under the low item quality condition, the test length of 60 items and sample size of 200 determined similar proportions. For the high item quality and sample size of 200, the mean of proportions were low as 0.907 for the G-DINA model, 0.924 for the DINO model, and test length of 15 items; 0.974 for the G-DINA model, 0.980 for the DINO model, and test length of 30 items; and 0.995 for the G-DINA model, 1.00 for the DINO model, and test length of 60 items. For the low item quality, the mean of proportions were low as 0.814 for the

G-DINA model, 0.830 for the DINO model with sample size of 800 and test length of 15 items; 0.908 for the G-DINA model, 0.912 for the DINO model with sample size of 400 and test length of 30 items; and 0.970 for both G-DINA and DINO models with sample size of 200 and test length of 60 items.

Figure 3.3 contains the mean of proportion of correctly classified individual attribute estimated with G-DINA, DINO and DINA models for the data generated with the A-CDM. As expected, due to the A-CDM is less restricted than DINA and DINO models, the G-DINA model determined higher proportions of correctly classified individual attribute than DINA and DINO models. It should be noted that, as the Figure 3.3 presents, the proportions were higher from the sample size of $N = 100$ under both high and low item quality conditions.

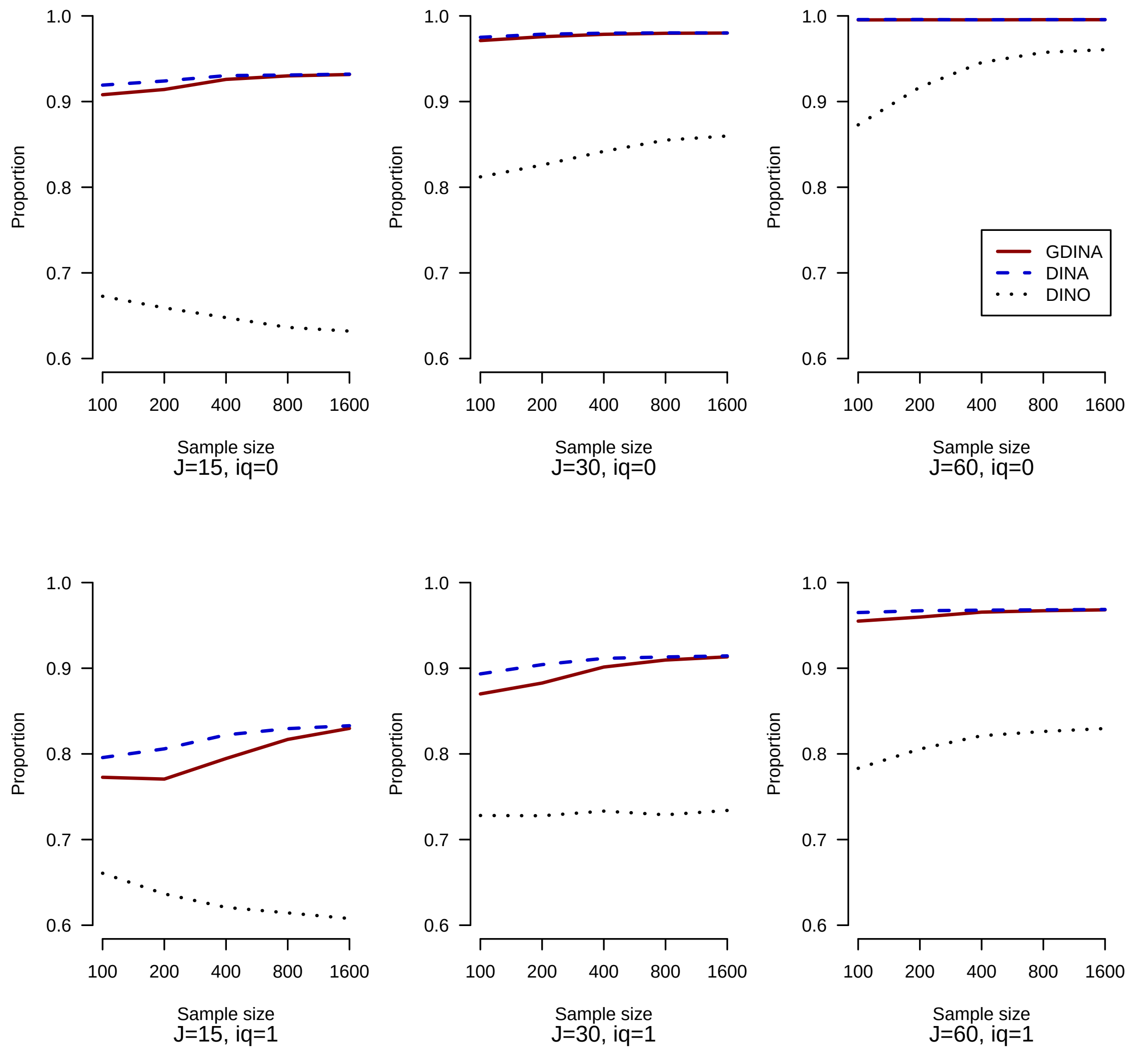

Figure 3.1. Proportion of correctly classified individual attribute for data generated with DINA model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

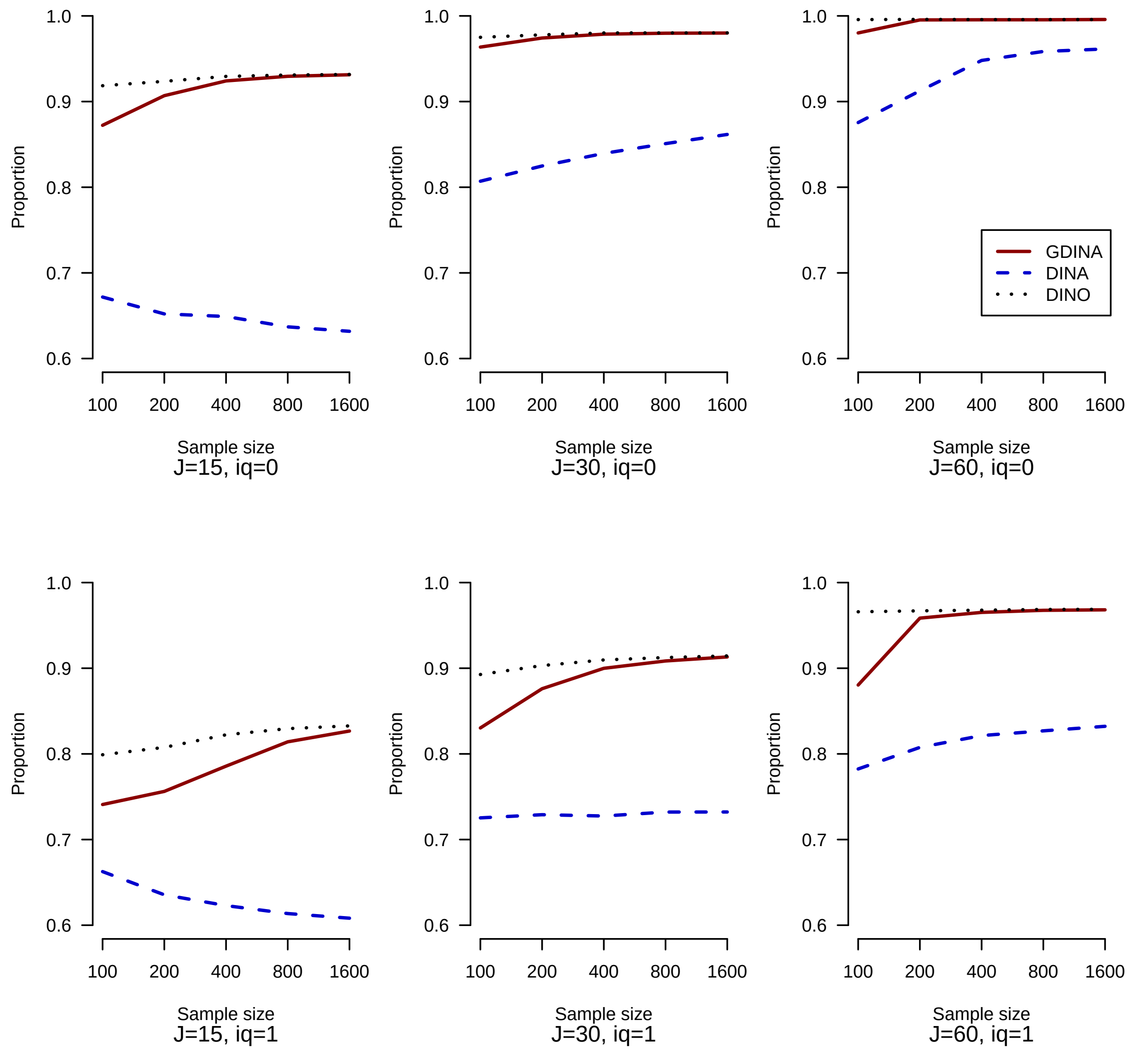

Figure 3.2. Proportion of correctly classified individual attribute for data generated with DINO model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

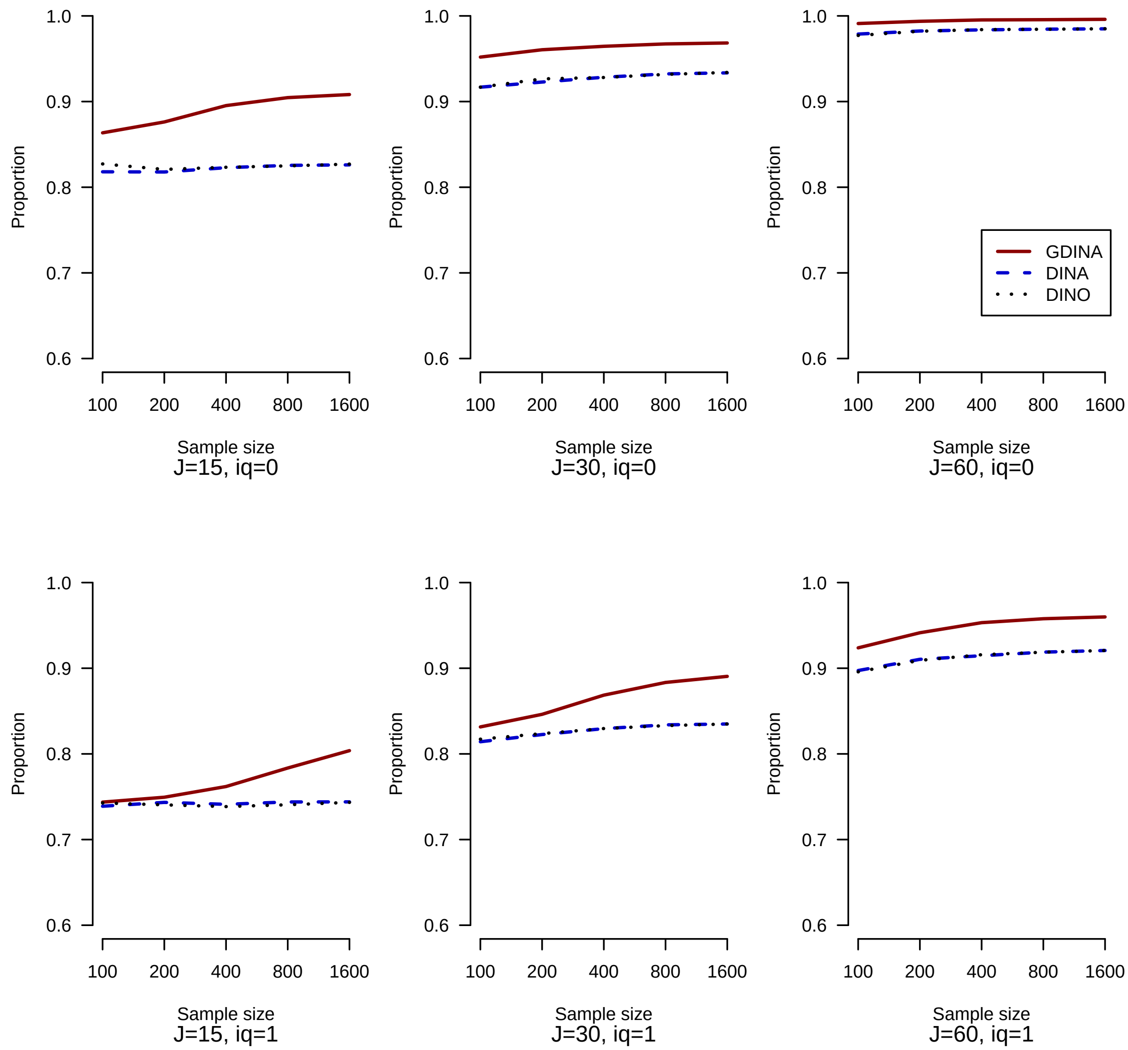

Figure 3.3. Proportion of correctly classified individual attribute for data generated with A-CDM model and $K = 5$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

Figures 3.4 through 3.6 summarize the results for each fitted model measuring

10 attributes. Figures 3.4, 3.5 and 3.6, contain the mean of proportion of correctly classified individual attribute estimated with G-DINA, DINO and DINA models for the data generated with the DINA model, DINO model and A-CDM, respectively. Looking at Figure 3.4, for each test length, item quality, and sample size, the G-DINA and DINA models determined similar degree of proportion. As a illustration, for the sample size of 400 and test length of 60 items, the mean of proportions was low as 0.979 for the G-DINA model, 0.980 for the DINA model when the high item quality were considered; whereas 0.901 for the G-DINA model, and 0.910 for the DINA model under the low item quality condition.

The Figure 3.5 in which the data were generated with DINO model, it should be noted that the minimum sample size of 400 is required to provide similar degree of proportion between DINO model and G-DINA model for each test length and item quality. For the sample size of 400 and test length of 60 items, the mean of proportions was low as 0.978 for the G-DINA model, and 0.980 for the DINO model when the high item quality were considered; whereas 0.894 for the G-DINA model, 0.910 for the DINO model under the low item quality condition.

Figure 3.6 illustrates proportions estimated from the data generated with A-CDM. Notice in this figure that G-DINA model provided higher proportions of attribute classification. In general, the proportion of correctly classified individual attribute increased as the test length and item quality increased.

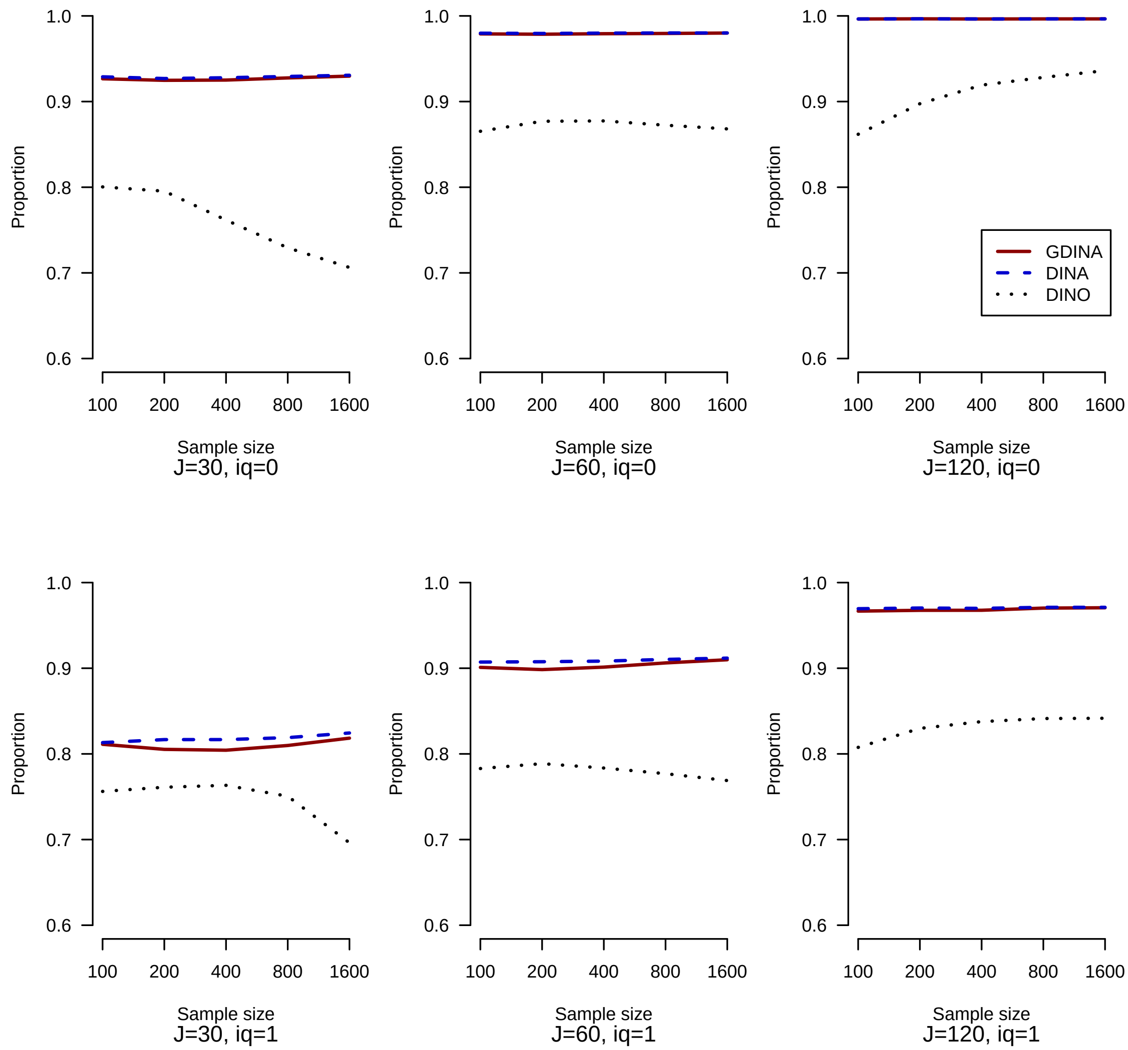

Figure 3.4. Proportion of correctly classified individual attribute for data generated with DINA model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

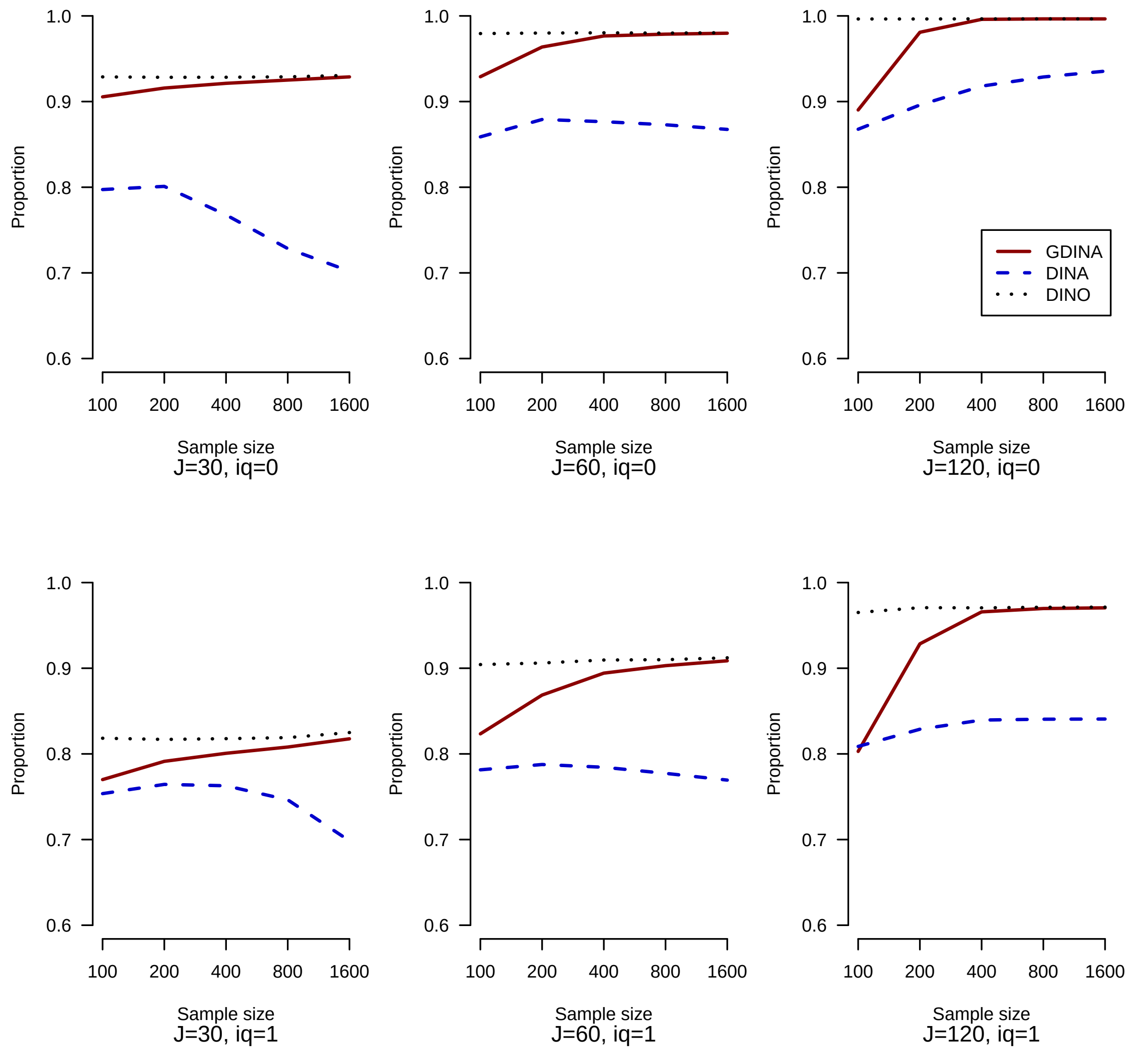

Figure 3.5. Proportion of correctly classified individual attribute for data generated with DINO model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

