Application of Ordered Latent Class Regression in Educational Assessment

Jisung Cha

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

Columbia University

2011

ABSTRACT

Application of Ordered Latent Class Regression in Educational Assessment

Jisung Cha

Latent class analysis is a useful tool to deal with discrete multivariate response data. Croon (1990) proposed the ordered latent class model where latent classes are ordered by imposing inequality constraints on the cumulative conditional response probabilities. Taking stochastic ordering of latent classes into account in the analysis of data gives a meaningful interpretation, since the primary purpose of a test is to order students on the latent trait continuum. This study extends Croon's model to ordered latent class regression that regresses latent class membership on covariates (e.g., gender, country) and demonstrates the utilities of an ordered latent class regression model in educational assessment using data from Trends in International Mathematics and Science Study (TIMSS). The benefit of this model is that item analysis and group comparisons can be done simultaneously in one model. The model is fitted by maximum likelihood estimation method with an EM algorithm. It is found that the proposed model is a useful tool for exploratory purposes as a special case of nonparametric item response models and cross-country difference can be modeled as different composition of discrete classes. Simulations is done to

evaluate the performance of information criteria (AIC and BIC) in selecting the appropriate number of latent classes in the model. From the simulation results, AIC outperforms BIC for the model with the order-restricted maximum likelihood estimator.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my advisor, Matthew S. Johnson, for his support, encouragement and the opportunities he gave me to complete this dissertation. I have enjoyed working with him. I would like to thank Arron M. Pallas, Young-Sun Lee, Hammou El Barmi, and Zhiliang Ying, the members of my thesis committee, for their helpful comments and insight.

I would like to thank my parents, Jung-Wook and Heesun for their dedicated love, and for the sacrifices they have made in order to give opportunities to their children. I would like to thank Dong-Kyu, my brother.

I would like to thank my girls, Jennifer, Erin and Mary. A special thank-you to my husband Donghoon. Without his companionship, and understanding I never would have been able to complete this work.

To Jung-Wook and Heesun

Chapter 1

Introduction

A number of statistical models have been developed in the area of educational measurement, in order to make inferences about tests and test takers. In education, test items aim to determine ability in a subject or content area. There are three important features of educational assessment data that need to be dealt with via latent variable models.

First, the statistical models include unobservable latent variables denoted by θ such as ability, aptitude, or attributes in which test items are designed to measure. These variables are latent because they are not directly observable. Since the target of the measurement is not directly observable, multiple measurements are necessary. These observed response variables denoted by $\mathbf{Y}$ are strongly related to one another because of the latent variable. The models attempt to explore the relationships among a set of the observed categorical variables by means of an underlying latent variable (Eaton et al., 1989; Hagenaars, 1979a; Meredith and Millsap, 1992).

Second, the models assume that the probability of a positive response to an item, j, denoted by $Pr(Y_j = y_j|\theta) = P_j(\theta)$ increases along a latent continuum, whereby test instruments contain categorical responses of individuals to the items.

The data can be binary (e.g., correct or incorrect responses) as with multiple-choice items or polytomous (e.g., correct, partially correct, or incorrect responses) as with open-ended items.

Third, then the observed variables are measured on an ordinal scale, the order of the categories should be taken into account properly in the statistical analysis. Given that the observed variables are ordinal, it is natural to assume that the underlying latent variable also has an order.

Description of Latent Variable Models

Latent variable models are prominent in social sciences such as education, sociology and psychology. A latent variable model is any model that assumes the existence of a latent variable that describes the interdependency among observed variables. Some examples of latent variable models include factor analysis, structural equation models, latent class models and item response theory (IRT) models.

Bartholomew (1983) classified latent variable models into four categories based on the scale types of the latent and observed variables; they are factor analysis, latent trait analysis, latent profile analysis, and latent class analysis. In latent class analysis and latent trait analysis models, observed variables (also called indicators) are dichotomous, ordinal, or nominal categorical variables and their conditional distributions are assumed to be binomial or multinomial, as shown in Table 1.1 (Bartholomew, 1987; Heinen, 1996). Latent class models differ from latent trait models where a continuous underlying latent variable is replaced by a discrete variable with "classes" that define homogeneous groups of individuals. The comparison of the two models was discussed in detail in Heinen (1993). This study focuses on categorical observed outcome variables such as response data to items or questions in a test or survey. Thus latent trait model (also called IRT model) and latent class analysis model are mainly discussed.

Table 1.1: Classification of Latent Variable Methods

	Observed Variable	
Latent Variable	Metrical	Categorical
Metrical	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

Educational measurement has mainly utilized IRT models. IRT models assume that the relationship between the probability of a positive response and the underlying latent variable is an increasing one; this functional relationship is called the item response function (IRF) (also called an item characteristic curve(ICC)).

Recently an interest in latent class models has increased, because a smaller number of ordered latent classes can approximate the continuous latent variable with sufficient accuracy (Heinen, 1996; Lindsay et al., 1991). Compared to latent class analysis models, the disadvantages of the IRT model are that it makes stronger assumptions about the latent distribution and that its estimation is computationally more intensive (Vermunt, 2001; Vermunt and Magidson, 2005). Ordered latent class models are constrained latent class models by imposing order restrictions to classify students into ordinal categories while a latent trait models order students along a continuous latent trait. Therefore the ordered latent class models are a feasible option for educational assessment data.

Ordering classes

Standard latent class models treat the response categories of observed variables and latent variables not as ordinal but as nominal. Since the primary purpose of a test is to order students on the latent trait continuum, accounting for stochastic ordering of latent classes in the analysis of data gives a meaningful interpretation.

To impose an order to latent classes on a single dimensional scale, the item response probabilities are posited to be nondecreasing functions of the latent class.

Two methods are specified depending on the type of restriction. One way is to use a parametric functional form for the relationship between the item response probability and the latent class as in IRT models. This method may provide order among classes, but it considers the latent classes interval-level measures, not ordinal-level measures. In fact, this implies that the distance between class 1 and class 2 is the same as that between class 2 and class 3. In many cases, this constraint can become too restrictive and inappropriate.

The other way is to impose inequality constraints directly on the item response probability parameters of a latent class model. In categorical data analysis, researchers usually treat categories as nominal or interval and truly ordinal models are rarely used (Agresti et al., 1986; Heinen, 1996; Vermunt, 2001). Croon (1990) proposed an ordered latent class analysis (OLCA) model where the latent classes are placed on an ordinal scale by imposing inequality constraints on item response probabilities given the latent classes, often referred to as item response step function (ISRF). He also developed an algorithm to obtain the maximum likelihood estimates of the ISRF parameters. This ordered latent class model can be viewed as a special case of nonparametric IRT models which restricts the item response probability, $Pr(Y_j \geq y_j|\theta)$ to be non-decreasing in θ for all y_j. In comparison to the parametric IRT model, the advantages of this model are that it directly estimates IRFs and it relaxes the assumptions such as the parametric assumptions on item response probability or the normal assumption on the ability distribution of examinees(van Onna, 2002). This study follows Croon's approach, treating the latent classes as ordinal without unnecessary constraints.

Assumptions of the Ordered Latent Class Analysis (OLCA) Models

Data are assumed to have the following structure. Let Y_j denote the observed response of an individual to item $j \in \{1, ..., J\}$ with K categories. The latent class is denoted by $\xi \in \{1, ..., T\}$. The distribution of latent classes is denoted as:

$$\Pr(\xi = x) = \pi_x.$$

The main assumptions of this model are homogeneity, local independence, unidimensionality, and monotonicity of this model.

Homogeneity (H)

The core assumption in latent class analysis is that the population consists of a set of mutually exclusive and homogeneous subgroups called classes. The individuals within a sub-group are homogeneous in the sense that the probability for a particular response on a particular item depends only on the latent class to which the individual belongs.

$$Pr(Y_j = k|\xi = x) = p_{jkx}. \tag{1.1}$$

Local Independence (LI)

Local independence assumes that a vector of the observed variables, Ys are related only due to the latent class ξ. Under this assumption, the joint probability of $\mathbf{Y}$

given ξ can be written as the product of probabilities of Y_j given the latent class x.

$$\begin{aligned} Pr(Y_1 &= y_1, Y_2 = y_2, ..., Y_J = y_J|\xi = x) \\ &= Pr(Y_1 = y_1|\xi = x)Pr(Y_2 = y_2|\xi = x)\cdots Pr(Y_J = y_J|\xi = x) \\ &= \prod_{j=1}^{J} Pr(Y_j = y_j|\xi = x) \\ &= \prod_{j=1}^{J} p_{j,y_j,x}. \end{aligned} \tag{1.2}$$

Unidimensionality (U)

The assumption of unidimensionality posits that the observed categorical variables Y are assumed to measure only one ability, attitude, trait, or attribute.

Monotonicity (M)

To obtain stochastic ordering among the latent classes within an item, Croon (1991) proposed an ordinal latent class model by imposing inequality restrictions directly on the IRFs:

$$\begin{aligned} \Pr(Y_j \leq k|\xi = x_1) &\geq \Pr\left(Y_j \leq k|\xi = x_2\right) \\ \sum_{m=1}^{k} p_{jmx_1} &\geq \sum_{m=1}^{k} p_{jmx_2} \end{aligned} \tag{1.3}$$

for all j and k, and for all x_1 and x_2 such that $x_1 < x_2$.

Three assumptions (LI, U, and M) are often referred to as the monotone homogeneity (MH) models (Holland and Rosenbaum, 1986).

Ordered Latent Class Regression (OLCR) Model

This study extends Croon's ordered latent class analysis (OLCA) model to ordered latent class regression (OLCR) models where latent class membership is regressed

on covariates. In ordered latent class models, the latent class probability is constant across individuals within each class; however, in ordered latent class regression models, the latent class probability is a function of additional explanatory variables such that the latent class probability is different across individuals. The covariates could be continuous, categorical, or product of covariates in ordered to model interactions as in ordinary linear regression models. Inclusion of covariates in an ordered latent class model is often useful in several ways:

1. The researcher may be interested in comparisons among groups such as gender, race, and socio-economic status.

2. Incorporating auxiliary information in the form of a covariate improves the inference about the students and the item. (Mislevy and Sheehan, 1984).

3. Item analysis and group comparison can be done simultaneously in a model. Therefore, assessing group differences does not require a secondary analysis.

Objectives of This Study

1. This study develops the algorithm to estimate the ordered latent class regression model.

2. This study demonstrates how to utilize the proposed model for item analyses and group comparisons in educational assessment applied to data from Trends in International Mathematics and Science Study (TIMSS).

3. Simulation studies are conducted to evaluate the procedure and inference of model selection by AIC and BIC given the proposed model.

Overview of This Study

Chapter 2 presents an overview of the IRT models, latent class models and ordinal latent class models. In addition, it illustrates parametric and nonparametric methods of imposing order constraints in the context of latent class models, and introduces the latent class models with covariate (e.g. gender) and differential item analysis based on the item response models. Last, model selection procedure is presented. Chapter 3 introduces an estimation procedure for the ordered latent class regression model. Chapter 4 illustrates the results of empirical data analysis using TIMSS data to investigate group differences on math performance in terms of latent class distribution and item response probabilities. Chapter 5 conducts two simulation studies to evaluate the model selection procedure in the ordered latent class model by information criteria (AIC and BIC). Chapter 6 discusses a summary of results, limitations and future directions of this study.

Chapter 2

Literature Review

2.1 Introduction

This chapter begins with an overview of IRT models including parametric IRT models and nonparametric IRT models in Section 2.2. Section 2.3 presents standard latent class models. Section 2.4 presents ordered latent class analysis models and methods for order constraints: parametric and nonparametric methods. Sections 2.5 and 2.6 illustrate the extensions of a standard latent class models and the differential item functioning, respectively. Section 2.7 discusses the information criteria for model selection.

2.2 Item Response Theory Models

Latent trait, or item response theory (IRT) models have been dominantly used in educational testing and research. The objective of these models is to assess the characteristics of the items and to make an inference about an examinee's proficiency in a content area being measured. IRT models typically assume unidimensionality (UN), local independence (LI) and monotonicity (M). The monotonicity

assumption determines the probabilistic relationship, $Pr(Y_j = 1|\theta) = P_j(\theta)$, between positive responses to items, j and latent proficiency, θ. $P_j(\theta)$ is also called the item response function (IRF). Thus, examinees with higher proficiency have a higher probability of answering an item correctly. A considerable number of models have been proposed for IRT models. They fall into two classes according to the estimation method of the IRF, $P_j(\theta)$: parametric and nonparametric methods. Parametric item response theory (PIRT) models fall into two parts according to the number of response categories. For binary response data the number of parameters determines the shape of the IRF and characterizes IRT models, as the Rasch model (or one-parameter model), the two-parameter model, and the three-parameter model. For polytomous response data (i.e., data with more than two response options), the models include the graded response model, partial credit model (including the generalized partial credit model), and the rating scale model. PIRT models are commonly fitted with maximum marginal likelihood estimation (Bock and Aitkin, 1981), for which is usually assumed that:

- item response functions (IRFs) are logistic; and
- the ability distribution of population follows a parametric distribution (i.e., normal or uniform).

Section 2.2.1 illustrates parametric IRT models and then nonparametric IRT models are discussed in Section 2.2.2.

2.2.1 Parametric Item Response Theory (PIRT) Models

Models for Binary Data

The two parameter logistic model assumes the logit of the item response function (IRF), $P_j(\theta)$, is a linear function of θ,

$$\text{logit}\{P_j(\theta)\} = \alpha_j(\theta - \beta_j) \tag{2.1}$$

$$P_j(\theta) = \frac{1}{1 + \exp\{\alpha_j(\beta_j - \theta)\}} \tag{2.2}$$

The parameter β is interpreted as the difficulty of an item. Larger values of β are associated with lower proportions of correct responses. The slope parameter, α, is called the discrimination parameter of an item. The parameter indicates the extent to which the item discriminates the abilities of the examinees. It is represented as the steepness of the item response function, and is item specific in the 2PL model. To make the model identified, ability parameter, θ is constrained to have a mean of 0 and a variance of 1.

The Rasch model (RM) is often referred to as the one-parameter logistic model (1PL). Unlike the 2PL model, the slope parameters, αs, are constant across all items j. When all items have the same discrimination, the IRFs do not intersect; this property is called the invariant item ordering property (IIO) (Sijtsma and Hemker, 2000; Sijtsma and Junker, 1996). In RMs, the item response functions (IRFs) do not intersect, and thus $P_j(\theta)$ of the item j is constantly larger than $P_k(\theta)$ of the item k for the entire range of θ. That is called the non-intersection assumption (NI). The NI assumption combined with monotonicity assumption is often called double monotonicity (Mokken, 1997, 2001; van Onna, 2002). Another well-known property is the specific objectivity that the comparison of two individuals is independent of the items on a test (Johnson, 2007; Molenaar, 1997; Sijtsma and Junker, 1996).

Birnbaum (1968) introduced the three parameter logistic model, which in-

cludes a guessing parameter. The IRF is:

$$P_j(\theta) = \gamma_j + \frac{1-\gamma_j}{1+\exp\{\alpha_j(\beta_j-\theta)\}}, \tag{2.3}$$

where $0 \leq \gamma_j \leq 1$, and γ_j is a guessing parameter that allows for a non-zero lower asymptote for the IRF. This parameter accounts that low ability students have a probability of at least γ_j to correctly answer the item.

Polytomous Item Response Models

For analyzing polytomous response data, several models have been used. Below this section reviews the graded response model (Samejima, 1969), the partial credit model (Masters, 1982) and the rating scale model.

The graded response model (GRM) assumes that the cumulative log odds for scoring $k \in \{1, 2, ..., K\}$ or higher on item j, is a linear function of θ:

$$\log\left(\frac{Pr(Y_j \geq k|\theta)}{Pr(Y_j < k|\theta)}\right) = \alpha_j(\theta - \beta_{jk}),$$

where k is response category $\{1 \leq k \leq K\}$, j is an item, and θ is the latent variable such as ability. The discrimination parameter, α_j, are fixed across item categories and that the item-category step parameters, β_{jk} are ordered by the category index j, $\beta_{j1} < \beta_{j2} < .. < \beta_{jK-1}$.

The partial credit model (PCM) assumes that the adjacent-categories logit is a linear function of θ:

$$\log\left(\frac{Pr(Y_j = k|\theta)}{Pr(Y_j = k-1|\theta)}\right) = \alpha_{jk}(\theta - \beta_{jk}),$$

which leads to the following item-category response function

$$\begin{aligned} P_{jk}(\theta) &= \Pr\{Y_{ij} = k|\theta\} \\ &= \frac{\exp\{\sum_{l=0}^{k}\alpha_j\left(\theta-\beta_{jl}\right)\}}{\sum_{r=0}^{K_j}\exp\{\sum_{l=0}^{r}\alpha_j\left(\theta-\beta_{jl}\right)\}}. \end{aligned}$$

PCM is a polytomous version of a Rasch model where α is constant across items. The generalized partial credit model (GPCM) generalizes PCM to allow for the discrimination parameter, α, to vary across items and was formulated by (Muraki, 1992). The PGM ad GPCM differ from GRM in that they belong to the Rasch family of models and β_{jk} are not necessarily ordered.

The rating scale model (RSM) is also an extension of the Rasch model, first presented by Rasch (1961) and was restructured by Andrich (1978). It assumes that the category scores are equally spaced and the continuation logit is a linear function of θ:

$$\log\left(\frac{Pr(Y_j \geq k|\theta)}{Pr(Y_j = k-1|\theta)}\right) = \alpha_j(\theta - \beta_{jk})$$

2.2.2 Nonparametric Item Response Theory (NIRT) Models

The logistic function on parametric item response probability in Equations (2.1-2.3) above is mathematically convenient, but it may also be too restrictive. When parametric item response theory (PIRT) models fit the data poorly (or are suspected to), restrictions on the models are not appropriate or there are violations of the unidimensional assumption. Misfit of parametric unidimensional IRT models does not necessarily address the violation of unidimensionality.

Thus many researchers have studied nonparametric approaches to relax the constraints. Research in nonparametric item response theory (NIRT) model has focused on relaxing two assumptions of PIRT: (1) normal assumption for the distribution of latent trait, (2) parametric functions of the response probability $P_j(\theta)$ to be a monotonically increasing function of θ. Woods and Thissen (2006) proposed a spline-based density estimation for the latent population distribution and found that this estimation method provides a flexible alternative to existing procedures that use a normal distribution. Junker and Sijtsma (2001) defined nonparametric

IRT as an ordinal measurement model with a minimal set of assumptions: unidimensionality, local independence, and monotonicity. The use of NIRT models are to examine the assumptions of IRT models and determine model fit. Techniques, such as kernel, splines, or isotonic regression have often been used for plotting the IRF in order to visually inspect IRFs and they are also used for equating tests (Johnson, 2007; Ramsay, 1991; Ramsay and Abrahamowicz, 1989). Johnson (2007) estimated $P_j(\theta)$ by using B-spline by Bayesian estimation.

