A general mixture model for cognitive diagnosis

A Thesis Proposal Presented to

The Faculty of the School of Statistics

University of the Philippines – Diliman

In Partial Fulfillment

of the Requirements for the Degree of

Master of Science in Statistics

2nd Semester A.Y. 2019-2020

by

Joemari Olea

0.1 Abstract

Although the Generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011) is a general cognitive diagnosis model (CDM), it does not account for the heterogeneity that is rooted from existing subgroups in the population of examinees. To address this, this study proposes the Mixture G-DINA model, a CDM that incorporates the G-DINA model within the finite mixture modeling framework. An Expectation-Maximization algorithm will be developed to estimate the Mixture G-DINA model. To determine the viability of the proposed model, an extensive simulation study will be conducted to examine the parameter recovery performance, model fit, and correct classification rates. Responses of students to a proportional reasoning assessment will be analyzed to further demonstrate the capability of the proposed model.

Keywords: *cognitive diagnosis, G-DINA, mixture model*

Contents

Chapter 1

Introduction

Educational assessments aid educators, policy makers, and students in coming up with decisions, which is why it is essential to make sure that proper assessment and analysis of examinees' responses are carried out. The two most popular types of educational assessments are the summative and formative assessments. Summative assessments are examinations that locate the students' achievements across a continuum, which oftentimes administered when the course is finished for the purpose of certifying students to advance in his/her degree or for professional practice (National Research Council, 2001). These types of assessments are analyzed through methods in classical test theory (CTT) and item response theory (IRT; Jansenn et. al., 2014).

On the other hand, formative assessments identify the students' strengths and weaknesses, commonly conducted in the middle of the course, to aid educators in improving classroom instruction (National Research Council, 2001). Cognitive diagnosis models (CDMs) are statistical models used to analyze response data from formative assessments with the purpose of knowing which skills examinees have and have not mastered (de la Torre and Minchen, 2014).

Over the past years, several CDMs have been developed, some noteworthy models are: deterministic inputs, noisy, “and” gate model (DINA; Haertel, 1989; Junker and Sitjsma, 2001); deterministic inputs, noisy, “or” gate model model (DINO; Templin and Henson, 2006); additive CDM (ACDM; de la Torre, 2011); linear logistic model (LLM; Maris, 1999); and reduced, reparameterized unified model (RRUM; DiBello, Roussos, and Stout, 2007; Hartz, 2002). All said models are shown to be special forms of the Generalized deterministic inputs, noisy ”and” gate model (G-DINA; de la Torre, 2011).

Although the G-DINA can be thought of as a general CDM, it does not account for the heterogeneity that is rooted from possible existing subgroups in the population. The heterogeneity in the data can be caused by items with differential item functioning (DIF), or items wherein the probability of correctly answering an item is unequal for examinees with the same attribute profile but are from different groups (Hou et. al., 2014). The heterogeneity can also be brought upon by different strategies in answering problem solving questions (Mislevy and Verhest, 1990; Wang and Xu, 2015). This study will introduce an extension of the G-DINA model by incorporating it with finite mixture modelling framework. Finite mixture models are a flexible way of modelling datasets that come from different subgroups of population (McLachlan and Basford, 1988). The proposed model, named *Mixture G-DINA model*, aims to provide better model fit of response data from examinees that come from multiple subgroups inherent in the population.

In the CDM literature, von Davier (2007) introduced a mixture CDM named Mixture Generalized Diagnostic Models (MGDM), which is an extension of the Generalized Diagnostic Models (GDM; von Davier, 2005). However, it can be noted that GDM and G-DINA model have some differences. The most evident is the parametrization of the model: GDM contains parameters similar to item difficulty and discrimination, which resemble those of Rasch item response theory models that uses the logit link, which is why it is considered to be a mixed Rasch model (Henson et. al., 2009). Furthermore, GDM is currently being applied only as a compensatory model, or a model which allows examinees "make-up" from lacking one required skill by mastering another (Henson et. al., 2009). On the other hand, the special cases of the G-DINA model can be considered as compensatory and non-compensatory models, and the G-DINA model can be utilized to have identity, logit, and log links (de la Torre, 2011). Hence, the Mixture G-DINA model is being proposed.

1.1 Significance of the Study

The main appeal of extending CDMs to include mixture modelling is that it can be used in determining whether different distributions must be assumed for different subgroups. This means that it can improve model fit and correct classification rate of examinees than using the usual G-DINA model for fitting datasets with inherent heterogeneity. Several accounts have also stated its viability in solving some popular problems in CDMs. According to von Davier (2007), it can be a stepping stone in determining differential item functioning. In item response theory (IRT), mixture modelling has already been used in uncovering different strategies in answering a problem-solving question (Mislevy and Verhest, 1990; Wang and Xu, 2015), which is also a certain problem in CDMs. Additionally, Aitkin and Tunnicliffe Wilson (1980) stated that mixture modelling can also be used in identifying outliers in the dataset, opening its practicality for identification of aberrant response patterns.

1.2 Objectives of the Study

This study proposes to employ EM Algorithm to obtain the maximum likelihood estimates for the parameters of the Mixture G-DINA model and test the capability of the proposed model in different aspects in comparison with the usual G-DINA model. To be specific, the objectives of the study are:

1. To examine the performance of Mixture G-DINA model in estimating the different item parameters from different subgroups
2. To analyze whether it can improve the correct classification of the attribute profile of students from different subgroups than the usual G-DINA model
3. To determine the performance of the model in correctly classifying the subgroups of the examinees
4. To identify what relative model fit criteria can be used in assessing the number of group components in the population, if the number of subgroups are unknown.
5. To apply the proposed model in an actual data.

1.3 Scope and Limitations

This study will only be limited in studying the advent of Mixture G-DINA model on the assumption that the attributes and responses are dichotomous in nature. This study only focuses in the performance of the proposed model and the proposed estimation procedure in estimating the model parameters from different subgroups with different distributions.The assessment of the proposed model in its viability to extract information about differential item functioning in the examination, multiple strategies in answering problem-solving questions, and aberrant responses from examinees are not included in the study.

Chapter 2

Review of Related Literature

2.1 On Cognitive Diagnosis Models

Cognitive diagnosis models (CDMs) or Diagnostic classification models (DCMs) are statistical models that elicit a set of finer-grained examinee's attributes or skills for diagnostic purposes (de la Torre and Minchen, 2014). Although CDMs originated from education with the purpose of analyzing formative assessments, CDMs have already been used to analyze clinical response data (de la Torre et al., 2015), and responses from competency based situational judgment tests (Garcia et. al., 2014, and Sorrel et. al., 2016).

In the CDM literature, the attributes of examinee i are represented by a vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})$, where K is the number of attributes in consideration for the entire exam. For the case wherein the attributes are dichotomous, $\alpha_{ik} = 1$ if examinee i mastered attribute k, and $\alpha_{ik} = 0$ otherwise. The required attributes in answering the items in the examination is represented by the $J \times K$ matrix called *Q-matrix*, conceptualized by Tatsuoka (1983), where J is the number of items in the exam. For the case of dichotomous attributes, $q_{jk} = 1$ if attribute k is required to answer item j, and $q_{jk} = 0$ otherwise.

The response data of the examinees are represented by the $N \times J$ matrix $\mathbf{X}$. On the assumption that the responses are dichotomous, $X_{ij} = 1$ if examinee i managed to answer item j correctly, $X_{ij} = 0$ otherwise.

For the past years, several CDMs have been developed; some of which are: deterministic inputs, noisy, "and" gate model (DINA; Haertel, 1989; Junker and Sitjsma, 2001); deterministic inputs, noisy, "or" gate model (DINO; Templin and Henson, 2006); additive CDM (A-CDM; de la Torre, 2011); linear logistic model (LLM; Maris, 1999); and reduced, reparametrized unified model (RRUM; DiBello, Roussos, and Stout, 2007; Hartz, 2002). All said models are shown to be special forms of the Generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011).

2.1.1 The G-DINA model

The G-DINA model framework is proposed to loosen the assumptions of the DINA model and allow the differing subsets of attributes to have different success probabilities for an item (de la Torre, 2011). In the DINA model, the population is partitioned into two groups per item: those who have all the required attributes, and those who lacks at least one required attribute in answering an item, while in the G-DINA model, the population is divided into $2^{K_j^*}$ groups, where $K_j^* = \sum_{k=1}^{K_j^*} q_{jk}$ is the number of required attributes for item j (de la Torre and Lee, 2013).