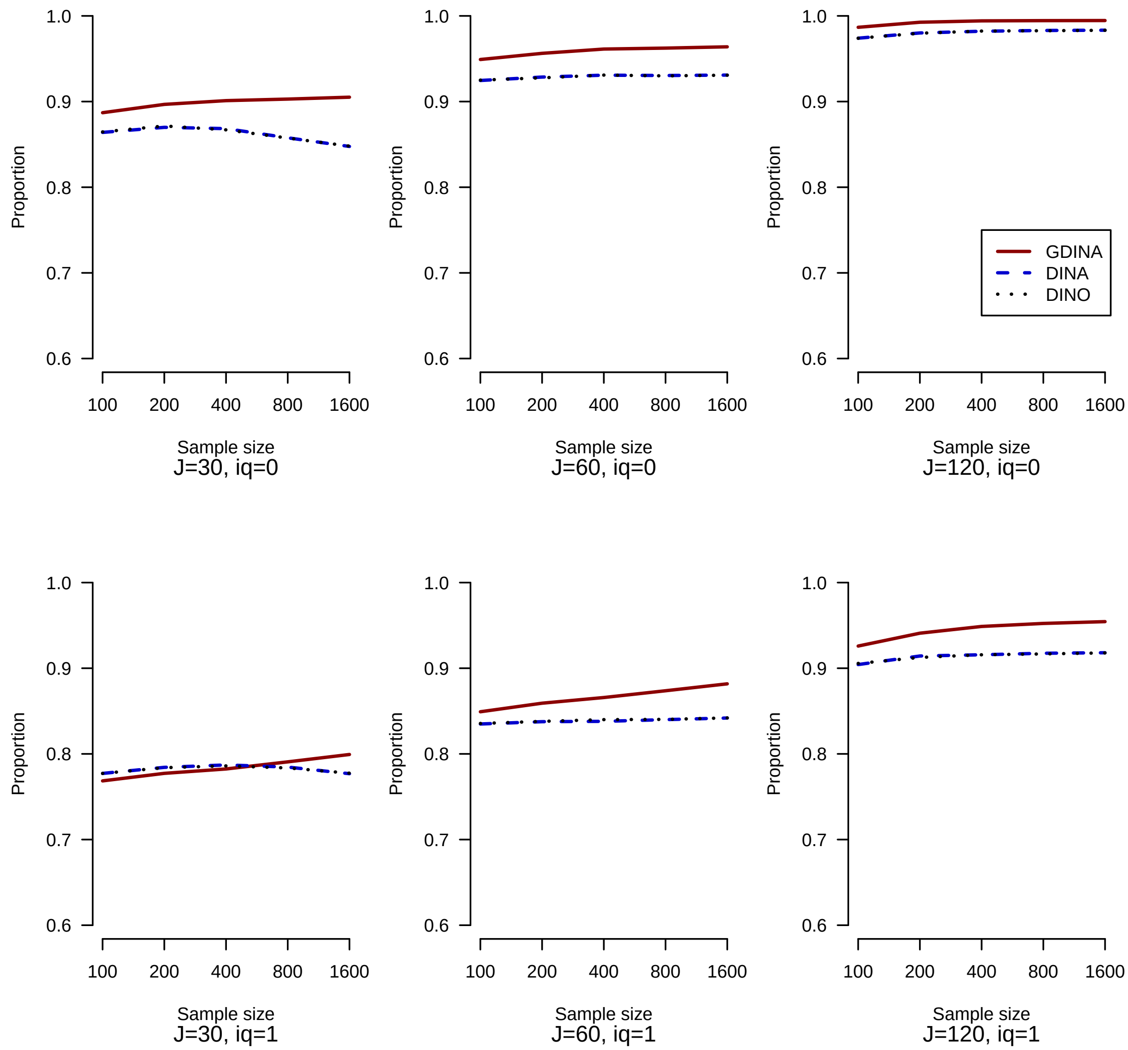

Figure 3.6. Proportion of correctly classified individual attribute for data generated with A-CDM model and $K = 10$. iq=0 high item quality, iq=1 low item quality. J represents number of item test. Legends correspond to fitted models.

Figures 3.7 and 3.8 display the mean of proportion of correctly classified attribute vectors for $N = 200$ and $N = 400$, respectively, as a function of test length with G-DINA, DINA and DINO models. Figures 3.7 and 3.8 summarize the results for each fitted model measuring 5 and 10 attributes, respectively. Each x-axis corresponds to data generating model.

The panels of both Figures 3.7 and 3.8 suggest that the proportion of correctly classified attribute vectors increased as the number of items increased. When the generating model was DINA, the ACA of G-DINA and DINA models show similar degrees of proportions at the vector level for each test length, especially when the test length is larger. Similar results were obtained when data were simulated with DINO and fitted with the G-DINA and DINO models. Particularly, when the number of attributes was $K = 5$, data were generated with the DINA model, item quality was high and number of items was 60, the mean of proportions was 0.979 for the G-DINA model and 0.980 for the DINA model. Also, similar results were obtained when $K = 10$ attributes were assessed, that is, the mean of proportions was 0.996 for both G-DINA and DINA model under the high item quality and test length of 120 items.

Finally, when the data were generated with the A-CDM and $K = 5$ attributes or $K = 10$ were used, the G-DINA model always had higher ACA. For example, the mean of proportions was 0.970 for the G-DINA model, 0.915 for the DINA model, and 0.913 for the DINO model under conditions of high item quality, $K = 5$ attributes and test length of 60 items.

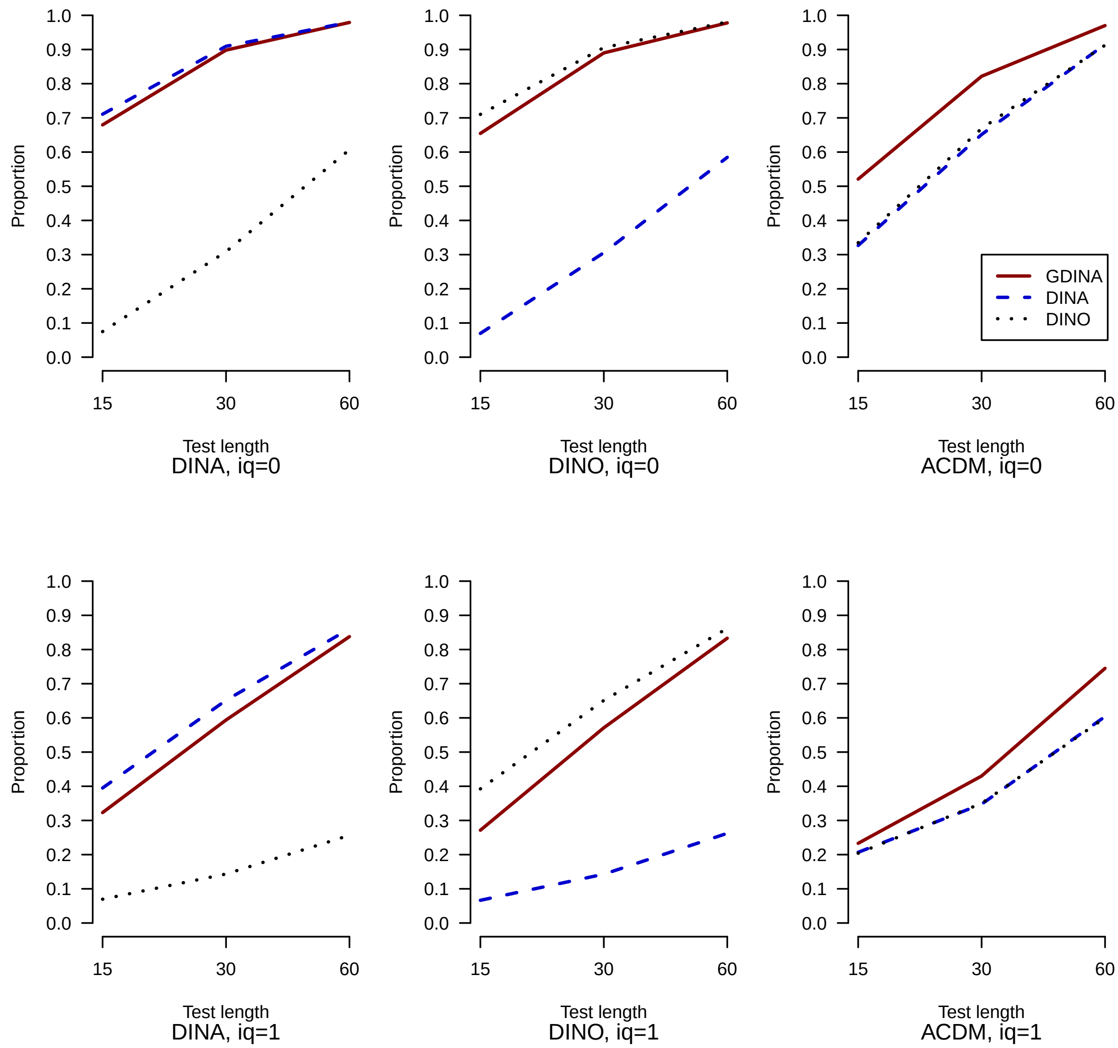

Figure 3.7. Proportion of correctly classified attribute vectors when $N = 200$ and $K = 5$. Legends correspond to fitted models and data generating models are shown on the x-axis. iq=0 high item quality, iq=1 low item quality.

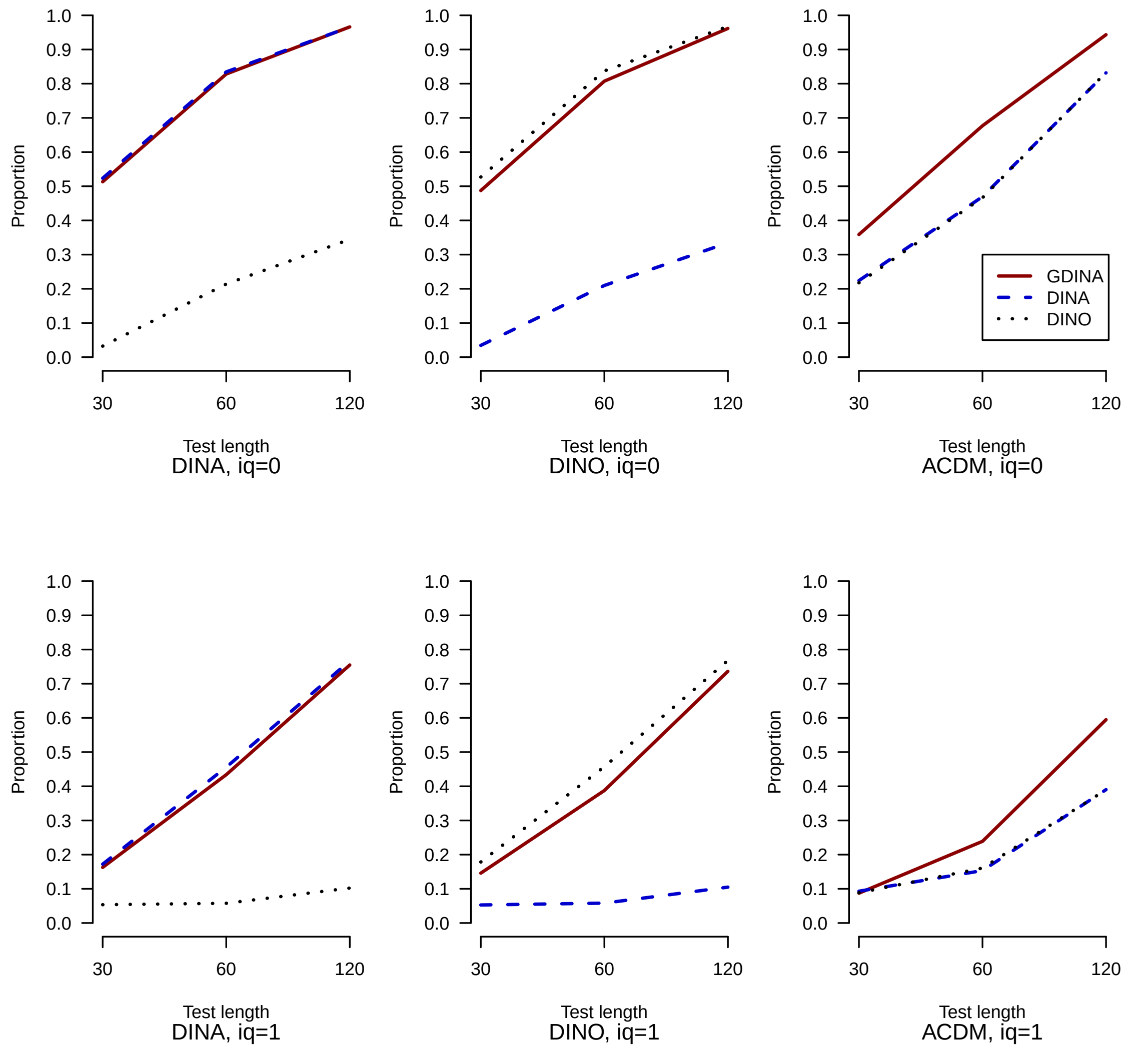

Figure 3.8. Proportion of correctly classified attribute vectors when $N = 400$ and $K = 10$. Labels correspond to fitted models and data generating models are shown on the x-axis. iq=0 high item quality, iq=1 low item quality.

Item parameter estimates

The item parameters recovery rates from the models were examined by the RMSE of the equation (3.3). For illustration purposes this section provides results pertaining to each generated model for condition of the number of required attributes. Tables 3.3 shows the mean of RMSE across replication of parameters recovery with G-DINA, DINA and DINO models as a function of sample size ($N = 200$), test length and item quality for the data generated with DINA model, A-CDM and DINO model. It should be realized to see Table 3.3 that G-DINA model is more general than DINA and DINO models.

For instance, when the data generating model was the DINA, and taking into account that the true model has better accuracy than G-DINA, it should be noted that under high item quality and at least 30 items, the means of RMSE of parameters recovery with DINA and G-DINA models were less than 0.05. In addition, as expected the DINO model have lower accuracy than DINA and G-DINA models.

Table 3.3. *Mean of RMSE of parameters recovery with GDINA, DINA and DINO models when $N = 200$ and $K = 5$*

			Fitted model		
Generating model	Test length	Item quality*	DINA	DINO	G-DINA
DINA	15	0	0.05	0.20	0.07
		1	0.09	0.18	0.15
	30	0	0.04	0.24	0.05
		1	0.06	0.19	0.09
	60	0	0.04	0.23	0.05
		1	0.05	0.19	0.07
DINO	15	0	0.20	0.05	0.01
		1	0.18	0.08	0.18
	30	0	0.24	0.04	0.00
		1	0.19	0.06	0.09
	60	0	0.24	0.04	0.00
		1	0.19	0.05	0.07
A-CDM	15	0	0.15	0.14	0.11
		1	0.14	0.14	0.19
	30	0	0.14	0.14	0.07
		1	0.13	0.13	0.12
	60	0	0.13	0.13	0.07
		1	0.12	0.12	0.09

* 0 indicates high item quality and 1 represents low item quality.

3.1.3 Conclusions

This study addressed a comparison involving four models which have been described within the background. The models covers different assumptions about how attributes combine or interact to produce an item response. The simulation study is intended to give a contribution in practical guidance to researchers and practitioners in selecting the appropriate CDMs when the sample size is relatively small.

Results from the simulation study indicated that if the sample size is small, the item parameters estimation with the G-DINA model is not as good as the true model, but it is also not the worst. It should be noted that item parameters estimated with G-DINA model is in the middle and close to the optimal estimate when sample size is large. The ACA using G-DINA model is not the best, when the true model is known (i.e., DINA, DINO, A-CDM), however ACA with G-DINA is the best when the true model is unknown and close to the optimal results. In addition, even if item parameter estimates are not stable, attribute classification is accurate especially when the test length is large.

In this study, each data set were generated under assumption that test items reflect a common underlying model. However, it can turn out that each item reflects a specific CDM. This issue can be part of future investigation to examine the extent to which it affects the ACA. It also turn out to note that a third of items were specified requiring 1 attribute. Due to this, some simulation study conditions provided high proportion of classified attribute when the true model is not known and the data were fitted with either DINA or DINO models.

3.2 Study II: An application of CDMs to Asperger Syndrome data

Typically, reported studies within measurement of psychological disorders have focused on reporting of a single summary score, X, or a single latent trait, θ, to decide whether or not a patient possesses the disorder of interest. In order to overcome this difficulty, the CDMs make it possible to evaluate the test by providing evidence for how well each item helps categorize individuals as well as giving a probability that each person has the skills profile on the skills measured by the test.

The present study aims to provide an alternative use of CDMs in psychological assessment. The study focuses on the analysis of the instrument Escala Autónoma (EA; Belinchon, Hernández, Martos, Sotillo, Márquez, and Olea, 2008) designed to assess behaviors and abilities indicative of Asperger Syndrome or High Functioning Autism among individuals above the age of 6. Another purpose of this study is to introduce psychologist and professionals to CDMs as a tool available for the analysis of tests.

3.2.1 Diagnosis of asperger syndrome

Asperger Disorder (AD) is currently categorized as a Pervasive Developmental Disorder in both DSM-IV-TR (American Psychiatric Association, 2000) and ICD-10 (World Health Organization, 2010) system. At date, there has been some controversy about whether AD is a distinct entity from High Functioning Autism (HFA, Howlin, 2003; Volkmar and Klin, 2005). AD is defined by three areas of impairment: in social development, communication, and imagination (Matson, 2008; Molloy and

Vasil, 2004). However last DSM-5 draft criteria are now proposing to include both groups in a single diagnostic category (Autism Spectrum Disorder, ASD). Current clinical definitions of AD/HFA include impairments in social and communication and restricted, repetitive and stereotyped patterns of behavior, interest, and activities with no history of cognitive or language delay. Recent works describe a prevalence of 1% of AD/HFA in children and young population (Gillberg, 2010), being higher the detection rate in men.

A variety of interview and structured scales have been designed in order to identify people with AD/HFA as early as possible, and to offer them and their families the supports they requires. Usually, the scales include items measuring clinical symptoms that must be responded by teachers, parents, and health practitioners. Matson and Boisjoli (2008) provide a diversity of specific measures of core symptoms, and detailed how they vary between cases, and what symptoms appear to be most salient for diagnosis. However, the most commonly method used is the application of questionnaires to obtain a single score and collect evidences of diagnostic validity. A problematic issue is to establish levels of sensitivity and specificity with different cut-points. CDMs represent an alternative approach by providing diagnostic information in the clinical assessment. Since diagnostic information of each individual is obtained, CDMs can improve the evaluation diagnostic by guiding practitioners.