A recent NIRT approach is the ordered latent class model. One may consider the ordered latent class model as a means of assessing the status of students on an underlying latent continuum. For example, two discret skill classes may be discerned the "non-master", and "master". On depression scale, one might consider ordered classes "not depressed", "slightly depressed", and "severely depressed". The latent classes are ordered, because the trait levels can be ordered from 'low' to 'high'. The response probability on an item is assumed to be the same for each in a latent class, therefore the IRFs for the ordered latent class model can be plotted as a discrete function. Figure 2.1 shows the item response function (IRF) estimated by parametric IRT model and the item step response function (ISRF) estimated by ordered latent class analys (OLCA) model. The ordered latent class model has been studied by both the Bayesian method (Hoijtink and Molenaar, 1997) and the maximum likelihood estimation (Croon, 1990; van Onna, 2002).

PIRT models are relatively easy to estimate parameters and provide useful interpretations to parameters. Therefore, if the assumptions hold, the gain is large. However if the assumptions do not hold, parameter estimates are biased due to misspecification of IRFs. Consequently, biased estimates lead to inaccurate inferences about the true value of a student's latent trait.

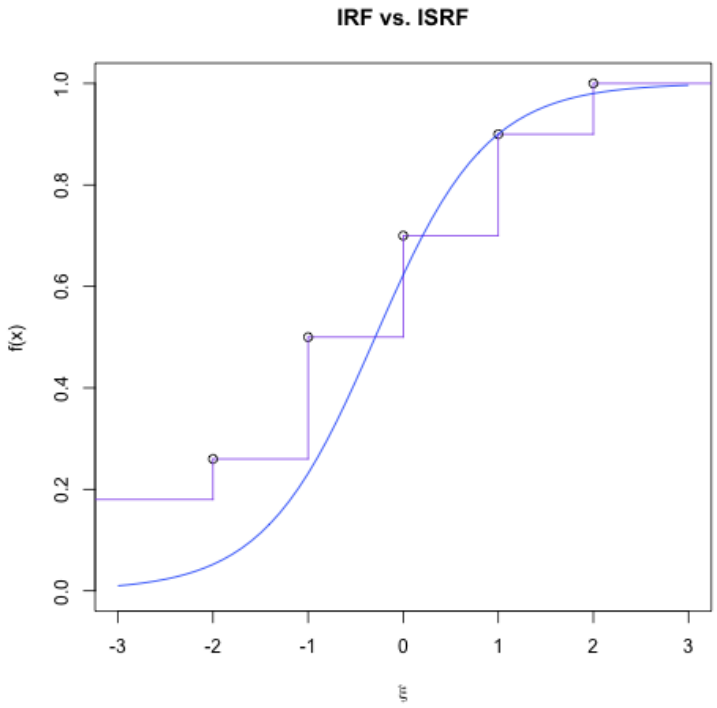

Figure 2.1: Item response function (IRF) and Item step response function (ISRF)

2.3 Latent Class Model

In social sciences the use of categorical variables is prevalent. Traditionally the relationships between categorical variables were studied by means of the contingency table analysis. Lazarsfeld and Henry (1968) proposed the latent class model to analyze the social attitudes of individuals via surveys. Since Goodman (1974) developed maximum likelihood estimation procedures, the latent class model has been applied in a wide range of areas, such as marketing (Dillon and Kumar, 1994), sociology (Rost and Langeheine, 1997), psychometrics (Eaton et al., 1989), medical research (Albert et al., 2002; Bandeen-Roche et al., 1999, 1997; Formann, 1996; Moustaki, 1996; Sullivan et al., 1998) , and psychosocial (Garrett and Zeger, 2000; Hudziak et al., 1998; Neuman et al., 2001).

Latent class models intend to explain the interdependency of the categorical observed variables $\mathbf{Y}$ by introducing the explanatory latent variable ξ. If ξ is observed, then a method for determining whether ξ actually explains the relationships among observed variables $\mathbf{Y}$ can be straightforward. But if the possible explanatory variable ξ is not observed (latent), then methods for explaining the relationship become more complex. Goodman (1974) found an iterative procedure in which the maximum likelihood equations could be solved for the latent class model, which is a latent variable in the model. It is a special case of the well-known EM algorithm, where E stands for the expectation step and M stands for the maximization step, and they generalized the method and developed the theory (Anderson, 1982).

Latent class model associates observed categorical variables with latent categorical variables. Let Y_{ij} denote the observed response of examinee i for item j. Y_{ij} has a discrete value $k \in \{1, 2, ..., K\}$. The latent class is denoted by $\xi \in \{1, 2, ..., T\}$. The population of examinees is assumed to consist of T mutually exclusive latent classes with a restriction such that $\sum_{x=1}^{T} Pr(\xi = x) = 1$. Assuming local independence, joint probabilities within latent class can be represented as a product of the

form:

$$Pr(Y_1 = y_1, \ldots, Y_J = y_J | \xi = x) = \prod_{j=1}^{J} \prod_{k=1}^{K} p_{jkx}^{\delta_{jk}},$$

where $\delta_{jk} = 1$ if $y_j = k$, otherwise 0. Then marginal distribution of $\mathbf{Y}$ is:

$$\begin{aligned} Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{x=1}^{T} Pr(\mathbf{Y} = \mathbf{y}, \xi = x) && (2.4) \\ &= \sum_{x=1}^{T} Pr(\xi = x) \prod_{j=1}^{J} Pr(Y_{ij} = y_{ij} | \xi = x) && (2.5) \\ &= \sum_{x=1}^{T} \pi_x \prod^{J} \prod^{K} p_{jkx} && (2.6) \end{aligned}$$

Given estimates $\hat{\pi}$ and $\hat{p}_{jkx}$, the posterior probability denoted by $p_{x|jk}$ shows that each examinee belongs to each class, given the observed data. The posterior probability is calculated using Bayes' formula:

$$Pr(\xi = x | \mathbf{Y}) = \frac{\hat{\pi_x} \hat{p_{jkx}}}{\sum_{x=1}^{T} \hat{\pi_x} \hat{p_{jkx}}} = p_{x|jk} \qquad (2.7)$$

2.4 Ordered Latent Class Model

It may be noted that, thus far, latent classes have been assumed to be unordered, having a nominal measurement level. A recent development is to put order restrictions on the response probability, p_{jkx}, to express that there is an ordering among the latent classes, such that students in a higher latent class have a higher probability, p_{jkx} of answering an item correctly. The ordered latent class model assumes stochastic ordering of the item response by latent class $Pr(Y_{ij} > y | \xi = x) \leq Pr(Y_{ij} > y | \xi = x + 1)$. Below discusses what this means in both the case of dichotomous and polytomous responses.

Dichotomous Case

Suppose that a set of latent classes is ordered such that the latent class 1 is the "lowest" and the latent class T is the "highest"; the latent class x is lower than the latent class $x+1$. Suppose all items are dichotomous with categories ($K = 2$), with $y = 0$ representing an incorrect response and $y = 1$ representing a correct response. The corresponding response probabilities for item j given latent class x are p_{j0x} and p_{j1x}. For dichotomous items, $p_{j0x} + p_{j1k} = 1$. If a set of latent classes are ordered along a latent continuum, it is reasonable to assume that the probability of getting a correct response increases or does not decrease as a latent class moves from low to high. The system of inequalities is expressed as follows:

$$p_{j11} \leq \cdots \leq p_{j1x} \leq p_{j1,x+1} \leq \cdots \leq p_{j1T}$$

Polytomous Case

For polytomous items with $K > 2$ response categories, the set of response categories $k \in \{1, ..., K\}$ is split into two sets: $\{1, ..., g-1\}$ and $\{g, ..., K\}$. For $1 \leq t \leq T$, and $1 \leq k \leq K$,

$$\sum_{k=g}^{K} p_{jkx} \leq \sum_{k=g}^{K} p_{jk,x+1}$$

For simplicity, in the case of $K = 3$ two sets of constraints should be satisfied as follows:

$$p_{j3x} \leq p_{j3,x+1}$$

$$p_{j2x} + p_{j3x} \leq p_{j2,x+1} + p_{j3,x+1}.$$

These constraints should be considered for each item j. The monotonicity constraints on cumulative response probabilities are related to the concept of regression dependence, which was described by Lehmann (1966). If the conditional probability

of Y is a nondecreasing function of x, then for all x and y

$$x_1 \leq x_2 \Rightarrow Pr(Y \leq y|x_1) \geq Pr(Y \leq y|x_2). \tag{2.8}$$

If the condition that $Pr(Y \leq y|x)$ is non-increasing in x holds, one can say that Y is positively regression dependent on ξ (Lehmann, 1966). Therefore, with inequality constraints of response probabilities it is assumed that the observed variables Y are positively regression dependent on the latent variable ξ.

Figure 2.2 shows the ISRFs of a standard (unconstrained) latent class model and an ordered latent class model. In the right panel of Figure 2.2 the four steps represent the ISRFs estimated by an ordered latent class model. The x-axis is for scores (1-3) and the y-axis is for the probability of getting a category lower than k given latent class x, Thus, lower lines represent for higher order classes. Specifically the blue line represents the response probabilities given class 4, which is the most proficient group and the green line represents the response probabilities given class 3, which is the second proficient group.

The red line represents the response probabilities given class 2 and the purple line represents the response probabilities given class 1, which is the least proficient group. Graphically the four lines do not cross, which means the condition, $Pr(Y \leq y|x_1) \geq Pr(Y \leq y|x_2) \geq Pr(Y \leq y|x_3) \geq Pr(Y \leq y|x_4)$, is satisfied for any k. Therefore the ordering from low to high is: $x_1 \leq x_2 \leq x_3 \leq x_4$. However, in the left panel of Figure 2.2, the four steps estimated by a standard latent class model cross one another. The probability of getting a category lower than 1 given class 3 represented by the green line is lower than the probability of a category lower 1 given class 2 represented by the red line, however the probability of category lower than 2 given class 3 is higher than the probability of category lower than 2 given class 2. It means that the inequality condition has not been met for all k, therefore stochastic ordering among the classes does not exist. Estimation procedures of ISRFs that satisfy order constraints are discussed in detail in Chapter 3.

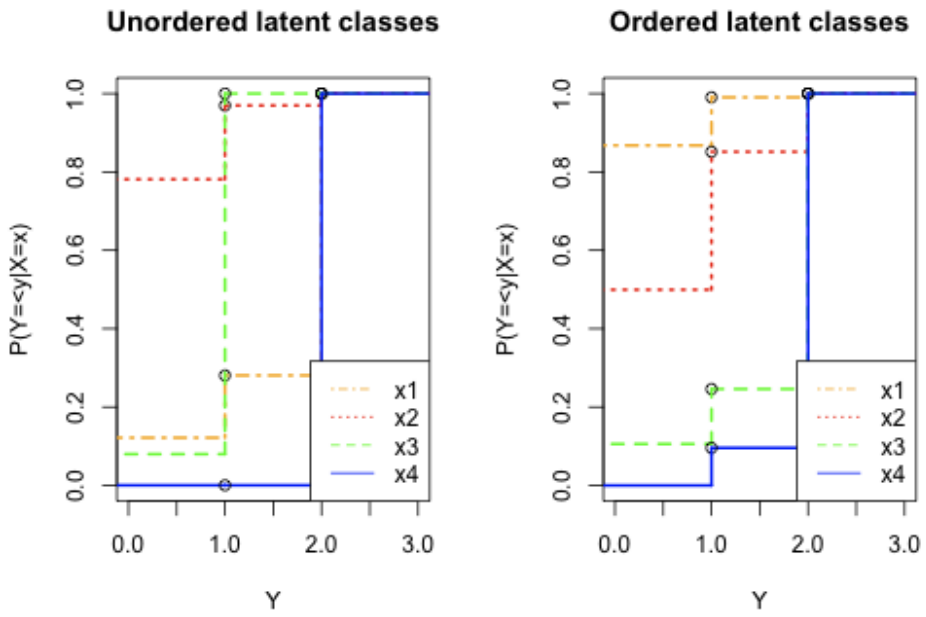

Figure 2.2: Stochastic Ordering among Latent Classes

Vermunt (2001) reviewed the methods of order constraints for ordinal categorical data and divided them into two: nonparametric and parametric way. A parametric approach is presented which is based on imposing linear equality constraints on response probabilities, however a nonparametric approach is based on imposing inequality constraint on response probabilities.

Parametric Order Constraints

The logit function is a very useful tool when an ordinal relationship between a categorical dependent variable and independent variables needs to be taken into consideration, in our case the relationship between Y and ξ. As Agresti (2003) illustrated in his book, there are several ways of defining log odds of response probabilities. Four types of logits are shown next, (see also (Heinen, 1993; Mellenbergh, 1995; van der Ark, 2001; Vermunt, 2001)).

1. Cumulative log odds for scoring k or higher on item j,

$$\log\left(\frac{Pr(Y_j \geq k|\xi = x)}{Pr(Y_j < k|\xi = x)}\right) = \alpha_j(x - \beta_{jk}),$$

2. Adjacent-categories log odds

$$\log\left(\frac{Pr(Y_j = k|\xi = x)}{Pr(Y_j = k-1|\xi = x)}\right) = \alpha_j(x - \beta_{jk}),$$

3. Continuation log odds

$$\log\left(\frac{Pr(Y_j \geq k|\xi = x)}{Pr(Y_j = k-1|\xi = x)}\right) = \alpha_j(x - \beta_{jk}),$$

where k is a response category $\{2 \leq k \leq K\}$, j is the item, and x is the latent class $\{2 \leq x \leq T\}$.

Four types of logits are shown next, (see also (Heinen, 1993; Mellenbergh, 1995; van der Ark, 2001; Vermunt, 2001)). IRT models are extended for polytomous data using these odd ratios as discussed in Section 2.2.1. The graded response model (Samejima, 1969) assumes the cumulative logits, and the partial credit model and the generalized partial credit model (Masters, 1982) assume the adjacent logits, and the rating scale model (RSM) assumes the continuation logits.

The corresponding item step response functions (ISRFs) are as follows:

$$ISRF^{cum} = Pr(Y \geq k|\xi = x) = \frac{1}{1 + \exp\{\alpha(\beta_k - x)\}}$$

$$ISRF^{adj} = \frac{Pr(Y = k+1|\xi = x)}{Pr(Y = k|\xi = x) + Pr(Y = k+1|\xi = x)} = \frac{1}{1 + \exp\{\alpha(\beta_k - x)\}}$$

$$ISRF^{conI} = \frac{Pr(Y \geq k|\xi = x)}{Pr(Y \geq k-1|\xi = x)} = \frac{1}{1 + \exp\{\alpha(\beta_k - x)\}}.$$

If ξ has an integer value corresponding to each latent class(e.g. 1,2, ..., T), the logits have equal slopes across categories. Treating a latent class as an integer results in various constraints on odds ratios between adjacent categories k and $k+1$.

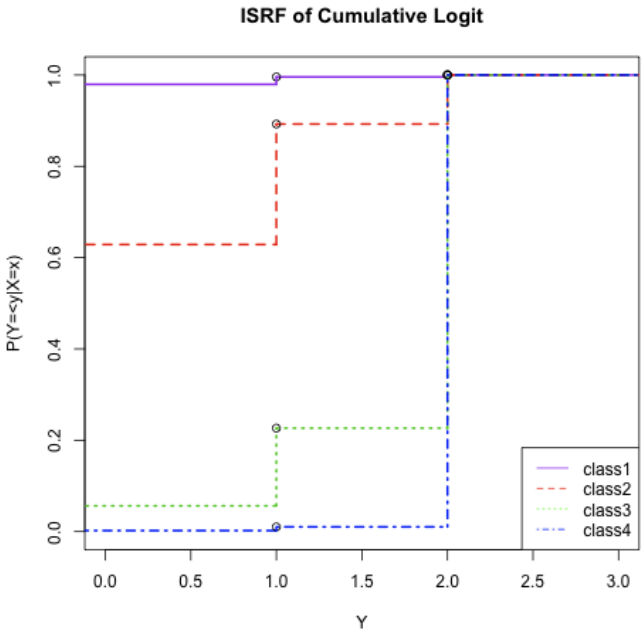

Figure 2.3: The estimates of ISRFs using cumulative logits

These parametric order restricted latent class models can be viewed as discrete analogues of IRT models. As the number of classes increases, the item step response function (ISRF) is a closer approximation to the item response function (IRF) (Vermunt and Magidson, 2005). Figure 2.3 shows the ISRFs estimated using parametric cumulative logit in the equation above. The x-axis is for scores (1-3) and the y-axis is for the probability of getting a category lower than k given latent class x, Thus, lower lines represent for higher order classes.

Nonparametric Order Constraints

As alternative methods to the models above, Coull and Agresti (2002) reviewed in details on inequality-constrainted methods for four types of odd ratios in contingency tables to treat two observed random variables (Y, X) as ordinal. Croon employed cumulative logit for the ordered latent class analysis (OLCA) model, and

noted that the cumulative logit is more flexible than the adjacent category logit. The OLCA model proposed by Croon does not have a functional form between observed variables and the unobserved latent class ξ. It only assumes that monotonicity, so it is a nonparametric approach. Nonparametric ordered latent class models where ISRFs, p_{jkx}, are directly estimated under the inequality constraints on p_{jkx} as follows:

$$\sum_{k=1}^{g} p_{jkx} \geq \sum_{k=1}^{g} p_{jk,x+1}$$

These constraints satisfy the monotonicity assumption that the probability of positive response increases as the latent class number increases. Compared to parametric models, these models are more flexible and less restrictive. Many researchers have studied this approach to incorporate into IRT models (Croon, 1990; Hoijtink and Molenaar, 1997; van Onna, 2002).

2.5 Latent Class Regression

Latent class regression model (LCR) generalizes latent class analysis(LCA) model by allowing for covariates to be related to latent class (Bandeen-Roche et al., 1997; Dayton and Macready, 1988; Heijden et al., 1996; Melton et al., 1994). In the LCA model, it is assumed that every individual has the same probabilities of being in a latent class; however, in the LCR model it is assumed that latent class probabilities differ by individuals depending on their observed covariates. It is called a concomitant latent class model or latent class model with covariates (Dayton and Macready, 1988). It is also closely related to latent class analysis for multiple groups which is called simultaneous latent class model by many researchers (Clogg and Goodman, 1984, 1986; Formann, 1985; Hagenaars, 1979b). There are two ways of relating covariates to latent class models:

1. Modeling the relationship between covariates and the latent class (Bandeen-

Roche et al., 1997; Dayton and Macready, 1988).

2. Modeling the relationship between covariates and item response probabilities of the measured indicators (Clogg and Goodman, 1984; Huang and Bandeen-Roche, 2004; Melton et al., 1994).

The former can estimate the effects of covariates on the latent variable and the latter can examine the local independence assumption of the model to see if response probabilities differ by characteristics of individuals and prevent possible misclassification of underlying variable categories.

Bandeen-Roche et al. (1997) fitted a regression extension of the latent class model that allowed direct effects of covariates not on item response probability but latent class membership. Huang and Bandeen-Roche (2004) extended it to the model for self-reported visual disability by including direct covariates effects on item response probability.