The reduced attribute vector $\boldsymbol{\alpha}_{lj}^* = (\alpha_{lj1}, \alpha_{lj2}, ..., \alpha_{ljK_j^*})$ will be used in place of the full attribute vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$, where the reduction is based on row j of the Q-matrix. As an example, if $K = 3$ and item j only requires the first and third attribute α_1 and α_3 (i.e. $\mathbf{q}_j$, the row j of the Q-matrix is equal to $(1, 0, 1)$), the reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$ is equal to $(\alpha_{lj1}^* = \alpha_1, \alpha_{lj2}^* = \alpha_3)$. Several link functions can be utilized for the G-DINA model such as the *identity, logit,* and *log* links. The item response function (IRF) of the G-DINA model under the identity link is given by:

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k\prime=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk\prime}\alpha_{lk}\alpha_{lk\prime} + ... + \delta_{j12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \tag{2.1}$$

where $P(\boldsymbol{\alpha}_{lj}^*)$ is the probability of a correct response for item j for examinees with $\boldsymbol{\alpha}_{lj}^*$ reduced attribute vector for item j, δ_{j0} is the probability of correct response if the examinee does not master any attribute required for item j, δ_{jk} is the change in probability of a correct response due to the mastery of α_k, $\delta_{jkk'}$ is the change in the probability of a correct response due to the mastery of α_k and $\alpha_{k'}$ that is over and above the additive impact of the mastery of the same two attributes, and $\delta_{j12...K_j^*}$ is the interaction effect due to $\alpha_1, ..., \alpha_{K_j^*}$ or or the change in the probability of the correct response due to the mastery all the attributes that is over and above the additive impact of the main and lower-order interaction effects. Several models can also be derived from the G-DINA model, which will be presented in the next sections.

2.1.2 The DINA Model

One of the first CDMs is the deterministic inputs, noisy, "and" gate (DINA) model, a parsimonious model that only has two parameters per item regardless of the number of attributes required to answer the item correctly: namely the guessing and slip parameters (de la Torre, 2009). The ideal response η_{ij}, is equal to 1 if examinee i has all the required attributes needed to answer item j correctly, 0 otherwise. The guessing parameter g_j is the probability of a correct response from an examinee that has not mastered all the required attributes for item j, that is $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$. The slip parameter s_j is the probability of a wrong response from an examinee that has mastered all the required attributes for item j, that is $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$. Based on this, the probability of a correct response from an examinee i with attribute pattern $\boldsymbol{\alpha}_i$ can be expressed as

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}. \tag{2.2}$$

In the DINA model, mastering other attributes do not compensate for the lack of the required attributes, which is why the DINA model is a non-compensatory model (Ravand, 2015). As an example, for the case that $K = 3$ and $\mathbf{q}_j = (1, 1, 1)$, then $P_j(\boldsymbol{\alpha}_i) = g_j$ if $\boldsymbol{\alpha}_i$ is equal to any of the following attribute patterns: (0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1), while $P_j(\boldsymbol{\alpha}_i) = 1 - s_j$ only if $\boldsymbol{\alpha}_i = (1, 1, 1)$.

If we set all the parameters in the G-DINA model except δ_{j0} and $\delta_{j12\ldots K_j^*}$ to be equal to 0, the IRF of the G-DINA model becomes

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \tag{2.3}$$

This IRF is equivalent to the IRF of the DINA model, wherein $g_j = \delta_{j0}$ and $1 - s_j = \delta_{j0} + \delta_{j12\ldots K_j^*}$.

2.1.3 The DINO Model

Similar to the DINA model, the deterministic inputs, noisy, "or" gate model or DINO model contains two parameters per item. Let $\eta_{ij} = 1$ if examinee i has at least one of the required attributes, 0 otherwise. The IRF of the DINO model is given by

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = {g'_j}^{1-\eta_{ij}}(1 - s'_j)^{\eta_{ij}}, \tag{2.4}$$

where g'_j is the probability of a correct response from an examinee that has at least one of the required attributes, and s'_j is the probability of an incorrect response from an examinee that has none of the required attributes. In contrast with the DINA model, all the examinees will have the same probability of correct response, except the examinees that don't have any of the required attributes in answering the item. The G-DINA model can be reduced to the DINO model by setting

$$\delta_{jk} = -\delta_{jk'k''} = ... = (-1)^{K_j^*+1}\delta_{j12...K_j^*}, \tag{2.5}$$

for $k = 1, ..., K_j^*, k' = 1, ..., K_j^*$, and $k'' > k', ..., K_j^*$. Therefore, aside from alternating the sign according to the order of the interaction, the magnitude of the main and interaction effects is also set to be equal to each other. Based on the parameters of the G-DINA model and using these constraints, $g'_j = \delta_{j0}$, and $(1 - s'_j) = \delta_{j0} + \delta_{jk}$.

2.1.4 Additive CDM

If we set the interaction effects of the G-DINA model to be equal to 0, the IRF reduces to

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}. \tag{2.6}$$

This is the IRF of the Additive Cognitive Diagnosis Model (A-CDM; de la Torre, 2011). In the A-CDM, mastering attribute α_k increases the success probability on item j by δ_{jk}, and the contribution of the attributes to the success probability are independent from each other. In comparison with the DINA and DINO model, which only have two parameters per item, there are $K_j^* + 1$ parameters per item in the A-CDM.

2.1.5 Linear Logistic Model

The Linear Logistic Model (LLM; see Maris, 1999) has the following IRF:

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \frac{\exp\left(\lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk}\right)}{1 + \exp\left(\lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk}\right)} \tag{2.7}$$

If we take the logit of the success probability, we will have the following

$$logit[P(\boldsymbol{\alpha}_{lj})] = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk}\alpha_{lk}, \tag{2.8}$$

We can see that the IRF of the LLM resembles that of A-CDM, but instead of using the identity link, LLM utilizes the logit link. Like the A-CDM, the LLM has no interaction term and has $K_j^* + 1$ parameters per item.

2.1.6 L-CDM: RRUM and G-NIDA

The Reduced Reparametrized Unified Model (RRUM; Hartz, 2002) has the following IRF:

$$P(\boldsymbol{\alpha}_{lj}) = \pi_j^* \prod_{k=1}^{K} r_{jk}^* \times \prod_{k=1}^{K} \left(\frac{1}{r_{jk}^*}\right)^{\alpha_{lk}} \tag{2.9}$$

If we let $\pi_j^* \prod_{k=1}^{K} r_{jk}^* = \prod_{k=1}^{K_j^*}$ and $r_{jk}^* = \dfrac{g_{jk}}{(1 - s_{jk})}$, the IRF is now equal to

$$P(\boldsymbol{\alpha}_{lj}) = \prod_{k=1}^{K_j^*} g_{jk} \times \prod_{k=1}^{K_j^*} \left(\frac{1 - s_{jk}}{g_{jk}}\right)^{\alpha_{lk}} \tag{2.10}$$

which is the IRF of the generalized noisy inputs, deterministic, "and" gate model (G-NIDA; Junker and Sijtsma, 2001). This means that R-RUM and G-NIDA are equivalent models that vary only on the way the models are parametrized. If we take the log of the IRF above, we will get the following

$$\log P(\boldsymbol{\alpha}_{lj}) = \sum_{k=1}^{K_j^*} \log g_{jk} + \sum_{k=1}^{K_j^*} \alpha_{lk} \log\left(\frac{1-s_{jk}}{g_{jk}}\right) = \nu_{j0} + \sum_{k=1}^{K_j^*} \nu_{jk}\alpha_{lk} \tag{2.11}$$

where $\nu_{j0} = \sum_{k=1}^{K_j^*} \log g_{jk}$ and $\nu_{jk} = \log\left(\frac{1-s_{jk}}{g_{jk}}\right)$ which is the IRF of the Log CDM (L-CDM; de la Torre, 2011). This resembles the G-DINA model with the log link wherein the interaction effects are equated to 0.

All A-CDM, LLM, and L-CDM are additive models. However, the three will not provide the same fit since the effect of additional skills mastered behaves differently for the three models. The mastery of an attribute has an additive impact to the probability of a correct response for the A-CDM, while it has a multiplicative impact for the L-CDM. Lastly, the mastery of an attribute has a multiplicative impact on the odds of a correct response for the case of the LLM.

2.1.7 Estimation of the G-DINA Model

In order to get the maximum likelihood estimates (MLE) for the parameters in the G-DINA model, de la Torre (2011) implemented an Expectation-Maximization (EM) Algorithm to maximize the log of marginalized likelihood given below:

$$l(\mathbf{X}) = \log L(\mathbf{X}) = \log \prod_{i=1}^{N} \sum_{l=1}^{L} L(\mathbf{X}_i|\boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l) \tag{2.12}$$

where the likelihood of $\mathbf{X}_i$, the response vector of examinee i, given the attribute vector $\boldsymbol{\alpha}_l$ denoted by $L(\mathbf{X}_i|\boldsymbol{\alpha}_l)$ is expressed as,

$$L(\mathbf{X}_i|\boldsymbol{\alpha}_\mathbf{l}) = \prod_{j=1}^{J} P(\boldsymbol{\alpha}_{lj}^*)^{X_{ij}} \left(1 - P(\boldsymbol{\alpha}_{lj}^*)\right)^{1-X_{ij}}, \tag{2.13}$$

and $p(\boldsymbol{\alpha}_l)$ is the prior probability that an examinee has attribute pattern $\boldsymbol{\alpha}_l$.