Several instruments have been developed that can be used as part of the diagnostic process of AD/HFA (see Campbell, 2005; Howlin, 2000; Matson and Boisjoli, 2008 for critical reviews). In Spain, Belinchon et al. (2008) developed the EA Spanish version with the primary objective of use it as a screening tool of AD/HFA.

Measurement of AD/HFA using CDMs must yield data reliable and valid for its intended purpose. Reliability, as internal consistency assesses the consistency of results across items of the test. Diagnostic validity refers to the extent to which the assessment protocol correctly identifies or classifies clinical cases. Validity is also determined by evaluating the extent to which the test results predict actual performance.

The accuracy of a test depends on how well it separates the clinical populations being tested into those with and without disorder in question. Therefore we focused our analysis on the instrument EA, which was developed to identify people whose psychological functioning pattern fits with that described for people with AD/HFA. In the case of AD/HFA, the attributes are considered to represent EA criteria of these conditions, with mastery of an attribute referred to as having satisfied a criterion and non mastery of an attribute referred to as having not satisfied a criterion. Under CDMs framework, a person who meets most of measured attributes would be predicted to be diagnosed as having AD/HFA.

3.2.2 Method

Instrument and participants

To demonstrate the use of CDMs framework in psychological measurement, this study focuses on the EA version which contains 50 items measuring six dichotomous latent variables specified in the Table 3.4. The data for our analysis were from a data described and used by Belinchon et al. (2008). The data were collected from 177 individuals (i.e., 68% males and 32% females) in Spain from three clinical populations of children and adolescents diagnosed with AD/HFA (33%), ADHD, (14%) and a Non developmentally delayed group (NDD, 53%). These data were responses of parents and teachers to 50 items involving 6 attributes. The Q matrix is given in Table 3.4.

The EA includes items that assess latent variables of the psychological functioning of individuals, not as clinical symptoms strictly. A multidisciplinary team of practitioners (i.e., researchers in developmental disorders, psychometricians, and specialists in psychological assessment) developed an initial item pool that sampled 6 psychological dimensions where people with AD typically fail or differ from other groups: social skills (SS), fiction and imagination (FI), cognitive processes (CP), mentalizing (M), language and communication (LC), and executive functions (EF). After an empirical selection of items, an EA reduced version was applied in Spain to three clinical populations diagnosed with AD/HFA, Attention deficit hyperactivity disorder (ADHD) and Non delayed development (NDD). Belinchon et al. (2008) collected evidences of validation such as unidimensional internal structure and high correlation (i.e., Pearson correlation .89) with ASDI Scale (Gillberg, Gillberg, Rastam and Wentz, 2001). A cutoff of 36 (maximum score 72) provided high mastery of diagnostic classification (i.e., sensitivity of 100%, specificity of 97% in the NDD group, and specificity of 72% in ADHD).

Table 3.4. *Q-matrix for the observed data*

Item	SS	FI	CP	M	LC	EF	Item	SS	FI	CP	M	LC	EF
1	1	1	0	0	0	1	26	1	1	0	0	1	0
2	1	1	0	0	0	1	27	1	1	0	0	0	0
3	0	1	1	0	0	0	28	1	0	0	0	1	0
4	1	1	0	0	0	0	29	1	1	0	0	0	1
5	0	0	0	1	1	1	30	1	1	0	0	0	1
6	1	0	0	0	0	1	31	1	1	0	0	0	0
7	1	0	0	0	0	0	32	1	1	0	0	0	0
8	0	0	0	0	1	0	33	0	0	0	0	0	1
9	0	1	1	0	0	0	34	1	1	0	0	0	1
10	1	1	0	0	0	0	35	0	1	1	1	0	0
11	1	1	0	0	0	1	36	0	0	0	1	0	0
12	1	1	0	0	0	0	37	0	0	0	1	0	0
13	1	1	0	0	0	1	38	1	0	0	0	0	1
14	1	1	0	0	0	0	39	1	1	0	0	1	1
15	0	0	0	0	0	1	40	0	0	0	0	0	1
16	0	0	0	1	1	0	41	1	1	0	0	0	0
17	1	1	0	0	0	1	42	0	0	0	1	1	0
18	1	0	0	1	1	0	43	1	1	0	0	1	0
19	1	1	0	0	0	0	44	0	0	0	1	1	0
20	1	1	0	0	0	1	45	1	1	0	0	0	1
21	1	1	0	0	0	0	46	0	0	0	1	1	0
22	0	0	0	0	0	1	47	1	0	0	1	1	0
23	1	1	0	0	0	1	48	1	0	0	0	0	1
24	1	1	0	1	1	1	49	0	0	0	1	1	0
25	1	1	0	0	0	0	50	0	0	0	1	1	0

Data analyses

The initial items used in the EA had 26 raters (researchers in developmental disorders, psychometricians, and clinical researchers) assessing each item on each of 6 attributes. A selected panel of experts evaluated each item by giving the item a rating 1 (for clearly measuring), -1 (clearly not measuring), or 0 (degree to which it measures the content area is unclear) for each attribute. After experts completed the evaluation, if at least half of the raters judged that item measures the attribute then the entry of the Q matrix was one.

Based on the Q matrix (see Table 3.4) there were 355 item parameters to be estimated when these data were analyzed using the G-DINA model. The model parameters were estimated using the MMLE algorithm, which was written in the computer program Ox (Doornik, 2009) by de la Torre (2009). Given the convergence of the algorithm, parameter estimates were interpreted. From the model we estimated posterior probabilities of satisfying each attribute for each individual. Probabilities that are close to either zero or one reflect strong evidence in support the absence or presence of a criterion, respectively. If an individual had a posterior probability of meeting an attribute was greater than .50, then the individual was classified as having that attribute.

3.2.3 Results

Results are presented into three parts: model fit comparison, individual diagnoses, and item level information results.

Model fit comparison

Table 3.5 presents the model fit indices of the G-DINA, DINA and DINO models. These indices belong to the Q matrix with lowest values on the information criteria. Based on the deviance estimates, results show that G-DINA model fits better than both DINA and DINO. A likelihood-ratio test can be implemented to test whether either DINA or DINO should be rejected in favor of G-DINA model. According to the likelihood-ratio statistic between G-DINA and DINA, which is the difference in deviances with degrees of freedom equal to the difference in number of parameters, the value was 840.6 (df = 192, $p < .001$). Moreover, the likelihood-ratio test statistic was 318.62 (df = 192, $p < .001$) between DINO and G-DINA models. Thus, comparison suggested that G-DINA was preferred model against both nested models. It also was analized the GDINA model absolute fit to the data using the z-score based on the log-odds ratio index ($z(l)$; Chen, de la Torre, & Zhang, 2012). The maximum $z(l)$ was 2.56, which was smaller than the critical value 4.46 at the nominal α level of 0.01 after the Bonferroni correction $\alpha^* = \alpha/J/(J-1)$, where J is equal to 50.

Table 3.5. *CDMs fit indices*

Model	Deviance	Number of parameters
G-DINA	5028.20	355
DINO	5346.82	163
DINA	5868.80	163

Individual diagnoses

The G-DINA model can also provide an estimate of the attribute prevalence in the population, which is the percentage of the sample that has shown attribute mastery. Among the attributes, the CP had the highest value (i.e., 57%). The remaining attribute prevalences for all the attributes were less than 45%.

Based on attribute mastery, we calculated the latent classes the examinees belong to. Since there were six attributes used for this study, there were 64 possible combinations for the mastery of each attribute. The results show that combinations 000000, 001000 and 111111 had highest posterior probability of 0.2. The remaining attribute patterns showed posterior probabilities less than or equal to .02.

Table 3.6 presents the percentage of classification of individuals by clinical population as a function of the posterior probabilities of meeting an attribute. Because of panel of experts have previously classified each individual into each clinical population, it was expected that a larger percentage of subjects with AD/HFA had posterior probabilities greater than or equal to .60. In contrast, a larger of individuals belonging to NDD had posterior probabilities less than or equal to .40. However, the 95% of individuals with NDD had posterior probabilities between .40 and .60 in the attribute CP.

Table 3.6. *Percentage of classification of individuals by group and posterior probabilities*

	Attribute																	
	SS			FI			CP			M			LC			EF		
Group	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
AD/HFA	5	0	95	10	0	90	10	10	80	0	0	100	3	0	97	22	0	78
ADHD	52	0	48	28	4	68	60	28	12	24	0	76	48	0	52	84	4	12
NDD	95	0	5	96	0	4	5	95	0	98	0	2	100	0	0	99	0	1

Note. p_1 indicates posterior probabilities less than .4; p_2 indicates values between .4 and .6; and p_3 represents probabilities greater than or equal to .6.

Because of five of the six attributes (i.e., SS, FI, M, LC and EF) have posterior probabilities less than .4 or greater than .60 in both AD/HFA and NDD the results presented in Table 3.6 give evidence in support the cutoff of .50 in the posterior probability of meeting an attribute. As alluded above, if an individual had a posterior probability of meeting an attribute was greater than .50, then the individual was classified as having that attribute, thus 1 referred to as having satisfied an attribute and 0 otherwise.

Table 3.7 shows the classification of individuals by clinical population as a function of the number of attributes. As expected, a larger of number of subjects belonging to NDD group has not possessed attributes. Six and three individuals satisfied one and two attributes, respectively. The common attributes among NDD satisfying one or two attributes were SS, and FI. It also turn out that most of individuals with AD/HFA meet three or more attributes and one subject satisfied two attributes (i.e., SS and IA). The common attributes among subjects with AD/HFA satisfying two or three attributes were M, SS, and LC. Looking at ADHD group, individuals were distributed from one to six attributes. In addition, a larger of percentage of individuals with ADHD had two attributes.

Table 3.7. *Classification of individuals by group and number of attributes*

Number of attributes	Number of individuals		
	AD/HFA	ADHD	NDD
0	0 (0)	0 (0)	84 (90.3) *
1	0 (0)	1 (4.0)	6 (6.5)
2	1 (1.7)	15 (60.0)	3 (3.2)
3	5 (8.5)	2 (8.0)	0 (0)
4	5 (8.5)	5 (20.0)	0 (0)
5	7 (11.9)	1 (4.0)	0 (0)
6	41(69.5)	1 (4.0)	0 (0)
Total	59 (100)	25 (100)	93 (100)

Note. *Percentage in parentheses.

Table 3.8 gives clinical attribute profiles for 4 individuals. In the illustration, individual A is most likely with AD/HFA than individual D. Individual A satisfied all attributes and the attribute profile indicates that individual A should be classified with AD/HFA.

Table 3.8. *Example of attribute pattern of individuals*

Individual	Attribute					
	SS	FI	CP	M	LC	EF
A	1.00	1.00	1.00	1.00	1.00	.80
B	1.00	.00	.00	1.00	1.00	1.00
C	1.00	.00	.50	1.00	.02	.00
D	.00	1.00	.10	.00	.00	1.00

Item level information

Table 3.9 presents three illustrative items in which helpful diagnostic information is obtained from the interpretation of the estimated item parameters. Item 22, He/She has peculiar voice, which measures EF, is discussed. Results show that individuals without EF have a 9% chance of satisfying the item; individual with EF have an 80% chance of satisfying the item.

Table 3.9. *Example of estimated item parameters*

	Attribute							
								SS
					SS	SS	FI	FI
Item	-	SS	FI	EF	FI	EF	EF	EF
20	.01	.07	.00	.00	.13	.50	.00	.75
22	.09	-	-	.80	-	-	-	-
38	.00	.11	-	.36	-	.86	-	-

Examine the results for item 38, *He/She talks in an overly formal, pedantic or intellectualized way*, measuring SS and EF, estimated item parameters indicate that individuals who have neither SS or EF have a 0% chance of endorsing the item; individuals with SS and EF have 11% and 36% chance of endorsing the statement, respectively; finally individuals who satisfy both SS and EF have 86% chance of endorsing the item.

Parameter estimates results of item 20, *He/She expresses stereotyped or peculiar social formulas in making conversation*, which measures attributes SS, FI and EF show that individuals with neither attribute have 1% chance of endorsing the item. In contrast, individuals who satisfy all the three attributes have 75% chance of endorsing the item.

3.2.4 Conclusions

This study is intented to provide a contribution in practical guidance to researchers and practitioners in applying a general CDM in psychological assessment, particularly when the AD/HFA is measured. The study also demonstrated that CDMs can be used for both diagnostic classification and scale development in psychological assessment. The methodology includes the EA (Belinchon et al., 2008) to assess behaviors and abilities indicative of AD/HFA in individuals above six years of age and the G-DINA (de la Torre, 2011) model to estimate the attribute patterns and the item parameters.

Criteria were assigned for each item of the EA based on latent variables. The Q matrix was constructed and validated according to the panel of expert evaluation. Based on the Q matrix attribute profile for each subject and item parameters were estimated. Thus, the diagnostic results suggest properly classification of each individual in each clinical population. That is to say, we obtained empirical evidences in support diagnostic validity. In addition, item parameter interpretation provided rich diagnostic information that can be used to make inference about its discriminate among individual and thus develop new items.

Potential limitations in the use of CDMs as a tool for measuring AD/HFA are in need of discussion. First, for example, construction of the Q matrix was specified based on a scale which was originally created under classical test theory. This issue can lead poor model fit. However, according to de la Torre and Chiu (2010) and Templin and Henson (2006) specification of the Q matrix is often unknown and they recommend that entries of the Q matrix can be empirically validated, and much work to be done in this area. Second, a concern considered by this study is regarding the classification of individual with NDD. It is expected that subjects with NDD should not possess any attribute; however results show nine individuals having one or two attributes. Thus, a verification of attributes yields SS as common attribute among individuals with NDD. Third, although interpretation of the item parameters should be made with caution because sample size (i.e., N=177) is relatively small when the G-DINA model is used as part of diagnostic classification, the estimated item parameters indicated that reduced models (i.e., DINA and DINO models) are not appropriated for items of EA scale (e.g., see Table 3.9).

Assessment of AD/HFA cannot be based on results from one scale only, but the use of the G-DINA model applied to EA improves detection of behaviors indicative of this condition in children and adolescents. It is our hope that the proposed methodology can improve the efficiency of the diagnostic evaluation by guiding clinicians toward criteria that require more versus less assessment.

Chapter 4

Differential Item Functioning

Once ACA has been exhaustively analyzed in both real and simulation contexts, the investigation now focus on DIF. As mentioned in the introduction, the ACA can be influenced by the invariance of item parameters due to the person parameters are estimated under the assumption that item parameter are known. Before proceeding to analyze ACA with data reflecting DIF in a particular model, it is argued in this project that a well defined DIF procedure is required. For this reason and taking into account that a few DIF methods have been established in the CDMs framework, this chapter focuses on the development of DIF procedure in the specific DINA model, which is one of the most tractable and interpretable CDMs. The information of this chapter is described as follows: by using generated data the DIF is defined and proposed in one of the specific models; a comparison between the proposed DIF procedures with other method is conducted; and an application of the proposed method to real data is described.

4.1 Study III: Detecting DIF in the DINA model

This study seeks propose two indices of DIF detection with its corresponding significance test in the context of the DINA model, one of the most tractable and interpretable CDMs (de la Torre, 2009). The methods are based on the exact area measures of Raju (1988, 1990) between two item response functions, in which the item parameters should be calibrated separately for the comparison groups, then groups differences in parameters are examined for each item to determine whether DIF exists. The study also aims to examine the effectiveness of the proposed methods in detecting both uniform and nonuniform DIF in the DINA model. The viability of the proposed methods will be explored through a simulation study, by documenting their empirical Type I error and power. The impact of factors such as sample size, item quality, DIF size, DIF type, and number of attributes per item are also considered. The study will also compare the performance of the proposed statistics against that of the Mantel-Haenszel method with attribute profiles as matching criterion (MHP; Zhang, 2006). Finally, the paper illustrates the computation of the proposed methods by using real data from the 2007 TIMSS fourth grade mathematics assessment.

4.1.1 New DIF statistics for the DINA model

In this section it is introduced the indices for detecting DIF in the DINA model. The two DIF detection measures are presented based on the following definition of DIF. It is assumed the existence of two groups of examinees. Let $f_0(X_j \mid \eta_{jl}, h = 0)$ and $f_1(X_j \mid \eta_{jl}, h = 1)$ represent two IRFs for groups $h = 0$ and $h = 1$, respectively for item j. The DIF is present if the probability of a correct response differs, $f_0(X_j \mid \eta_{jl}, h = 0) \neq f_1(X_j \mid \eta_{jl}, h = 1)$, for those two groups of examinees of equal component η_{jl}, that is, when the IRFs in the group $h = 0$ and $h = 1$ are not equal.

In the DINA model, those two IRFs $f_0(X_j \mid \eta_{jl}, h = 0)$ and $f_1(X_j \mid \eta_{jl}, h = 1)$ can be expressed as

$$f_0(X_j = 1 \mid \eta_{jl}, h = 0) = g_{j0}^{(1-\eta_{jl})}(1 - s_{j0})^{\eta_{jl}} \tag{4.1}$$

$$f_1(X_j = 1 \mid \eta_{jl}, h = 1) = g_{j1}^{(1-\eta_{jl})}(1 - s_{j1})^{\eta_{jl}} \tag{4.2}$$

where s_{j0}, s_{j1}, g_{j0} and g_{j1} are item parameters for item j, which are defined as

$$s_{j0} = P(X_j = 0 \mid \eta_{jl} = 1, h = 0),$$

$$s_{j1} = P(X_j = 0 \mid \eta_{jl} = 1, h = 1),$$

$$g_{j0} = P(X_j = 1 \mid \eta_{jl} = 0, h = 0), \text{and}$$

$$g_{j1} = P(X_j = 1 \mid \eta_{jl} = 0, h = 1).$$

It is important to point out that $1 - s_{j0}$ and $1 - s_{j1}$ represent the probability of correctly answering the item j for groups 0 and 1, respectively.

Let $P_{j0} = (g_{j0}, s_{j0})$ and $P_{j1} = (g_{j1}, s_{j1})$ represent DINA model parameters for the same item j for groups $h = 0$ and $h = 1$, respectively. The difference ΔP_j between the probabilities of a correct response to the item j can be expressed as $\Delta P_j = (\delta_{j0}, \delta_{j1})$ where

$$\delta_{j0} = g_{j1} - g_{j0} \tag{4.3}$$

$$\delta_{j1} = (1 - s_{j1}) - (1 - s_{j0}) = s_{j0} - s_{j1} \tag{4.4}$$

Because of P_{0j} and P_{1j} are obtained from two independent sample, the standard error of δ_{j0} and δ_{j1} can be expressed as

$$SE(\delta_{j0}) = \sqrt{SE^2(g_{j0}) + SE^2(g_{j1})} \text{ and}$$

$$SE(\delta_{j1}) = \sqrt{SE^2(s_{j0}) + SE^2(s_{j1})} \text{, respectively.}$$

To implement the two DIF detection measures based on the difference $\Delta P_j = (\delta_{j0}, \delta_{j1})$, there are nine possible combinations of δ_{j0} and δ_{j1} indicating the presence of either uniform or nonuniform DIF, and no DIF presence: (C0) Both δ_{j0} and δ_{j1} are equal to zero; (C1) Both δ_{j0} and δ_{j1} are negative; (C2) Both δ_{j0} and δ_{j1} are positive; (C3) δ_{j0} is negative and δ_{j1} is equal to zero; (C4) δ_{j0} is positive and δ_{j1} is equal to zero; (C5) δ_{j0} is negative and δ_{j1} is positive; (C6) δ_{j0} is positive and δ_{j1} is negative; (C7) δ_{j0} is is equal to zero and δ_{j1} is positive; and (C8) δ_{j0} is is equal to zero and δ_{j1} is negative. Combination C0 indicates that the DIF is not present, C1 and C2 represent uniform DIF, and combinations C3-C8 indicate nonuniform DIF. These C0-C8 combinations are central to understanding and interpreting DIF studies, and different statistical detection procedures may be needed depending on whether a potential uniform or nonuniform DIF is present.