To avoid confusion, a latent class regression model is referred to as the first case in this study. As in ordinary multiple regression, covariates can be either continuous or categorical variables. For a case with categorical variables, one can consider grouping variables such as race, gender, or SES to examine the difference between groups.

There are two parts in the latent class model regression: (1) item response probabilities that add up to 1, p_{jkx}, and (2) latent class probabilities, π_x (e.g., prior probabilities). π_x is reparameterized as follows:

$$\begin{aligned} \ln(\pi_{2i}/\pi_{1i}) &= Z_i\gamma_2 \\ \ln(\pi_{3i}/\pi_{1i}) &= Z_i\gamma_3 \\ &\vdots \\ \ln(\pi_{Ti}/\pi_{1i}) &= Z_i\gamma_T \end{aligned}$$

The first latent class serves as a reference category.

$$Pr(\mathbf{Y} = k|\mathbf{Z}) = \sum_{x=1}^{T} Pr(\xi = x|\mathbf{Z}) Pr(\mathbf{Y} = k|\xi = x),$$

where Z_i represents the covariates for individual i and γ is a vector of coefficients to the latent class $(\xi = x)$.

$$\pi_{xi} = Pr(\xi_i = x|\mathbf{Z}_i) = \frac{\exp(\mathbf{Z}_i \boldsymbol{\gamma}_x)}{\sum_{x=1}^{T} \exp(\mathbf{Z}_i \boldsymbol{\gamma}_x)}.$$

In a simple case of a three latent class model with one covariate, the probability of being in x class is written in a logistic form that depends on the covariate (Dayton and Macready, 1988):

$$\pi_{1|z_i} = \frac{1}{1 + \exp(\gamma_{02} + \gamma_{12} z_i) + \exp(\gamma_{03} + \gamma_{13} z_i)} \tag{2.9}$$

$$\pi_{2|z_i} = \frac{\exp(\gamma_{02} + \gamma_{12} z_i)}{1 + \exp(\gamma_{02} + \gamma_{12} z_i) + \exp(\gamma_{03} + \gamma_{13} z_i)} \tag{2.10}$$

$$\pi_{3|z_i} = \frac{\exp(\gamma_{03} + \gamma_{13} z_i)}{1 + \exp(\gamma_{02} + \gamma_{12} z_i) + \exp(\gamma_{03} + \gamma_{13} z_i)} \tag{2.11}$$

The latent class regression model can easily be expanded to p covariates as in ordinary multiple regressions.

$$\log\left(\frac{\pi_{x|z_i}}{\pi_{1|z_i}}\right) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_p z_p$$

2.6 Differential Item Functioning

Differential item functioning (DIF) is an important issue in large scale standardized testing. It refers to the unexpected difference in item performance among groups of equally proficient examineer (Lord,1980). Thus, its presence would threaten the validity of inferences drawn from a test. A variety of methods for assessing DIF have been developed. Pontenza and Dorans (1995) overviewed the methods for finding DIF items and classified them on the basis of whether they define DIF with

respect to an observed variable or a latent variable, and whether or not a parametric form describe the relationship between item scores and the matching variables (e.g., number-right score).

DIF can be defined in the context of IRT models which provides a class of models describing the relationship between item response function and the latent trait (Hambleton and Swaminathan, 1985). Item response function (IRF) discribes the item response probability of getting an item correctly given the latent trait level. If an item does not display DIF, then its IRF should be the same for all groups under consideration. On the other hand if an item displays DIF, the IRFs will be different across groups. More technically, DIF is said to occur whenever the IRF, $P_j(\theta)$, of getting correct response differs for the two groups. For item j the null DIF hypothesis is

$$H_0 : p_{g1,j}(\theta) = p_{g2,j}(\theta), \text{for all } \theta.$$

The IRF is characterized by item parameters (e.g., α, β). DIF falls into two categories: uniform DIF and non uniform DIF. In the two-parameter logistic model when $\alpha_{g1,j} = \alpha_{g2,j}$, but $\beta_{g1,j} \neq \beta_{g2,j}$, the two IRFs are parallel and there is a location shift due to different group membership. The uniform DIF describes DIF against the same group for all ability levels regardless of whether the magnitude of DIF is constant or not constant as ability level varies. Uniform DIF occurs when $p_{g1}(\theta) \geq p_{g2}(\theta)$ or $p_{g1}(\theta) \leq p_{g2}(\theta)$ for all θ. The amount and the direction of DIF can vary at different ability levels. Non uniform DIF occurs when the discrimination parameters differ,$\alpha_{g1,j} \neq \alpha_{g2,j}$ across groups, therefore the probability of getting an item right for the two groups changes sign over the ability range. In IRT terms, non uniform DIF is indicated by two crossing item response functions while uniform DIF is displayed by two non-crossing item response functions.

Thissen et al. (1988) introduced model comparison measures which is implemented by a likelihood ratio test to detect DIF. They used marginal maximum

likelihood (MML). The likelihood for maximizing is marginalized with respect to θ and has the form,

$$L(\alpha, \beta) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} (1 - P_j)(\theta)^{(1-y_{ij})} dF(\theta_i)$$

The likelihood ratio test involves the comparison of two models, a reduced model under H_0 and a full model under H_1. The test statistics is

$$G^2 = -2 \log \left[\frac{\max_{\omega_0} L(\alpha, \beta)}{\max_{\omega_1} L(\alpha, \beta)} \right],$$

where ω_0 is the parameter space under H_0 and ω_1 is the parameter space under H_1. Its p-value is obtained by chi-square distribution.

Zwinderman (1991, 1997) modeled the relationship between observed variables and the latent trait in order to yield more accurate results for inference about the latent trait using just a few items. DIF analysis based on the IRT model examines whether IRFs perform differently for manifest variables such as race or gender by allowing the item parameters to vary for each group. DIF could be viewed as a misfit of the IRT model because the existence of DIF implies that there is another factor that influences the item response probabilities (Glas, 2001; Thissen et al., 1993). Also using background variables could be a useful way to investigate why an observation occurs. For example, differences of item response functions between races could be explained by social-economic status (Glas, 2001; Rogers et al., 1999).

Muthen (1985, 1989) extended the item response model to a structural equation model with observed covariates called multiple indicator multiple covariates (MIMIC). Skrondal and Rabe-Hesketh (2004) applied it to DIF analysis. First, they considered a model without item bias, a MIMIC model where a factor is regressed on covariates. Second, they considered a model incorporated with item bias, a MIMIC model where the covariates have a direct effect on items and the factor. Since two models are nested each other, likelihood-ration tests can be performed. In Chapter 4, this study shows the DIF analysis with real data in a similar way.

One of recent approach is mixture IRT models (Rost, 1990, 1991). In these models, the IRFs not only depend on the continuous latent variable θ, but also on latent class membership ξ, however the OLCA models assume that the IRFs are conditionally independent within an ordered class. Since the mixture IRT models fit an IRF within a latent class that has nominal measurement level, the item parameters (α, β) vary by latent classes. This approach is viewed as DIF modeling with unobserved grouping variables.

2.7 Model Selection

In fitting latent class models to data, determining the number of latent classes remains a challenge to analysts (Bandeen-Roche et al., 1997; Nylund et al., 2007). In general, latent class model proceeds by starting the most parsimonious model and fitting successive models with an increasing number of classes to determine the most parsimonious model that provides an adequate fit to the data (Lin and Dayton, 1997; Nylund et al., 2007). As criteria to the decision of the number of classes in mixture modeling, Akaike's information criterion (AIC) and the Bayesian Information Criterion (BIC) are widely used (Anderson, 1982).

AIC is a measure of the goodness of fit of a model that considers the number of model parameters (p) being estimated in the model.

$$AIC = -2 \cdot \ln L + 2 \cdot p.$$

AIC is a information criterion for ordering alternate models for data. The individual AIC values are not meaningful and are much affected by sample size. Only those differences in AIC are interpretable as to the strength of evidence.

$$\triangle_i = AIC_i - AIC_{min},$$

where AIC_{min} is the minimum of the possible AIC_i values. Some rules of thumbs are often useful in assessing the relative merits of models in the set: Models having

$\triangle_i \leq 2$ have substantial supports, those in which $4 \leq \triangle_i \leq 7$ have considerably less support, and the models having $\triangle_i \geq 10$ have no support (Anderson, 1982)

BIC is a measure of the goodness of fit of a model that considers the number of parameters (p) and the number of observations(N).

$$BIC = -2 \cdot \ln L + (\ln N) \cdot p.$$

As with AIC, the model with the smallest value of BIC among all possible models is selected.

The BIC applies larger penalties per parameters of $\ln(N)$ than AIC , thus other factors being equal, BIC tends to select simpler models than AIC.

In addition to selecting the number of latent classes, another consideration for model comparison is the inference on group effect. By comparing the OLCA model with OLCR model fitted to the same data set, the effects of covariates are investigated. Since two models are completely nested, likelihood ratio test can be used as well as AIC and BIC.

Model selection bias and model selection uncertainty are important issues that deserve a further investigation. Monte Carlo simulation study is conducted to assess the probability of finding true models in Chapter 5.

In this chapter a literature review of measurement models suitable for educational assessment and their limitations and strengths was discussed. The next chapter discusses model estimation procedure for the OLCR model by maximizing likelihood subject to inequality constraints.

Chapter 3

Inference in the Ordered Latent Class Regression Models

3.1 Introduction

This study focuses on an extension of Croon's ordinal latent class analysis (OLCA) model to a regression model that explains the relationship between a covariate and latent class probability. Section 3.2 describes the algorithm to estimate the ordered latent class regression (OLCR) model, and Section 3.3 provides the outline of data analysis by means of this model.

3.2 Estimation of Ordered Latent Class Regression Models

The log-likelihood function for the OLCR model is identical to that of the OCLA model except that the function $\Pr(\xi_i = x|Z_i)$ replaces $\Pr(\xi = x)$. OLCR models estimate all the parameters, π and p simultaneously by maximizing the log-likelihood

function as follows:

$$\ln L = \sum_{i=1}^{N} \log \sum_{t=1}^{T} \Pr(\xi_i = x|Z) \prod_{j=1}^{J} \prod_{k=1}^{K_j} Pr(Y_{ij} = k|\xi_i = x)^{\delta_{ijk}}$$

$$\delta_{ijk} = \begin{cases} 1 & \text{if } y_{ij} = k \\ 0 & \text{if } y_{ij} \neq k \text{ or } y_{ij} \text{ is missing} \end{cases}$$

$$\pi_x = P(\xi = x|Z_i) = \frac{\exp(Z_i\gamma_x)}{\sum_{x=1}^{T} \exp(Z_i\gamma_x)},$$

where π_x is the prior probability that a randomly chosen individual with covariate Z_i belongs to class x; p_{jkx} is the probability of a response in category k on item j for a person in latent class x.

The complete likelihood denoted by $\log L^c$ can be written as:

$$\log L^c(\theta|Y, \xi) = \sum_{v} \sum_{x} n_{vx} \ln p_{vx},$$

where $n_{\nu x}$ denotes the number of subjects in the sample who have patterns ν and fall in class x and p_{vx} denotes the unobserved probability of falling simultaneously in the categories denoted by vector ν and the latent class x. Since x is not observable, $n_{\nu x}$ is not an observable quantity.

$$n_{\nu x} = n_x \cdot p_{\nu|x}, \tag{3.1}$$

where n_x is the number of individuals in class x, and $p_{\nu|x}$ is the conditional probabilities of response pattern ν given class x. The complete likelihood can be split into two parts: conditional response probabilities p_{jkx} and the latent class proportions π_x, so that they can be optimized independently :

$$\begin{aligned} \log L^c(\theta|Y, \xi) &= \sum_{\nu} \sum_{x} n_{\nu x} \log p_{\nu x} \\ &= \sum_{\nu} \sum_{x} n_{\nu x} \log [\pi_x + p_{jkx}] \\ &= \sum_{\nu} \sum_{x} n_{\nu x} \ln \pi_x + \sum_{\nu} \sum_{x} n_{vx} \ln p_{jkx}, \end{aligned}$$

where $n_{\nu x}$ denotes the number of subjects in the sample who have patterns ν and fall in class x.

The log-likelihood is maximized by the following EM algorithm to solve model parameters.

1. In the E-step,the expected values of the sufficient statistics of the unobservable complete data are computed given the observed complete data and parameter estimates.

$$Q(\theta|\theta^{t-1}, x) = E_{x|y,\theta^{t-1}}[\log L^c(\theta|x, y)],$$

   where θ is parameter vectors and x is unobserved latent class and y is observed data.

2. In the M step, parameters are updated by maximizing the likelihood of the complete data, considering them as if they are observed. The item response parameters p_{jkx} and the regression parameters γ are separately estimated.

$$Q(\theta^t) = \arg\max_{\theta} Q(\theta|\theta^t)$$

3. Using these new model estimates, another E-step can be performed to obtain new estimates for the complete data, and so on. The algorithm iterates until convergence.

Latent Class Regression Parameters, γ

In a latent class regression model, π_x is reparameterized as shown so that multinomial logit model can be fitted using the Newton-Raphson method (Bock, 1975; Heijden et al., 1996):

$$\pi_x = Pr(\xi = x|Z_i) = \frac{\exp(Z_i\gamma_x)}{\sum_{x=1}^{T}\exp(Z_i\gamma_x)},$$

where i is an index for individual and γ is a parameter relating the explanatory variable Z to the latent class ξ.

Newton-Raphson Method Newton's method, or the Newton-Raphson method, is a quadratic numerical approximation method to find the value of θ that minimizes $f(\gamma)$. Taylor series approximations are the first place that the Newton method is derived.

It can be shown that as h goes to 0 the higher-order terms in the equation above go to 0 much faster than h goes to 0. This means that

$$f(x+h) \approx f(x) + f'(x)h$$

This is referred to as a first - order Taylor approximation of f at x. First, one constructs a quadratic approximation to the function of interest around some initial values for parameters. Next, one adjusts the parameter value to that which maximizes the quadratic approximation. Then, one iterates until it converges:

$$\hat{\gamma_{new}} = \hat{\gamma_{old}} + (-D_\gamma^2 \log L)^{-1} D_\gamma \log L,$$

where D_γ is the gradient and D_γ^2 is the Hessian matrix with respect to γ.

Due to practical considerations, the Hessian matrix H is sometimes substituted by an estimated $\tilde{H}$ that is an approximation to the Hessian matrix by using the negative of the identity matrix for the maximization problem, and this technique is called a quasi-Newton method.

Item Response Probabilities, p_{jkx}

The formulation for p_{jkx} in the standard latent class model (or unconstrained model) is as follows:

$$\hat{p}_{jkx} = \frac{n_{jkx}}{n_{j+x}} \tag{3.2}$$

However, the parameters, p_{jkx} of the OLCA model are estimated subject to inequality constraints. As for the parameters, p_{jkx}, there are two kinds of constraints: the

T-1 equality constraints, noted as:

$$\sum_{k}^{K} p_{kt} = 1 \tag{3.3}$$

the (T-1)(M-1) inequality constraints,for all $x : 1 \leq x \leq T$

$$\sum_{k=1}^{K-1} p_{kt} \leq \sum_{k=1}^{K-1} p_{k,t-1} \tag{3.4}$$

for $k : 1 \leq k \leq K-1$ and $x : 1 \leq x \leq T-1$. If cumulative response probabilities q_{kt} are defined as:

$$q_{kt} = \sum_{g=1}^{k} p_{gt}$$

then, the set of inequalities is equivalent to the following one:

$$q_{kt} \leq q_{k,t-1}$$

If the estimates $\hat{p}_{jkx}$ obtained as in Equation (3.2) meet the inequality constraints, $\hat{p}_{jkx}$ become our estimates. Otherwise the estimates $\hat{p}_{jkx}$ that satisfy the constraints are computed. The objective function is

$$f(p) = \sum_{k=1}^{K} \sum_{t=1}^{T} n_{kx} \ln p_{kt}.$$

$f(p)$ is maximized subject to the constraint conditions $g(p)$.

Difficulties often arise when one wishes to maximize or minimize a function subject to constraints. The method of Lagrange multipliers is used to incorporate these constraints into the new objective function.

El Barmi and Dykstra (1994) formulated how to maximize a multinomial likelihood under these two restrictions (see Equations 3.3 and 3.4.). Dystra's algorithm can be applied to our case so that T latent classes are assumed independent multinomial random variables under order restriction. El Barmi and Johnson (2006) described the detailed procedure on how constrained MLE estimates of p_{kx}

can be computed using Lagrange multipliers. Figure 3.1 and Figure 3.2 displayed two cases that the stochastic orderings of latent classes do not hold: within an item and between items.

Ordering Classes within an Item To fulfill the monotonicity assumption of the model, IRFs are estimated subject to equality constraints. As the order of the latent class increases, the probability of getting a response less than k decreases for all k values. Figure 3.1 shows a case in which the unconstrained estimates of IRFs do not satisfy ordering among classes within an item. The right panel shows how the unconstrained estimates are changed to be the constrained estimates.

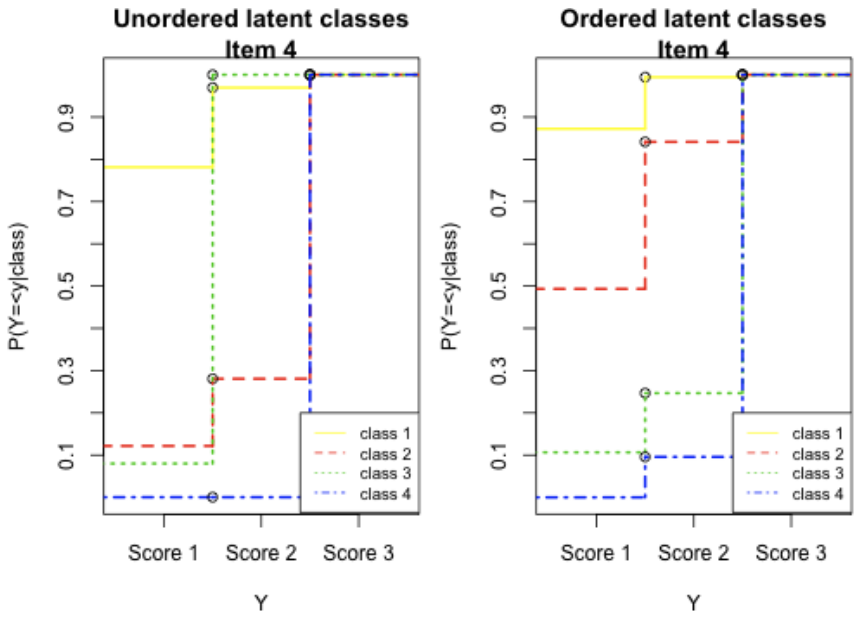

Figure 3.1: Comparison of ISRFs within an Item

Ordering Classes across Items Figure 3.2 (See pp. 37) shows a case in which ordering among classes is not held across items. In the figure, four step lines do not cross one another in Item 5 and Item 6, but the order of classes in Item 5 is

different from that of item 6. For example, the green line indicates Class 2 in Item 5 and Class 3 in Item 6.