By equating the derivative of $l(\mathbf{X})$ and with respect to $P(\boldsymbol{\alpha}^*_{lj})$ to 0 and a few algebraic manipulations, the marginal maximum likelihood estimator for $P(\boldsymbol{\alpha}^*_{lj})$ is shown by de la Torre (2011) to be given by

$$\hat{P}(\boldsymbol{\alpha}^*_{lj}) = \frac{R_{\boldsymbol{\alpha}^*_{lj}}}{N_{\boldsymbol{\alpha}^*_{lj}}} \tag{2.14}$$

where $N_{\boldsymbol{\alpha}^*_{lj}} = \sum_{i=1}^{N} p(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i)$ is the expected number of examinees in the latent group $\boldsymbol{\alpha}^*_{lj}$, $R_{\boldsymbol{\alpha}^*_{lj}} = \sum_{i=1}^{N} p(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i) \times X_{ij}$ is the expected number of examinees in the latent group $\boldsymbol{\alpha}^*_{lj}$ that will answer item j correctly, and $p(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i)$ represents the posterior probability that examinee i is in the latent group $\boldsymbol{\alpha}^*_{lj}$

In order to convert the success probability parameter estimates $\hat{P}(\boldsymbol{\alpha}^*_{lj})$ to the parameter components of the G-DINA model (such as $\delta_{j0}, \delta_{jk}, \delta_{jkk'}, ..., \delta_{j12...K^*_j}$), the design matrix $\mathbf{M}_j$ is needed. Let $\mathbf{A}_j = \alpha_{lk}$ be a $2^{K^*_j} \times K^*_j$ matrix of the possible combinations of the required attributes for item j be defined as

$$\mathbf{A}_j = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

For the saturated models, the row l of the saturated design matrix $\mathbf{M}^{(s)}_j$ can be generated from row l of $\mathbf{A}_j$. It has 1 as the first element, followed by α_{lk} for $k = 1, ..., K_j^*$, then by $\alpha_{lk} \times \alpha_{lk'}$, for $k = 1, ..., K_j^* - 1$ and $k' = k+1, k+2, ..., K_j^*$ and so forth; the last element of this vector is $\prod_{k=1}^{K_j^*} \alpha_{lk}$. To illustrate, let $K_j^* = 3$. The saturated design matrix is

$$\mathbf{M}_{j[8\times 8]}^{(s)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

To get the least squares estimates for the G-DINA components with the identity link $\boldsymbol{\delta}_j = (\delta_{j0}, \delta_{j1}, ..., \delta_{jK_j^*}, \delta_{j12}, ..., \delta_{j12K_j^*})$ given $\hat{\mathbf{P}}_\mathbf{j} = (\hat{P}(\boldsymbol{\alpha}_1^*), \hat{P}(\boldsymbol{\alpha}_2^*), ..., \hat{P}(\boldsymbol{\alpha}_{L_j}^*))$, we have the following expression:

$$\hat{\boldsymbol{\delta}}_j = \left(\mathbf{M}_j^{(s)\prime}\mathbf{M}_j^{(s)}\right)^{-1}\mathbf{M}_j^{(s)\prime}\hat{\mathbf{P}}_\mathbf{j} \tag{2.15}$$

To get the model components of the G-DINA model with the logit and log link, simply change $\hat{\mathbf{P}}_\mathbf{j}$ above with $logit\left(\hat{\mathbf{P}}_\mathbf{j}\right)$ and $\log\left(\hat{\mathbf{P}}_\mathbf{j}\right)$ respectively.

2.2 On Mixture Models

Finite mixture modelling is a flexible way of modelling heterogenous data that came from different groups inherent in the population (McLachlan and Peel, 2000). Mixture modelling has been applied in several areas such as biometrics, genetics, medicine, and marketing (Früwirth-Schnatter, 2006). McLachlan and Basford (1988) showed a maximum likelihood approach in estimating the parameters of a finite mixture model.

Define H to be the number of subgroups in a population. Let π_h (oftentimes called the mixing proportion), be the proportion of population that belong in group h, where $\sum_{h=1}^{H} \pi_h = 1$, and $\pi_h \geq 0$ for all $h = 1, 2, ..., H$. The density function of a vector of observations $\mathbf{X}_i$ is given by

$$f(\mathbf{X}_i; \boldsymbol{\phi}) = \sum_{h=1}^{H} \pi_h f_h(\mathbf{X}_i | \boldsymbol{\theta}_h), \tag{2.16}$$

where $f_h(\mathbf{X}_i|\boldsymbol{\theta}_h)$ is the density function of the random vector $\mathbf{X}_i$ on group h, $\boldsymbol{\theta}_h$ is a vector of unknown parameters associated with group h, $\boldsymbol{\phi}$ is the vector of unknown parameters consisting of the model parameters associated with group h and the mixing parameters, i.e., $\boldsymbol{\phi} = (\boldsymbol{\pi'}, \boldsymbol{\theta'})'$ where $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_H)'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, ..., \boldsymbol{\theta}'_H)'$.

In order to determine the group classification of the observations, Bayes' Rule is commonly utilized (Bayes, 1763). The probability that element i of the population is in group h given the random vector $\mathbf{X}_i$ is given by

$$\tau_{ih} = \frac{\pi_h f_h(\mathbf{X}_i|\boldsymbol{\theta}_h)}{\sum_{h=1}^{H} \pi_h f_h(\mathbf{X}_i|\boldsymbol{\theta}_h)} \tag{2.17}$$

2.2.1 Estimation of Mixture Models

The advent of Expectation-Maximization (EM) Algorithm has increased the interest in modelling heterogenous data using finite mixture models. It simplifies the estimation of the parameters, because the groupings inside the population can be treated as missing data, in which EM algorithm excels at (McLachlan and Peel 2000). The EM Algorithm for the estimation of parameters from mixture models was presented by McLachlan and Basford (1988).

Let $L(\mathbf{X}) = \prod_{i=1}^{N} f(\mathbf{X}_i; \boldsymbol{\phi})$, where $L(\mathbf{X})$ is the likelihood function of the entire dataset $\mathbf{X}$ which is has N random vectors. To get the estimates $\hat{\boldsymbol{\phi}}$ for the model that maximizes this likelihood function, one can get the derivative of $L(\mathbf{X})$ with respect to $\boldsymbol{\phi}$ and equate the derivative with respect to 0. Define the grouping variables for random vector i to be $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{iH})$ to be equal to

$$Z_{ih} = \begin{cases} 1, & \text{if } \mathbf{X}_i \text{ is in group h} \\ 0, & \text{if } \mathbf{X}_i \text{ is not in group h} \end{cases}$$

where $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_N$ are independent and identically have multinomial distribution with parameters $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_{H-1})$. $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$ and $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_N$ are assumed to be conditionally independent with the following log density:

$$l(\mathbf{X}_i) = \sum_{h=1}^{H} Z_{ih} \log f_h(\mathbf{X}_i|\boldsymbol{\theta}_h). \tag{2.18}$$

The log likelihood for the complete data, $\mathbf{X}_1, \mathbf{X}_1, ..., \mathbf{X}_N$ and $\mathbf{Z}_1, \mathbf{Z}_1, ..., \mathbf{Z}_N$ is given by

$$L_c(\boldsymbol{\phi}) = \sum_{h=1}^{H} \sum_{i=1}^{N} Z_{ih}[\log \pi_h + \log f_h(\mathbf{X}_i|\boldsymbol{\theta}_h)]. \tag{2.19}$$

For the E-step, we need to calculate

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(0)}) = E(L_c(\boldsymbol{\phi})|\mathbf{X}; \boldsymbol{\phi}^{(0)}) = \sum_{h=1}^{H} \sum_{i=1}^{N} \tau_{ih}[\log \pi_h + \log f_h(\mathbf{X}_i|\boldsymbol{\theta}_h)] \tag{2.20}$$

For the M-step, we need to choose the value of $\boldsymbol{\phi}$ that maximizes $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(0)})$. In mixture modelling, to maximize the likelihood, it solves the following equations (McLachlan and Basford, 1988).

$$\hat{\pi}_h = \sum_{i=1}^{N} \frac{\hat{\tau}_{ih}}{N} \tag{2.21}$$

$$\sum_{h=1}^{H} \sum_{i=1}^{N} \hat{\tau}_{ih} \frac{\partial \log f_h(\mathbf{X}_i|\hat{\boldsymbol{\theta}}_h)}{\partial \hat{\boldsymbol{\theta}}_h} = 0 \tag{2.22}$$

The E and M steps are alternated repeatedly until the convergence criterion is met.

2.2.2 Mixture Models used in CDMs

Mixture modelling was also utilized in item response theory and cognitive diagnosis models. In IRT, mixture models have been used by researchers to analyze the identification of different strategies in problem solving (Mislevy and Verhest, 1990; Wang and Xu, 2015), detection of test speededness of examinees (Bolt et. al., 2002), detection of differential item functioning (Cho and Cohen, 2010; Cohen and Bolt, 2005), and measurement of individual growth (Cho et.al., 2010).