Figure 4.1 contains an example of two IRFs of two items exhibiting DIF when $g_{j0} = s_{j0} = 0.2$. For each figure it is assumed the responses of two different groups to the same item. A difference between the IRFs suggest that examinees from the two groups, with the same attribute pattern, do not have the same probability of success on the item. The left panel of Figure 4.1 shows an item exhibiting uniform DIF, and the right panel displays an item representing nonuniform DIF. In the item with uniform DIF is evident the both differences δ_{j0} and δ_{j1} are positive (i.e., combination C2); whereas in the item with nonuniform DIF the difference δ_{j0} is positive and δ_{j1} is negative (i.e., combination C6).

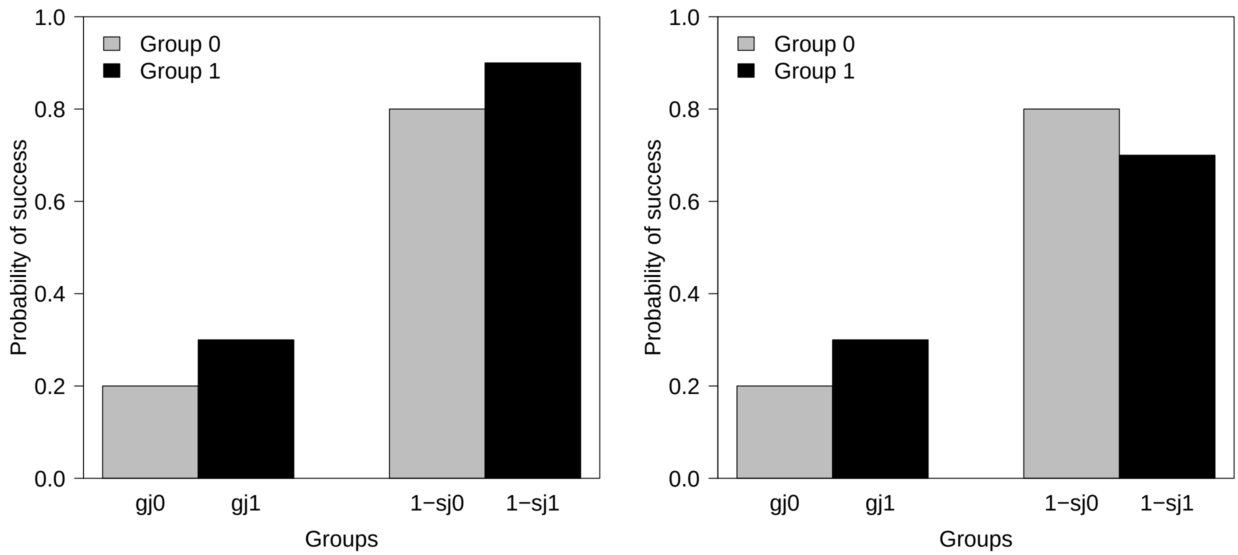

Figure 4.1. Example of uniform and nonuniform DIF for one item when $g_{j0} = s_{j0} = 0.2$.

The signed difference

Recall δ_{j0} and δ_{j1} the differences between estimated item parameters, and its corresponding standard errors $SE(\delta_{j0})$ and $SE(\delta_{j1})$. Let $Z_1 = \frac{\delta_{j0}}{SE(\delta_{j0})}$ and $Z_2 = \frac{\delta_{j1}}{SE(\delta_{j1})}$ be independent random variables. Based on the definition of the sum of Z_n independent random variables, the sequence $Y = \frac{1}{\sqrt{n}}(Z_1 + \cdots + Z_n)$ is asymptotically normally distributed. This means that $Z = \frac{1}{\sqrt{2}}(Z_1 + Z_2)$ follows a normal distribution, and $Z^2 = \frac{1}{2}(Z_1 + Z_2)^2$ is distributed according to the χ^2 distribution.

The signed difference (SDI_j) index for item j is defined as

$$SDI_j = \left(\frac{1}{\sqrt{2}} \left(\frac{\delta_{j0}}{SE(\delta_{j0})} + \frac{\delta_{j1}}{SE(\delta_{j1})} \right) \right)^2 = \left(\sum_{l=0}^{1} \frac{1}{\sqrt{2}} \frac{\delta_{jl}}{SE(\delta_{jl})} \right)^2, \qquad (4.5)$$

where SDI_j is square of standardized sum of differences. Under the null hypothesis of no DIF that

$$H_0 : \Delta P_j = 0 \qquad \text{or} \qquad \delta_{j0} = \delta_{j1} = 0,$$

SDI_j is asymptotically χ^2_ν distributed with $\nu = 1$ degree of freedom.

The unsigned difference

Examination of SDI_j statistic of the equation (4.5) shows that if $\frac{\delta_{j0}}{SE(\delta_{j0})}$ is equal to $-\frac{\delta_{j1}}{SE(\delta_{j1})}$, then a cancellation of DIF effect could exist. In addition to this, combinations (C5) and (C6) described above can produce undesirable results for DIF detection. A way to improve this issue can be provided by the following index. For the item j the unsigned difference (UDI_j) index between two IRFs is defined as

$$UDI_j = \left(\frac{\delta_{j0}}{SE(\delta_{j0})} \right)^2 + \left(\frac{\delta_{j1}}{SE(\delta_{j1})} \right)^2 = \sum_{l=0}^{1} \left(\frac{\delta_{jl}}{SE(\delta_{jl})} \right)^2, \qquad (4.6)$$

where UDI_j is sum of squared standardized differences. Because of $\frac{\delta_{j0}}{SE(\delta_{j0})}$ and $\frac{\delta_{j1}}{SE(\delta_{j1})}$ are independent random variables, the sum of their squares is approximately χ^2 distributed. Therefore, under the null hypothesis of no DIF, UDI_j is distributed according to χ^2_ν distribution with $\nu = 2$ degree of freedom.

4.1.2 Method

Factors manipulated

In the simulation study, data were generated using a fixed number of attributes ($K = 5$) and test length (J=30), with four factors manipulated: sample size, the item quality of the reference group parameter values, DIF size and DIF type. The reference group size, N_0 was fixed to 1000, while the focal group size, N_1, was either 500 or 1000. For the reference group, N_1, all the slip and guessing parameters are equal to 0.1, 0.2 or 0.3. When the guessing and slip parameters for the reference group (i.e., g_{j0} and s_{j0}) were equal to 0.2 and 0.3, two DIF sizes of 0.05 and 0.1 were assessed, defined as the differences of the guessing parameters or the slip parameters between the two groups. When the slip and guessing parameters for the reference group were equal to 0.1, only one DIF size of 0.05 were evaluated.

The joint distributions of attribute patterns are generated with equal probabilities from a multinomial distribution. The Q-matrix ($K = 5$) used in this simulation study, which represents a subset of the 32 possible attribute patterns, can be found in Table 4.1. This Q-matrix was constructed such that each attribute appears alone, in a pair, or in a triple the same number of times as other attributes.

Table 4.1. *Q-matrix for the simulated data*

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	11	1	1	0	0	0	21	1	1	1	0	0
2	0	1	0	0	0	12	1	0	1	0	0	22	1	1	0	1	0
3	0	0	1	0	0	13	1	0	0	1	0	23	1	1	0	0	1
4	0	0	0	1	0	14	1	0	0	0	1	24	1	0	1	1	0
5	0	0	0	0	1	15	0	1	1	0	0	25	1	0	1	0	1
6	1	0	0	0	0	16	0	1	0	1	0	26	1	0	0	1	1
7	0	1	0	0	0	17	0	1	0	0	1	27	0	1	1	1	0
8	0	0	1	0	0	18	0	0	1	1	0	28	0	1	1	0	1
9	0	0	0	1	0	19	0	0	1	0	1	29	0	1	0	1	1
10	0	0	0	0	1	20	0	0	0	1	1	30	0	0	1	1	1

The DIF was simulated from nine combinations of $g_{j1} - g_{j0} = d_{j0}$, and $s_{j1} - s_{j0} = d_{j1}$, where $d_{j0} = \{-0.05, 0, 0.05\}$ or $d_{j0} = \{-0.1, 0, 0.1\}$, and $d_{j1} = \{-0.05, 0, 0.05\}$ or $d_{j1} = \{-0.1, 0, 0.1\}$. Uniform DIF is produced when the slip parameter is increased and the guessing parameter is decreased, or when the slip parameter is decreased and the guessing parameter is increased; nonuniform DIF is produced when both the slip and guessing parameters are simultaneously increased or decreased, and when either slip or guessing parameter are increased or decreased. The DIF is not present when the differences were equal to zero, that is, when both guessing and slip were not increased or decresead. The details are presented in Table 4.2.

By taking into account that the guessing and slip parameters for the reference group were equal to 0.1, 0.2 or 0.3, the combinations produced nine DIF types, which can be interpreted as follows: (C0) both guessing and slip are equal to zero; (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only. Combination C0 indicates that the DIF is not present, C1 and C2 represent uniform DIF, and combinations C3-C8 indicate nonuniform DIF.

Table 4.2. *Summary of simulation conditions*

Factors in the Study	Details
Sample size	$N_0 = 1000,\ N_1 = 500$
	$N_0 = 1000,\ N_1 = 1000$
Reference group parameter values	$g_{j0} = 0.1, s_{j0} = 0.1$
	$g_{j0} = 0.2, s_{j0} = 0.2$
	$g_{j0} = 0.3, s_{j0} = 0.3$
DIF size	$\lvert g_{j1} - g_{j0}\rvert = \lvert s_{j1} - s_{j0}\rvert = .05$
	$\lvert g_{j1} - g_{j0}\rvert = \lvert s_{j1} - s_{j0}\rvert = .1$
DIF type (e.g., DIF size = 0.05)	C0: $g_{j1} - g_{j0} = 0,\quad s_{j1} - s_{j0} = 0$
	C1: $g_{j1} - g_{j0} = -0.05, s_{j1} - s_{j0} = 0.05$
	C2: $g_{j1} - g_{j0} = 0.05,\quad s_{j1} - s_{j0} = -0.05$
	C3: $g_{j1} - g_{j0} = -0.05, s_{j1} - s_{j0} = 0$
	C4: $g_{j1} - g_{j0} = 0.05,\quad s_{j1} - s_{j0} = 0$
	C5: $g_{j1} - g_{j0} = -0.05, s_{j1} - s_{j0} = -0.05$
	C6: $g_{j1} - g_{j0} = 0.05,\quad s_{j1} - s_{j0} = 0.05$
	C7: $g_{j1} - g_{j0} = 0,\quad s_{j1} - s_{j0} = -0.05$
	C8: $g_{j1} - g_{j0} = 0,\quad s_{j1} - s_{j0} = 0.05$

Data generation and analyses

For each condition, five hundred datasets have been simulated and analyzed. All of the test items were generated with DIF. The Type I error and power of both proposed indices using the significance level .05 was the focus of this study. A range from 0.025 to 0.75 of the liberal criterion (Bradley, 1978) was used to examine the Type I error rate for the nominal level of 0.05. All conditions where $g_{j1} - g_{j0} = 0$ and $s_{j1} - s_{j0} = 0$ result in the no DIF test condition were used to assess Type I error rates. For the rest of the conditions power was evaluated. The item parameters were estimated via MMLE. The data generation, parameter estimation, and the DIF computation were written in Ox (Doornik, 2003).

4.1.3 Results

The results of the simulation study are presented in two parts: the Type I error and power study. Each of those parts contains the performance of the SDI and UDI statistics as a function of the sample size, DIF size, DIF type, reference parameters values, and the number of required attribute to correctly answer an item.

Type I Error study

Type error I rate is defined as the percentage of DIF detection for the item out of the number of replications (i.e., 500) when the no DIF condition is generated. Table 4.3 illustrates the Type I error rate study results for SDI and UDI as a function of the reference item parameter values, sample size and the number of attributes required for correctly responding to the item at the nominal α level of 0.05. As shown in Table 4.3, because of the smaller sample size, the larger standard error of the item parameter values, the Type I error rates of SDI and UDI decreased as sample size increased.

According the criterion of Bradley (1978), it should be noted that the Type I error rates of the SDI statistic were not inflated and those rates were very close to the nominal rate of 0.05. The reference item parameter values had impact on the Type I error rates of the UDI statistic. Particularly, the larger values of reference parameters, the larger Type I error rates. Consistent with the low level of discrimination of the item with high guessing and slip parameters values, the Type I error rates of the UDI statistic were inflated when the item guessing and slip parameters were equal to 0.3.

The number of required attributes to correctly answering the item influenced the Type I error rates of the UDI index. The Type I error rates of UDI index decreased as number of attributes increased. The Type I error rate of UDI was inflated when only one attribute was assessed, the sample size was 1500, and guessing and slip parameters were equal to 0.2.

Table 4.3. *Summary of Type I error rates by indices ($\alpha = 0.05$)*

	Reference Parameter Values †	Sample Size							
		$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
		$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
SDI	0.1	0.054	0.047	0.060	0.054	0.053	0.047	0.049	0.050
	0.2	0.056	0.053	0.051	0.053	0.053	0.053	0.045	0.050
	0.3	0.073	0.066	0.060	0.066	0.054	0.049	0.047	0.050
UDI	0.1	0.058	0.049	0.061	0.056	0.059	0.051	0.053	0.054
	0.2	0.078	0.062	0.063	0.067	0.071	0.060	0.054	0.062
	0.3	0.138	0.119	0.100	0.119	0.122	0.091	0.083	0.098

Note. $^{\dagger}g_{j0} = s_{j0}$.

Power study

Power rate is defined as the proportion of correctly identified DIF items out of the total 500 replications by the methods. Power rates results are presented for both SDI and UDI. A cutoff of 0.90 was used to indicate excellent power rates, and moderate if the power rates were between 0.80 and 0.90.

SDI results

Tables 4.4 through 4.6 summarize the results of the power rates of SDI calculated using the χ^2_1 distribution. As expected, the power rates of SDI increased as sample size and DIF size increased. The value of the reference item parameters had effect on the power rates of the SDI statistic. Looking at reference group parameter values, the power rates decreased as the values of item parameters increased. In addition, for each DIF type (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; and (C6) larger guessing and slip, the power rates of SDI increased as the number of attributes required to correctly answering the item increased, but this relationship is reverse for the conditions (C7) smaller slip only and (C8) larger slip only, where the power rates decreased as the number of required attributes increased.

The DIF size had impact on the power rates for the SDI index, that is, the larger DIF size, the higher power rates across DIF type and sample sizes. For instance, when the DIF size was 0.05, the SDI statistic had power rates high as 0.70 for the reference item parameter values equal to 0.2, and 0.473 for the reference parameter values equal to 0.3; whereas the DIF size was 0.1, sample size of 1500 or 2000, and reference parameter values equal to 0.2, the DIF type C1 and C2 had power rates very close to 1; for the reference parameter values equal to 0.3, the C1 and C2 DIF types had power rates overall and power rates by number of attributes greater than 0.8.

Power rates of SDI statistic varied across the DIF types. According to Tables 4.4 through 4.6 the power rates of SDI were higher when the DIF type condition was smaller guessing but larger slip or larger guessing but smaller slip (i.e., C1 or C2), that is to say, when the uniform DIF was generated. For those C1 and C2 conditions the power rates were very similar. For instance, in Table 4.4, the power rates overall with highest values were C1 and C2. For the sample size of 1500, the power rates overall was 0.853 for the C1 DIF type, and 0.796 for the C2 condition. When the sample size was 2000, the power rates overall was 0.954 for C1, and 0.924 for C2.

The power rates of the SDI statistic varied as the involved number of attributes to correctly answering an item varied. For instance, in Table 4.4, for sample size of 2000, and number of attributes $K_j = 3$, the power rate of the DIF type smaller guessing only (i.e., combination C3) was 0.807. In Table 4.5, for the reference parameter values equal to 0.2, DIF size of 0.1, sample size of 2000, and number of attributes $K_j = 3$ or $K_j = 2$, the power rates of DIF type C3-C4 were ranged from 0.828 to 0.983. For the sample size of 1500, the power rates of C3 was 0.863 when $K_j = 2$ attributes were used, and 0.923 for $K_j = 3$. In Table 4.6, for sample size of 2000, and number of attributes $K_j = 3$, the power rates of DIF types C1, C2 and C3 were 0.979 and 0.952, 0.904, respectively.

In summary, when uniform DIF was generated, the SDI statistic yielded moderate to excellent power rates for the studied sample sizes and reference item parameter values. However, when the nonuniform DIF was introduced into the data, the SDI tended to have lower power rates, but it should be noted that among the nonuniform DIF conditions, the DIF types smaller guessing only or larger guessing only offered moderate to excellent power rates when the number of attributes were $K_j = 2$ or $K_j = 3$, items had moderate level of discrimination, and larger sample size and DIF size.

Table 4.4. *Summary of SDI power rates by DIF type ($\alpha = 0.05$) when $g_{j0} = s_{j0} = 0.1$*

DIF Size	DIF Type*	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
		$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.836	0.870	0.853	0.853	0.945	0.961	0.955	0.954
	C2	0.821	0.812	0.756	0.796	0.941	0.930	0.899	0.924
	C3	0.399	0.585	0.660	0.548	0.497	0.727	0.807	0.677
	C4	0.228	0.365	0.436	0.343	0.349	0.520	0.613	0.494
	C5	0.068	0.179	0.352	0.200	0.043	0.197	0.410	0.217
	C6	0.042	0.103	0.199	0.115	0.054	0.153	0.298	0.168
	C7	0.414	0.258	0.131	0.268	0.499	0.302	0.180	0.327
	C8	0.235	0.128	0.088	0.150	0.344	0.196	0.124	0.221

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

Table 4.5. *Summary of SDI power rates by DIF type ($\alpha = 0.05$) when $g_{j0} = s_{j0} = 0.2$*

DIF Size	DIF Type*	$N_0 = 1000, N_1 = 500$ $K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$N_0 = 1000, N_1 = 1000$ $K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.509	0.521	0.534	0.521	0.664	0.688	0.700	0.684
	C2	0.488	0.509	0.495	0.497	0.643	0.669	0.655	0.656
	C3	0.192	0.301	0.360	0.284	0.251	0.399	0.473	0.374
	C4	0.157	0.239	0.315	0.237	0.203	0.328	0.410	0.314
	C5	0.063	0.113	0.191	0.122	0.057	0.125	0.245	0.142
	C6	0.060	0.105	0.162	0.109	0.064	0.129	0.218	0.137
	C7	0.199	0.130	0.091	0.140	0.244	0.135	0.103	0.161
	C8	0.150	0.082	0.072	0.101	0.205	0.114	0.083	0.134
0.1	C1	0.981	0.991	0.993	0.988	0.998	0.999	1.000	0.999
	C2	0.975	0.979	0.971	0.975	0.997	0.997	0.997	0.997
	C3	0.641	0.863	0.923	0.809	0.757	0.952	0.983	0.897
	C4	0.399	0.648	0.771	0.606	0.549	0.828	0.907	0.761
	C5	0.072	0.331	0.615	0.339	0.055	0.413	0.773	0.413
	C6	0.107	0.287	0.491	0.295	0.081	0.345	0.638	0.355
	C7	0.639	0.385	0.235	0.420	0.762	0.482	0.292	0.512
	C8	0.382	0.203	0.131	0.238	0.559	0.298	0.171	0.342

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

Table 4.6. *Summary of SDI power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.3$

		Sample Size							
DIF	DIF	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
Size	Type*	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.305	0.326	0.352	0.328	0.398	0.443	0.473	0.438
	C2	0.309	0.337	0.353	0.333	0.374	0.414	0.442	0.410
	C3	0.130	0.206	0.249	0.195	0.153	0.257	0.351	0.254
	C4	0.169	0.226	0.283	0.226	0.126	0.225	0.297	0.216
	C5	0.065	0.105	0.167	0.112	0.061	0.117	0.217	0.132
	C6	0.146	0.202	0.235	0.194	0.073	0.142	0.232	0.149
	C7	0.137	0.094	0.070	0.100	0.136	0.088	0.068	0.097
	C8	0.131	0.094	0.076	0.100	0.113	0.076	0.064	0.085
0.1	C1	0.810	0.865	0.897	0.857	0.921	0.964	0.979	0.955
	C2	0.787	0.826	0.852	0.822	0.906	0.932	0.952	0.930
	C3	0.375	0.645	0.769	0.596	0.484	0.789	0.904	0.726
	C4	0.338	0.529	0.648	0.505	0.337	0.617	0.784	0.579
	C5	0.072	0.262	0.496	0.277	0.075	0.344	0.650	0.356
	C6	0.235	0.372	0.526	0.378	0.156	0.392	0.644	0.397
	C7	0.368	0.206	0.126	0.233	0.449	0.228	0.143	0.274
	C8	0.258	0.157	0.114	0.176	0.276	0.131	0.091	0.166

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

UDI results

Because the power rates can be increased by the inflated Type I error rates in the condition of higher values of the reference item parameter, in order to make meaningful power comparisons among simulation conditions, the power rates of UDI were calculated using the empirical distributions. The significant values of the empirical UDI statistic distribution by reference item parameter values and sample sizes at a level of 0.05 are presented in Table 4.7. The cutoff values decreased as both sample size and number of attributes increased.