3.2.1 Local Identifiability

The standard latent class model has $T - 1 + \sum_j^J T(K_j - 1)$ parameters, where T is the number of latent classes, J is the number of items, and K is the maximum categories of responses.

Including P covariates increases the number of parameters to $(T-1)(P+1)+\sum_j^J T(K_j-1)$. For latent class regression model, $(T-1)\times(P+1)$ parameters are used for π_x instead of $T-1$. The restriction for model identification is

$$T(1 + M - J) < Q + 1$$

$$T < \frac{Q+1}{1+M-J},$$

where J is the number of items. K_j is the number of categories on item j. $M = \prod_j K_j$ is the possible response patterns. Q is the number of unique response patterns. T is the number of latent classes.

A necessary condition for the model to be identifiable is that the number of degrees of freedom is not negative. The number of parameters to be estimated cannot be greater than the number of observed frequencies. However, this condition is nevertheless insufficient. It is a sufficient condition for local identifiability in the latent class model. Testing local identifiability can be carried out by calculating the rank of the information matrix (or its inverse - e.g., the estimated asymptotic variance-covariance matrix) for the parameters in the model. It should be of full column rank (Goodman, 1974; Heinen, 1993, 1996). Formann (1985) stated that if all eigenvalues of the matrix of second order derivatives are negative, the model is locally identifiable. That is the observed information matrix should be positive

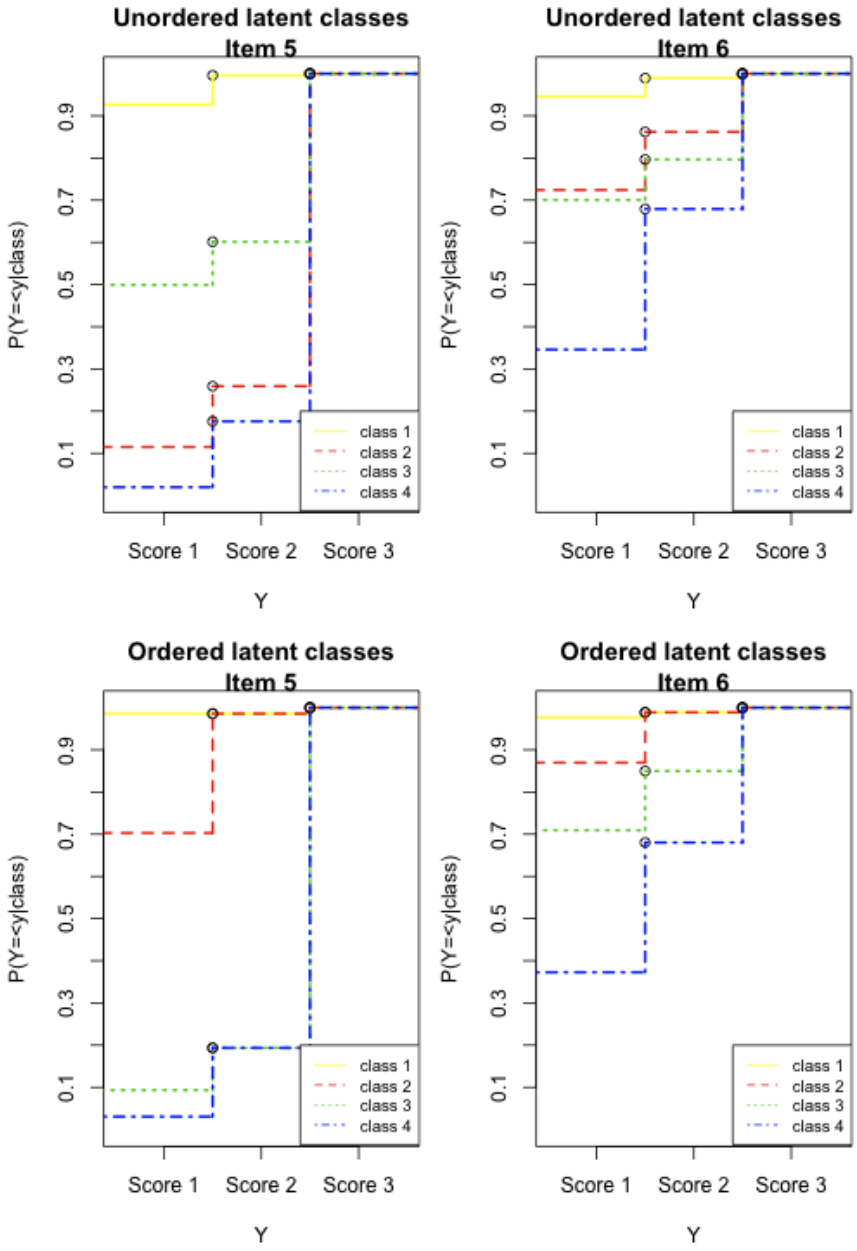

Figure 3.2: Comparison of ISRFs between Items 5 and 6

definite for the identification of the model (i.e., all eigenvalues of this matrix should be greater than zero).

Chapter 4

Real Data Analysis

4.1 Introduction

The ordered latent class regression is applied to real data from the Trends in International Mathematics and Science Study (TIMSS) to investigate the differences of groups and item analysis. In this study, gender and country differences are explained by the difference in latent class probability across genders and country using the ordered latent class analysis model.

One of the advantages of this approach is that group comparison and item analysis can be done simultaneously. An alternate estimation procedure has three steps: estimate the basic latent class model, calculate the predicted posterior class membership probabilities, and then regress these values as the dependent variable on the covariates. Bolck et al. (2004) showed the alternate procedure yields biased coefficients.

Another advantage is that the model can separate group differences from item effects that otherwise might be confounded. Differential item function (DIF) detection can be confounded when the distributions of the groups are different. For example, suppose that the distribution of men's math ability has a bigger variance

and a higher mean than the distribution of women's ability; then, items can be determined to be DIF items due to group differences even if they are good items. The error associated with deeming an item as DIF is known as Type I error.

This chapter presents an analysis of TIMSS data, answering the following research questions.

1. Which items are reliable indicators of the underlying latent variable? Compared to GPCM, what are the contributions of OLCA for analysis of items ?

2. Is incorporating covariates (e.g. gender or country) into OLCA model useful to find group differences in term of the measured proficiency?

3. Do items perform differently across groups after taking into account the difference in the distribution of the latent variable ?

To illustrate the utility of the proposed ordered latent class model (OLCA), Section 4.2 and 4.3 present an application of the OLCA model fitted to TIMSS data to examine group differences and conduct item analyses. The results are compared with the results from fitting the generalized partial credit model (GPCM). This allows us to illustrate the merits of using the OLCA and examine the model fit. Since these two models are not nested, likelihood ratio tests are inappropriate. Model fit is determined by AIC and BIC; models with lower AIC (or BIC) are considered better. Section 4.2 concentrates on group comparison using both GPCM and OLCA. Group effects are determined by the likelihood ratio test (LRT), and AIC. Section 4.3 focuses on item analysis.

Data Description

TIMSS provides data on the mathematics and science achievement of U.S. fourth- and eighth-grade students compared to that of students in other countries. TIMSS

is designed to align broadly with mathematics and science curricula in the participating countries. The data were collected in 2007 for the fourth time: results from 36 countries at grade 4 and 48 countries at grade 8 were collected.

The sample used in this study consists of 2,032 eighth graders from six countries (Chinese Taipei, Korea, Singapore, U.S., Australia, and England) who completed booklet 14 from the 2007 TIMSS. Each booklet contains 30 items on mathematics consisting of 16 multiple choice items and 14 constructed-response (CR) items. Eight CR items are dichotomous and six CR items are polytomous, scored with three categories (0 - 2). This analysis focuses on the six polytomous items. Table 4.1 provides the cognitive skills and topic of the six items that are analyzed in this study.

Table 4.1: Six Items Used for the Analysis

Items	ID	Cognitive Skill	Topic
1	M022232	Applying	Numbers
2	M022234A	Applying	Geometry
3	M022234B	Applying	Numbers
4	M042302A	Applying	Numbers
5	M042302B	Applying	Numbers
6	M042302C	Reasoning	Numbers

4.2 Analysis for Comparisons of Countries

International studies in education provide the information on how students from different countries with similar and dissimilar educational environment perform on a test, and researchers investigate the factors that influence students achievement. Figure 4.1 displays the mean raw scores of male and female students by country. Male students outperformed female students on average in the U.S. and Australia,

whereas female students outperformed male students on average in Singapore, Korea, Chinese Taipei, and England.

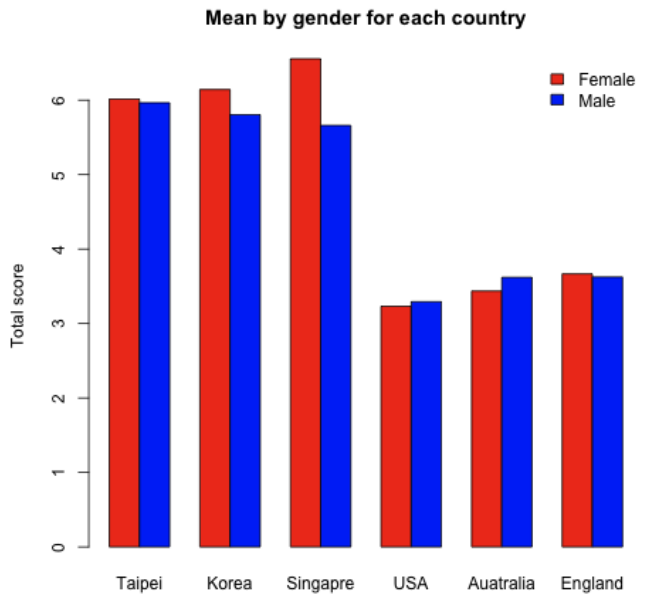

Figure 4.1: Mean Raw Score by Gender Within Country

Cross-country comparisons based on TIMSS typically are reported in terms of means and variances of the scale scores by country, and the ranks are reported as mean scores of national achievements (see Olson, Martin, and Mullis, 2008). However, it is worth investigating why the average achievement of the U.S. students is lower than that of students in other countries ranked above the U.S. It could be due to an absence of the best-performing students or an abundance of poorly performing students. Betts and Grogger (2003) conducted a quantile regression and found that imposing higher standards has a positive effect on the higher test scores but a negative effect on the lower test scores. The proposed model serves as an alternative to quantile regression in that it considers the whole range of distributions of ability across the grouping variables (e.g., gender, country) so that a comparison by level

of proficiency can be drawn. This model provides specific and useful information to policy makers and curriculum developers on student achievement and the extent to which the U.S. educational system meets the standard.

Objectives of the Analysis of Group Comparisons

The proposed model is applied to investigate how the math performance of U.S. students differs from that of students from other countries and to examine whether there is a gender difference in terms of math ability measured by six polytomous items.

First, the analysis is performed by country to see whether there is a gender effect within that country. Then, the data sets are combined to compare countries and gender in a model. The combined data set consists of six countries with a sample size of 2,032.

Investigation of Gender Effect on Math Achievement within Each Country

This section illustrates the results of the OLCA and the GPCM to examine a gender effect within each county based on the AIC, BIC, and the likelihood ratio test (LRT) and compares these two models in terms of model fit. It starts with U.S. and then goes through other counties.

U.S.: As displayed in Table 4.2, we fit the GPCM and the OLCA with gender as a covariate. First the OLCA model is fitted to the data and the number of latent classes is varied to determine how many latent classes are considered adequate to describe the data (AIC = 4075.04). Then the ordered latent class regression (OLCR) is fitted to the data using the gender covariate. Female is coded as a reference. The addition of the covariate does not improve the model fit (AIC=4,078.65).

BIC also favors the same model, 3-class OLCA with a value of 4237.35, compared with 3-class OLCR (BIC =4,249.49). Therefore, we found no gender effect with the U.S. data based on AIC and BIC using the OLCA. According to the table, the OLCA model without gender fits better for 3 and 4 latent classes based on AIC; however, for 5 latent classes, the model with gender fits better than without gender. As the number of latent classes increases, the impact of the gender effect could change. Moreover, as the number of latent classes increases, the number of components to be tested also changes. That is, the number of parameters for the class probabilities, π, in the 5-latent class model is 8, whereas for the 3-latent class and the 4-latent class models, have 4 and 6 parameters respectively.

In addition to the OCLA, a gender effect was examined using the GPCM with and without gender as shown in Table 4.2. The first model takes into account the mean difference of ability distributions between gender groups. The second model ignores the difference.. The GPCM with gender treats the gender variable as an item. This model requires recalculation of the likelihood to condition on groups (i.e., gender) by subtracting a constant from the full likelihood. Both AIC and BIC favor the model without gender shown in Table 4.2.

AIC and BIC are inconclusive when comparing the GPCM and the OLCA model. AIC for the best GPCM is 4,081.87 and AIC for the best OLCA model is 4,075.04. The change in AIC for the OLCA model (a decrease of 6.7 for an increase of 20 parameters) could indicate that the GPCM does not fit the data well. Conversely, the BIC value of 4,158.75 for the GPCM is smaller than the BIC value of 4237.35 for the OLCA model. Compared to AIC, BIC penalizes more on the number of parameters and tends to favor a simpler model. The number of parameters in the GPCM is 18 versus 38 parameters in the 3-latent class model, more than twice as many as the GPCM parameters.

Table 4.2: U.S.

Model	N class	N parm	log-likelihood	AIC	BIC
OLCA	T=2	25	-2,041.63	4,133.27	4,240.04
OLCA	T=3	38	-1,999.52	4,075.04	4,237.35
OLCA	T=4	51	-1,990.38	4,082.75	4,300.57
OLCA	T=5	64	-1,985.67	4,099.35	4,372.69
OLCR	T=3	40	-1,999.33	4,078.65	4,249.49
OLCR	T=4	54	-1,989.16	4,086.31	4,316.95
OLCR	T=5	68	-1,980.12	4,096.24	4,386.67
GPCM	no gender	18	-2,022.93	4,081.87	4,158.75
GPCM	gender	19	-2,022.91	4,083.82	4,164.97

Other Countries As with U.S. data shown in Table 4.2, the same analyses are executed for the other five countries; the results are presented in the Appendix. First, the analysis involves examining the gender effect in the OLCA model and the GPCM. Second, model fit between the OLCA model and the GPCM are compared.

Table 4.3 gives a summary of results of the best-fitted models based on the AIC criteria of the GPCM and the OLCA model. The best-fitted models for Korea, Singapore and Australia were the ordered latent class regression (OLCR) models (i.e., OLCA model with gender covariate). The best-fitted models for the remaining countries were the ordered latent analysis (OLCA) models (i.e., without the gender covariate). Gender differences were found in three countries (Korea, Singapore, and Australia). Since the original parameters (γs) are on the logit scale, they were converted into class probabilities by gender (see Table 4.4). For Korea, the 3-latent class model fits the data the best. The probability of being in Class 3 is higher for female students than male students; the probability of being in the Class 1 is also higher for female students than for male students. For Singapore, the 5-latent class model fits the data best. The probability of being in the highest class, Class 5, is

Table 4.3: Model Comparison by AIC

Country	OLCA				GPCM		
	Class	Covariate	AIC	N Parm	Covariate	AIC	N Parm
Chinese Taipei	3		2,636.27	38		2,640.01	18
Korea	3	gender	2,892.63	40		2,894.12	18
Singapore	5	gender	3,045.38	68	gender	3,059.92	19
U.S.	3		4,075.04	38		4,081.86	18
Australia	4	gender	2,314.30	54		2,314.41	18
England	3		2,393.03	38		2,425.60	18
ALL			17,356.65	276		17,415.92	109

higher for female students than for male students, and the probability of being in the lowest class is higher for male students than female students, which means the difference favors female students. For Australia, the 4-latent class model fits the data the best. The probability of being in the highest class is higher for female students than for male students and the probability of being in the lowest class is lower for male students than for female students.

Using the GPCM, Table 4.3 shows that the model with gender fits significantly better for Singapore, whereas the model without gender fits better for Chinese Taipei, Korea, U.S., England, and Australia. As mentioned earlier, the GPCM with gender consider only the difference in the means of male and female students. As shown in Figure 4.1, the mean difference between gender is the biggest for Singapore. Thus, using the GPCM leads to the conclusion that a gender difference was found only in Singapore on math achievement. Because the OLCA model considers the distribution of proficiency by groups, it is more likely to detect the group differences.

Table 4.5 shows the model selections by BIC. There is no gender effect found

Table 4.4: Class Probability by Gender

Country	Gender	Class 1	Class 2	Class 3	Class 4	Class 5
Korea	F	0.32	0.17	0.50		
	M	0.29	0.30	0.40		
Singapore	F	0.13	0.14	0.28	0.16	0.26
	M	0.21	0.21	0.11	0.30	0.14
Australia	F	0.34	0.26	0.22	0.15	
	M	0.29	0.18	0.41	0.10	

for all countries by means of either of GPCM or OLCA. In the table BIC indicates that 3-class OLCA model fits the best for only The U.S. and Korea. Table 4.3 and Table 4.5 shows the discrepancy of AIC and BIC model selection. For example, from AIC model selection, the best fitted model was 5 class model for Singapore whereas from BIC model selection only 2 class model was the best fitted model for Singapore.

In addition to AIC and BIC, Table 4.6 gives the likelihood ratio test (LRT) for the gender effect, γs for each country. Under the null, all γs are equal to zero, which implies that the class probabilities are the same between genders. Table 4.6 shows that the test statistics, G^2 are greater than the critical values for Singapore and Australia, suggesting a significant gender effect. When comparing the model fit of the OLCA model and the GPCM, the AIC values for the OLCA are smaller than those of the GPCM in all five countries except Australia, indicating that the OLCA fits the data better than the GPCM. For Australia, the difference between the AIC values of two models is small (0.11), therefore, the GPCM is negligibly better.

Table 4.3 lists the best-fitted models by each country by the GPCM and the OLCA. In the table the AIC value of OLCA labeled as ALL was obtained by

Table 4.5: Model Comparison by BIC

Country	OLCA				GPCM		
	Class	Covariate	BIC	N Parm	Covariate	BIC	N Parm
Chinese Taipei	2		2740.11	25		2705.82	18
Korea	3		3034.29	38		2961.15	18
Singapore	2		3203.56	25		3129.71	18
U.S.	3		4237.34	38		4258.74	18
Australia	2		2455.16	25		2380.59	18
England	2		2525.30	25		2491.79	18
ALL			18195.78	176		17927.8	108

Table 4.6: Likelihood Ratio Test for Regression Parameters

Country	df	G^2	CV ($\alpha = 0.05$)	p-value
Chinese Taipei	2	1.28	5.99	0.52
Korea	2	4.16	5.99	0.12
Singapore	4	14.81	9.48	0.01
U.S.	2	0.39	5.99	0.82
Australia	3	8.22	7.81	0.04
England	2	0.47	5.99	0.79

adding all AIC values for the best-fitted models of all countries (2636.27 + 2892.63 + 3045.38 + 4075.04 + 2314.30 + 2393.03). The AIC value for the GPCM was obtained similarly (2640.01 + 2894.12 + 3059.92 + 4081.86 + 2314.41 + 2425.60). The AIC value of the OLCA is smaller than that of the GPCM by the 159.3 as shown in Table 4.3.