In the CDM literature, von Davier (2007) introduced Mixture Generalized Diagnostic Models (MGDM), which is an extension of the Generalized Diagnostic Models (GDM; von Davier, 2005). Assuming that the attributes are dichotomous, the IRF for the Generalized Diagnostic Model is

$$P(\mathbf{X}_i = \mathbf{x}|\boldsymbol{\beta}_i, \mathbf{q}_i, \boldsymbol{\gamma}_i, \boldsymbol{\alpha}) = \frac{\exp(\beta_{xi} + \sum_{k=1}^{K} \gamma_{xik} h_i(q_i k, \alpha_k))}{1 + \sum_{y=1}^{m_i} \exp(\beta_y i + \sum_{k=1}^{K} \gamma_{yik} h_i(q_i k, \alpha_k))} \tag{2.23}$$

with some necessary conditions on $\sum_{k=1}^{K} \gamma_{yik}$ and $\sum_{y=1}^{m_i} \beta_{yi}$ to make the model identifiable. Here $h_i(.,.)$ maps α_k and q_{ik} to the real numbers and defines how the Q-matrix entries and skills interact. For the case of polytomous attributes, β_{xi} can be thought of as item difficulties, and γ_{xik} can be thought of as slope parameters. On the other hand, the IRF of the Mixture Generalized DIagnostic Model is given below

$$P(\mathbf{X}_i = \mathbf{x}|\boldsymbol{\beta}_i, \mathbf{q}_i, \boldsymbol{\gamma}_i, \boldsymbol{\alpha}, h) = \frac{\exp(\beta_{xih} + \sum_{k=1}^{K} \gamma_{xikh} h_i(q_i k, \alpha_k))}{1 + \sum_{y=1}^{m_i} \exp(\beta_y ih + \sum_{k=1}^{K} \gamma_{yikh} h_i(q_i k, \alpha_k))}, \tag{2.24}$$

which resembles the IRF for the GDM. The only difference is that the item parameters are indexed by h, indicating the possibility of distinct item parameters per group. von Davier proposed an estimation procedure for the parameters in MGDM using EM algorithm. To prove its viability in real-life application, von Davier (2007) then applied MGDM to analyze responses from an English language test (TOEFL iBT pilot data), the 2002 12th Grade National Assessment of Educational Progress in reading and mathematics, and an adult literacy assessment.

Aside from the obvious differences in the parametrization of the two general CDMs, G-DINA and GDM have other differences. The IRF of GDM resemble those of Rasch item response theory models that uses the logit link, which is why it is considered to be a mixed Rasch model (Henson et. al., 2009). Furthermore, GDM is currently being applied only as a compensatory model, or a model which allows examinees “make-up” from lacking one required skill by mastering another (Henson et. al., 2009). On the other hand, the special cases of the G-DINA model can be considered as compensatory and non-compensatory models, and the G-DINA model can be utilized to have identity, logit, and log links (de la Torre, 2011).

Another finite mixture cognitive diagnosis model is proposed by Zhan (2020), which is called the DINMix models. The commonly used IRF for the relationship between observed and ideal responses can be expressed as follows:

$$P(X_{ij} = 1|g_j, s_j, \eta_{ij}) = g_j + (1 - s_j - g_j)\eta_{ij}, \tag{2.25}$$

where X_{ij} is the observed response of examinee i for item j, η_{ij} is the ideal response of the examinee i to item j, g_j is the guessing parameter of item j, and s_j is the slip parameter of item j. Under the DINA model,

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{2.26}$$

Under the DINO model,

$$\eta_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} . \tag{2.27}$$

Under the deterministic-inputs, noisy ratio (DINR; Song et. al., 2012) model,

$$\eta_{ij} = \frac{\sum_{k=1}^{K} \alpha_{ik} q_{jk}}{\sum_{k=1}^{K} q_{jk} q_{jk}} \tag{2.28}$$

The IRF of the DINMix model can be stated as

$$P(X_{ij} = 1 | g_j, s_j, \psi_{ij}) = g_j + (1 - s_j - g_j)\psi_{ij}. \tag{2.29}$$

where ψ_{ij} is the mixture ideal response of examinee i to item j, which can be further defined as

$$\psi_{ij} = \sum_{h=1}^{H} \tau_{ih} \eta_{ijh} \tag{2.30}$$

where τ_{ih} is as stated earlier in equation 2.17. Notice that the number of components can change per item as H_j is indexed by j, but it can be assumed that $H = H_j$ for all $j = 1, 2, ..., J$. In the study by Zhen (2020), the components were limited only to the mixtures of DINA, DINO, and DINR where the mixture ideal response of examinee i to item j for $H = 3$ is given by

$$\psi_{ij} = \tau_{j1} \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} + \tau_{j2} \left(1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} \right) + \tau_{j3} \frac{\sum_{k=1}^{K} \alpha_{ik} q_{jk}}{\sum_{k=1}^{K} q_{jk} q_{jk}} \tag{2.31}$$

Zhang (2020) used the Markov chain Monte Carlo (MCMC) method in estimating the item parameters in the DINMix model. Another difference of the DINMix model with the proposed mixture G-DINA model is that it was the ideal responses, not the response probabilities that were weighted. Also, the weighting is done at the examinee level, not the usual case wherein the weighting is done at the level of the subgroups. Based on the simulation study by Zhang (2020), the DINMix model parameters can recover the item parameters well especially when there is a larger number of examinees, longer test length, and higher item quality. Additionally, the DINMix model can accurately identify the subgroups in the population, either existing simultaneously in a single item or separately across different items.

2.3 On Model Fit Criteria

Several model fit criteria will be used for relative fit evaluation and determining the number of clusters in the population. This section presents these criteria. In relative fit evaluation, we compare several models to select a model with the best fit, while absolute fit evaluation is mostly concerned in conducting a test of hypothesis whether the model itself fits the data well (Chen et. al., 2013). This paper will solely focus on relative fit evaluation.

2.3.1 A Survey of Several Model Fit Criteria

The Akaike information criterion (AIC; Akaike, 1974), selects the model that minimizes

$$AIC = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + 2d \tag{2.32}$$

where d is the number of unknown model parameters and $L(\hat{\boldsymbol{\phi}}|\mathbf{X})$ is the likelihood function of the entire dataset. If the number of parameters d is large relative to the sample size n, a small-sample version (AIC_c; Hurvich and Tsai, 1989) is given below

$$AIC_c = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + \frac{2dn}{n-d-1} \tag{2.33}$$

The Bayesian information criterion (BIC; Schwarz, 1978) picks the model with the lowest

$$BIC = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + d \log N \tag{2.34}$$

where N is the number of observations in the dataset. Cavanaugh (1996) has developed an asymptotic unbiased estimator for the Kullback information criterion (KIC) given by

$$KIC = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + 3(d+1) \tag{2.35}$$

The bias correction of KIC (KIC_c; Seghouane and Maiza, 2004) is expressed as

$$KIC_c = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + \frac{2(d+1n)}{n-d-2} - n\psi\left(\frac{n-d}{2}\right) + n \log\left(\frac{n}{2}\right) \tag{2.36}$$

where $\psi(.)$ is the digamma or psi function. The approximation of the KIC ($AKIC_c$;

Seghouane et. al., 2005) is given by

$$AKIC_c = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + \frac{(d+1)(3n-d-2)}{n-d-2} + \frac{d}{n-d} \tag{2.37}$$

The large sample approximation of the Integrated Classification Likelihood (ICL-BIC; Biernacki et al, 1998) selects the model with the minimum

$$ICL - BIC = -2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + 2EN(\hat{\boldsymbol{\tau}}) + 3(d+1) \tag{2.38}$$

where $\hat{\boldsymbol{\tau}}$ is a $N \times H$ matrix in which $\hat{\boldsymbol{\tau}} = (\tau_{ih})$ and the entropy term $EN(\hat{\boldsymbol{\tau}})$ is defined as

$$EN(\hat{\boldsymbol{\tau}}) = -\sum_{i=1}^{N} \sum_{h=1}^{H} \tau_{ih} \log() \tau_{ih} \tag{2.39}$$

The entropy term $EN(\hat{\boldsymbol{\tau}})$ assesses how well the observations were correctly categorized into groups: the lower the entropy, the better the classification of observations (Celeux and Soromenho, 1996).

The Classification Likelihood Criterion (CLC; Biernacki and Govaert, 1997) is given by

$$CLC = 2 \log L(\hat{\boldsymbol{\phi}}|\mathbf{X}) + 2EN(\hat{\boldsymbol{\tau}}) \tag{2.40}$$

The model with the lowest CLC is considered to be the best model.

The Approximate Weight of Evidence (AWE; Banfield and Raftery, 1993) is expressed as

$$AWE = -2 \log L_c + 2d \left(\frac{3}{2} + \log(n) \right) \tag{2.41}$$

where $\log L_c = logL(\hat{\boldsymbol{\phi}}|\mathbf{X}) + EN(\hat{\boldsymbol{\tau}})$. $\log L_c$ is based on the Hathaway (1986) mixture logarithmic likelihood. The model with the lowest AWE score can be considered to be the best model.

The entropy function $EN(\hat{\boldsymbol{\tau}})$ cannot be used to get the number components in a mixture model, since it automatically increases when the number of component increases, which is why it was normalized by Biernacki et. al. (1999) and created the Normalized Entropy Criterion (NEC) shown below

$$NEC_g = \frac{EN(\hat{\boldsymbol{\tau}})}{logL(\hat{\boldsymbol{\phi}}|\mathbf{X}) - logL(\hat{\boldsymbol{\phi}}^*|\mathbf{X})} \tag{2.42}$$

where $\hat{\boldsymbol{\phi}}^*$ is the likelihood estimator for $\boldsymbol{\phi}$ assuming that there are only one subgroup in the population. The NEC is used to know the number of clusters in the subgroup, assuming that the number of clusters is greater than or equal to 2.