Table 4.7. *Significant values of the empirical UDI distribution by reference item parameter value and sample size ($\alpha = 0.05$)*

Reference	Sample Size							
Parameter	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
Values †	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.1	6.33	6.37	6.15	6.31	6.09	6.17	6.12	6.12
0.2	7.11	6.55	6.68	6.78	6.89	6.78	6.44	6.66
0.3	9.89	8.39	8.01	8.80	8.26	7.76	7.59	7.94

Note. $^{\dagger}g_{j0} = s_{j0}$.

Tables 4.8 to 4.10 summarize the results of the power rates calculated using the empirical distributions of the UDI statistic. As expected, the power rates of UDI increased as sample size and DIF size increased. The reference item parameter values had impact on the average empirical power rates of the UDI statistic. For each

DIF type (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; and (C6) larger guessing and slip, the power rates decreased as the values of item parameters increased. The impact of the number of attributes required to correctly answering an item on the empirical power rates of the UDI index was reflected in each DIF type. For each C1 to C6 DIF type the power rates increased as the number of attributes required to correctly answering the item increased, but the power rates of (C7) smaller slip only and (C8) larger slip only conditions decreased as the number of required attributes increased.

The DIF size had impact on the power rates for the UDI index, that is, the larger DIF size, the higher power rates across DIF type and sample sizes. For instance, in Table 4.8, for the DIF size of 0.05, reference parameter values equal to 0.1, the conditions C7 and C8 presented the lowest power rates overall among the eight DIF types. When the sample size was 2000, the power rates overall with highest values were C1-C6, and ranged from 0.819 to 0.965. For the sample size of 1500, the power rate overall was 0.817 for the C1 DIF type, and 0.880 for the C5 condition, whereas the remaining DIF types obtained power rates overall lower than 0.8. Further, in Table 4.9 and 4.10, when the DIF size was 0.05, the UDI statistic had power rates high as 0.696 for the reference item parameter values equal to 0.2, and 0.418 for reference parameter values equal to 0.3.

The DIF type had an effect on the empirical power rates of the UDI statistic. According to Tables 4.8 through 4.10 the power rates overall of UDI were lowest when the DIF type condition were C7 and C8. For instance, in Table 4.9, six out of eight DIF types (i.e., C1-C6) provided power rates overall varying from 0.876 to

0.999 when the item parameter values were equal to 0.2, sample size of 2000 and DIF size was 0.1. By using a sample size of 1500, for the DIF size of 0.1, five out of eight DIF types (i.e., C1-C3, C5 and C6) obtained power rates overall ranged from 0.807 to 0.990 under condition of item parameter value equal to 0.2.

The number of attributes to correctly answering an item had impact on the empirical power rates of the UDI index. In Table 4.8, for the number of attributes $K_j = 3$, sample size of 2000, the power rates of DIF types C1-C6 were higher than 0.80, varying from 0.819 to 0.971. In Table 4.9, for the reference parameter values equal to 0.2, sample size of 2000, DIF size of 0.1, and number of attributes $K_j = 2$ or $K_j = 3$, the power rates of DIF types C1-C6 were close to 1. When the number of attributes was $K_j = 1$, the power rates of DIF types C1-C3 and C5-C7 ranged from 0.881 to 0.998. In addition, for sample size of 1500, number of required attributes $K_j = 3$, DIF size of 0.1, the DIF type C1-C6 had power rates greater than 0.9. For number of attributes $K_j = 2$, the C1-C6 DIF types kept the power rates above 0.8, varying from 0.828 to 0.994. For number of attributes $K_j = 1$, the C1, C2, and C5 DIF types obtained power rates of 0.936, and 0.928, and 0.984, respectively. In Table 4.10, for the sample size of 2000, DIF size of 0.1, and number of attributes $K_j = 3$, the power rates of DIF types C1-C6 were higher than 0.90, varying from 0.910 to 0.984. Further, for the reference parameter values equal to 0.3, sample size of 1500, DIF size of 0.1, and number of attributes $K_j = 3$, the power rates of DIF types C1-C3, and C5 were higher than 0.80, and ranged from 0.804 to 0.904. For number of attributes $K_j = 2$, only the C5 DIF type kept the power rates above 0.8.

In summary, when both uniform and nonuniform DIF was generated, the UDI statistic yielded excellent power rates for the studied sample sizes and reference item parameter values equal to 0.2; whereas the UDI had excellent power rates when $K_j = 3$ attributes were assessed, and moderate to excellent for $K_j = 2$ attributes for the reference item parameter values equal to 0.3. In general, the DIF types smaller slip only or larger slip only produced the lowest power rates overall, but it should be highlighted that a moderate power rate was presented when $K_j = 1$ attribute was measured, larger sample size, reference item parameter values were 0.2, and larger DIF size.

Table 4.8. *Summary of UDI power rates by DIF type ($\alpha = 0.05$) when $g_{j0} = s_{j0} = 0.1$*

		Sample Size							
DIF	DIF	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
Size	Type*	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.744	0.837	0.874	0.817	0.899	0.949	0.970	0.940
	C2	0.730	0.709	0.704	0.711	0.893	0.884	0.873	0.885
	C3	0.545	0.765	0.851	0.717	0.699	0.906	0.953	0.854
	C4	0.299	0.493	0.593	0.456	0.497	0.719	0.819	0.681
	C5	0.862	0.891	0.893	0.880	0.956	0.967	0.971	0.965
	C6	0.535	0.581	0.645	0.583	0.784	0.827	0.840	0.819
	C7	0.557	0.341	0.180	0.358	0.707	0.405	0.229	0.449
	C8	0.302	0.140	0.096	0.178	0.490	0.260	0.155	0.304

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

Table 4.9. *Summary of UDI power rates by DIF type ($\alpha = 0.05$) when $g_{j0} = s_{j0} = 0.2$*

		Sample Size							
DIF	DIF	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
Size	Type*	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.353	0.444	0.501	0.407	0.515	0.594	0.688	0.584
	C2	0.334	0.405	0.431	0.364	0.488	0.542	0.603	0.528
	C3	0.209	0.380	0.468	0.329	0.291	0.509	0.636	0.463
	C4	0.164	0.304	0.371	0.257	0.252	0.418	0.564	0.395
	C5	0.410	0.505	0.513	0.451	0.543	0.629	0.696	0.607
	C6	0.278	0.354	0.392	0.318	0.435	0.500	0.583	0.490
	C7	0.218	0.156	0.106	0.146	0.288	0.157	0.111	0.178
	C8	0.159	0.095	0.067	0.096	0.240	0.130	0.103	0.150
0.1	C1	0.936	0.983	0.995	0.969	0.993	0.999	1.000	0.997
	C2	0.928	0.954	0.963	0.942	0.987	0.993	0.997	0.992
	C3	0.766	0.968	0.991	0.905	0.900	0.996	0.999	0.965
	C4	0.492	0.828	0.919	0.732	0.702	0.948	0.987	0.876
	C5	0.984	0.994	0.996	0.990	0.998	1.000	0.999	0.999
	C6	0.701	0.845	0.914	0.807	0.881	0.961	0.987	0.941
	C7	0.771	0.520	0.296	0.506	0.898	0.616	0.376	0.615
	C8	0.482	0.260	0.145	0.280	0.711	0.382	0.231	0.431

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

Table 4.10. *Summary of UDI power rates by DIF type ($\alpha = 0.05$) when $g_{j0} = s_{j0} = 0.3$*

		Sample Size							
DIF	DIF	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
Size	Type*	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.130	0.199	0.271	0.150	0.234	0.318	0.401	0.293
	C2	0.138	0.210	0.258	0.154	0.220	0.285	0.356	0.261
	C3	0.085	0.179	0.272	0.131	0.149	0.278	0.391	0.248
	C4	0.135	0.224	0.281	0.166	0.148	0.258	0.357	0.232
	C5	0.153	0.238	0.278	0.172	0.264	0.333	0.418	0.314
	C6	0.242	0.306	0.320	0.235	0.249	0.327	0.397	0.301
	C7	0.091	0.073	0.062	0.059	0.135	0.084	0.062	0.085
	C8	0.113	0.092	0.075	0.073	0.133	0.088	0.071	0.088
0.1	C1	0.530	0.751	0.885	0.671	0.789	0.923	0.980	0.890
	C2	0.515	0.692	0.804	0.611	0.766	0.871	0.947	0.849
	C3	0.308	0.710	0.873	0.579	0.529	0.880	0.976	0.787
	C4	0.358	0.628	0.775	0.532	0.400	0.729	0.914	0.667
	C5	0.637	0.827	0.904	0.747	0.842	0.948	0.984	0.919
	C6	0.539	0.670	0.771	0.615	0.620	0.790	0.910	0.763
	C7	0.307	0.181	0.120	0.173	0.490	0.253	0.140	0.279
	C8	0.297	0.199	0.131	0.179	0.335	0.181	0.115	0.198

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

Comparison of UDI ***and*** SDI ***with MHP detection***

The Type I error rates of MHP are reported in Table 4.11. According to the three levels of item quality, the performance of the MHP method offered Type I errors rates very close to the studied nominal value of 0.05, when the level of discrimination of items were moderate or high, regardless the sample size of the reference and focal groups. Regarding to the low level of item discrimination the Type I error rates were controlled when items involved three attributes.

Table 4.11. *Summary of Type I error rates of* MHP *(*$\alpha = 0.05$*)*

	Reference	Sample Size							
	Parameter	$N_0 = 1000, N_1 = 500$				$N_0 = 1000, N_1 = 1000$			
	Values †	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
MHP	0.1	0.051	0.038	0.041	0.044	0.057	0.045	0.041	0.048
	0.2	0.083	0.058	0.045	0.062	0.075	0.059	0.046	0.060
	0.3	0.203	0.113	0.062	0.126	0.177	0.092	0.060	0.110

Note. $^{\dagger}g_{j0} = s_{j0}$.

Table 4.12 displays the power rates for MHP across the eight DIF types for the items with moderate level of discrimination. The power rates of the MHP procedure had a similar tendency as the SDI index. Powers rates produced by the MHP were moderate to excellent in the presence of uniform DIF, but power rates decreased in conditions where the nonuniform DIF was generated.

In terms of the number of attributes per item, as can be seen in Table 4.12, for the combinations C1 and C2 the power rates provided by MHP were moderate to excellent regardless the sample size. Power rates were ranged from 0.817 to 1, but these values were lower than the rates provided by SDI, in which the power rate were slightly smaller than one. Nevertheless, it was noted that under the nonuniform DIF presence the MHP required increasing the number of attribute and sample size to produce power rates as good as UDI or SDI for larger DIF. Particularly, the MHP required three attributes per item to produce moderate power rates for combinations C3, C4 and C5.

Table 4.12. *Summary of MHP power rates by DIF type* ($\alpha = 0.05$) *when* $g_{j0} = s_{j0} = 0.2$

DIF Size	DIF Type*	$N_0 = 1000, N_1 = 500$: $K_j = 1$	$K_j = 2$	$K_j = 3$	Overall	$N_0 = 1000, N_1 = 1000$: $K_j = 1$	$K_j = 2$	$K_j = 3$	Overall
0.05	C1	0.482	0.549	0.596	0.542	0.622	0.724	0.797	0.714
	C2	0.773	0.751	0.763	0.762	0.701	0.657	0.692	0.683
	C3	0.448	0.659	0.774	0.627	0.544	0.786	0.887	0.739
	C4	0.383	0.580	0.679	0.547	0.489	0.688	0.799	0.659
	C5	0.028	0.507	0.693	0.410	0.048	0.623	0.828	0.500
	C6	0.131	0.382	0.592	0.368	0.096	0.477	0.706	0.426
	C7	0.460	0.148	0.067	0.225	0.564	0.216	0.084	0.288
	C8	0.355	0.139	0.079	0.191	0.441	0.169	0.081	0.230
0.1	C1	0.970	0.989	0.998	0.986	0.998	1.000	1.000	0.999
	C2	0.826	0.807	0.818	0.817	0.998	0.999	0.998	0.999
	C3	0.466	0.726	0.828	0.673	0.582	0.837	0.915	0.778
	C4	0.421	0.634	0.739	0.598	0.534	0.748	0.845	0.709
	C5	0.029	0.552	0.762	0.447	0.048	0.684	0.870	0.534
	C6	0.175	0.418	0.638	0.410	0.131	0.522	0.761	0.471
	C7	0.486	0.149	0.067	0.234	0.610	0.211	0.082	0.301
	C8	0.383	0.149	0.083	0.206	0.479	0.185	0.087	0.250

Note. * (C1) smaller guessing but larger slip; (C2) larger guessing but smaller slip; (C3) smaller guessing only; (C4) larger guessing only; (C5) smaller guessing and slip; (C6) larger guessing and slip; (C7) smaller slip only; and (C8) larger slip only.

4.1.4 Conclusions

A new method for assessing DIF in the CDMs framework was introduced in this study. The proposed statistics, namely SDI and UDI are based on the item parameter estimates of the DINA model, and the signed and unsigned area formulas between two IRFs of Raju (1988, 1990). In practical implementation, the method assumes a validated Q-matrix, which specifies the attributes measured by each item. Then, for each group, the item parameters are estimated separately for the item j. Once the four item parameters have been estimated, it is computed the differences between the probabilities of correctly answering the item. These differences are divided by its corresponding standard error to obtain two standardized differences, which give the SDI and UDI statistics. Finally, the two SDI and UDI indices are tested using the χ^2 distribution with one and two degrees of freedom, respectively.

A simulation study has been carried out to evaluate its performance in detecting uniform and nonuniform DIF in terms of Type I error and power. The new statistics offers several theoretical advantages over the previous studies developed in the context of the DINA model. First, it is used separate item parameters calibration to avoid potential bias in estimating the item parameters and attribute patterns. Second, each index is proposed with its corresponding significance test that can be used for examining whether an observed difference is significantly different from zero. Third, under the conditions examined in the study, these new indices controlled the Type I error rates and power rates reasonably well. Fourth, in general, for larger DIF size, the SDI performed very well in detecting uniform DIF regardless of sample size, whereas the UDI was sensitive to detect both uniform and nonuniform DIF. It also was clear that the statistics UDI had higher power rates in detecting nonuniform DIF than the modified MHP. Fifth, the influence of the sample size in the power rates of both SDI and UDI were higher when the two comparison groups had equal sample size. Additionally, the power rates increased with increase the number of required attributes to correctly answering the item.

The different factors manipulated in the simulation study (i.e., sample size, reference item parameter values, DIF size, DIF type, and number of attributes per item) were used to assess the performance of both SDI and UDI statistics in detecting DIF. Each factor affected the Type I error rates and power rates of the statistics. For the SDI statistic the Type I error rates overall were controlled at the studied nominal level. In contrast, for the UDI statistic, when the item parameter were higher, Type I error rates overall were inflated, thus results reinforced the need to identify cutoff according to the empirical distribution of the UDI statistic, and the probability distribution would need to be adjusted or redefined.

Although the results are generally supportive of the two statistics, there are some potential areas to be investigated with regards to DIF analysis in the CDMs framework. Issues such as a generalization of the proposed method and how it can be applied to other CDMs would need to be investigated. Moreover, a study using a relatively small sample sample size may need to be systematically conducted.

4.2 An example of empirical data analysis using SDI and UDI statistics

This example serves as an illustration on DIF detection with the DINA model. The purpose of the illustration focused only in the computation of potential DIF, thus it is not concern the interpretation of the sources of the gender differences in mathematics.

Data and Analysis

The data for this illustration were taken from booklets 4 and 5 of TIMSS 2007 fourth grade mathematics assessment of the data originally described and used by Lee, Park and Tayland (2011). This study analyzed the students' responses from the United States and the two benchmark states to detect DIF in the DINA model using the 25 items involving 15 attributes or skills as prescribed by Lee, Park and Tayland (2011). The data from 823 students were used. This includes 389 male students and 434 female students. Female students served as the reference group, whereas male examinees served as focal group. Given the validated Q-matrix by Lee, Park and Tayland (2011), the item parameters were estimated separately via MMLE using the code written in Ox (Doornik, 2003) by de la Torre (2009). The SDI and UDI statistics were utilized to detect the existence of items showing uniform and nonuniform DIF.

Results

The item parameter estimates and its corresponding standard errors (SE) are presented in Table 4.13 for both groups. Also, given in this table are the SDI and

UDI statistics. Results of Table 4.13 show different conditions systematically studied in the simulation. For instance, the item 4 represents the larger guessing only (C4) DIF type, whereas the item 15 indicates the larger guessing but smaller slip (C2).

The item parameter estimates and their corresponding standard errors (SE) are presented in Table 4.13 for both groups. Also, given in this table are the SDI and UDI statistics. Results of Table 4.13 show different conditions systematically studied in the simulation. For instance, item 4 represents the larger guessing only (C4) DIF type, whereas the item 11 indicates the larger guessing but smaller slip (C2).

According to the nominal α criterion level of .05, it is highlighted three items displaying DIF. Both SDI and UDI have detected two common items exhibiting DIF (i.e., items 3 and 11). For those common items displaying statistically significant DIF, the item 3 has nonuniform DIF, and item 11 has been identified with uniform DIF. In addition, looking at the item parameter estimates of the item 22 in both male and female groups, it can be seen that item behave as nonuniform DIF.

The index SDI detected the item 11 as exhibiting DIF. Because the value of SDI_{11} exceeds the critical value, it would be concluded that the item functions different for the two groups. The index UDI detected the items 3, and 22 as exhibiting DIF. It turned out that item 22 was only identified by the UDI statistic.

Finally, the real data illustration results suggested the practical implementation of the proposed statistics in detecting DIF, and the the high extent of agreement between statistics in determining items that function different across groups. Moreover, both SDI and UDI have been capable to detect potential DIF in similar conditions as in the conducted simulation study.