Model Comparison of the Combined Data

The prior analysis was done by country. Now attention is turned to analyze the data in aggregate form. The analysis was aimed at examining the effect of gender and country by fitting the five models (M1-M5) listed below. The inferences were drawn using AIC.

- M1: the OLCR without covariates

  $\log(\frac{\pi_{ti}}{\pi_{1i}}) = \beta_0$

- M2: the OLCR with gender

  $\log(\frac{\pi_{ti}}{\pi_{1i}}) = \beta_0 + \beta_1 \times \text{Gender}$

- M3: the OLCR with country

  $\log(\frac{\pi_{ti}}{\pi_{1i}}) = \beta_0 + \beta_2 \times \text{Country}$

- M4: the OLCR with gender and with all countries

  $\log(\frac{\pi_{ti}}{\pi_{1i}}) = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{Country}$

- M5: the OLCR with gender, countries, and interaction

  $\log(\frac{\pi_{ti}}{\pi_{1i}}) = \beta_0 + \beta_1 \times \text{Gender} + \beta_2 \times \text{Country} + \beta_3 \times \text{Interaction}$

From Table 4.7, it is clear that the AIC favors M4 that is the 5-class OLCR model with gender and country and without the interaction of these two variables. BIC indicated that M3 with 3 class OLCR model with country only describes the data

best. Except M1, BIC indicates 3 class solution as the best from M2 to M5. For gender effect, female is coded as a reference; for country effect Chinese Taipei is coded as a reference.

The estimates of regression parameters, γs and the standard errors of M4 and M5 can be found in the appendix. Figure 4.2 was derived from the estimates of Table A.1 in Appendix displaying the estimated class probabilities, $\hat{\pi}_t$, that an individual belongs to each classes. Substantial differences were found in the estimated class probabilities across countries. Unlike Figure 4.1, which compares only the mean differences by group, Figure 4.2 allows us to compare them by levels: ordered latent classes. Singapore has the highest proportion in Class 5, followed by Korea. The U.S. has the highest proportion in Class 1, followed by England and Australia. The majority of Chinese Taipei students belong to Class 4 in Figure 4.2. Compared to the cross-country difference, gender difference seems to be relatively small. The proportion of female students is larger than that of male students in classes 1 and 3 across countries.

Comparison between TIMSS Report and our OLCA Estimation Result

As a way of interpreting the achievement result, TIMSS uses four points on the scale as International benchmarks and describe achievement at those benchmarks in relation to students' performance on the test question (Mullis et al., 2008). According to TIMSS 2007 report, each of the anchoring points corresponds to the 90th percentile, the 75th percentile, the 50th percentile and 25th percentile, respectively, on the international benchmark. The four anchoring points in the TIMSS report correspond to imposing five classes on the scale of proficiency score derived from an IRT model (see Table 4.8). According to this classification, different countries have different probabilities of belonging to each class. For example, only 6% of U.S. students belong to the top class, a figure that is considerably lower than

Table 4.7: Model 1 to Model 5

Model	N class	N parm	log-like	AIC	BIC
M1	2	25	-9,098.07	18,246.14	18,386.56
	3	38	-8,711.25	17,498.50	17,711.81
	4	51	-8,660.29	17,422.58	17,708.86
	5	64	-8,643.76	17,415.52	17,774.78
M2	2	26	-9,097.82	18,247.65	18,393.69
	3	40	-8,861.40	17,802.81	18,027.48
	4	54	-8,811.05	17,730.11	18,033.41
	5	68	-8,790.89	17,717.78	18,099.72
M3	2	30	-8,991.54	18,043.09	18,211.59
	3	48	-8,673.83	17,443.67	17,713.27
	4	66	-8,637.24	17,406.48	17,777.19
	5	84	-8,579.49	17,326.99	17,798.80
	6	102	-8,560.95	17,325.91	17,898.83
	7	120	-8,544.90	17,329.81	18,003.82
M4	2	31	-8,991.31	18,044.63	18,218.75
	3	50	-8,669.26	17,438.53	17,719.37
	4	69	-8,633.86	17,405.72	17,793.28
	5	88	-8,572.93	17,321.87	17,816.15
	6	107	-8,555.12	17,324.25	17,925.24
	7	126	-8,537.91	17,327.83	18,035.54
M5	2	36	-8,988.34	18,048.68	18,250.88
	3	60	-8,666.13	17,452.26	17,789.27
	4	84	-8,623.96	17,415.92	17,887.73
	5	108	-8,567.61	17,351.22	17,957.83
	6	132	-8,546.49	17,356.99	18,098.40
	7	156	-8,526.00	17,364.00	18,240.22

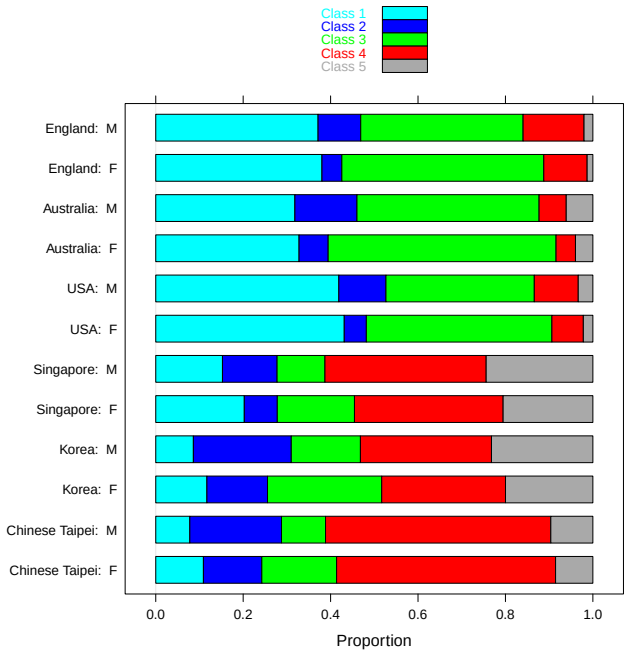

Figure 4.2: Class Probabilities by gender and country in M4

the benchmark of 10 %. However, as many as 45 % of students in Chinese Taipei belong to this top class.

Comparing the classes derived from the OLCA model in Table 4.9, Class 1 from the OLCA model roughly corresponds to the bottom two intervals (Below low and Low-Intermediate), Class 2 corresponds to the second interval (Intermediate-High), Class 3 corresponds to the third interval (High-Advanced) and Class 4 and 5 correspond to the top interval combined.

Given that the OLCA classes attempts to distinguish students across their level of proficiency in discrete way using all available information on response patterns of students and items, one could argue that the classes estimated in the OLCA model are better positioned than the arbitrary intervals used in the TIMSS report. In this example, the second interval (Low -Intermediate) and the bottom interval (Below Low) are not significantly distinguishable from each other; this separation

Table 4.8: Percentages of Students Reaching International Benchmarks by Country

Country (cut-points)	Below Low ≤ 400	Low-Inter 400	Inter-High 475	High-Advanced 550	Above Advanced 626
Chinese Taipei	5	9	15	26	45
Korea	2	8	19	31	40
Singapore	3	9	18	30	40
U.S.	8	25	36	25	6
England	9	22	34	27	8
Australia	11	28	37	18	6

Table 4.9: Class Proportion from OLCA by Country

Country	Class 1	Class 2	Class 3	Class 4	Class 5
Chinese Taipei	7	19	14	51	9
Korea	9	17	23	29	21
Singapore	17	11	15	35	21
U.S.	41	10	38	9	2
England	32	9	49	5	5
Australia	35	11	41	12	1

of classes would have been better used to divide the top interval into two distinct classes.

Estimation of discrete classes in OLCA models would have an advantage over an arbitrary separation of students into classes using certain ad hoc choices of intervals when a researcher tries to separate students for cross-country comparisons or to create a selection threshold for admission purposes.

4.3 Item Analysis

Objectives of the Analysis

In the last section, the group differences in the class probabilities were investigated. This section focuses on detecting the group differences in the IRFs. This section highlights the application of the GPCM, OLCA and OLCR models to TIMSS data to assess group differences in IRFs. The two fitted OLCR models as follows:

1. An OLCR model with a covariate in both the class probabilities and the IRFs, denoted by M_d

2. An OLCR model with a covariate relating only to the class probabilities, denoted by M_s

Figure 4.3 graphically represents the mean scores of the six items for six countries. It is interesting to note that the profiles of Chinese Taipei, Korea, and Singapore are similar to one another and the profiles of the U.S., Australia, and England are similar to one another. This finding suggests that cultural or environmental factors might influence whether students in the groups answer the items correctly.

Table 4.10 shows the raw mean scores for the items by country. Based on the total mean scores of items, the rank of the items in the order of difficulty is given by: 4-5-2-3-6-1, with Item 4 being the easiest and Item 1 being the most difficult item.

Item Analysis without a Group Covariate in IRF

This section illustrates the results of the GPCM and OLCR without a group difference in the IRF and compares the results.

Table 4.10: Mean Scores of Items by Country

Country	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Chinese Taipei	0.52	0.95	1.16	1.42	1.38	0.55
Korea	0.49	1.01	0.96	1.25	1.40	0.86
Singapore	0.60	1.08	1.04	1.31	1.26	0.78
U.S.	0.34	0.40	0.28	1.00	0.91	0.32
Australia	0.28	0.37	0.25	1.17	1.03	0.43
England	0.40	0.54	0.24	1.10	0.96	0.40
Total	0.43	0.70	0.62	1.19	1.13	0.53

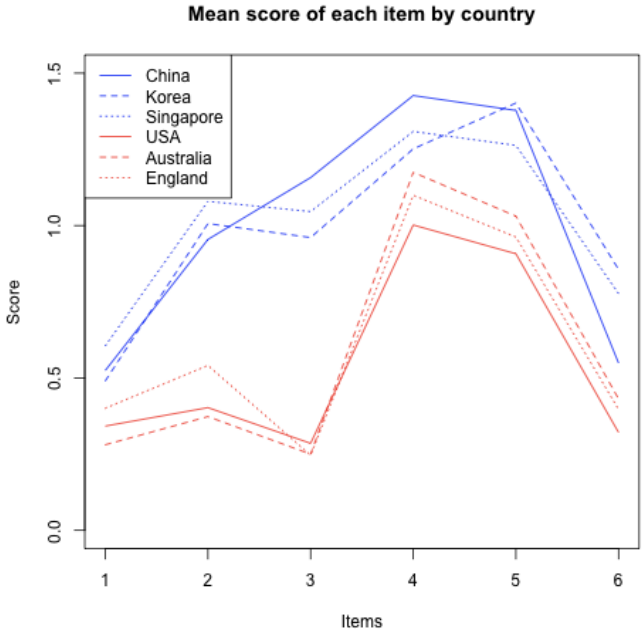

Figure 4.3: Profile of Mean Scores of Six Items

Table 4.11: Item parameter estimates of GPCM and Category Proportion

Item	GPCM Parameters			Proportions of Positive Response in Each Category		
	β_1	β_2	α	Category 1	Category 2	Category 3
1	4.57	-2.20	0.78	0.77	0.02	0.20
2	0.95	0.06	1.42	0.58	0.13	0.28
3	2.87	-1.62	1.20	0.67	0.02	0.30
4	-0.16	-0.41	1.62	0.31	0.17	0.50
5	0.15	-0.51	1.75	0.37	0.12	0.50
6	1.88	0.09	0.86	0.66	0.13	0.20

GPCM

Table 4.11 provides the parameter estimates of the GPCM. The GPCM provides two aspects of item characteristics: thresholds, β (interpreted as difficulty) and discrimination, α. Since the analyzed items have three categories, the GPCM estimates one discrimination and two thresholds parameters per item.

$$P_{jk}(\theta) = \frac{\exp \sum_{c=0}^{k} \alpha_j(\theta - \beta_{jc})}{\sum_{r=0}^{m_j} \exp \sum_{c=0}^{r} \alpha_j(\theta - \beta_{jc})}$$

where $\sum_{c=0}^{0} \alpha_i(\theta - \beta_{ic}^{*}) = 0$.

For the discrimination parameter, α, Item 1 and Item 6 have lower values than the rest of items. β_k values are not necessarily ordered such that $\beta_{i1} < \beta_{i2} < \cdots < \beta_{im}$, because the parameter represents the relative magnitude of the two adjacent probabilities. For threshold parameters, the estimate of Threshold 1 for Item 1 is 4.57 , which is bigger than the rest of parameter estimates. It corresponds to the proportion of Category 1 on Item 1 which is the biggest proportion as shown in the Table 4.11. Category 1 is an incorrect response on the item; therefore, an item having the biggest proportion of Category 1 responses is the most difficult item.

OLCA

Table 4.12 provides the estimates of IRFs of M4. The first column contains the category of the response and the second column contains the latent class. The IRF of getting Category 1 to Item 1 given Class 1 is 0.97 ($p_{111} = 0.97$). The italicized estimated response probabilities are retained to be the same by forcing the order relation on the latent classes in Table 4.12. To satisfy the order condition, $6 \times 2 \times 3$ inequality constraints are imposed on these item response parameters. In the right panel of Figure 4.5, the y-axis represents the cumulative response probabilities that students scored category 2 or category 3 given latent class and the x-axis represents the six items. From Table 4.12 and Figure 4.5, the five latent classes are ordered along the latent continuum, with Class 1 representing the most 'poorly performing' group and Class 5 representing the most 'competent' group within the entire set of latent classes. For Class 3 the IRF vector is (Item 1=0.18, Item 2=0.15, Item 3=0.08, Item 4=0.77, Item 5=0.74, Item 6=0.19). This class appears to contain the students who have problems with Items 1, 2, 3, and 6 but find Items 4 and 5 relatively easy. For Class 4 students, the IRF vector is (Item1=0.33, Item 2=0.63, Item 3=0.83, Item 4=0.77, Item 5=0.75, Item 6=0.19). The students who belong to Class 4 have problems with Items 1 and 6 but find the rest of the items relatively easy. The IRFs for Item 6 show a big difference between Class 5 and Class 4. According to the TIMSS report, Item 6 belongs to reasoning on the cognitive domains; the rest of the items belong to applying. Therefore, members of Class 5 students can be assumed to be more proficient in reasoning than members of the Class 4. The differences among the six items shown with respect to the way in which their IRFs vary as a function of the ordinal latent classes.

Similar to the IRT models, two important aspects of this functional relationship should be pointed out: its overall level and its steepness. For example, to compare the way in which the IRFs of Item 1 and 2 vary as a function of the ordinal

Table 4.12: IRF estimates for M4

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.97	0.96	0.99	0.80	0.95	0.97
	2	0.86	0.76	0.92	0.56	0.50	0.74
	3	0.79	0.73	0.91	0.08	0.12	0.61
	4	0.64	0.08	0.11	0.08	0.12	0.61
	5	0.35	0.06	0.11	0.00	0.00	0.06
2	1	0.01	0.02	0.00	0.15	0.03	0.03
	2	0.03	0.12	0.00	0.39	0.37	0.12
	3	0.03	0.12	0.01	0.15	0.14	0.20
	4	0.03	0.29	0.06	0.15	0.14	0.20
	5	0.02	0.12	0.02	0.11	0.01	0.06
3	1	0.01	0.02	0.01	*0.05*	0.02	0.00
	2	0.11	0.11	*0.08*	*0.05*	0.13	0.14
	3	0.18	0.15	*0.08*	*0.77*	0.74	*0.19*
	4	0.33	0.63	0.83	*0.77*	0.75	*0.19*
	5	0.63	0.82	0.86	0.89	0.98	0.88

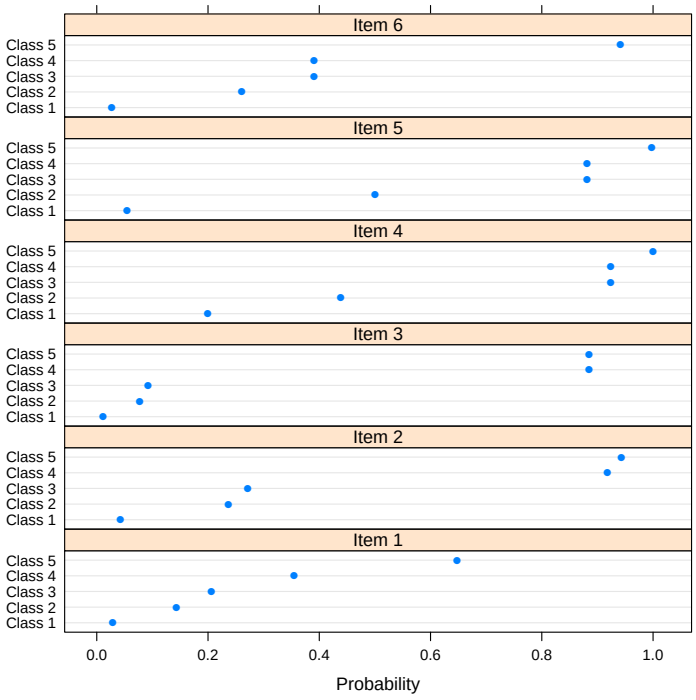

Figure 4.4: ISRFs of Item 6 in the Best-Fit model for Each Country

latent variable, we observe that, irrespective of which latent class we consider, Item 1 is responded to in a more negative way than Item 2.

The second aspect, the steepness of the functional relationship, has to do with the discriminatory power of the item. For example, to compare the cumulative response probabilities of Item 6 are not only higher than Item 5 in this respect we observe that the cumulative response probabilities of Item 5 are not only a higher level than those of Item 1, but they also change more drastically when one moves along the ordinal latent continuum.

The IRF of getting Category 3 to Item 1 given Class 5 is only 0.63 ($p_{135} = 0.63$), which is much smaller than the probabilities of the other items given Class 5 ($p_{235} = 0.82$, $p_{335} = 0.86$, $p_{435} = 0.89$, $p_{535} = 0.98$, $p_{635} = 0.88$). This finding means that the group of students in the highest level has only 0.63 probability of getting Category 3 on Item 1. The GPCM result is consistent with this finding showing that the lowest discrimination parameter estimate of 0.78 among those six items.

Compared to the GPCM, the flexibility of the OLCA provides different aspects regarding the IRF. For Items 6 the GPCM yields a low discrimination parameter estimate of 0.86; however, the result of the OLCA shows that it discriminates well between Class 4 and Class 5 as shown in Figure 4.4 and Table 4.12. The IRF of Item 6 given Class 5 is 0.88; however, given Class 4, the figure is sharply increased by 0.69. For Item 3, the GPCM yields the discrimination parameter estimate of 1.2, which is a medium size relative to the other items. But the OLCA estimates of IRFs show that Item 3 classifies the respondents into only two, not five groups (see Table 4.12). This finding provides evidence that the proficiency measured by Item 3 is assumed to be discrete rather than to be normally distributed. The results of the OLCA show that the items discriminate between the classes differently due to the full parameterization of IRFs.