2.3.2 Performance of Model Fit Criteria on CDM and Mixture Models

According to many authors, AIC tends to overfit models and overestimate the number of components in mixture modelling (Soromenho, 1993; Celeux and Soromenho, 1996).

Roeder and Wasserman (1997), Campbell et. al. (1997), and Dasgupta and Raftery (1998) have reported from their simulation studies that BIC performs well in determining the number of components in a normal mixture model wherein the density is estimated nonparametrically. Meanwhile, Celeux and Soromenho (1996) stated that BIC tends to underestimate the number of subgroups in the population if the sample size is small, and the component densities are estimated well. Meanwhile, if the component densities is not valid, Biernacki et. al. (1998) have stated that BIC tends to overestimate the number of components, if the component densities were not estimated well.

Biernacki, Celeux, and Govaert (1999) stated that CLC works well in estimating the number of components in the population when the mixing proportions π_h's are equal, but it tends to overestimate the number of components when there are no restrictions for the mixing proportions.

According to Biernacki et. al. (1998), the $ICL - BIC$ performs as well as the ICL when the sample size is large. Based on the simulation studies on three conditions conducted by McLachlan and Peel (2001), $ICL-BIC$ performed better than AIC, BIC, and CLC. In another simulation conducted by Akogul and Erisoglu (2016), the KIC performed better than AIC, AIC_c, $AKIC_c$, AWE, BIC, CLC, KIC_c, and NEC in determining the number of components in the population. However, these simulations were only restricted to multivariate normal distributions, therefore the criteria might behave differently when used in the Mixture G-DINA model.

For the case of the G-DINA model, the relative fit criteria AIC and BIC were investigated in determining the best fitting CDM by Chen et. al. (2013). Based on their simulation study, AIC and BIC performed well in determining model misspecifications, but BIC is shown to be superior than AIC.

Chapter 3

Methodology

This study will introduce the Mixture G-DINA model, an extension of the G-DINA model by incorporating it with finite mixture modelling framework. The Mixture G-DINA model is hypothesized by the researcher to be more flexible than the usual G-DINA model when there are inherent subgroups in the population that affect the probabilities of correct response. This chapter provides the methodology for this proposed study. In Section 3.1., the proposed statistical model and its parameters are laid out. In Section 3.2., the maximum likelihood estimator was derived, and the steps for the proposed estimation procedure that utilizes EM Algorithm is stated. In Sections 3.3. and 3.4., the design for the simulation study and the planned real-data application of the Mixture G-DINA model is expressed.

3.1 The Proposed Model

Let K be the number of attributes under consideration for the entire examination, K_j be the number of required attributes for item j, $P_{h\boldsymbol{\alpha}^*_{lj}}$ be the probability that the examinee with $\boldsymbol{\alpha}^*_{lj}$ reduced attribute vector will answer the question j correctly if the examinee is in subgroup h, and π_h (oftentimes called the mixing proportion) be the proportion of examinees in subgroup h, $h = 1, 2, ..., H$. The reduced attribute vector $\boldsymbol{\alpha}_{lj}^* = \left(\alpha_{lj1}, \alpha_{lj2}, ..., \alpha_{ljK_j^*}\right)$ will be used in place of the full attribute vector $\boldsymbol{\alpha}_{lj} = (\alpha_{lj1}, \alpha_{lj2}, ..., \alpha_{ljK})$. The information from the Q-matrix is already summarized in the reduced attribute vector. The interpretations of the model parameters are based on de la Torre (2011). Using the identity link, the item response function of the Mixture G-DINA model is given by

$$P_{h\boldsymbol{\alpha}_{lj}^*} = \delta_{j0h} + \sum_{k=1}^{K_j^*} \delta_{jkh}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'h}\alpha_{lk}\alpha_{lk'} + ... + \delta_{j12...K_j^*h} \prod_{k=1}^{K_j^*} \alpha_{lk}, \tag{3.1}$$

where δ_{j0h} is the probability of correct response, if the examinee in group h does not master any attribute required for item j, δ_{jkh} is the change in probability of a correct response due to the mastery of α_k, if the examinee is in group h, $\delta_{jkk'h}$ is the change in the probability of a correct response due to the mastery of α_k and α_k' that is over and above the additive impact of the mastery of the same two attributes, if the examinee is in group h, $\delta_{j12...K_j^*h}$ is the interaction effect due to $\alpha_1, \alpha_2, ..., \alpha_{K_j^*}$, or the change in the probability of the correct response due to the mastery of all the attributes that is over and above the additive impact of the main and lower-order interaction effects, if the examinee is in group h. Note that if $H = 1$, then the Mixture G-DINA model reduces to the usual G-DINA model.

3.2 Estimation Procedure

Let $L(\mathbf{X})$ be the likelihood of the entire response dataset $\mathbf{X}$, a matrix of dimension $N \times J$, where N is the number of examinees, and J is the number of items in the exam. Define

$L(\mathbf{X})$ to be

$$L(\mathbf{X}) = \prod_{i=1}^{N} L(\mathbf{X}_i) \tag{3.2}$$

under the assumption that the response vectors are independent, where $L(\mathbf{X}_i)$ is the likelihood function for the response vector of examinee i with dimension $1 \times J$. In finite mixture modelling, $L(\mathbf{X}_i)$ can be expressed to be equal to

$$L(\mathbf{X}_i) = \sum_{h=1}^{H} \pi_h L(\mathbf{X}_i|h) \tag{3.3}$$

where $L(\mathbf{X}_i|h)$ is the likelihood function for the examinee i assuming that it is in group h, and π_h is the probability that an examinee is in group h, $h = 1, 2, ..., H$. $L(\mathbf{X}_i|h)$ can be further defined to be equal to

$$L(\mathbf{X}_i|h) = \sum_{l=1}^{L} L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)P(\boldsymbol{\alpha}_l|h), \tag{3.4}$$

where $P(\boldsymbol{\alpha}_l|h)$ is the probability that an examinee has skill vector $\boldsymbol{\alpha}_l$ is in group h, and $L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)$ is the likelihood function for examinee i assuming it is in group h and has skill vector $\boldsymbol{\alpha}_l$. Assuming that the responses are dichotomous in nature, $L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)$ can be expressed as

$$L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h) = \prod_{j=1}^{J} P_{h\boldsymbol{\alpha}^*_{lj}}^{X_{ij}} \left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)^{1-X_{ij}}, \tag{3.5}$$

where $P_{h\boldsymbol{\alpha}^*_{lj}}$ is the probability that examinee i answer the item j correctly assuming that examinee i is in group h and X_{ij} is the response of examinee i on item j. To obtain the MLE for $P_{hj} = P_{h\boldsymbol{\alpha}^*_{lj}}$ for $lj = 1, 2, ..., 2^{K_j}$ we need to maximize

$$l(\mathbf{X}) = \log L(\mathbf{X}) = \log \prod_{i=1}^{N} L(\mathbf{X}_i) = \sum_{i=1}^{N} \log L(\mathbf{X}_i)$$

Taking its derivative with respect to P_{hj},

$$\frac{\partial l(\mathbf{X})}{\partial P_{hj}} = \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \frac{\partial L(\mathbf{X})}{\partial P_{hj}} = \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \sum_{h=1}^{H} \sum_{l=1}^{L} \pi_h P(\boldsymbol{\alpha}_l|h) \frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)}{\partial P_{hj}}$$

Now, getting the derivative in the right hand side, we have

$$\frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)}{\partial P_{hj}} = L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}}$$

Therefore,

$$\begin{aligned}
\frac{\partial l(\mathbf{X})}{\partial P_{hj}} &= \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \sum_{h=1}^{H} \sum_{l=1}^{L} \pi_h P(\boldsymbol{\alpha}_l|h) L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}} \\
&= \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{\pi_h L(\mathbf{X}_i|h)}{L(\mathbf{X}_i|h)} P(\boldsymbol{\alpha}_l|h) L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}} \\
&= \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \sum_{h=1}^{H} \sum_{l=1}^{L} \pi_h L(\mathbf{X}_i|h) \frac{P(\boldsymbol{\alpha}_l|h) L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)}{L(\mathbf{X}_i|h)} \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}} \\
&= \sum_{i=1}^{N} \frac{1}{L(\mathbf{X}_i)} \sum_{h=1}^{H} \sum_{l=1}^{L} \pi_h L(\mathbf{X}_i|h) P(\boldsymbol{\alpha}_l|\mathbf{X}_i, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}}
\end{aligned}$$

where $P(\boldsymbol{\alpha}_l|\mathbf{X}_i, h) = \dfrac{P(\boldsymbol{\alpha}_l|h) L(\mathbf{X}_i|\boldsymbol{\alpha}_l, h)}{L(\mathbf{X}_i|h)}$ is the posterior probability that examinee i is in group h. Continuiung the manipulations,

$$\begin{aligned}
\frac{\partial l(\mathbf{X})}{\partial P_{hj}} &= \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{\pi_h L(\mathbf{X}_i|h)}{L(\mathbf{X}_i)} P(\boldsymbol{\alpha}_l|\mathbf{X}_i, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}} \\
&= \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{l=1}^{L} \tau_{ih} P(\boldsymbol{\alpha}_l|\mathbf{X}_i, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}_{lj}^*}}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] \frac{\partial P_{h\boldsymbol{\alpha}_{lj}^*}}{\partial P_{hj}}
\end{aligned}$$

where $\tau_{ih} = \dfrac{\pi_h L(\mathbf{X}_i|h)}{L(\mathbf{X}_i)}$ is the posterior probability that examinee i is in group h given the response pattern $\mathbf{X}_i$. We know that for a subset of attribute patterns, $\boldsymbol{\alpha}_l$ can be reduced to $\boldsymbol{\alpha}^*_{lj}$ (that is, $\boldsymbol{\alpha}_l \rightarrow \boldsymbol{\alpha}^*_{lj}$). Also, we have the following expression

$$P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h) = \sum_{\forall l \in \boldsymbol{\alpha}_l \rightarrow \boldsymbol{\alpha}^*_{lj}} P(\boldsymbol{\alpha}_l|\mathbf{X}_i, h)$$