Table 4.13. *Item parameters estimates for both groups*

	Group					
	Reference		Focal			
Item	g_R	s_R	g_F	s_F	SDI	UDI
1	0.00 (0.10)	0.07 (0.02)	0.00 (0.12)	0.02 (0.01)	4.23	8.35
2	0.00 (0.03)	0.16 (0.03)	0.00 (0.04)	0.09 (0.03)	1.38	2.76
3	0.29 (0.03)	0.21 (0.04)	0.26 (0.03)	0.01 (0.02)	9.60	26.04
4	0.31 (0.04)	0.00 (0.02)	0.46 (0.04)	0.00 (0.02)	3.45	6.91
5	0.42 (0.04)	0.05 (0.02)	0.42 (0.03)	0.00 (0.03)	0.96	1.82
6	0.89 (0.02)	0.00 (0.02)	0.81 (0.03)	0.00 (0.07)	2.24	4.48
7	0.32 (0.03)	0.00 (0.03)	0.46 (0.03)	0.00 (0.05)	4.24	8.47
8	0.25 (0.03)	0.00 (0.04)	0.26 (0.03)	0.01 (0.04)	0.00	0.12
9	0.00 (0.10)	0.05 (0.01)	0.00 (0.07)	0.02 (0.01)	1.22	2.48
10	0.26 (0.04)	0.21 (0.03)	0.31 (0.04)	0.08 (0.03)	8.58	11.06
11	0.18 (0.03)	0.48 (0.04)	0.30 (0.03)	0.36 (0.05)	12.24	13.03
12	0.52 (0.04)	0.00 (0.03)	0.63 (0.04)	0.05 (0.02)	0.26	5.66
13	0.43 (0.04)	0.01 (0.01)	0.61 (0.03)	0.01 (0.01)	6.92	11.37
14	0.35 (0.03)	0.13 (0.04)	0.26 (0.03)	0.00 (0.04)	0.03	10.41
15	0.49 (0.05)	0.14 (0.02)	0.51 (0.04)	0.06 (0.02)	4.81	7.97
16	0.07 (0.03)	0.22 (0.03)	0.11 (0.03)	0.21 (0.03)	0.80	1.02
17	0.00 (0.03)	0.11 (0.05)	0.00 (0.04)	0.16 (0.04)	0.22	0.44
18	0.33 (0.03)	0.03 (0.02)	0.41 (0.03)	0.00 (0.02)	4.21	4.28
19	0.15 (0.03)	0.01 (0.02)	0.15 (0.03)	0.06 (0.02)	1.09	2.52
20	0.52 (0.03)	0.11 (0.03)	0.53 (0.03)	0.08 (0.03)	0.62	0.80
21	0.15 (0.02)	0.65 (0.05)	0.12 (0.02)	0.63 (0.04)	0.32	1.19
22	0.45 (0.03)	0.38 (0.05)	0.36 (0.03)	0.16 (0.05)	0.93	15.74
23	0.21 (0.04)	0.19 (0.02)	0.26 (0.04)	0.15 (0.02)	2.19	2.19
24	0.15 (0.05)	0.47 (0.03)	0.29 (0.05)	0.50 (0.03)	1.01	4.25
25	0.59 (0.04)	0.00 (0.03)	0.48 (0.04)	0.00 (0.05)	1.95	3.89

Note. † Standard error in parentheses.

Chapter 5

A Computer Software for calibrating CDMs

The motivation for computer program selection and use can be influenced by features such as free and open source software, programming language and environment, and documentation. Recently, some commercial software, such as MPLUS (Muthén & Muthén, 2006) can provide item parameters estimates and attribute patterns of different CDMs under the LCDM approach (Henson, Templin & Willse, 2009; Rupp & Templin, 2008). Also, a free software such as R (R Core Team, 2013) has implemented methods for CDM estimation. To use the R program, the package called Cognitive Diagnosis Modeling (Robitzsch, Kiefer, George, and Uenlue, 2013) needs to be installed. The package can estimate models such as DINA (Junker & Sijtsma, 2001) and DINO (Templin & Henson, 2006) models. Another software is the Ox (Doornik, 2003) code developed by de la Torre (2009) that can be used to fit same CDMs. This code can be obtained from de la Torre upon request.

The programs MPLUS, R, and Ox console were originally created to perform different statistical analyses, and CDMs calibration has been implemented by typing commands to enter data and do analyses. Although users can find program documentation with detailed description, typical users are more familiar with graphical user interface (GUI) rather that environments that require programming language to do analyses.

This study aimed at providing special purpose software in the CDMs context. Specifically, it was designed and developed to perform CDMs estimation under the GUI environment. The software is called winCDM version 0.1 for windows, which includes different applications for CDMs estimation. The winCDM 0.1 program has been created to be used even by individuals without any programming skills.

5.1 Availability

WinCDM is writen in C/C++ and runs on Windows operating system. A copy of winCDM can be obtained upon request to the author by e-mail at guaner.rojas@yahoo.com or guaner.rojas@ucr.ac.cr.

5.2 Description

The first version of winCDM is a freeware windows GUI application that implements marginal maximum likelihood estimation, specifically EM method to estimate the item parameters and expected a posteriori to classify the individuals. At present, winCDM can handle the DINA and DINO models. The program provides item parameters estimates and the corresponding standard errors. In a separate output file, the attribute classifications for all individuals are provided. Moreover, winCDM also gives the posterior probability of each latent class, relative fit indices AIC and BIC, and prevalence of each attribute. In addition, winCDM has a simulation environment that can generate responses based on DINA, DINO, and A-CDM (*additive* CDM; de la Torre, 2011) models.

5.2.1 Interface Characteristics

Input files for running winCDM are Q-matrix and item responses file. Both Q-matrix and item responses should be dichotomous data in a file with extension “*.txt” in a tabulated format. They have to be placed in where the path to the winCDM executable is found. Once the program has been executed, then a windows as in Figure 5.1 shows a menu in which the following options are possible: File, Simulation, Calibration, Output, Help.

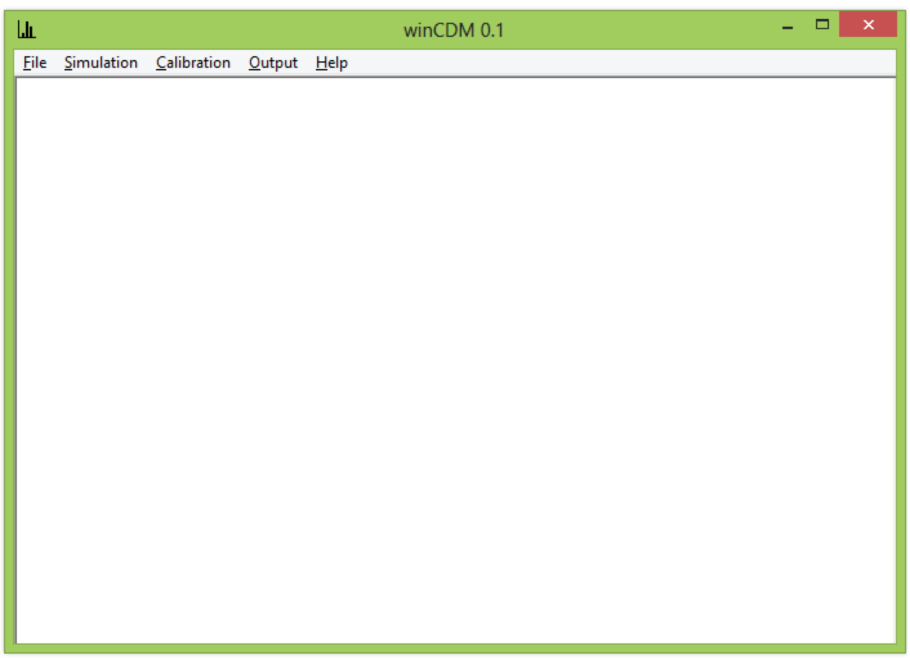

Figure 5.1. winCDM window

File* → *Exit

Allows users exit application.

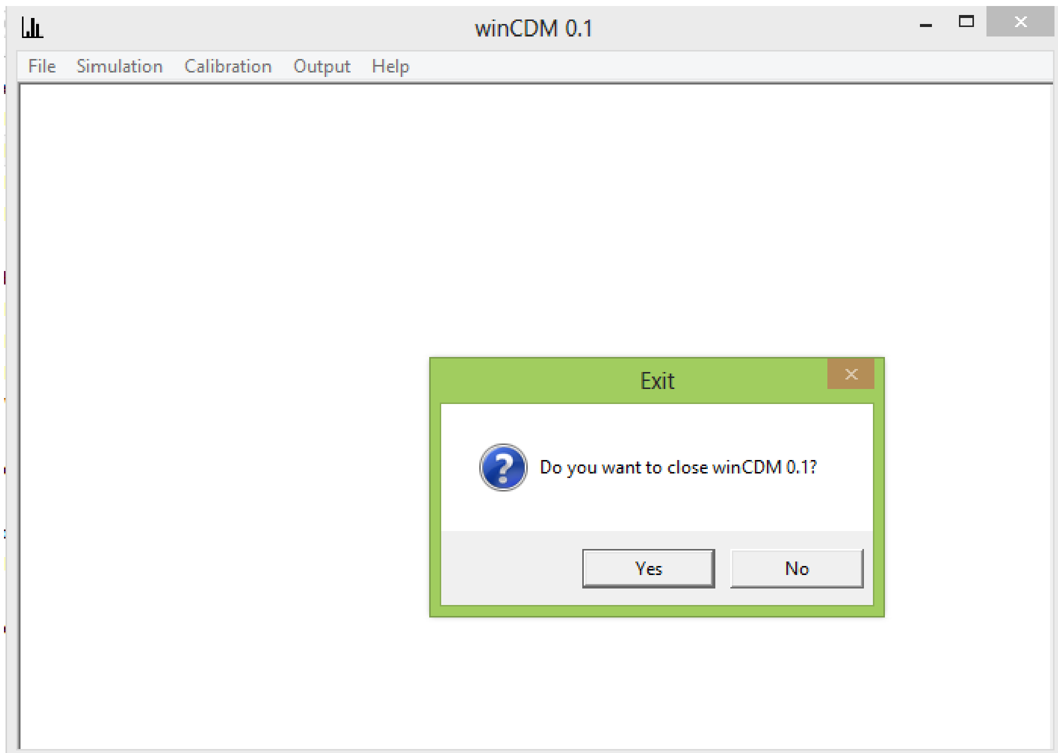

Figure 5.2. Exit Window

Simulation* → *Model Specification

Creates an item response data set based on DINA, DINO or A-CDM models. If the DINA model is selected, the default name for the generated item responses data is "sample_ dina.txt". The default name for the attributes pattern will be "alpha_ dina.txt". These files will be generated and placed in where the path to the winCDM executable is found.

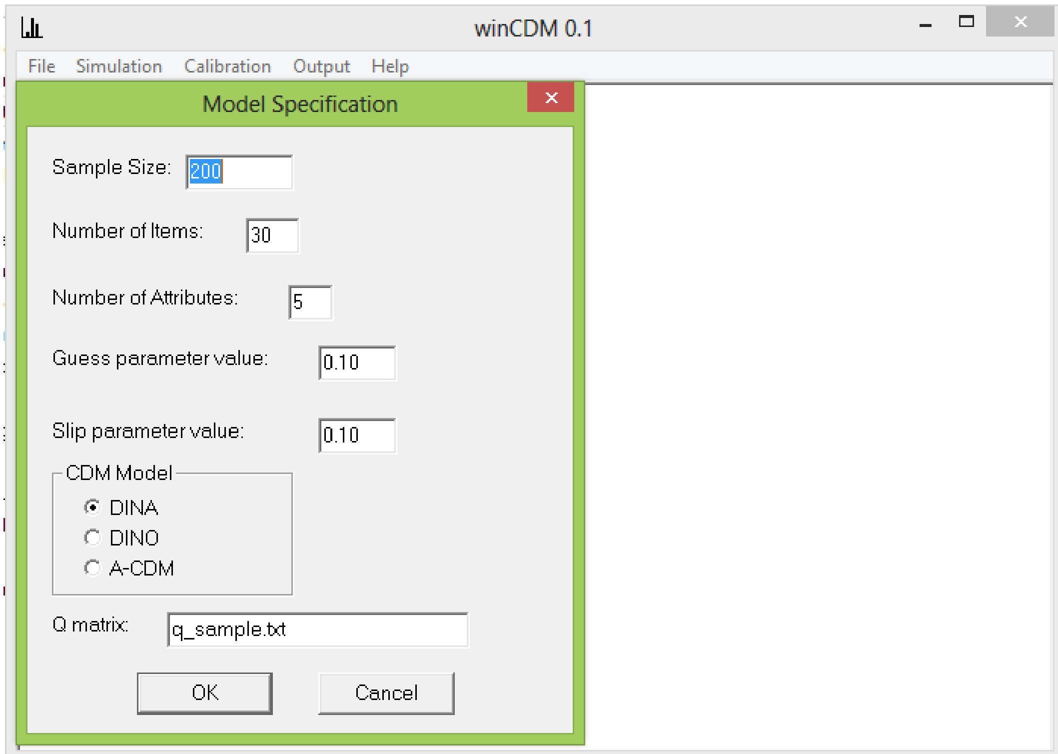

Figure 5.3. Model Specification Window

Calibration → Model Selection

The model selection option, as in Figure 5.4, allows the user to specify sample size, test lenght, number of attributes, model, name of Q-matrix and response data files used in the analysis. Users should remember that the files containg Q-matrix and item responses need to be placed in where the path to the winCDM executable is found.

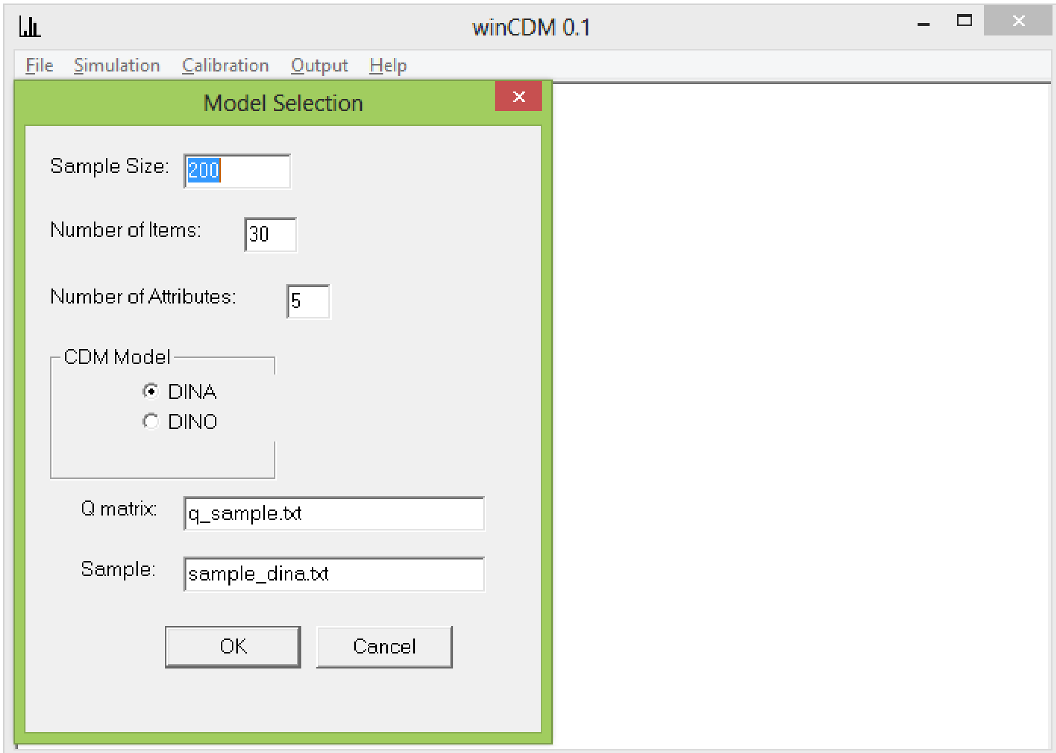

Figure 5.4. Model Selection Window

Once the model selection has been specified, a message as in Figure 5.5 will ask user for confirmation of model selection. Users should click on “OK” and they should wait for the model parameters estimation. When the program has finished, users will receive a message as in Figure 5.6.

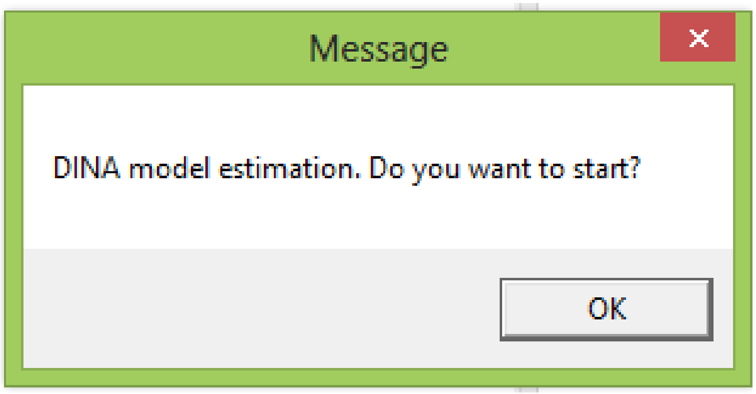

Figure 5.5. Model Selection start message Window

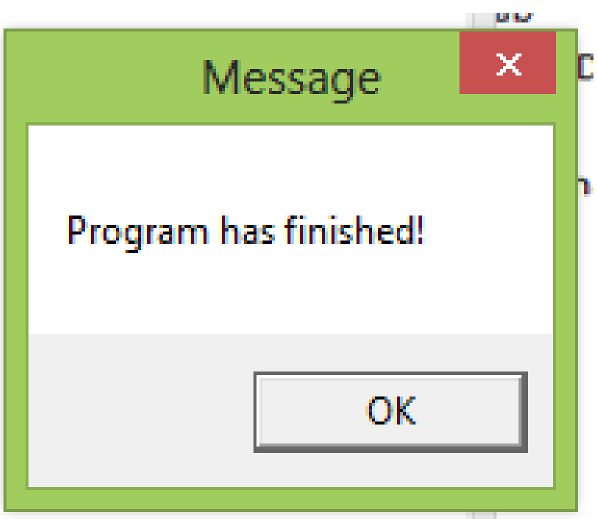

Figure 5.6. Model Selection finish message Window

Output → Go

Open the output files of calibration process. Output files contain the relative fit indices, item parameters estimates, attribute classification, attribute prevalence, and latent classes and its posterior probabilities. The output files are shown in Figure 5.7 through 5.11.

Figure 5.7 displays the relative fit indices of the DINA model calibration. Given any two estimated models, the model with the lower value of AIC adn BIC is preferred.

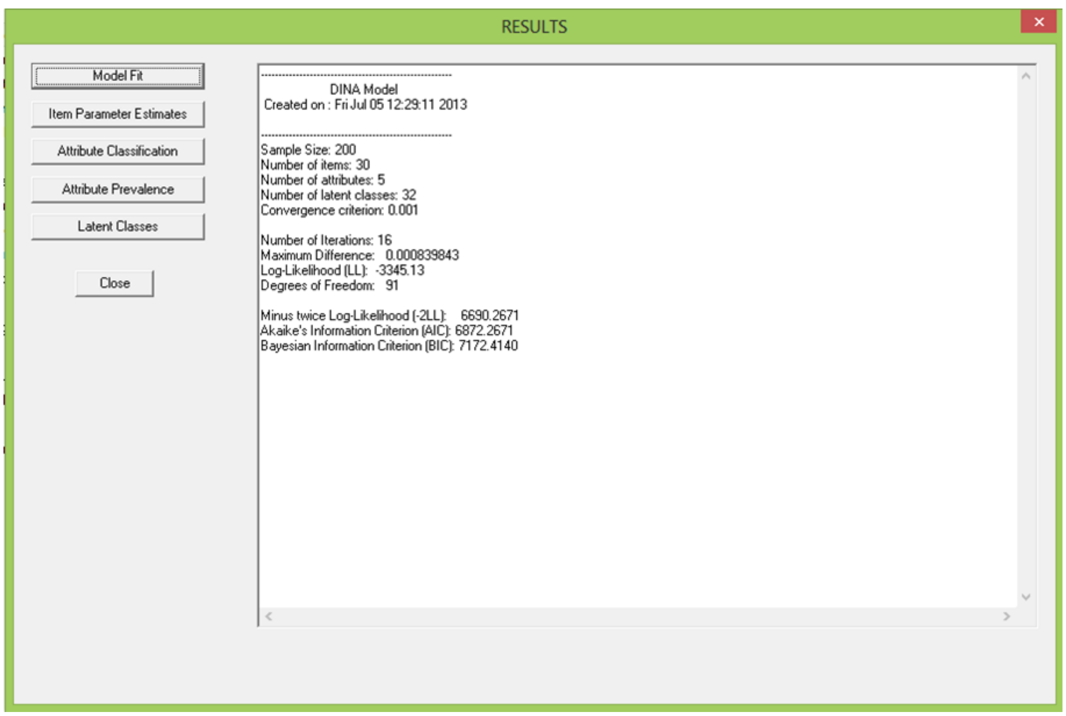

Figure 5.7. Relative Fit Indices Output Window

Figure 5.8 displays the item parameter estimates of the DINA model. For instance, the model specifies that, for item 10, only examinees who have mastered all the required attributes will have probability of success equal to $1 - 0.1319$ or 86.81%, whereas all other examinees will have a chance of success equal to 15.48%.