Figure 4.5 displays the IRF estimates of the best-fitted model of each country

on Item 6. The last plot in this figure labeled ALL is the IRF estimates of M4, which was fitted with the combined data. For the U.S., Chinese Taipei, Korea, and England, 3-class OLCA were found to describe the data adequately. For Singapore and Australia, 5-class and 4-class OLCA respectively described the data adequately. This figure reveals that the classes are defined differently across countries so that comparison of performance are made of each class across groups.

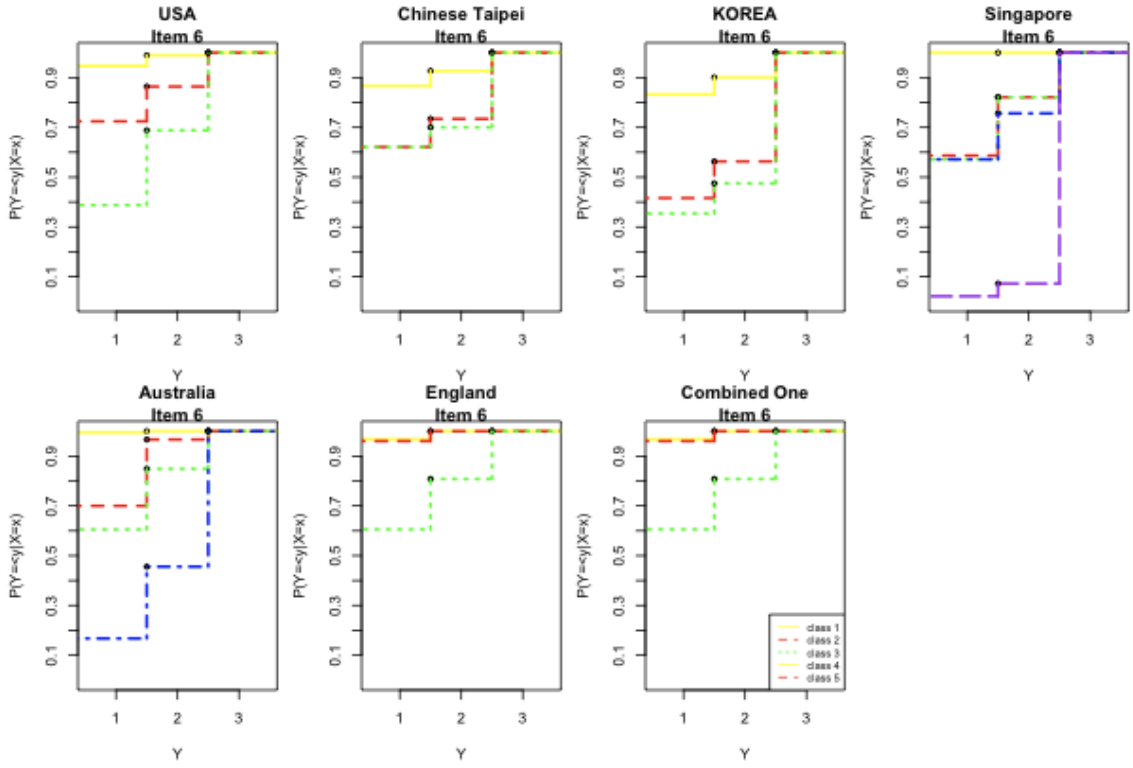

Figure 4.5: ISRFs of Item 6 in the Best-Fit model for Each Country

Group Differences in IRFs

So far we have investigated the merits of the OLCA for item analysis over the GPCM. The purpose of this section to demonstrate the use of OLCA to answer

this question, "Do the items in the test perform differently across groups ? ". Typically it is called the item bias or differential item functioning (DIF).

If a set of items favors one group over another, it is a potential threat of validity of test scores. Our analysis is based on the validation approach introduced by Muthen (1985, 1989). Diagrams of models M_s and M_d are shown as below. M_s is specified as a model with a coefficients of group effects on item responses set to zero. M_d incorporating item bias is specified as a model where the covariates have direct effects on item responses and in addition to indirect effect via latent class. Since M_s is clearly nested in M_d, likelihood-ratio tests can be performed as well as AIC and BIC.

As mentioned earlier, Figure 4.3 shows that the mean profiles of Chinese Taipei, Korea and Singapore are similar to one another and that those of U.S., Australia, and England are similar to one another. Using the OLCA and the GPCM, this analysis attempts to identify the source of the differences between the two groups to determine whether the class probabilities or the IRFs are different. For the analysis of item bias, we divided the six countries into two groups and investigated the source of the differences by comparing two models:

- Group1 : Chinese Taipei, Korea, Singapore
- Group2: the U.S., Australia, England

OLCA

Model M_s allows class probabilities, π, to be related to the covariate, z.

$$\Pr(Y_{i1} = y_1, ..., Y_{iJ} = y_J | z_i) = \sum_{t=1}^{T} \Pr(\xi_i = t | z_i) \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{jkt}^{y_{jk}}$$

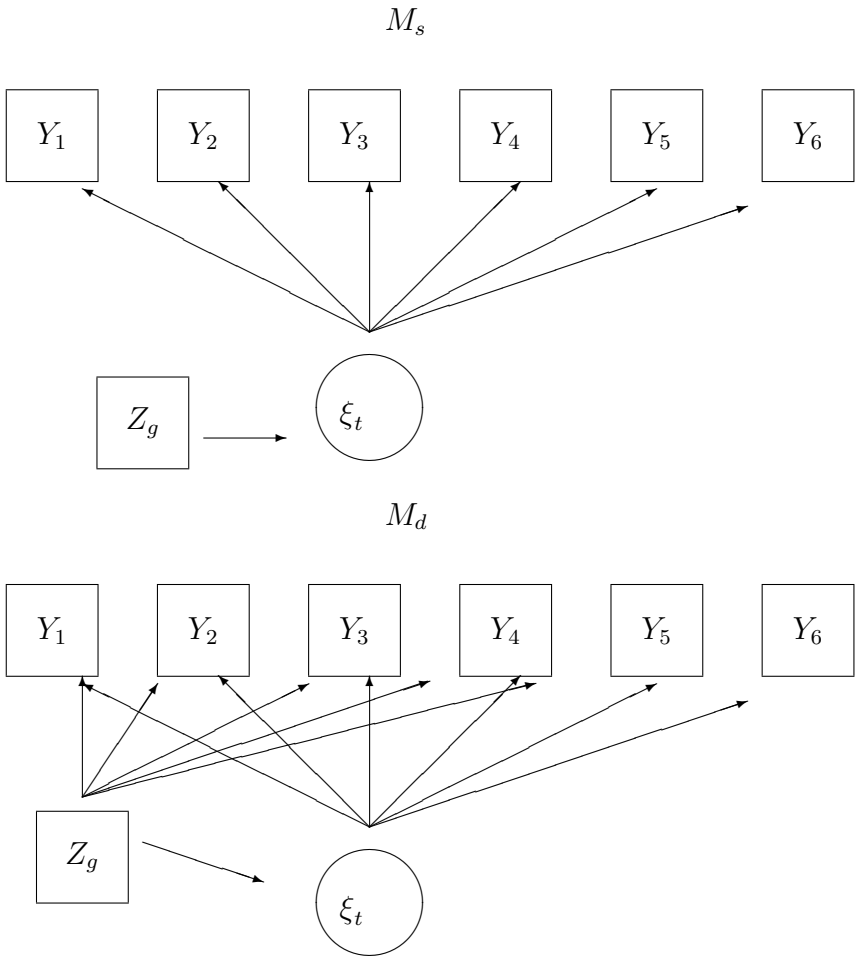

Figure 4.6: Diagram

Table 4.13: Model Comparisons for IRFs

Model	N Class	N Parm	Log-likelihood	AIC	BIC
M_s:	4	54	-8,648.90	17,405.81	17,709.11
	5	68	-8,598.76	17,333.53	17,715.47
	6	82	-8,583.64	17,331.29	17,791.87
M_d:	4	102	-8,533.52	17,271.06	17,772.80

Model M_d allows both class probabilities, π and IRFs, p_{jkt}, to be related to covariate, z.

$$\Pr(Y_{i1} = y_1, ..., Y_{iJ} = y_J | z_i) = \sum_{t=1}^{T} \Pr(\xi_i = t | z_i) \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{jkt}^{y_{jk}}(z_i)$$

The IRFs, p_{jkt}, are constrained to be the same across two groups in Model M_s, but to be different in Model M_d. The two models are compared by AIC and BIC as in Table 4.13. It shows that AIC indicates M_d fits better than M_s. It means that the group difference were found in the IRFs as well as in the class probabilities. However BIC indicates M_s with 4 latent class fit better than M_d.

In addition to the AIC and BIC, LRT for M_d and M_s was conducted using the values from Table 4.13, since two models are nested. A test statistic denoted by G^2 is computed by twice the difference between the log-likelihood values of the two models. Given that G^2 (230.75) was bigger than the critical value of 65.170, the null hypothesis of equal IRFs between the two groups was rejected. Both of the likelihood ratio statistics, the LRT test and AIC yielded the same conclusion. The model allowing different IRFs across two groups fits the data better.

$$H_0 : p_{(jkt,g=1)} = p_{(jkt,g=2)}$$

$$H_A : p_{(jkt,g=1)} \neq p_{(jkt,g=2)}$$

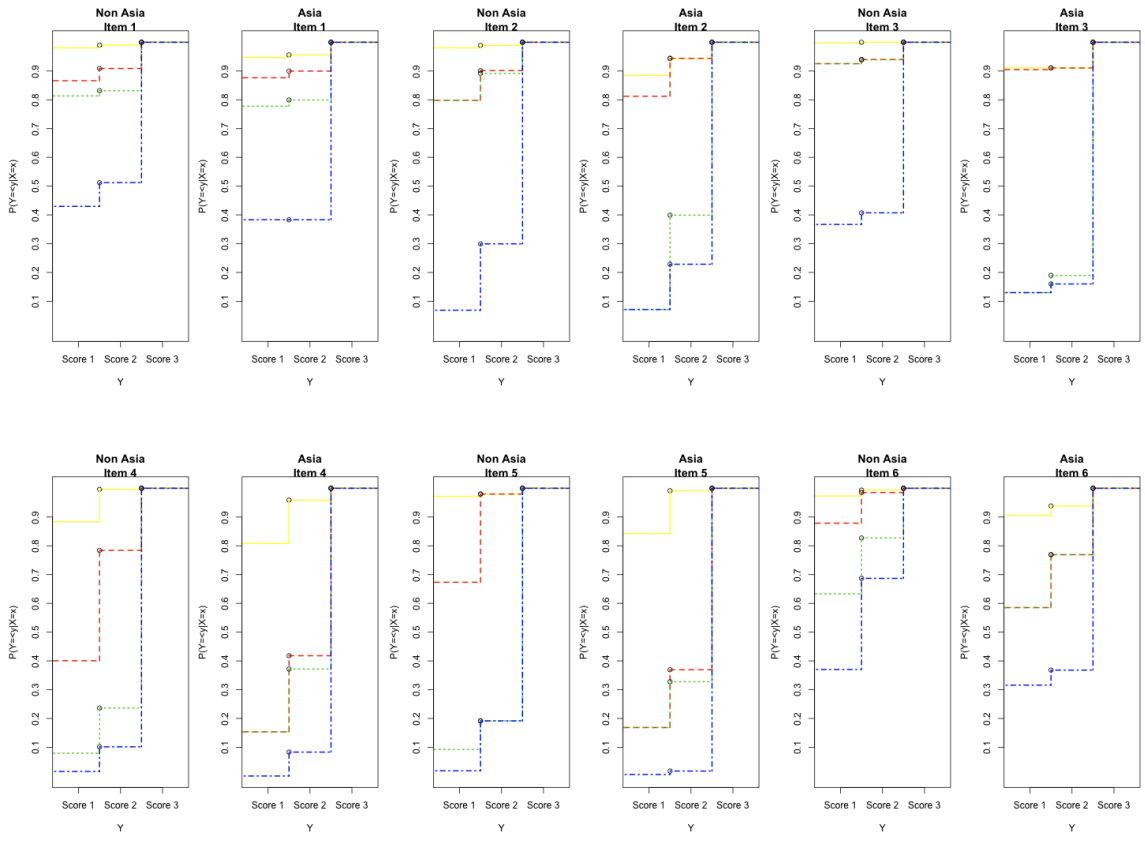

Figure 4.7: Comparison of the IRFs by items between two groups

To examine where the difference comes from, we attempted to quantify the differences by items.

- The IRF, p_{jktg}, of getting a response k or less in j item for group g given t class were compared across groups.
- To quantify those differences and compare them across the items, d was computed for each item by obtaining the absolute differences of IRFs between the groups.

$$d_j = \sum_{k=1}^{K} \sum_{t=1}^{T} |\hat{p}_{(jtk,g1)} - \hat{p}_{(jkt,g2)}|$$

Table 4.14: Difference of Item Response Probability between Two Groups

	Item 1	Item 2	Item 3	Iteme 4	Item 5	Item 6
d_j	0.43	1.87	2.32	1.19	2.12	1.47

To examine the difference between each item, Figure 4.7 provides visual comparisons of the IRFs of the two groups. In this figure, Group 1 is labeled "Non Asia", and Group 2 is labeled "Asia". Two step functions for Item 1 are the most similar and those for Item 3 are the most different. By computing d for each item, Table 4.14 shows that the differences of the IRFs for Item 3 are the biggest and for Item 1 are the least.

Using the GPCM also led us to the same conclusion. We fit the GPCM by allowing the item parameters for all six items vary across two groups (AIC = 17430.86). Compared to the GPCM with the same IRFs across the groups (AIC = 17711.89), the AIC decreased by 281.09. Next, we fitted the GPCMs by allowing the different item parameters for only one item vary across group at a time, holding those for the rest of the items the same. Table 4.15 shows the AICs of six models.

The model with different item parameters for Item 3 has the lowest AIC of 17583.12, compared to the rests which implies that the size of the difference of item parameters is the biggest in Item 3. This finding is consistent with the result of the OLCA.

Comparison of M_s and M_d

Table 4.16 displays the estimates of the class probabilities of M_d and M_s respectively. Class probabilities are more differentiated between groups in M_s, which restricts the IRFs to be equal across groups; the flexibility in M_d accounts for group differences by way of class probabilities or group-specific item response functions. As a result, M_s explained the group differences in Figure 4.3 by means of class probabilities, whereas M_d accounted for the group differences more by IRFs.

To evaluate the precision of model selection based on the AIC, we conducted 100 Monte Carlo simulations where the simulated data were drawn based on the model with the estimates M_d, then we fitted the simulated data by both of the correct model, M_d, and the competing alternative model, M_s, and chose the model based on the AIC. The more often that the correct model, M_d is chosen against the alternative model, M_s, the more reliable this model selection based on the AIC could be said to be. Table 4.17 shows the outcome of this competition. The correct model (M_d) was selected only 32 times out of 91, and the alternative model, M_s was chosen 59 times, which indicates that the measure of the AIC for the model

Table 4.15: Six GPCMs Were Fitted by Allowing the Item Parameters for One Item to Vary across the Groups

Item with different parameters per group	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
AIC	17688.72	17686.48	17583.12	17615.85	176865.52	17676.73

Table 4.16: Class Probability Estimates for Models M_s and M_d

	Model M_s			
	Class 1	Class 2	Class 3	Class 4
Group 1	0.23	0.17	0.05	0.53
Group 2	0.45	0.32	0.20	0.01
	Model M_d			
	Class 1	Class 2	Class 3	Class 4
Group 1	0.21	0.21	0.29	0.28
Group 2	0.29	0.20	0.34	0.15

Table 4.17: AIC Model Selection Between M_s and M_d

Generated Model	Fitted Models	
	M_d	M_s
M_d	32	59

selection in this case did not perform well. Nine cases were not converged out of 100 replications.

Residuals

Stone and Zhang (2003) argued that uncertainty in an individual's proficiency estimation in IRT models is responsible for deviations in the approximation of the goodness-of-fit statistics to the null distributions, particularly in shorter length tests. They considered a fit statistic based on posterior probabilities to account for uncertainty in proficiency estimation. Therefore, we considered the expected scores of individuals that are computed from taking an expectation of posterior probabilities. Residuals denoted by ϵ_{ij} represent how an expected score from the model deviates from the observed data; the process is similar to residuals in a regression

context helping to identify the source of model misfit (Kingston and Dorans, 1985; Molennar and Hoijtink, 1990; Orlando and Thissen, 2000; Rost and von Davier, 1994; von Davier and Molennar, 2003). For the OLCA model, the expected score of individual i for item j, category k given t latent class is given by:

$$E(y_{ij}) = \sum_{1}^{K}(\sum_{1}^{T} \hat{\pi}_{t|v} \times \hat{p}_{jtk}) \times k$$

where $\hat{\pi}_{t|v}$ is a vector of the posterior probabilities given a response pattern, v and $\hat{p}_{jtk}$ are estimated from the model. The standardized residuals, ϵ_{ij}, were computed by taking the difference between the observed score and the expected score for an item and dividing it by the square root of the variance of $E(y_{ij})$.

$$\epsilon_{ij} = \frac{y_{ij} - E(y_{ij})}{\sqrt{V(E(y_{ij}))}}$$

These quantities are not literally interpretable but the relative magnitude of these values can be examine either the item fit or the person fit. Poor item fit indicates that item parameters of a measurement model have questionable validity; poor person fit indicates that a specific response pattern is inconsistent with a scaling model (Reise, 1990; Sijtsma and Meijer, 2005).

Table 4.18 provides the overall means of model residuals of the OLCA, OLCR(M4), and OLCR(M_d). The mean value of the OLCR is smaller than that of the OLCA. Figure 4.8 shows the standardized residuals by each country for three models.

In summary, this chapter showed that OLCA is a complementary method to parametric item response model and an alternative way for group comparisons and it can be used to find DIF items in a test.

In the next chapter, simulation studies are conducted to study properties of the proposed procedure and to evaluate the parameter estimates obtained by the main model, and to examine how well AIC and BIC select the correct model.