This will lead to simplification that

$$\frac{\partial l(\mathbf{X})}{\partial P_{hj}} = \sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{l=1}^{L_j} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h) \left[\frac{X_{ij} - P_{h\boldsymbol{\alpha}^*_{lj}}}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}}$$

where $L_j = 2^{K_j}$. This reduces the number of inside of terms from 2^K to 2^{K_j}. By rearranging the terms we have,

$$\begin{aligned}
\frac{\partial l(\mathbf{X})}{\partial P_{hj}} &= \sum_{h=1}^{H}\sum_{lj=1}^{L_j} \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}} \left[\frac{1}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] \sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)[X_{ij} - P_{h\boldsymbol{\alpha}^*_{lj}}] \\
&= \sum_{h=1}^{H}\sum_{lj=1}^{L_j} \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}} \left[\frac{1}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] \sum_{i=1}^{N} \left[\tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)X_{ij} - \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)P_{h\boldsymbol{\alpha}^*_{lj}}\right] \\
&= \sum_{h=1}^{H}\sum_{lj=1}^{L_j} \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}} \left[\frac{1}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] \left[\sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)X_{ij} - \sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)P_{h\boldsymbol{\alpha}^*_{lj}}\right] \\
&= \sum_{h=1}^{H}\sum_{lj=1}^{L_j} \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}} \left[\frac{1}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] \left[\sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)X_{ij} - P_{h\boldsymbol{\alpha}^*_{lj}} \sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)\right] \\
&= \sum_{h=1}^{H}\sum_{lj=1}^{L_j} \frac{\partial P_{h\boldsymbol{\alpha}^*_{lj}}}{\partial P_{hj}} \left[\frac{1}{P_{h\boldsymbol{\alpha}^*_{lj}}\left(1 - P_{h\boldsymbol{\alpha}^*_{lj}}\right)}\right] [R^*_{hlj} - P_{h\boldsymbol{\alpha}^*_{lj}N^*_{hlj}}]
\end{aligned}$$

where $N^*_{hlj} = \sum_{i=1}^{N} \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)$ is the expected number of examinees with reduced attribute pattern $\boldsymbol{\alpha}^*_{lj}$ in group h, and $R^*_{hlj} = \sum_{i=1}^{N} X_i j \tau_{ih} P(\boldsymbol{\alpha}^*_{lj}|\mathbf{X}_i, h)$ is the expected number of examinees with reduced attribute pattern $\boldsymbol{\alpha}^*_{lj}$ in group h that answered item j correctly. By expanding the summation, the derivative can be written as

$$\frac{\partial l(\mathbf{X})}{\partial P_{hj}} = \frac{\partial P_{1\boldsymbol{\alpha}_1^*}}{\partial P_{hj}} \left[\frac{1}{P_{1\boldsymbol{\alpha}_1^*}\left(1 - P_{1\boldsymbol{\alpha}_1^*}\right)} \right] [R_{11}^* - P_{1\boldsymbol{\alpha}_1^*} N_{11}^*] + \frac{\partial P_{1\boldsymbol{\alpha}_2^*}}{\partial P_{hj}} \left[\frac{1}{P_{1\boldsymbol{\alpha}_2^*}\left(1 - P_{1\boldsymbol{\alpha}_2^*}\right)} \right] [R_{12}^* - P_{1\boldsymbol{\alpha}_2^*} N_{12}^*]$$

$$+ ... + \frac{\partial P_{1\boldsymbol{\alpha}_{L_j}^*}}{\partial P_{hj}} \left[\frac{1}{P_{1\boldsymbol{\alpha}_{L_j}^*}\left(1 - P_{1\boldsymbol{\alpha}_{L_j}^*}\right)} \right] [R_{1L_j}^* - P_{1\boldsymbol{\alpha}_{L_j}^*} N_{1L_j}^*] + ... + \frac{\partial P_{H\boldsymbol{\alpha}_1^*}}{\partial P_{hj}} \left[\frac{1}{P_{H\boldsymbol{\alpha}_1^*}\left(1 - P_{H\boldsymbol{\alpha}_1^*}\right)} \right] [R_{H1}^* - P_{H\boldsymbol{\alpha}_1^*} N_{H1}^*]$$

$$+ \frac{\partial P_{H\boldsymbol{\alpha}_1^*}}{\partial P_{hj}} \left[\frac{1}{P_{H\boldsymbol{\alpha}_2^*}\left(1 - P_{H\boldsymbol{\alpha}_2^*}\right)} \right] [R_{H2}^* - P_{H\boldsymbol{\alpha}_2^*} N_{H2}^*] + ... + \frac{\partial P_{H\boldsymbol{\alpha}_{L_j}^*}}{\partial P_{hj}} \left[\frac{1}{P_{H\boldsymbol{\alpha}_{L_j}^*}\left(1 - P_{H\boldsymbol{\alpha}_{L_j}^*}\right)} \right] [R_{HL_j}^* - P_{H\boldsymbol{\alpha}_{L_j}^*} N_{HL_j}^*]$$

For the lj^{th} derivative,

$$\frac{\partial P_{h\boldsymbol{\alpha}_{lj'}^*}}{\partial P_{h\boldsymbol{\alpha}_{lj}^*}} = \begin{cases} 0 & \text{if } j' \neq j \\ 1 & \text{if } j' = j \end{cases}$$

Therefore, for $P_{hj} = P_{h\boldsymbol{\alpha}_{lj}^*}$, the derivative for the entire log likelihood can be written as

$$\frac{\partial l(\mathbf{X})}{\partial P_{hj}} = \left[\frac{1}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] [R_{hlj}^* - P_{h\boldsymbol{\alpha}_{lj}^*} N_{hlj}^*].$$

Equating the expression above to 0,

$$\frac{\partial l(\mathbf{X})}{\partial P_{hj}} = \left[\frac{1}{P_{h\boldsymbol{\alpha}_{lj}^*}\left(1 - P_{h\boldsymbol{\alpha}_{lj}^*}\right)} \right] [R_{hlj}^* - P_{h\boldsymbol{\alpha}_{lj}^*} N_{hlj}^*] = 0$$

$$\rightarrow [R_{hlj}^* - P_{h\boldsymbol{\alpha}_{lj}^*} N_{hlj}^*] = 0$$

$$\rightarrow \hat{P}_{h\boldsymbol{\alpha}_{lj}^*} = \frac{R_{hlj}^*}{N_{hlj}^*}$$

Below are the steps of the EM Algorithm in estimating the parameters for the Mixture G-DINA model.

1. At iteration 0, assign initial values $P_{1\boldsymbol{\alpha}_1^*}^{(0)}, P_{2\boldsymbol{\alpha}_1^*}^{(0)}, ..., P_{H\boldsymbol{\alpha}_1^*}^{(0)}, ..., P_{1\boldsymbol{\alpha}_{L_j}^*}^{(0)}, P_{2\boldsymbol{\alpha}_{L_j}^*}^{(0)}, ..., P_{H\boldsymbol{\alpha}_{L_j}^*}^{(0)}$,

$\tau_{11}^{(0)}, \tau_{12}^{(0)}, ..., \tau_{1H}^{(0)}, ..., \tau_{N1}^{(0)}, \tau_{N2}^{(0)}, ..., \tau_{NH}^{(0)}$.

2. Compute for $\pi_1^{(0)}, \pi_2^{(0)}, ..., \pi_H^{(0)}$ using the $\tau_{11}^{(0)}, \tau_{12}^{(0)}, ..., \tau_{1H}^{(0)}, ..., \tau_{N1}^{(0)}, \tau_{N2}^{(0)}, ..., \tau_{NH}^{(0)}$ where $\pi_h^{(0)} = \sum_{i=1}^N \tau_{ih}^{(0}, h = 1, 2, ..., H$.