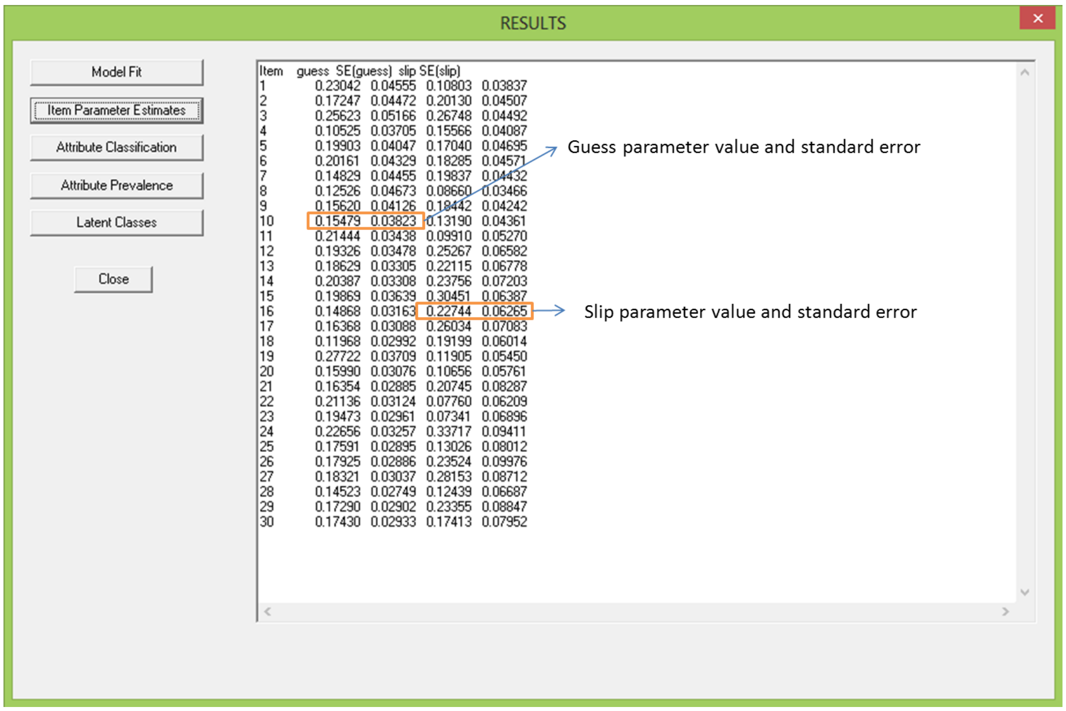

Figure 5.8. Item parameter estimates Output Window

Figure 5.9 displays the attribute classification of the DINA model. Examinee 23 has high probability of having the second and fourth attribute, but not the first, third, and fifth. By taking a cutoff point of 0.5, the attribute pattern is $\alpha_{23} = (0, 1, 0, 1, 0)$.

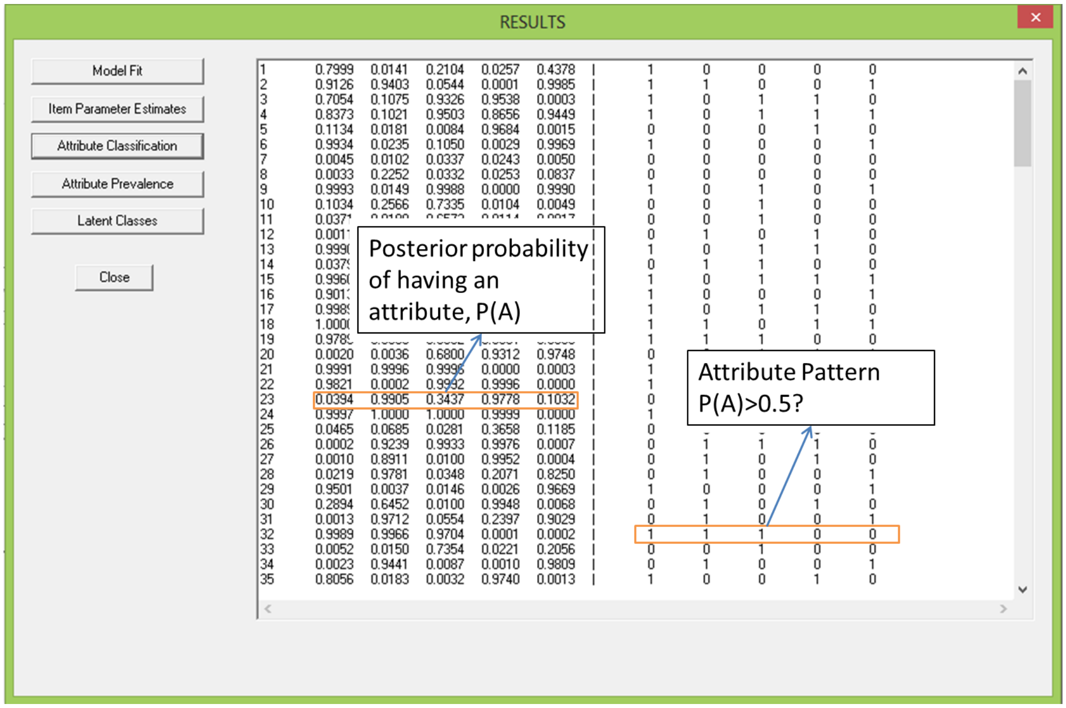

Figure 5.9. Attribute Classification Output Window

Figure 5.10 displays the attribute prevalences of the DINA model. For instance, 56.43% of examinees had the Attribute 3.

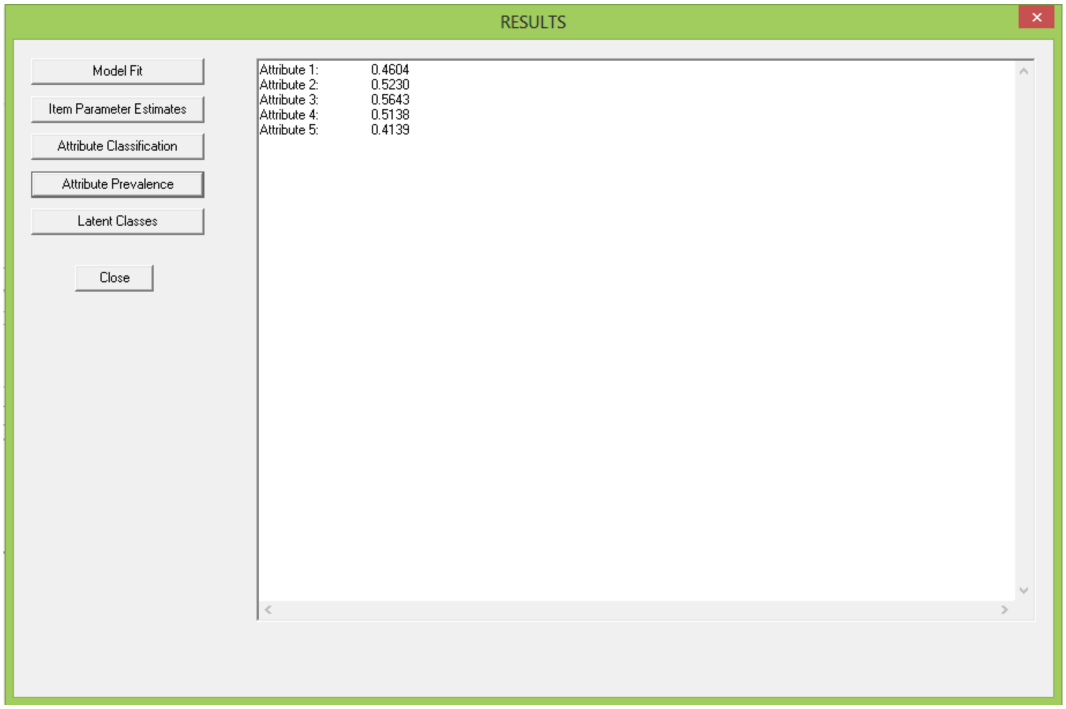

Figure 5.10. Attribute Prevalences Output Window

Figure 5.11 displays the latent classes and its posterior probabilities of the DINA model. The probability that a randomly selected examinee belongs to latent class 00000 is 0.0556.

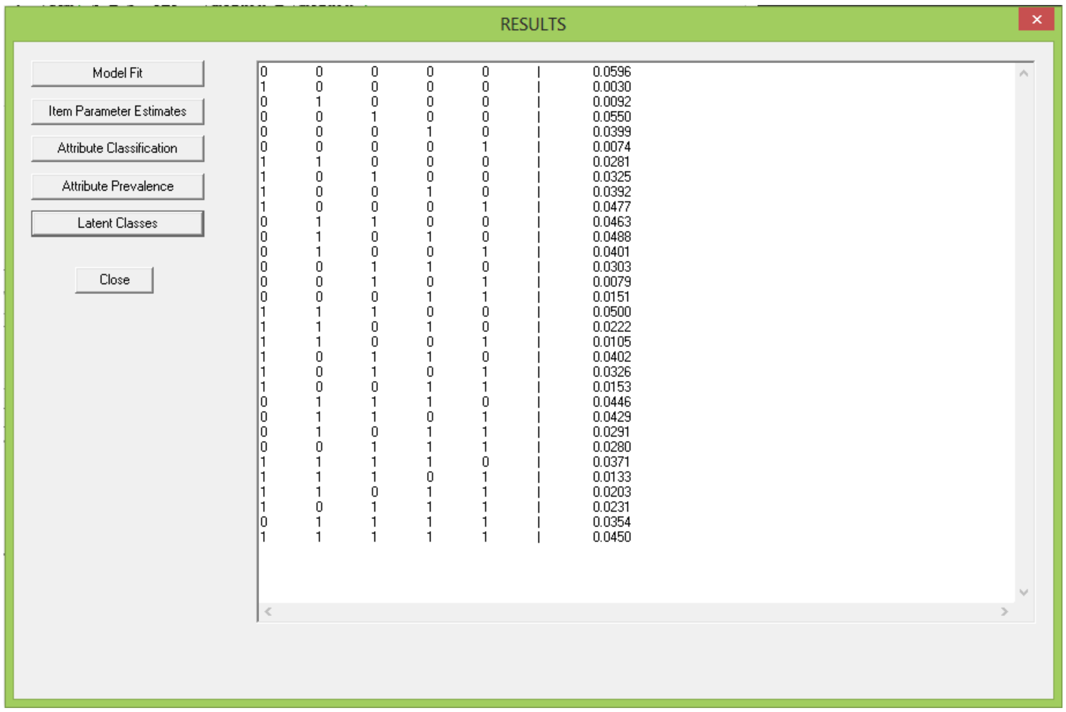

Figure 5.11. Latent Classes and its Posterior Probabilities Output Window

Chapter 6

Discussion

It is particularly clear that relative benefits of the studies developed are dependent on the needs and goals of researchers and practitioners working on psychological measurement and testing. Issues such as the sample size, test length , item quality, attribute classification, selection of a statistical model, design of a Q matrix are a few of the essential variables that must be considered at the process of developing a test. Among those variables mentioned above, a practical and theoretical implication derived from the developed project focuses on the interest of a researcher about the use of CDMs. One can be interested in obtain information with high value of interpretation in two ways, person and items. These ways can be explored through conditions that an assessment tool commonly reflects. Thus, the studies described here demonstrated several conditions that seem to be involved in testing.

6.1 Attribute Classification

In principle, one could know the correct model specification from the test design, and then use data from samples to estimate the attribute patterns and item parameters. As a result, it should be recognized that the ACA using the correct model is always more effective than employing the inappropriate model. In contrast, some caution must be exercised in applying CDMs to data with unknown model because choosing an appropriate model is essential to accomplish the benefits of the CDMs relevant to examine ACA and DIF.

Because the focus of the CDMs is to provide individual feedback to examinees regarding each of the attributes measured by the assessment, new specific and general CDMs have been proposed (de la Torre, 2011; Junker & Sijtsma, 2001; Templin & Henson, 2006; Henson, Templin & Willse, 2009; von Davier, 2005) with detail description of crucial issues such as model estimation, model fit and families of models. However, there is less understanding as to how accurately examinees are classified in real settings such as the sample size requirements.

The simulation study attempts to address the ACA in a structure test defined by the Q matrix and varying underlying model to the generated data. Using DINA, DINO, A-CDM and G-DINA models, it was argued that the G-DINA model produces high ACA as the DINA and DINO models when the sample size is small. The A-CDM play a rol as a model with more complexity than both DINA and DINO models. This A-CDM allows studies to show the performance of the G-DINA model in the ACA. It also was argued that the G-DINA model may be needed to characterize Asperger Syndrome data, and it was illustrated how the CDMs can be used to analyze tests and estimate person parameters.

Results of the simulation analysis suggested that G-DINA model was more accurate than both DINA and DINO models with small sample size in terms of ACA, when it was not the underlying model to the data and the number of attributes varied from five to ten. As reported in the simulation study, the ACA of CDMs increases with the number of items as well as the item quality. The number of items to have high PCA and PCV can vary, depending of the item discrimination. Items with high level of discrimination contributed to high PCA and PCV.

6.2 Differential Item Functioning

This dissertation also focused on the DIF. In the CDMs, particularly, in the DINA model the parameters are assumed to be invariant (de la Torre, 2009). Such an assumption could be annulled in the presence of DIF. A test with items displaying DIF can result in attribute profiles that are biased. A particular attempt of this dissertation was to provide researcher a new DIF approach by using the differences between to IRFs in the DINA model. This procedure was motivated by those studies showing that a distinct method need to be developed based on the CDMs paradigm. The DIF procedure was compared with Mantel-Haenszel method, and a study with real data illustrated the use of the indices of DIF detection.

It was proposed a perspective for DIF detection in the DINA as an attempt to cover weaknesses of previous procedures used in the studies of Li (2008) and Zhang (2006), where the known methods SIBTEST and Mantel-Haenszel were adapted to the DINA model using attribute vectors as matching criterion. The study of Li (2008) also proposed the marginalized differences in probabilities of success of an item, but it was explored in the higher-order DINA model (de la Torre & Douglas, 2004).

The new procedure concentrated on the item parameter estimates for two groups. Standardized differences were obtained based on the differences between the probabilities of success of an item, that is, it was computed and combined in statistics the difference between the guessing parameters and the discrepancy between slip parameters. Such an approach attempts to allow measurement specialists to study DIF in variety of conditions. For example, the UDI statistic would reveal whether or not the non-uniform DIF is present when the SDI could fail.

A relevant reason to begin defining DIF in the DINA model relies on keeping the data analysis as simple as possible. Thus, as long as no essential features are missed, simplicity is one of the best DIF modeling strategies. Moreover, the DIF method based on the standardized differences in the DINA model can be extended to the G-DINA model, in which by fixing different constraints a variety of specific CDMs can be obtained. This might provide a flexible approach taking into account that the model which describes the data is not always known.

It is also important to point out the observation that under other approaches, as in IRT, are needed an anchor set of unbiased items in order to link the scales of the two comparison groups. This issue is addressed by the Q matrix, which is the element used to link the scales of two groups. The generated data did not require to be transformed to the same scale with the estimated parameters due to the guessing and slip parameters are both in the invariant probability scale.

The simulation study showed that the proposed approach supports the detection of items exhibiting DIF and has the advantage of known asymptotic distributions of the statistics. It also was demonstrated that indices for DIF detection in the DINA model produced better control over Type I error and power rates than traditional methods such as Mantel-Haenszel.

6.3 Applications

Regarding to the implemented CDM in the test expressly intended to detect individuals with Asperger Syndrome; it was found that the classification of persons in the predefined groups (i.e., AD/HFA, ADHD and NDD) was high in terms of the number of attributes. This means that individual with AD/HFA had attribute profiles containing probabilities of having a criterion close to one, whereas persons classified as NDD group had attribute patterns with probabilities close to zero.

One aspect of the studies, concerns the analysis of TIMSS 2007 fourth grade mathematics assessment. In this study, DIF indices were able to detect potential DIF in similar conditions as in the conducted simulation study. Indeed, the third item which exhibited DIF has been reported in the study of Lee, Park and Tayland (2011). This item had a high value of slip parameter and students tended to choose one of the distractors instead of the correct answer.

The software in the CDMs context was designed and presented for calibrating the DINA and DINO models. One of the reasons to develop the winCDM software was motivated by the reduced number programs available to be used with graphical user interfaces that could make attractive to practitioners with minor experience in programming (Rupp & Templin, 2008a).

6.4 Limitations and Future Work

Although the studies reported here were exhaustively analyzed, two question are raised. The two simulation studies used an optimized Q matrix that have been also adopted in others investigations reported through this dissertation. The fixed Q matrix contains ten items measuring only one attribute; the next ten items need two attributes and the last ten required three attributes, that is, the Q matrix had attribute combinations in items ordered and well organized. An advantage of incorporating the use of an optimized Q matrix is that it allows specialists to control over the manipulated factors. However, in real situations Q matrices can be unstructured, unless the test design defines the Q matrix as structured format. Empirical studies should be made in supporting the ACA with other Q matrices.

Studies usually report a cutoff point of 0.5 to convert the posterior probabilities of having an attribute into dichotomous format of zeros and ones. Indeed, it was used in our simulation studies. Nevertheless, establishing a cutoff point of 0.5 may affect the attribute profiles when the probabilities are very close to 0.5, therefore is reasonable to think that instead of taking the common cutoff point of 0.5, one possible approximation might focus on ROC curves to check specificity and sensitivity, and then establish cutoff values.

A comparison among DIF indices with other methods created under the CDM approach in detecting DIF for dichotomous items is of the great interest. Future studies may also be conducted to investigate the effects of the manipulated factors of the present project on the performances of CDMs-based and other non-CDMs based procedures in detecting DIF with both simulated and real data. Finally, a systematic comparison between the performance of winCDM and other programs model parameter estimates can be part of a future simulation study. Also, other CDMs could be incorporated into the program.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19*, 716–723.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Press, Inc.

Belinchon, M., Hernández, J., & Sotillo, M. (2008). *Personas con síndrome de asperger: funcionamiento, detección y necesidades [people with asperger syndrome: functioning, detection and needs]*. España: Gráficas Flora, S.L.

Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical*, *31*, 144–152.

Campbell, J. (2005). Diagnostic assessment of aspergerś disorder: A review of five third party rating scales. *Journal of Autism and Develomental Disorders*, *35*, 25–35.

Chen, J., de la Torre, J., & Zhang, Z. (2012). *Relative and absolute fit evaluation in cognitive diagnosis modeling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education: Vancouver.

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 1–21.

de la Torre, J., & Chiu, C. Y. (2010). *General empirical method of Q-matrix validation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education: Denver.

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

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47(2)*, 227–249.

de la Torre, J., & Lee, Y. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47(1)*, 115–127.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). A review of cognitively diagnostic assessment and a summary of psychometric models. In C. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, p. 979 - 1030). Elsevier. Available from `http://www.sciencedirect.com/science/article/pii/S0169716106260310` http://dx.doi.org/10.1016/S0169-7161(06)26031-0

Doornik, J. A. (2003). Object-oriented matrix programming using Ox version 3.1. [Computer software manual]. London: Timberlake Consultants Press.

Doornik, J. A. (2008). Object-oriented matrix programming using Ox version 5.1 [Computer software manual]. London: Timberlake Consultants Press.

Embretson, S. (2010). *Measuring psychological constructs: advances in model-based approaches.* USA: American Psychological Association.

Gillberg, C. (2010). The ESSENCE in child psychiatry: Early symptomatic syndromes eliciting neurodevelopmental clinical examinations. *Research in Developmental Disabilities*, *31*, 1543–1551.

Gillberg, C., Gillberg, M., C.and Rastam, & Wentz, E. (2001). The asperger syndrome (and high-functioning autism) diagnostic interview (asdi): A preliminary study of a new structured clinical interview. *Autism*, *5*, 57–66.

Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Doctoral dissertation, Champaign, IL: University of Illinois.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Howlin, P. (2000). Assessment instruments for asperger syndrome. *Child Psychology and Psychiatric Reviews*, *5*, 120–129.