Table 4.18: Sum of Standardized Residuals for OLCR and OLCA

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Total
OLCA		-0.033	-0.0382	-0.0357	0.0023	0.0034	-0.0350	-0.1371
OLCR	M4	-0.026	-0.0257	-0.0318	-0.0002	0.0003	-0.0332	-0.1172
OLCR	M_d	-0.029	-0.0322	-0.0294	-0.0030	0.0046	-0.0276	-0.1172

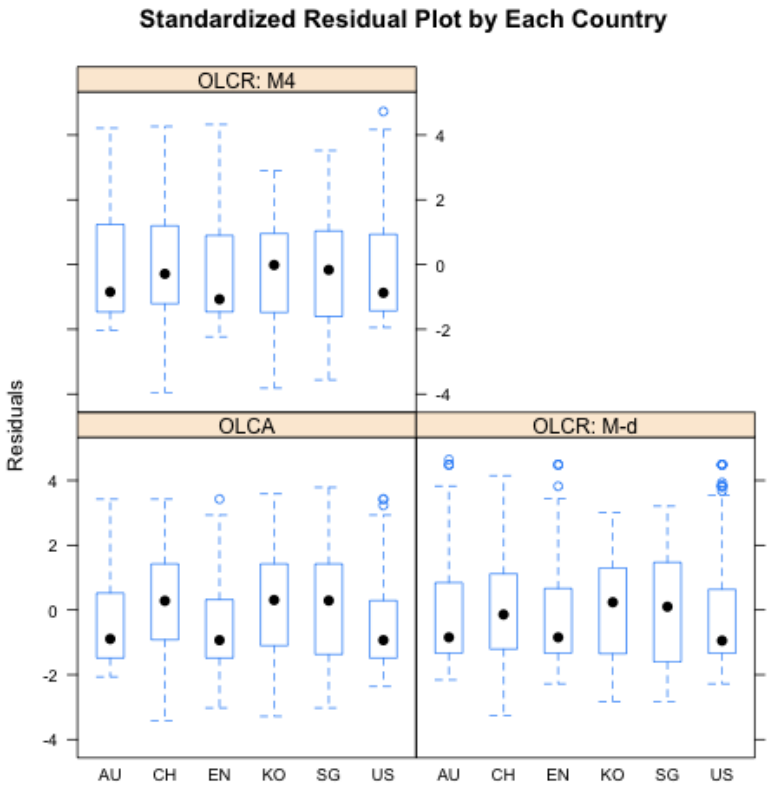

Figure 4.8: Residuals by Country

Chapter 5

Simulation

5.1 Introduction

This chapter conducts simulation study to examine the hypothesis testing procedure done in the real data analysis by the information criteria (AIC and BIC). It consists of two parts: Study 1 examines whether the information criteria choose the number of latent classes appropriately. Study 2 examines whether the information criteria and LRT test the effect of covariates correctly.

The simulation studies have three aims:

1. Evaluate the performance of an estimator by checking parameter estimate bias.

2. Evaluate the performance of information criteria (AIC and BIC) in selecting the number of latent classes.

3. Evaluate the performance of information criteria (AIC and BIC) and likelihood ratio test (LRT) on regression parameters for group differences.

Aim 1 The performance of an estimator involves checking parameter estimate bias (i.e., the difference between the average of the parameter estimates and the population parameter values). MSE is also used to assess the accuracy of the parameter estimates. In simulation studies, analysis models must be able to accurately recover the population parameters. If the models chosen to analyze the simulated data can not provide estimates that are close to the population parameters when specified correctly, it is difficult to make inferences on parameters. The estimated parameter values should be close to the population parameters that generated the data. The parameter recovery is evaluated in Study 1 and Study 2.

Aim 2 The performance of information criteria (AIC and BIC) in selecting the number of latent classes is also evaluated. The main measure is how many times the AIC and BIC select the correct model. In Study 1 the performance of the information criteria in selecting the number of latent classes is examined. There are many studies that explore the issue of determining the number of classes in latent class modeling (Muthen, 2006; Nylund et al., 2007; Yang, 2006). It is important to select the number of latent classes correctly, otherwise, an incorrect specification can result. Monte Carlo investigations on empirical performances of the AIC and BIC are needed.

Aim 3 Type I error rate of the information criteria on regression parameters, γs is examined in Study 2.

Measures of evaluating simulation results Simulation results are evaluated in terms of parameter recovery and classification. Bias and MSE are reported to establish the accuracy of the parameter recovery. As for classification, Pcc, Pcc.bias and Pcc.mse as defined next, are reported.

- Bias: The deviation in an estimate from the true value, which can indicate

the performance of the methods being assessed.

$$Bias = E[\hat{\theta} - \theta]$$

- Assessment of accuracy: The MSE provides a useful measure of the overall accuracy , as it incorporates both measures of both bias and variability.

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2],$$

- Pcc: Percent of correct classification of the model.

$$Pcc = \frac{\sum_{i=1}^{N} \delta_i}{N},$$

where $\delta = 1$ if the true class equals the predicted class, otherwise 0. The predicted class is obtained by the mode of posterior probabilities.

In addition to Pcc, this study also considers bias of Pcc and MSE of Pcc. Problems of correct classification rates were discussed by Haberman (2006). The idea is that a problem arises because Pcc does not account for the difference between when a probability of predicting the true class is 0.51 versus when it is 0.99. Another problem is that Pcc ignores how closely the predicted latent class is estimated to the true class because it considers only a hit rate. Pcc.bias and Pcc.mse represent the quality of classification.

$$Pcc.bias = \frac{1}{N} \sum_{i=1}^{N} \sum_{\tau=1}^{T} (\xi_i - \tau) \times \hat{p}_{i,\tau} \tag{5.1}$$

$$Pcc.mse = \frac{1}{N} \sum_{i=1}^{N} \sum_{\tau=1}^{T} (\xi_i - \tau)^2 \times \hat{p}_{i,\tau} \tag{5.2}$$

where ξ is the true class, $\tau = \{1, 2, .., T\}$, i is an individual, T is the number of classes, and $\hat{p}_{i,\tau}$ is a posterior probability that an individual i belongs to class τ.

The number of failures that occur should be recorded to gauge how likely it could happen in practice and determine whether the applied statistical procedure can be used reliably in situations being investigated.

5.2 Study 1

5.2.1 Purpose

The purpose of Study 1 is to evaluate the performance of the AIC and BIC in selecting the number of latent classes in the ordered latent class analysis (OLCA) models.

5.2.2 Design

Table 5.1 summarizes the conditions for the study. Each row in the table represents the models from which the data set was generated. The columns represent the models selected by the AIC.

The actual simulation procedure was performed in four steps.

1. For each condition (GPCM and 2-5 class OLCA), 100 data sets were generated. Population parameters were chosen from the U.S. data analysis results.

2. Each dataset was estimated repeatedly by the GPCM and OLCA with 2-5 class. In this way, models with correct and incorrect numbers of latent classes could be fitted to the data sets. Maximum log-likelihoods were obtained.

3. The values of information criteria (AIC and BIC) were calculated based on the maximum log-likelihoods. Information criteria were used for model selections to choose the model with the smallest values among competing models. The "best" models, according to each information criterion's suggestion, were counted and compared with true models in the original simulation.

4. The accuracy rates of information criteria were computed.

If the information criteria perform perfectly at selecting the correct model. Table 5.1 would be 100 % down the diagonal and zero everywhere else.

Table 5.1: Simulation Study Design

Generating Model	Calibrating Model(Fitted Model)				
(True Model)	GPCM	2 OLCA	3 OLCA	4 OLCA	5 OLCA
GPCM	Power	Type I	Type I	Type I	Type I
2 OLCA	Type I	Power	Type I	Type I	Type I
3 OLCA	Type I	Type I	Power	Type I	Type I
4 OLCA	Type I	Type I	Type I	Power	Type I
5 OLCA	Type I	Type I	Type I	Type I	Power

Population parameters for five conditions were drawn on the estimates of the GPCM and OLCA model without covariates using the U.S. data.

- Generating models : GPCM, 2-latent class model, 3-latent class model, 4-latent class model, and 5- latent class model.
- Sample size: 529
- Number of items: 6
- Population parameters were chosen from the estimates from the USA results in Table 4.2 in the previous section.
- Number of replications: 100

5.2.3 Result of Simulation Study 1

All computations of the estimation and model fit were done by an algorithm written in R. More computation time was needed as the number of latent classes of the fitted model increased. Estimating a misfit model took longer than estimating the true model from which data were generated.

Parameter Recovery

Parameter recovery was examined by using the bias and the mean squared error (MSE). The appendix shows that the population parameters for the five conditions from which the data were generated as well as, the bias and the MSE obtained from 100 replications.

- For the 2-Class OLCA model, the range of bias was from -0.0049 to 0.0058 and the range of MSE was from 0.0001 to 0.0011.
- For the 3-class OLCA model, the range of bias was from -0.0180 to 0.0110 and the range of MSE 0.0001 to 0.0088.
- For the 4-class OCLA model the range of bias was from -0.2814 to 0.2717 and the range of MSE was from 0.001 to 0.1198.
- For the 5-class OLCA model the range of bias was from -0.1598 to 0.1096 and the range of MSE was from 0.0001 to 0.0554.

The overall recovery for OLCA model seems to be good.

Power

Table 5.2 shows frequencies of the "best" models suggested by the AIC out of 100 replications. For the 2-class OLCA condition, the AIC selects the true model 97 % of the time and the 3- latent class model 3 %. For the 3-class OLCA condition, the true model was chosen for 91 % of the time by the AIC and the 4-latent class and the 2-latent class and the 5-latent class models were chosen for the 5 %, 2%, 1 % of the time respectively. For the 4-class OLCA condition, the values on the diagonal drop sharply from 91 to 25 and the most frequently chosen model was the 3-latent class model with 69 %. For 5- class OLCA, the result is similar to that of

Table 5.2: Frequency of Selected Models by AIC

	GPCM	2 OLCA	3 OLCA	4 OLCA	5 OLCA	N parm
GPCM	100	0	0	0	0	18
2 OLCA	0	97	3	0	0	25
3 OLCA	1	2	91	5	1	38
4 OLCA	4	1	69	25	1	51
5 OLCA	0	4	59	29	8	64

4-class OLCA condition. The 3 latent class model is chosen the most frequently as the best fitted model with 59 %.

In Table 5.2 displays that the percent of selecting the corrected model by the AIC is very low in 4 and 5 latent classes. One possible rationale is in the results of the USA data analysis. As shown in Table 4.2, the 3-class OLCA fits the USA data the best. The parameter values of the IRFs for both the 4-class OLCA and the 5-class OLCA model do not distinguish the latent classes well enough, therefore the data generated from 4-class and the 5-class OLCA models fit most adequately with the 3-latent class model. As shown in Table 5.4 and Table 5.5, many of the IRFs for the 4-class and the 5-class OLCA models are found very small.

In Table 4.2 BIC indicates that the GPCM fits the data the best (BIC = 4,158.7456). Table 5.3 contains the frequencies of the "best" models suggested by the BIC out of 100 replications for five conditions. For the GPCM condition, the BIC selected the true model 100 % of the time. For the 2-class OLCA condition, BIC selected the true model 84 % of the time. For the 3-class OLCA condition, the BIC selected the true model 1% and the GPCM 99 % of time. For the 4-class OLCA condition, the BIC selected the GPCM 100 % of the time. For the 5-class OLCA condition, the BIC selected the 2-class model and the 3-class model 46 % and 54 % of the time respectively. Even if the data analysis in Chapter 4 showed

Table 5.3: Frequency of Selected Models by BIC

	GPCM	2 OLCA	3 OLCA	4 OLCA	5 OLCA	N parm
GPCM	100	0	0	0	0	18
2 OLCA	16	84	0	0	0	25
3 OLCA	99	0	1	0	0	38
4 OLCA	100	0	0	0	0	51
5 OLCA	0	46	54	0	0	64

that the BIC prefers the model with a smaller number of parameters than the AIC, the results from Table 5.3 are inconclusive across the five conditions.

Many simulation studies have explored the issue of deciding how many classes to include in the mixture modeling. The AIC has been shown to overestimate the correct number of classes in finite mixture models (Celeux and Soromenho, 1996; Lin and Dayton, 1997; Soromenho, 1993; Yang, 2006). The BIC has been reported to perform well (Roeder and Wasserman, 1997). Jedidi et al. (1997) found that among commonly used model selection criteria, the BIC picked the correct model most consistently in the finite mixture structure equation model. Also, Lin and Dayton (1997) found that the AIC is more appropriate than the BIC when the models are complex. The results of Study 1 showed that AIC performed better than BIC in terms of selecting the number of latent classes appropriately in the OLCA models. One reason is that the OLCA model is a complex model with many parameters. Another reason is in a way of counting the number of the OLCA model to compute AIC and BIC. As discussed earlier, the OLCA model estimates the parameters of IRFs by imposing inequality constraints directly on the parameters. Therefore, even if the parameter space of this model is smaller than that of the standard latent class model, AIC and BIC were computed with the number of parameters the same as the standarad latent class model, which means that this

Table 5.4: Estimates of IRFs for the 4-class Ordered Latent Class model

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.93	0.96	1.00	0.87	0.99	0.98
	2	0.92	0.84	0.90	0.49	0.70	0.87
	3	0.82	0.79	0.90	0.11	0.09	0.71
	4	0.37	0.08	0.30	0.00	0.03	0.37
2	1	0.04	0.04	0.00	0.12	0.00	0.01
	2	0.03	0.04	0.03	0.35	0.28	0.12
	3	0.02	0.09	0.02	0.14	0.10	0.14
	4	0.05	0.23	0.04	0.10	0.16	0.31
3	1	0.03	0.00	0.00	0.01	0.01	0.01
	2	0.06	0.12	0.07	0.16	0.01	0.01
	3	0.16	0.12	0.09	0.75	0.81	0.15
	4	0.58	0.69	0.66	0.90	0.81	0.32

way is overcounting the number of parameters. Given that the BIC penalizes more on the number of parameters than the AIC, it was found that the BIC was a poorer selector than the AIC for ordered latent class models.

The appendix contains the tables for the parameters values, the bias and the MSE obtained from 100 replications for each condition.

Table 5.5: Estimates of IRFs for the 5- class Ordered Latent Class Model

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.93	0.96	1.00	0.87	0.98	0.99
	2	0.93	0.87	0.91	0.58	0.78	0.85
	3	0.92	0.83	0.91	0.12	0.12	0.75
	4	0.43	0.46	0.80	0.08	0.12	0.58
	5	0.39	0.09	0.19	0.00	0.02	0.36
2	1	0.03	0.04	0.00	0.10	0.01	0.00
	2	0.04	0.02	0.03	0.40	0.21	0.14
	3	0.01	0.06	0.00	0.16	0.15	0.11
	4	0.03	0.27	0.11	0.07	0.07	0.28
	5	0.05	0.17	0.00	0.12	0.17	0.28
3	1	0.93	0.96	1.00	0.87	0.98	0.99
	2	0.93	0.87	0.91	0.58	0.78	0.85
	3	0.92	0.83	0.91	0.12	0.12	0.75
	4	0.43	0.46	0.80	0.08	0.12	0.58
	5	0.39	0.09	0.19	0.00	0.02	0.36

Table 5.6: Classification

	2-Class OLCA	3-Class OLCA	4-Class OLCA	5-Class OLCA
Pcc	0.96	0.90	0.74	0.70
Pcc.bias	0.00	0.00	0.13	0.18
Pcc.mse	0.06	0.13	0.36	0.56

Classification

Pcc is the proportions of perfect match of the predicted latent class by the mode of posterior probabilities with the true class from which the response vector was generated from. The values in Table 5.6 were obtained by averaging the values of Pcc, Pcc.bias, and Pcc.mse over 100 replications. For the 2-class OLCA, Pcc is very high with a value of .96. However, as the number of latent classes increases, Pcc decreases (3 class =0.90, 4 class = 0.74, 5 class = 0.70), which means the prediction accuracy of the model predicted latent class, $\hat{\xi}$, gets lower. Conversely Pcc.bias and Pcc.mse are computed by accounting for how much the predicted latent class of the model deviates from the true class. They also use the posterior probabilities as weights. Therefore, a smaller number indicates more accurate classification. For 2-class and 3-class OLCA models, Pcc.bias is nearly zero. From the 4-class OLCA model on, it gets bigger. This is also related to the finding that AIC indicated 3 latent class describe the data best shown in Table 5.2. If Pcc.bias and Pcc.mse are zero, the models predict the true class ξ with a probability nearly equal to 1. Pcc takes into account only the hit rate, but Pcc.bias and Pcc.mse consider the degree of deviation of the predicted class from the true class. Therefore, they represent the quality of classification better. Pcc.mse is the variance of predicted latent class of the model.

5.3 Study 2

5.3.1 Purpose

Study 2 aims to assess the Type I error regarding group differences. The true models where the data were generated from do not assume group differences in the class probabilities. We fit both models - without and with group differences - and checked the error rates that reject the null when it is true.

5.3.2 Design

Population parameter values were chosen from the same estimates as in Study 1 for five conditions (GPCM and 2 to 5 class OLCA). Data were generated with 100 replications for each condition. Each data set is fit by an OLCA model and an OLCR model and then the two models were compared by the AIC and the BIC. Table 5.7 gives the estimated Type I error (i.e., the probability of incorrectly rejecting a true model). If these tests worked well - that is, if they correctly identified the OLCA model- we would expect the values in the cells on the off-diagonal to be small. They could be thought of a Type I error rate. As well LRT test is added to examine Type I error rate. Since OLCA models is nested in OLCR model as in a special case with regression parameters set to 0, LRT test is valid in this case.

5.3.3 Result of Simulation Study 2

Parameter Recovery

The parameter recovery of the misfit models which are the OLCR with a group covariate was examined. The same parameters were used as in Study 1. Therefore, only the tables for the bias and the MSE of parameter estimates are attached in the appendix. The class probabilities were obtained by marginalizing the estimates of

the OLCR and compared them with the true parameters for the class probabilities.

- For the 2-class OLCA model, the range of the bias was from -0.0066 to 0.0087 and the range of the MSE was from 0.0001 to 0.0011.
- For the 3-class OLCA model, the range of the bias was from -0.1440 to 0.1157 and the range of the MSE 0.0004 to 0.0649.
- For the 4-class OCLA model the range of bias was from -0.2326 to 0.2387 and the range of MSE was from 0.0001 to 0.1091.
- For the 5-class OLCA model the range of bias was from -0.1860 to 0.1289 and the range of MSE was from 0.0001 to 0.0669.

The estimates of the OLCR models are very close to the parameter values even if the models are not specified correctly. The tables of bias and MSE for the five conditions can be found in the appendix.

Type I Error

Table 5.7 provides the frequency of selecting the wrong models where a gender effect is assumed to exist. The error rates of AIC are 13 %, 22 %, 12%, 12%, and 36% for GPCM, 2-class, 3-class, 4-class, and 5-class models respectively. The error rates of BIC are 0 % , 3 %, 0 %, 1 %, and 1 % for GPCM, 2, 3, 4, and 5-class models respectively. The error rates of LRT are 4 %, 12 %, 10 %, 11 %, and 23 % for GPCM, 2-class, 3-class, 4-class, and 5-class models respectively.

The BIC outperformed the AIC for all the conditions in the Table 5.7. Looking at that table, one can notice that for the GPCM condition the BIC has 0 % Type I error rate and the AIC has 13 % error rate. This shows that BIC prefers simple models and AIC prefer complex models. For the OLCA, the BIC has consistently lower error rates than the AIC. LRT test shows only 4 % error rate for

Table 5.7: The percent of selecting the wrong models

Generated Model	GPCM (no gender)	2 OLCA	3 OLCA	4 OLCA	5 OLCA
Selected Model	GPCM (with gender)	2 OLCR	3 OLCR	4 OLCR	5 OLCR
AIC	13	22	12	12	36
BIC	0	3	0	1	1
LRT	4	12	10	11	23
Non.converge	0	0	12	14	33

GPCM model, however, it shows more than 10 % error rates for all OLCA models As the number of latent classes increases, the convergence slows down and non-converged cases take a place. For the 5-class OLCA only 67 pairs were compared because 33 data sets out of 100 replications did not converge in either the OLCA or the OLCR.