3. At iteration t, compute for $({R_{11}^*}^{(t)}, {N_{11}^*}^{(t)})$, $({R_{12}^*}^{(t)}, {N_{12}^*}^{(t)}), ...,$ $({R_{1H}^*}^{(t)}, {N_{1H}^*}^{(t)}), ...,$ $({R_{L_j1}^*}^{(t)}, {N_{L_j1}^*}^{(t)})$, $({R_{L_j2}^*}^{(t)}, {N_{L_j2}^*}^{(t)}), ...,$ $({R_{L_jH}^*}^{(t)}, {N_{L_jH}^*}^{(t)})$, using $P_{1\boldsymbol{\alpha}_1^*}^{(t-1)}$, $P_{2\boldsymbol{\alpha}_1^*}^{(t-1)}$, ..., $P_{H\boldsymbol{\alpha}_1^*}^{(t-1)}$, ..., $P_{1\boldsymbol{\alpha}_{L_j}^*}^{(t-1)}$, $P_{2\boldsymbol{\alpha}_{L_j}^*}^{(t-1)}$,..., $P_{H\boldsymbol{\alpha}_{L_j}^*}^{(t-1)}$ and $\tau_{11}^{(t-1)}$, $\tau_{21}^{(t-1)}$, ..., $\tau_{H1}^{(t-1)}$, ..., $\tau_{N1}^{(t-1)}$, $\tau_{N2}^{(t-1)}$, ..., $\tau_{NH1}^{(t-1)}$ where

$$ {N_{hlj}^*}^{(t)} = \sum_{i=1}^N P^{(t-1)}(\boldsymbol{\alpha}_{lj}^* | \mathbf{X}_i, h)\tau_{ih}^{(t-1)} $$

$$ {R_{hlj}^*}^{(t)} = \sum_{i=1}^N X_{ij} P^{(t-1)}(\boldsymbol{\alpha}_{lj}^* | \mathbf{X}_i, h)\tau_{ih}^{(t-1)} $$

Update $P_{1\boldsymbol{\alpha}_1^*}^{(t)}$, $P_{2\boldsymbol{\alpha}_1^*}^{(t)}$, ..., $P_{H\boldsymbol{\alpha}_1^*}^{(t)}$, ..., $P_{1\boldsymbol{\alpha}_{L_j}^*}^{(t)}$, $P_{2\boldsymbol{\alpha}_{L_j}^*}^{(t)}$, ..., $P_{H\boldsymbol{\alpha}_{L_j}^*}^{(t)}$ using

$$ P_{h\boldsymbol{\alpha}_{lj}^*}^{(t)} = \frac{{R_{hlj}^*}^{(t)}}{{N_{hlj}^*}^{(t)}}. $$

4. At iteration t, update $\tau_{11}^{(t)}$, $\tau_{12}^{(t)}$, ..., $\tau_{1H}^{(t)}$, ..., $\tau_{N1}^{(t)}$, $\tau_{N2}^{(t)}$, ..., $\tau_{NH}^{(t)}$ using

$$ \tau_{ih}^{(t)} = \frac{\pi_h^{(t-1)} L^{(t-1)}(\mathbf{X}_i | h)}{\sum_{h=1}^H \pi_h^{(t-1)} L^{(t-1)}(\mathbf{X}_i | h)}. $$

At the same time, compute for $\pi_1^{(t)}, \pi_2^{(t)}, ..., \pi_H^{(t)}$ using the $\tau_{11}^{(t)}$, $\tau_{12}^{(t)}$, ..., $\tau_{1H}^{(t)}$, ..., $\tau_{N1}^{(t)}$, $\tau_{N2}^{(t)}$, ...,$\tau_{NH}^{(t)}$ where $\pi_h^{(t)} = \sum_{i=1}^N \tau_{ih}^{(t)}, h = 1, 2, ..., H$ (Mclachlan and Basford, 1988).

5. Repeat steps 3 and 4 until convergence.

The algorithm will stop iterating when the changes in the model parameters $P_{1\boldsymbol{\alpha}_1^*}, P_{2\boldsymbol{\alpha}_1^*}, ..., P_{H\boldsymbol{\alpha}_1^*}, ..., P_{1\boldsymbol{\alpha}_{L_j}^*}, P_{2\boldsymbol{\alpha}_{L_j}^*}, ..., P_{H\boldsymbol{\alpha}_{L_j}^*}$, and $\tau_{11}, \tau_{12}, ..., \tau_{1H}, ..., \tau_{N1}, \tau_{N2}, ..., \tau_{NH}$ do not exceed 0.0001.

3.3 Simulation Design

In order to answer objectives 1, 2, 3, and 4, two simulation studies will be conducted. Simulation study 1 aims to answer objectives 1, 2, and 3 which is all about the viability of the Mixture G-DINA model in fitting heterogenous response data. Simulation study 2 will answer objective 4, which is about the estimation of the number of components in the population using relative model fit criteria.

3.3.1 Simulation Study 1: Parameter Recovery

The factors under considerations in the proposed simulation study 1 are test length, number of examinees, the mixing proportions, item quality, different generating models, and the number of groups. The factor levels for each of the factors are given in Table 3.1.

Table 3.1: **Factor levels for the simulation study 1**

Factors	**Factor Levels**
Number of items	$J = 15$ and $J = 30$
Number of examinees	$N = 1500$ and $N = 3000$
Number of groups	$H = 2$ and $H = 3$
Mixing proportions	Even 0.5:0.5 for two groups 1/3:1/3:1/3 for three groups Uneven 0.2:0.8 for two groups 0.2:0.3:0.5 for three groups
Item quality	Combinations of: Good item quality $(g, s \in unif[0.08, 0.12])$ Moderate item quality $(g, s \in unif[0.18, 0.22])$ Bad item quality $(g, s \in unif[0.28, 0.32])$
Generating model	Combinations of: G-DINA, DINA, DINO, ACDM, LLM, and RRUM

For the number of items, we will consider $J = 15$ and $J = 30$. For the number of examinees, we will study the case wherein $N = 1500$ and $N = 3000$. For the number of groups, we will consider $H = 2$ and $H = 3$. The performance of the Mixture G-DINA model will also be compared across the different partitioning of the mixing proportions: even and uneven. For two groups with even partitioning, 50% will be generated from group 1 and 2. For two groups with uneven partitioning, 20% will be generated from group 1, and 80% will be generated from group 2. For three groups with even partitioning, 1/3 will be generated from groups 1, 2 and 3. For uneven partitioning of the three groups, 20% will be generated from group 1, 30% will be generated from group 2, and 50% will be generated from group 3. For item quality, we will consider combinations of good, moderate, bad, and mixed item quality for a single test. Items generated to have guessing and slip parameters equal to 0.1, 0.2, and 0.3 are considered to be items with good, moderate, and bad quality respectively. Lastly, the generating model will also be manipulated for the simulation design. Combinations of different generating models: G-DINA, DINA, DINO, ACDM, LLM, and RRUM will be considered in the study.

The number of attributes in consideration is fixed at $K = 5$. The Q-matrix used in the study is from de la Torre (2011). The Q-matrix for $J = 30$ is given in Table 3.4. For $J = 15$, the Q-matrix is the rows 1 to 15 of the Table 3.2. For every treatment combinations, 100 replicated data will be generated. As stated, this study is only limited to analyzing dichotomous response and dichotomous attributes.

To answer objective 1: verifying whether the Mixture G-DINA model recaptures the parameters where the responses where generated, the bias of the estimates will be measured across different conditions. The bias of the estimates is given by the following formula:

$$BIAS = \sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{lj}^{L_j} \frac{|P_{h\boldsymbol{\alpha}_{lj}^*} - \hat{P}_{h\boldsymbol{\alpha}_{lj}^*}|}{\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{lj}^{L_j} 1} \tag{3.6}$$

where $P_{h\boldsymbol{\alpha}_{lj}^*}$ is the generated success probability for item j for examinees with $\boldsymbol{\alpha}_{lj}^*$ attribute vector and $\hat{P}_{h\boldsymbol{\alpha}_{lj}^*}$ is the estimated value of $P_{h\boldsymbol{\alpha}_{lj}^*}$.

For objectives 2 and 3: knowing the viability of Mixture G-DINA model in classifying the examinees correctly into the latent classes and the inherent subgroups in the population, the correct latent class classification rates, both vector-wise (CCL_{vec}) and element-wise (CCL_{el}), and the correct subgroup classification rates were measured (CCG). All of these measures are given below:

$$CCG = \sum_{i=1}^{N} \frac{I_{i[h=\hat{h}]}}{N} \tag{3.7}$$

$$CCL_{vec} = \sum_{i=1}^{N} \frac{I_{i[\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}]}}{N} \tag{3.8}$$

$$CCL_{el} = \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{I_{i[\alpha_k=\hat{\alpha}_k]}}{K \times N} \tag{3.9}$$

where $I_{i[h=\hat{h}]} = 1$ if the estimated grouping of examinee i, denoted by$\hat{h}$ is equal to h, 0 otherwise, $I_{i[\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}]} = 1$ if the estimated attribute vector $\hat{\boldsymbol{\alpha}}$ is equal to the actual attribute vector $\boldsymbol{\alpha}$ of examinee i, 0 otherwise, and $I_{i[\alpha_k=\hat{\alpha}_k]} = 1$ if the estimated attribute k $\hat{\alpha}_k$ is equal to the real attribute k α_k, 0 otherwise.

The correct attribute pattern rate of the Mixture G-DINA will be compared against the case wherein we only used the usual G-DINA model, to know whether it improves the correct attribute classification. For the estimation of the attribute vector, expected a posteriori, modal a posteriori, and the maximum likelihood estimation will be considered (Huebner and Wang, 2014).