Howlin, P. (2003). Outcome in high functioning adults with autism with and without early language delays: Implications for the differentiation between autism and asperger syndrome. *Journal of Autism and Developmental Disorders*, *33*, 3–13.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Researh and Evaluation*, *15*(3), 1–7. Available from `http://pareonline.net/getvn.asp?v=15&n=3`

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.

Lee, Y., Park, Y., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the

TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Leighton, J., & Gierl, M. (2003). *Cognitive assessment for education: theory and applications*. New York: Cambridge University Press.

Li, F. (2008). *A Modified Higher-order DINA Model for detecting Differential Item Functioning and Differential Attribute Functioning*. Doctoral dissertation, University of Georgia.

Maris, E. (2010). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Matson, J. (2008). *Clinical assessment and intervention for autism spectrum disorders*. USA: Elsevier Inc.

Matson, J., & Boisjoli, J. (2008). Strategies for assessing aspergerś syndrome: A critical review of data based methods. *Research in Autism Spectrum Disorders*, *2*, 27–248.

Milewski, G., & Baron, P. (2002). *Extending DIF methods to inform aggregate report on cognitive skills*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, Lousiana.

Molloy, H., & Vasil, L. (2004). *Asperger syndrome, adolescence, and identity: Looking beyond the label*. London: Jessica Kingsley Publishers.

Muthén, L., & Muthén, B. (1998-2006). M-plus user's guide (4th ed.) [Computer software manual]. Los Angeles: Muthén, L.K., & Muthén.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org/`

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53(4)*, 495–502.

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response funtions. *Applied Psychological Measurement*, *14(2)*, 197–207.

Robitzsch, A., Kiefer, T., George, A., & Uenlue, A. (2013). CDM: Cognitive diagnosis modeling [Computer software manual]. Available from `http://CRAN.R-project.org/package=CDM` (R package version 1.5-12)

Roussos, L., Templin, J., & Henson, R. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44(4)*, 293–311.

Roussos, L., Templin, J., & Henson, R. (2009). Measuring psychological constructs: advances in model-based approaches. In (chap. Skills diagnosis for Education an Psychology with IRT-based parametric latent class models). USA: American Psychological Association.

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

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

Rupp, A., Templin, J., & Henson, R. (2010). *Cognitive assessment for education: theory and applications*. USA: The Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconception based on item response theory. *Journal of Education Statistic*, *20*, 345–354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Tech. Rep. Nos. Research Report RR–05–16). Princeton: Educational Testing Service.

World Health Organization. (2010). *International classification of diseases* (10th ed.). Geneve, Switzerland: Author.

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

Apéndice A

Introducción

Las pruebas psicológicas juegan un papel importante en distintos ámbitos, como la educación, la psicología clínica y la organizacional. Por ejemplo, los educadores utilizan tests con puntuaciones para determinar quién será admitido a la universidad, los psicólogos clínicos los usan para ayudar en el diagnóstico de trastornos psicológicos y los psicólogos del trabajo y organizacionales, para procesos de selección.

Dos de los modelos de medición más comúnmente utilizados son la Teoría Clásica de los Tests (TCT) y los modelos de variables latentes. La primera teoría se centra en el concepto de valor esperado a partir de una puntuación observada, mientras que el segundo modelo conceptualiza atributos teóricos como variables latentes. A diferencia de la primera, la teoría de variables latentes utiliza modelos estadísticos ajustados para estimar las puntuaciones de los sujetos a partir de los datos observados.

En los modelos estadísticos comunes, tales como el Análisis Factorial Confirmatorio (AFC) y la Teoría de Respuesta al Ítem (TRI) uni y multidimensional, se asume que las puntuaciones de los sujetos en las variables latentes son continuas. Basándose en esta teoría, se puede clasificar las puntuaciones asignadas a los sujetos en distintos niveles, en función de los puntos de corte identificados por los investigadores en una escala continua latente.

A pesar de la popularidad de ambas teorías, TCT y TRI, en la actualidad, los modelos conocidos como Modelos de Diagnóstico Cognitivo (MDC) están cobrando una mayor relevancia en la literatura reciente sobre medición (de la Torre, 2011; 2009; de la Torre & Lee, 2010; Junker & Sijtsma, 2001; Henson, Templin & Willse, 2009; Huebner, 2010; Rupp, Templin & Henson, 2010; von Davier, 2005) y en las conferencias internacionales más importantes, como la conferencia de la Sociedad Psicométrica y el Consejo Nacional de Educación. La mayoría de las investigaciones con MDC se han centrado en la formulación y estimación de nuevos modelos.

Los MDC son modelos multidimensionales y confirmatorios, desarrollados específicamente para el diagnóstico de presencia o ausencia de diferentes atributos para resolver ítems de un test. El término atributo hace referencia a la variable latente que se asume como discreta. La naturaleza multidimensional de los MDC implica que existen varios atributos que se pueden medir en un mismo test, mientras que su característica confirmatoria significa que un modelo puede asociarse a una estructura previa basada en una teoría concreta. Sin embargo, la clave conceptual de los MDC se centra en una matriz con las especificaciones de los atributos, llamada Q-matrix. La Q-magtrix es fundamental para estimar los parámetros de un modelo, puesto que describe qué ítem está relacionado con cada atributo.

Además de las características multidimensional y confirmatoria de estos modelos, las respuestas a los ítems se modelan a partir de los parámetros que definen un ítem y patrones de atributos. El número de parámetros de un ítem depende del grado de complejidad del modelo utilizado para describir los datos. Por ejemplo, en el modelo DINA (Junker & Sijtsma, 2001), que es un modelo parsimonioso, se deben estimar únicamente dos parámetros para cada ítem. Sin embargo, en otros modelos más complejos, como por ejemplo, el modelo de G-DINA (de la Torre, 2011), el número de parámetros depende del número de atributos que conforman un ítem. Independientemente del MDC utilizado, se debe estimar un vector o patrón de atributos que contenga la probabilidad de que cada uno de estos atributos esté presente. Este vector de atributos se expresa normalmente con ceros y unos. Por lo tanto, la probabilidad más cercana a 1 se transforma en 1 e indica que el sujeto posee dicho atributo.

El objetivo principal de los MDC es el de clasificar a los individuos en un conjunto de categorías predefinidas o clases latentes. Estas categorías provienen del número de atributos medidos por un test. Utilizando los MDC como herramienta de medición, se obtiene un perfil detallado de cada individuo. A partir de estos perfiles, los investigadores, profesores y psicólogos pueden desarrollar planes de acción en el ámbito de la educación y de la psicología. Por ejemplo, en el ámbito de la psicología clínica, los patrones de atributos pueden proporcionar a los clínicos información relevante que les ayude a tratar algún tipo de trastorno (Templin & Henson, 2006). Igualmente, en el ámbito educacional, estos modelos permiten diseñar actividades de instrucción o de aprendizaje concretas para grupos definidos a partir de los perfiles obtenidos con el modelo (DiBello, Roussos & Stout, 2006).

Otro de los objetivos más importantes en una medición, es obtener estimaciones válidas y precisas de los sujetos examinados en las variables latentes de interés para el investigador. La puntuación de un individuo se expresa a través de la clasificación de sus atributos dentro del patrón de atributos. La estimación de una clasificación de atributos se ve afectada por distintas condiciones, tales como, número de atributos, tamaño de la muestra, calidad del ítem y longitud del test (e.g., de la Torre, Hong & Deng, 2010; Rupp & Templin, 2008a; Rupp & Templin, 2008b). Los estudios de simulación de von Davier (2004) y de la Torre y Douglas (2004) muestran que los MDC, como el modelo de diagnóstico generalizado y el modelo DINA, pueden ofrecer una precisión de la clasificación de un atributo individual dentro de un patrón de atributos superior al 90 %, siempre y cuando el modelo que subyace a los datos sea correcto. Sin embargo, no hay respuestas definitivas con respecto al tamaño de la muestra necesario cuando los investigadores escogen un modelo para fines de clasificación de atributos.

Una segunda cuestión estadística y metodológica que surge del paradigma de los MDC es el sesgo del ítem, para el cual hasta ahora se ha dedicado poca investigación (Rupp & Templin, 2008a; Li, 2008; Zhang, 2006). Sabiendo que cada ítem debería contribuir a la discriminación entre clases latentes y que las probabilidades de cada atributo se estiman asumiendo parámetros conocidos de los ítems, la cuestión de la no-invariancia del ítem es importante para la clasificación de atributos en los subgrupos de sujetos. La no-invariancia del ítem puede investigarse a través del funcionamiento diferencial del ítem (DIF). La presencia de DIF podría influenciar las estimaciones de los parámetros de los ítems y, por lo tanto, afectar a la clasificación de atributos.

Entre los beneficios de la implementación de los MDC se encuentra el software para estimaciones de los MDC. Algunos programas como R (R Core Team, 2013) u Ox (Doornik, 2003) utilizan un código de programación para ajustar los MDC.

Los investigadores, profesionales o desarrolladores de test están más acostumbrados a interfaces de usuario sencillas (point and click software) y normalmente tienen menos experiencia en entornos que requieran conocimientos de lenguaje de programación para realizar dichos análisis. Por lo tanto, los programas basados en una interfaz gráfica de usuario pueden facilitar a los investigadores la realización de estimaciones con MDC, sin necesidad de programar.

Esta investigación se centra en dos cuestiones metodológicas importantes relacionadas con la clasificación de atributos en la medición: la precisión de la clasificación de atributos y el funcionamiento diferencial en el contexto de los MDC. Por lo tanto, se han planteado tres objetivos principales en esta tesis: el primero, ha consistido en comparar sistemáticamente el impacto de una muestra pequeña en la precisión de la clasificación de atributos, en modelos generales y específicos de los MDC; el segundo objetivo, ha sido introducir un nuevo procedimiento para identificar el funcionamiento diferencial del ítem, en el contexto de los MDC; y, el tercer objetivo, ha sido desarrollar un programa informático, con una interfaz sencilla para el usuario, que sirva para calibrar los parámetros de ítems y sujetos para MDC.

La tesis se divide en siete capítulos. En el capítulo dos, se introduce el marco teórico de los MDC. En el capítulo tres, se describe un estudio de simulación, implementado para comparar la precisión de clasificación de atributos de un modelo general y tres específicos de los MDC. Además, se describe en este capítulo los detalles de la implementación y del análisis de una herramienta clínica, en el contexto de los MDC. El capítulo cuatro propone y analiza sistemáticamente un nuevo método para la detección del funcionamiento diferencial del ítem en el modelo DINA. En el capítulo seis, se describe el programa informático desarrollado específicamente para calibrar los MDC. En el capítulo siete, se exponen las conclusiones y consideraciones para futuras investigaciones.

Apéndice B

Discusión

Las ventajas relativas de los estudios desarrollados en esta tesis dependen de las necesidades y los objetivos de los investigadores y profesionales que trabajan en medición. Algunas de las variables esenciales que se deben considerar en el proceso de desarrollo de un test podrían ser cuestiones como el tamaño de la muestra, número de ítems de un test, calidad de los ítems, clasificación de atributos, selección de un modelo estadístico y diseño de una matriz Q. Entre las variables mencionadas, una implicación práctica y teórica que se deriva del proyecto desarrollado se concentra en el interés de un investigador sobre el uso de los MDC. Uno se puede interesar en obtener la información con un valor alto de interpretación tanto para personas como ítems. Los sujetos e ítems se pueden explorar mediante las condiciones que un instrumento de evaluación refleja. Así, en los estudios descritos se mostraron varias condiciones que podrían estar implicadas en las pruebas.

B.1 Clasificación de Atributos

En principio, a partir del diseño de un test uno podría conocer la especificación del model correcto y así usar datos de muestras para estimar los patrones de atributos y parámetros de los ítems. Por consiguiente, se debería reconocer que la precisión de la clasificación de atributos (ACA) con el modelo correcto siempre es más efectiva que el uso de un modelo inadecuado. En contraste, un poco de cuidado se debe tener en la aplicación de los MDC a datos cuando el modelo es desconocido porque la elección de un modelo apropiado es esencial para lograr las ventajas de los MDC para examinar ACA y DIF.

Diversos autores (de la Torre, 2011; Junker & Sijtsma, 2001; Templin & Henson, 2006; Henson, Templin & Willse, 2009; von Davier, 2005) han propuesto MDC específicos y generales con descripciones detalladas de cuestiones cruciales como la estimación, ajuste y familias de modelos. Esto con el objetivo de proporcionar información a los examinados en cuanto a cada uno de los atributos medidos por una herramienta de evaluación, sin embargo, hay poca atención a la forma de clasificar a los examinados en contextos reales que involucran distintos tamaños de muestras.

El estudio de simulación ha intentado abordar la ACA mediante una estructura de un test definida por la matriz Q y variando el modelo subyacente a los datos generados. Mediante el uso de los modelos DINA, DINO, A-CDM y G-DINA se ha argumentado que el modelo G-DINA produce ACA tan alta como los modelos DINA y DINO cuando el tamaño de la muestra es pequeño. El modelo A-CDM sirvió como modelo con más complejidad que los modelos DINA y DINO. El modelo A-CDM ha permitido mostrar el desempeño del modelo G-DINA en la ACA. También, se ha sostenido que el modelo G-DINA puede ser necesario para caracterizar datos de Síndrome de Asperger, por lo que se ilustró cómo los MDC pueden usarse para analizar pruebas y estimar los parámetros de la personas.

Los resultados del estudio de simulación sugirieron que la ACA del modelo G-DINA es más exacta que los modelos DINA y DINO cuando el tamaño de la muestra es pequeño, el modelo G-DINA no es el modelo subyacente a los datos y el número de atributos varió de cinco a diez. Tal como se reportó en el estudio de simulación, la ACA de los MDC aumenta con el incremento de la cantidad y calidad de los ítems. El número de ítems para tener PCA y PCV altas puede variar dependiendo de la discriminación de los ítems. Los ítems con nivel de discriminación alto contribuyeron a PCA y PCV altas.

B.2 Funcionamiento Diferencial del Ítem

Además de la clasificación de atributos, el proyecto de tesis se concentró en el DIF. En los MDC, en particular, en el modelo DINA se supone que los parámetros son invariantes (de la Torre, 2009). Tal supuesto se podría anular en la presencia de DIF. Un test con ítems que muestran DIF puede causar perfiles de atributos sesgados. Un intento particular de esta tesis era proporcionar a los investigadores un nuevo enfoque de DIF mediante el cálculo de las diferencias entre dos funciones de respuesta del ítem en el modelo DINA. Este procedimiento fue motivado por aquellos estudios que evidenciaron la necesidad de desarrollar un método distinto basado en el marco de los MDC. El procedimiento de detección de DIF se comparó con el método de Mantel-Haenszel. Además se realizó un estudio con datos empíricos para ilustrar el uso de los índices del detection de DIF.

Se propuso una perspectiva para la detección DIF en el modelo DINA como un intento de cubrir debilidades de los procedimientos usados en los estudios de Li (2008) y Zhang (2006), donde los métodos SIBTEST y Mantel-Haenszel se adaptaron al modelo DINA usando los vectores de atributos como el criterio de contraste. El estudio de Li (2008) propuso las diferencias marginales en las probabilidades de responder correctamente a un ítem, pero dicho procedimiento se exploró en el modelo "higher-order DINA (de la Torre & Douglas, 2004).

El nuevo procedimiento se concentró en las estimaciones separadas de los parámetros de los ítems para dos grupos. Las diferencias estandarizadas se obtuvieron basadas en las diferencias entre las probabilidades de responder correctamente a un ítem, es decir se calculó y combinó en dos estadísticos la diferencia entre los parámetros de adivinación y la discrepancia entre los parámetros de desliz. Tal enfoque intenta permitir que especialistas de medición estudien DIF en una variedad de condiciones. Por ejemplo, el estadístico UDI revelaría si el DIF no uniforme está presente en los ítems cuando el SDI podría fallar.

Una razón relevante para comenzar a definir DIF en el modelo DINA recae en manterner el análisis de datos lo más simple posible. Así, mientras ninguno de los rasgos esenciales se pierdan, la simplicidad es una de mejores estrategias de modelado del DIF. Además, el método DIF basado en las diferencias estandarizadas en el modelo DINA se podría extender al modelo G-DINA, en el cual fijando diferentes restricciones se pueden obtener una variedad de MDC específicos. Esta generalización podría proporcionar un enfoque flexible teniendo en cuenta que por lo general no se conoce el modelo que subyace a los datos.

También, es importante indicar la observación que bajo otros enfoques, como en la TRI, es necesario un conjunto de ítems de anclaje para poner en la misma métrica los parámetros de los dos grupos de comparación. Esta cuestión es solventada por la matriz Q, la cual es el elemento usado para ajustar la métrica de los dos grupos. Los parámetros estimados de los datos generados no requirieron transformarse a la misma escala debido a que los parámetros de adivinación y desliz están en la escala invariante de probabilidad.

El estudio de simulación mostró que el enfoque propuesto fundamenta la detección de ítems que presentan DIF y tiene la ventaja de tener distribuciones asintóticas conocidas. También, se demostró que los índices para la detección de DIF en el modelo DINA produjeron mejor control de las tasas de error del Tipo I y de potencia que los métodos tradicionales como Mantel-Haenszel.

B.3 Aplicaciones

En cuanto al CDM puesto en práctica en la prueba creada para detectar individuos con el Síndrome de Asperger; se encontró que la clasificación de personas en los grupos predefinidos (es decir, AD/HFA, ADHD y NDD) es alta en términos de número de atributos. Esto significa que los individuos con AD/HFA tuvieron perfiles de atributos con probabilidades de tener un criterio muy cercanas a uno, mientras que las personas del grupo de NDD presentaron patrones de atributos con probabilidades cercanas a cero.

Un aspecto de los estudios concierne al análisis de la evaluación de TIMSS del 2007 en cuarto grado. En este estudio, los índices de DIF fueron capaces de detectar ítems con potencial DIF en condiciones similares a el estudio de simulación. En efecto, el tercer ítem que presentó DIF ha sido reportado en el estudio de Lee, Park y Tayland (2011). Este ítem tiene un valor alto en el parámetro de desliz y los estudiantes tendieron a elegir uno de los distractores en vez de la respuesta correcta.

El software en el contexto de los MDC se diseñó y se presentó para calibrar los modelos DINA y DINO. Uno de los motivos para desarrollar el software winCDM fue motivado por el cantidad reducida de programas disponibles para usarse con interfaces gráficas de usuario que podrían ser más atractivas para profesionales con poca experiencia en la programación (Rupp & Templin, 2008a).

B.4 Limitaciones y líneas futuras de investigación

Aunque los estudios reportados se analizaron exhaustivamente, surge una preocupación relacionada con la matriz Q. Los dos estudios de simulación usaron una matriz Q optimizada que ha sido adoptada en varias aplicaciones reportadas en esta tesis. Esta matriz Q contiene diez ítems que miden sólo un atributo; los diez siguientes ítem requieren dos atributos y los últimos diez involucraron tres atributos, es decir la matriz Q tenía combinaciones de atributos bien organizados y ordenados. Una ventaja de incorporar el uso de una matriz Q optimizada consiste en que permite a los especialistas controlar los factores manipulados. Sin embargo, en situaciones reales las matrices Q tienen formatos no estructurados, a menos que el diseño de prueba defina la matriz Q con el formato estructurado. Los estudios empíricos deberían enforcarse en fundamentar la ACA con otras matrices Q.

Los estudios por lo general reportan un punto de corte de 0.5 para convertir en el formato dicotómico de ceros y unos las probabilidades posteriores de tener un atributo. En efecto, el punto de corte de 0.5 se empleó en los estudios de simulación.

Sin embargo, es razonable pensar que en lugar de tomar el punto de corte común de 0.5, una aproximación posible se podría concentrar en las curvas ROC para comprobar la especificidad y la sensibilidad para establecer valores de corte.

Es de gran interés que se realice una comparación entre los indices de detección DIF con otros métodos creados bajo el enfoque de los MDC. Los estudios futuros podrían investigar en datos empíricos y simulados, los efectos de los factores manipulados en esta tesis en aquellos procedimientos de detección de DIF que se basan o no en los MDC. Finalmente, una comparación sistemática entre el desempeño de winCDM y otros programas de estimación de parámetros podría ser parte de un estudio de simulación. Además, otros MDC se podrían incorporar en el programa.