3.3.2 Simulation Study 2: Estimation of the number of components

For simulation study 2, some of the factors are the same with the simulation study 1. The factors that will be manipulated are: number of items, number of examinees, number of generating groups, number of assumed groups, mixing proportions, item quality and generating model. The factors manipulated are given in table 3.2.

Table 3.2: **Factor levels for the simulation study 1**

Factors	**Factor Levels**
Number of items	$J = 15$ and $J = 30$
Number of examinees	$N = 1500$ and $N = 3000$
Number of generating groups	$H = 1$, $H = 2$ and $H = 3$
Number of assumed groups	$H = 1$, $H = 2$, $H = 3$, and $H = 4$
Mixing proportions	Even 0.5:0.5 for two groups 1/3:1/3:1/3 for three groups Uneven 0.2:0.8 for two groups 0.2:0.3:0.5 for three groups
Item quality	Combinations of: Good item quality ($g, s \in unif[0.08, 0.12]$) Moderate item quality ($g, s \in unif[0.18, 0.22]$) Bad item quality ($g, s \in unif[0.28, 0.32]$)
Generating model	Combinations of: G-DINA, DINA, DINO, ACDM, LLM, and RRUM

The factor levels for simulation study 2 is the same with simulation study 1 for the following factors: number of items, number or examinees, mising proportions, item quality and generating model. For the number of generated groups, the researcher will consider the case wherein $H = 1$ (or the population is not heterogenous), $H = 2$, and $H = 3$. In answering objective 4: testing the capabilities of the different model fit criteria in estimating the number of components in the population, the different generated responses will be fitted assuming that the number of groups are equal to $H = 1$, $H = 2$, $H = 3$, and $H = 4$. The correct selection rates of the criteria will be measured. These model fit criteria are AIC, BIC, KIC, AKIC, ICL-BIC, CLC, AWE, and NEC. The formula for the correct selection rate CSR is given below.

$$CSR = \sum_{m=1}^{M} \frac{I_{m[H=\hat{H}]}}{M} \tag{3.10}$$

where $I_{m[H=\hat{H}]} = 1$ if the estimated number of components in the population for replicate m, $\hat{H}$ is equal to H, 0 otherwise, and M is the number of data replicates.

The same as simulation study 1, the number of attributes are fixed at $K = 5$, and the number of replicates is fixed at $M = 100$. The same Q-matrix for $J = 15$ and $J = 30$ will be used for simulation study 2.

3.4 Real Data Application

To demonstrate the applicability of the Mixture G-DINA model, responses of middle schools students in the United States and Hong Kong to a 31-item proportional reasoning assessment (Tjoe and de la Torre, 2013) will be employed. Both the traditional and mixture G-DINA models will be fitted to the data and model fit indices will be compared.

Table 3.3 shows the final attributes for the proportional reasoning assessment from Tjoe and de la Torre (2013). There are six main attributes identified, and attributes A2 and A3 are further subdivided to two, bringing a total of eight attributes. Table 3.5 shows the Q-matrix for the 31-item proportional reasoning exam devised by Tjoe and de la Torre (2013).

Table 3.3: **List of proportional reasoning attributes**

Attribute	**Description**
A1	Prerequisite skills and concepts required in proportional reasoning
A2a	Ordering fractions
A2b	Comparing fractions
A3a	Constructing ratios
A3b	Constructing proportions
A4	Identifying a multiplicative relationship between sets of values
A5	Differentiating a proportional relationship from a non-proportional relationship
A6	Applying algorithms in solving proportional reasoning problems

Table 3.4: **Simulation Study Q-Matrix**

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

Table 3.5: **Q-Matrix for proportional reasoning assessment.** The HK Q-matrix is shown in baseline entries and those with * show entries in the US Q-matrix

Item	**Attribute**							
	A1	A2a	A2b	A3a	A3b	A4	A5	A6
1	1	0	0	0	0*	0	0	1
2	1*	1	0	1	0	0	0	0
3	1	0	0	0	1	1	1	0
4	1	1	0	1	0	0	0	0
5	1	1	0	1	0	0	0	0
6	1*	0	0	0	1	1*	1*	0
7	1	0	0	0	0	0	0	1
8	1*	0	1	1	0	0	0	0
9	1*	0	0	0	0	1	0	0
10	0	0	1	0	0	0	0	0
11	1	0	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0
13	1	0	0	0	1*	1*	1	0
14	1	1	0	1	0	0	0	0
15	0	0	0	0	1	0	0	0
16	0	0	0	0	1	0	0	0
17	1	0	1	0	0	0	0	0
18	1	0	0	0	1	0	1	1
19	1	0	0	0	1	1	1	0
20	0	0	0	0	1	1	1	0
21	1	1*	0	1	0	0	0	0
22	1	0	1	0	0	0	0	0
23	1	1*	0	1	0	0	0	0
24	1	0	0	0	1	0	1	1
25	1	1	0	1	0	0	0	0
26	0	1	0	1	0	0	0	0
27	1	0	0	0	0	1	1	0
28	1	0	0	0	1	0	1	1
29	1	0	0	0	1	1	1	0
30	1*	0	0	0	0	1	0	1*
31	1	0	1	0	0	0	0	0

Chapter 4

Bibliography

Aitkin, M. and Tunnicliffe Wilson, G. (1980). Mixture models, outliers, and the EM Algorithm. *Technometrics, 22,* 325-331.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6),* 716-723.

Akogul, S. and Murat, E. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications 21,* 34; doi:10.3390/mca21030034

Banfield, J.D.; Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics 49,* 803–821.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Reprinted in *Biometrika 45,* 1958, 296–315.

Biernacki, C. Celeux, G., and Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. *Technical Report No. 3521.* Rhone-Alpes: INRIA.

Biernacki, C.; Celeux, C.; Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognit. Lett., 20,* 267–272.

Biernacki, C.; Govaert, G. (1997) Using the classification likelihood to choose the number of clusters. *Comput. Sci. Stat., 29,* 451–457.

Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331–348.

Campbell, J.G., Fraley, C., Murtagh, F., Raftery, A.E. (1997), Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters, 18,* 1539-1548.

Cavanaugh, J.E., (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters, 44,* 333-344.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195–212

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in

cognitive diagnosis modeling. *Journal of Educational Measurement, 50,* 123-140.
Cho, S.-J., and Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336–370.
Cho, S.-J., Cohen, A. S., Kim, S.-H., and Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement, 34,* 483–504.
Cohen, A. S., and Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42,* 133–148.
Dasgupta, A. and Raftery, A.E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association, 93,* 294-302.
de la Torre, J., and Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50,* 355-373.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.
de la Torre, J., and Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicologia Educativa, 20(2),* 89–97.
de la Torre, J., van der Ark, L., and Rossi, G. (2015). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development.* DOI: 10.1177/0748175615569110
DiBello, L., Roussos, L., and Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics (pp. 979-1030).* Amsterdam, Netherlands: Elsevier.
Früwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* New York: Springer.
Garcia, P., Olea, J., and de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psichothema, 26,* 372-377.
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333-352.
Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.
Hathaway, R.J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Stat. Probab. Lett., 4,* 53–56.
Henson, R. A., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74,* 191-210.
Huebner, A., and Wang, C. (2014). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71(2),* 407-419.
Huo, L., de la Torre J., and Ratna, N. (2014). Differential item functioning assessment in cognitive diagnosis modeling: application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51(1),* 98-125.
Hurvich, C.M.; Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika, 76,* 297–307.
Jansenn G., Meier V., and Trace J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Columbian Applied Linguistics Journal, 16(2),* 167-184.

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

Mislevy, R. J., and Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

McLachlan, G., and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Deckker, Inc.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley & Sons, Inc.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. In J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.), *Board on Testing and Assessment, Center for Education.* Washington, DC: National Academy Press.

Ravand, H. (2015). Application of a Cognitive Diagnostic Model to a High-Stakes Reading Comprehension Test. *Journal of Psychoeducational Assessment,* 1-18.

Roeder, K. and Wasserman, L. (1997). Practical density estimation using mixtures of normals. *Journal of the American Statistical Association, 92,* 894-902.

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

Seghouane, A-K.; Bekara, M.; Fleury, G. (2005). A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal Process, 85,* 1405–1417.

Seghouane, A.-K.; Maiza, B. (2004). A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE Trans. Signal Process, 52,* 3314–3323.

Song, L., Dai, H., Wang, W., and Ding, S. (2012). Assessing the performance of EM algorithm for parameter estimation of the R-DINA model. *Psychological Exploration, 32,* 410-413.

Soromenho, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics, 9,* 65-78.

Sorrel, M., Olea, J., Abad, F., de la Torre, J., Aguado, D., and Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods. doi: 10.1177/1094428116630065*

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

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

Tjoe, H. and de la Torre J. (2013). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 19(2),* 237-255.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* (Educational Testing Service Reasearch Report No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2007). *Mixture Distribution Diagnostic Models.* (Educational Testing Service Reasearch Report No. RR-07-32). Princeton, NJ: ETS.

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. British *Journal of Mathematical and Statistical Psychology, 68,* 456–477.

Zhang, P. (2020). *Cognitive diagnostic modeling for adaptively identifying coexisting condensation rules.* ResearchGate.