Classification

Each data set was fitted by the OLCA and OLCR. Pcc, Pcc.bias, and Pcc.mse were computed as in Study 1. Table 5.8 shows that misfitting OLCR to the data with a group difference does not significantly affect the classification accuracy.

Table 5.8: Classification Rate

	2 class		3 class		4 class		5 class	
	OLCA	OLCR	OLCA	OLCR	OLCA	OLCR	OLCA	OLCR
Pcc	0.96	(0.96)	0.90	(0.90)	0.76	(0.75)	0.70	(0.68)
Pcc.bias	0	(0)	0	(0)	0.11	(0.12)	0.08	(0.07)
Pcc.mse	0.06	(0.06)	0.13	(0.13)	0.35	(0.36)	0.53	(0.55)

In summary, AIC outperformed BIC in terms of selecting appropriately the

number of latent classes in the OLCA model. For Type of I error on regression parameters, only LRT performs well with GPCM model which is a parametric IRT model, however it does not perform well with OLCA models.

Chapter 6

Summary and Discussion

This chapter summarizes a number of interesting findings presented in the previous chapters, discusses the implications of results and limitations of this study, and gives suggestions for future research.

6.1 Summary and Discussion

Throughout this dissertation, the utility of ordered latent class regression for group comparisons is addressed. Chapter 2 reviewed two classes of measurement models in psychometrics-latent class models and item response theory models-and noted their advantage and limitations. Chapter 3 introduced the extension of the ordered latent class analysis (OLCA) model where latent classes are regressed by covariates. As well the model estimation procedure is described. Chapter 4 presented the results of the real data analysis to demonstrated how ordered latent class regression (OLCR) model can be applied to educational assessment data. Chapter 5 conducted the simulation studies to evaluate the parameter estimates obtained by the main model, and to examine how well AIC and BIC select the correct model.

6.1.1 Data Analysis

This study exemplified a comparison of the six countries in terms of math performance on the six polytomous items showing that the students' proficiency distributions of the six countries are different over latent classes.

In contrast to TIMSS report using benchmark cut points, the ordered latent class analysis model splits the scores into a finite classes in a certain "optimal" way, therefore, one does not have to worry about what threshold points to use to divide classes. This study also demonstrated the use of this model for differential item functioning (DIF) by allowing for group effects on item response functions (IRFs) and results indicate that IRFs are not homogeneous across two groups: three Asia counties and three English speaking nations. There might be a factor to influence the latent class and IRFs, which could be educational environment factors or cultural factors. The result can be useful information for educators, curriculum developers, and policy makers in three English speaking countries when they review math curriculums and construct test items.

The model employs nonparametric methods to estimate IRFs. By obtaining the full parametrization of IRFs, the visualization of the shape of IRFs helps make a diagnosis whether the parametric IRT model does not fit the data. Specifically, Item 3 classified only two classes rather than five distinctive classes and this finding suggests that the skill measured by this item may not be normally distributed. Item 6 was estimated as an item with low discrimination parameter in the parametric IRT model, however, Item 6 was found to discriminate well between class 4 and class 5 in the OCLA model. The estimated item response probabilities of this item suggest that an exponential form of constraint on IRF may not be appropriate for this item. It is important to note that the lack of fit of a particular unidimensional parametric IRT model does not necessarily mean a violation of unidimensionality because it could fail to fit for the parametric form of the response variables. By

taking a nonparametric IRT approach, one can address whether the violation of unidimensionality exists given the data.

Since the OLCA model assumes that the latent variable is discrete instead of continuous, it can provide interpretations of latent classes in accordance with the characteristics of items. For example, as shown in Chapter 4, the Item 6 contributes to separation between Class 4 and Class 5. Item 6 belongs to the reasoning domain and other five items belong to the applying domain. Therefore, the students in Class 5 can be considered as a group with high reasoning skill relative to the other four classes. This approach is similar to cognitive diagnostic model (CMD) that pre-defines skills or attributes of the items required for correct responses and estimates the parameters imposing equal constraints on the item response probabilities containing the same attributes. The constraints on item response probabilities provide a diagnosis of a class. Unlike cognitive diagnostic models, the OLCA approach does not require a predefined Q-matrix defining skills and items, which is frequently criticized due to its subjectivity so that it can serve as an useful exploratory tool.

6.1.2 Simulation Study

Through simulation studies, the performance of the AIC and the BIC in selecting the appropriate number of latent classes was examined. The ordered latent class analysis (OLCA) model imposes order constraints on the parameters directly. To compute AIC values, the same number of parameters as the standard latent class model is used. This results in that AIC penalizes o the number of parameters the OLCA model against GPCM more than it should do due to overcount of the number of parameters. Since the OLCA model estimates the restricted parameters by inequality constraints, it uses smaller parameter space than the standard latent class model. In addition, given that the BIC penalizes more on the number of

parameters than the AIC does, it was found that the BIC was a poor selector for ordered latent class models.

6.2 Limitations

First, the TIMSS assessments include both constructed response items and multiple choice items, which is very common in educational standardized tests. But the data analysis of this study is limited to polytomous constructed response items scored from 0 to 2.

Second, the simulation studies showed that as the number of latent classes increased, the model frequently failed to converge to fit the data. During the estimation, one or more latent class probabilities tend to 0 or 1. This hindered the iterative process to converge.

Third, for models with order restricted parameters, specifying the number of model parameters is an unsolved issue. This study employed the AIC and the BIC for model selection which are designed for models with fixed number of parameters without inequality constraints. Anraku (1999) applied an AIC approach to a simple restricted inference and proposed an order restricted information criterion (ORIC). Further research around the issue of competing information measures for order restricted latent models is needed.

6.3 Future Directions

Usability of the ordered latent class model

The substantive work in this dissertation focused on the issue of cross-country comparisons on math achievement. However there are other sets of comparison to be investigated.

Annual Progress of student achievement

The ordered latent class model can be applied to an annual comparison of student achievements. More specifically, a comparison of performances between 4th graders of the current year and the previous year can be made using the proposed approach, as long as common items exist between two tests. In 2001, the federal government enacted the *No Child Left Behind* (NCLB) legislation which required the evidence of progress of students' proficiencies through an assessment system (Bhola et al., 2003; Guskey, 2007). These assessments are to be administered every year in third grade through eighth grade, and at least once in high school. Adequate yearly progress is the metric used to evaluate school and district performance under NCLB. The current methods to report an yearly report are 1) to estimate the scaled score and 2) to make group comparisons by the mean values of the scaled score or total score by group. However, these approaches lead to biased estimates of the strength of the association between the latent variables and covariates, since it treats the unobserved variables as if observed (Bolck et al., 2004). Current methods require secondary analyses and equating and linking procedures which are not necessary when the ordered latent class regression model is employed.

Effect of Cognitive test scores on labor market

the model is on examining the white-black wage gap. Neal and Johnson (1996) and Heckman et al. (2006), studied the black-white wage gap controlling the difference of pre-market skills and argued that cognitive test scores are a better summary measure of pre-market human capital. Cognitive test scores show differences in the quality of education and also account for any learning that has taken place outside of a formal classroom structure. However, test scores are measured with errors. Ignoring the errors may result in biased coefficients. Many social scientists do not model the measurement error inherent in the test score. Unlike the typical

errors-in-variables Anderson (1984) model which uses classic test theory to model homoskedastic measurement error by ability, we can use the ordered latent class model to model heteroskedastic measurement error by ability to examine whether the pre-market skill contributes to the black-white wage gap. National longitudinal survey of youth 1979 (NLSY79) contains both cognitive and non-cognitive test scores for this purpose.

Extension of the ordered latent class model

A multidimensional extension of the ordered latent class models can be investigated. Since the ordered latent class models used in my dissertation assume unidimensionality, the latent classes are ordered on a single dimension. This model assumes that the responses of students to items are characterized by one ability. However, as an extension, a bi-dimensional counterpart of the monotone homogeneity model can be formulated. Applied to the simple example, this model assumes that individuals are characterized by two abilities: e.g., the arithmetic and the verbal abilities. The model would consider the probability that a person gets an item j correct given that he is in the latent class r on the arithmetic dimension and the latent class q on the verbal dimension. Multidimensional ordered latent class model is motivated from a desire to understand how sensitive students' responses can be to the content and skill components of items. It would help specifying Q-matrix of cognitive diagnostic models.

Bibliography

Agresti, A. (2003). *Categorical Data Analysis*. Hoboken, NJ: Wiley.

Agresti, A., Chuang, C., and Kezouh, A. (1986). Order-restricted score parameters in association models for contingency tables. *Journal of the American Statistical Association*, 82:619–623.

Albert, P., McShane, L., and Shih, J. (2002). Latent class modeling approaches for assessing diagnostic error without a gold standard. *Biometrics*, 57:610–619.

Anderson, E. (1982). Latent structure analysis: A survey. *Scandinavian Journal of Statistics*, 9:1–12.

Anderson, T. (1984). Estimating linear statistical relationships. *The Annuals of Statistics*, 12(1):1–45.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43:561–573.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, 86(1):141–152.

Bandeen-Roche, K., Huang, G. H., Munoz, B., and Rubin, G. S. (1999). Determination of risk factor associations with questionnaire outcomes: a methods case study. *American journal of epidemiology*, 150(440):1165–1178.

Bandeen-Roche, K., Miglioretti, D., Zeger, S., and Rathouz, P. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 93(440):1375–1386.

Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics*, 22:229–243.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. New York, NY: Griffin.

Betts, J. and Grogger, J. (2003). The impact of grading standards on students achievement, educational attaintment and entry-level earnings. *Economics of education review*, 22:343–352.

Bhola, D., Impara, J., and Buckendahl, C. (2003). Aligning test with states' content standards: Methods and issues. *Educational Measurement: Issues and practice*, 22:21–29.

Birnbaum, A. (1968). Some latent traits models and their use in inferring an examinee's ability. In Lord, F. and Novick, M., editors, *Statistical theories of mental test scores*, pages 307–479. Boston, MA: Addison-Wesley.

Bock, R. (1975). *Multivariate statistical methods in behavioral resesarch*. New York, NY: McGraw-Hill.

Bock, R. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorith. *Psychometrika*, 46:443–459.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables. *Political Analysis*, 12:3–27.

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

Clogg, C. and Goodman, L. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79:782–771.

Clogg, C. and Goodman, L. (1986). On scaling models applied to data from several groups. *Psychometrika*, 51:123–135.

Coull, B. A. and Agresti, A. (2002). The analysis of contingency tables under inequality constraints. *Journal of Statistical Planning and Inferences*, 107:45–73.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, 43:171–192.

Croon, M. (1991). Investigating mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, 44:315–331.

Dayton, C. and Macready, G. (1988). Concomittant-variable latent class models. *Journal of the American Statistical Association*, 83:173–178.

Dillon, W. and Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview. In Bagozzi, R., editor, *Advanced methods of marketing research*, chapter 9, pages 352–388. Cambridge, UK:Blackwell.

Eaton, W., Dryman, A., Sorenson, A., and McCutcheon, A. (1989). Dsm-iii major depressive disorder in the community- a latent class analysis of data form the nimh epidemiologic catchment-axea program. *British Journal of Psychiatry*, 115:48–54.

El Barmi, H. and Dykstra, R. (1994). Restricted multinomial maximum likelihood estimation based upon french duality. *Statistics & Probability Letters*, 21:121–130.

El Barmi, H. and Johnson, M. (2006). A unified approach to testing for and against a set of linear inequality constraints in the product multinomial setting. *Journal of Multivariate Analysis*, 97:1894–1912.

Formann, A. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, 38:87–111.

Formann, A. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5:179–211.

Garrett, E. and Zeger, S. (2000). Latent class model diagnosis. *Biometrics*, 56:1055–1067.

Glas, C. (2001). Differential item functioning depending on general covariates. In Boomsma, A., van Duijn, M., and Snijders, T., editors, *Essays on Item Response Theory*, chapter 7, pages 131–148. New York, NY: Springer-Verlag.

Goodman, L. (1974). Exploratory latent structure using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Guskey, T. (2007). Multiple sources of evidene : An analysis of stakeholders' perceptions of various indicators of student learning. *Educaitonal measurement: issues and practice*, 26(1):19–27.

Haberman, S. (2006). Bias in estimation of misclassification rates. *Psychometrika*, 10:387–394.

Hagenaars, J. (1979a). *Categorical longitudinal data: loglinear panel, trend and cohort analysis.* Newbury Park, CA: Sage.

Hagenaars, J. (1979b). *Loglinear Model with Latent Variables.* Thousand Oaks, CA: Sage.

Hambleton, R. and Swaminathan, H. (1985). *Item Response Theory.* Holland: Kluwer.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Heijden, P., Dessens, J., and Bockenholt, U. (1996). Estimating the concomitant variable latent class model with em algorithm. *Journal of Educational and Behavioral Statistics*, 21:215–229.

Heinen, T. (1993). *Discrete Latent Variable Models.* Tilburg University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousands Oaks, CA: Sage.

Hoijtink, H. and Molenaar, I. (1997). A multidimensional item response model: Constrained latent class analysis using gibbs sampler and posterior predictive check. *Psychometrika*, 62:171–189.

Holland, P. and Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14:1523–1543.

Huang, G. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrikw*, 69:5–23.

Hudziak, J., Heath, A., Madden, E., Reich, W., Bucholz, K., Slutske, W., Bierut, L., Neuman, R., and Todd, R. (1998). Latent class and factor analysis of dsm-iv adhd: A twin study of female adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 3:848–857.

Jedidi, K., Jagpal, H., and DeSarbo, W. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16:39–59.

Johnson, M. (2007). Modeling dichotomous item responses with free-knot splines. *Computational Statistics and Data Analysis*, 61:4178–4192.

Junker, B. and Sijtsma, K. (2001). Nonparametric irt in action: An overview of the special issue. *Applied Psychological Measurement*, 25:211–220.

Kingston, N. and Dorans, N. (1985). The analysis of item-ability regressions: an exploratory irt model fit tool. *Applied Psychological Measurement*, 9:281–288.

Lazarsfeld, P. and Henry, N. (1968). *Latent Structure Analysis.* Boston, MA: Houghton Muffilin.

Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153.

Lin, T. and Dayton, C. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 2:249–264.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47:147–174.

Mellenbergh, G. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19:91–100.

Melton, B., Liang, K., and Pulver, A. (1994). Extended latent class approach to the study of familial/sporadic forms of a disease: Its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology*, 11:311–327.

Meredith, W. and Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57:289–311.

Mislevy, R. and Sheehan, K. (1984). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 49:359–381.

Mokken, R. (1997). Nonparametric models for dichotomous responses. In van der Linden and Hambleton, R., editors, *Handbook of modern item response theory*, pages 351–367. New York, NY: Springer-Verlag.

Mokken, R. (2001). *A theory and procedure of scale analysis with application in political research.* Berlin, Germany: Walker de Gruyter, Mouton.

Molenaar, I. (1997). Nonparametric models for polytomous responses. In van der Linden and Hambleton, R., editors, *Handbook of modern item response theory*, pages 369–380. New York, NY: Springer-Verlag.

Molennar, I. and Hoijtink, H. (1990). Many null distributions of person fit indices. *Psychometrika*, 55:75–106.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49:313–334.

Mullis, I., Martin, M., , and Foy, P. (2008). Timss 2007 international mathematics report. Technical report, PIRLS International Study Center, Boston College, Boston, MA.

Muraki, E. (1992). A generalized partial credit model: Application of an em algrithm. *Applied Psychologicval Measurement*, 16:159–176.

Muthen, B. (1985). A method of studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10:121–132.

Muthen, B. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika*, 54:385–396.

Muthen, B. (2006). Should substance use disorders be considered as categorical or dimensional? *Addition*, 101:6–16.

Neal, D. and Johnson, W. (1996). The role of pre-market factors in black-white wage differences. *Journal of Political Economy*, 104:869–895.

Neuman, R., Heath, A., Reich, W., Bucholz, K., Madden, P., Sun, L., Todd, R., and Hudziak, J. (2001). Latent class analysis of adhd and comorbid symptoms in a population sample of adolescent female twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42:933–942.

Nylund, K., Asparouhov, T., and Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, 14:535–569.

Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory model. *Journal of Educational and Bebavioral Statistics*, 24:50–64.

Pontenza, M. and Dorans, N. (1995). Dif assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19:23–37.

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630.

Ramsay, J. and Abrahamowicz, M. (1989). Binominal regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, 84:906–915.

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

Rogers, H., Swaminathan, H., and Egan, K. (1999). A multi-level approach for investigating differential item functioning. Paper presented at the annual meeting of the NCME.

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

Rost, J. (1991). A logistic mixture distribution model for polytochotomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44:75–92.

Rost, J. and Langeheine, R. (1997). *Applications of Latent Trait and Latent Class Models in the Social Sciences.* Munster, Germany: Waxmann.

Rost, J. and von Davier, M. (1994). Conditional item-fit index for rasch models. *Applied psychological measurement*, 18:171–182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, No.:17.

Sijtsma, K. and Hemker, B. (2000). A taxonomy of irt models for ordening persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 49:391–415.

Sijtsma, K. and Junker, B. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49:79–105.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling.* Champman and Hall, A CRC press company.

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

Stone, C. and Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40:331–352.

Sullivan, R., Kessler, R., and Kendler, K. (1998). Latent class analysis lifetime depressive symptoms in national comorbidity survey. *American Journal of Psychiatry*, 155:1398–1406.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In *Test Vailidty*. Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of irt models. In *Differential Item Functioning*. New York, NY: Springer-Verlag.

van der Ark, L. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, 25:273–282.

van Onna, M. (2002). Bayesian estimation and model selection in ordered latent class model for polytomous items. *Psychometrika*, 67:519–538.

Vermunt, J. (2001). The use of restricted latent class models for defining and testing non-parametric and parametric irt models. *Applied Psychological Measurement*, 25:283–294.

Vermunt, J. and Magidson, J. (2005). Factor analysis with categorical indicators: A comparison between traitional and latent class approach. In Van der Ark, A., Croon, M., and Sijtsma, K., editors, *New developments in Categorical Data Analysis for the socialand behavioral sciences*, pages 41–62. Mahwah: Erlbaum.

von Davier, M. and Molennar, I. (2003). A person-fit index for polytomous rasch models, latent class models and their mixture generalizations. *Psychometrika*, 68:213–228.

Woods, C. M. and Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71:281–301.

Yang, C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistical and Data Analysis*, 50:1090–1104.

Zwinderman, A. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, 56:589–600.

Zwinderman, A. (1997). Response models with manifest predictors. In van der Linden, W. and Hambleton, R., editors, *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Appendix A

Appendix

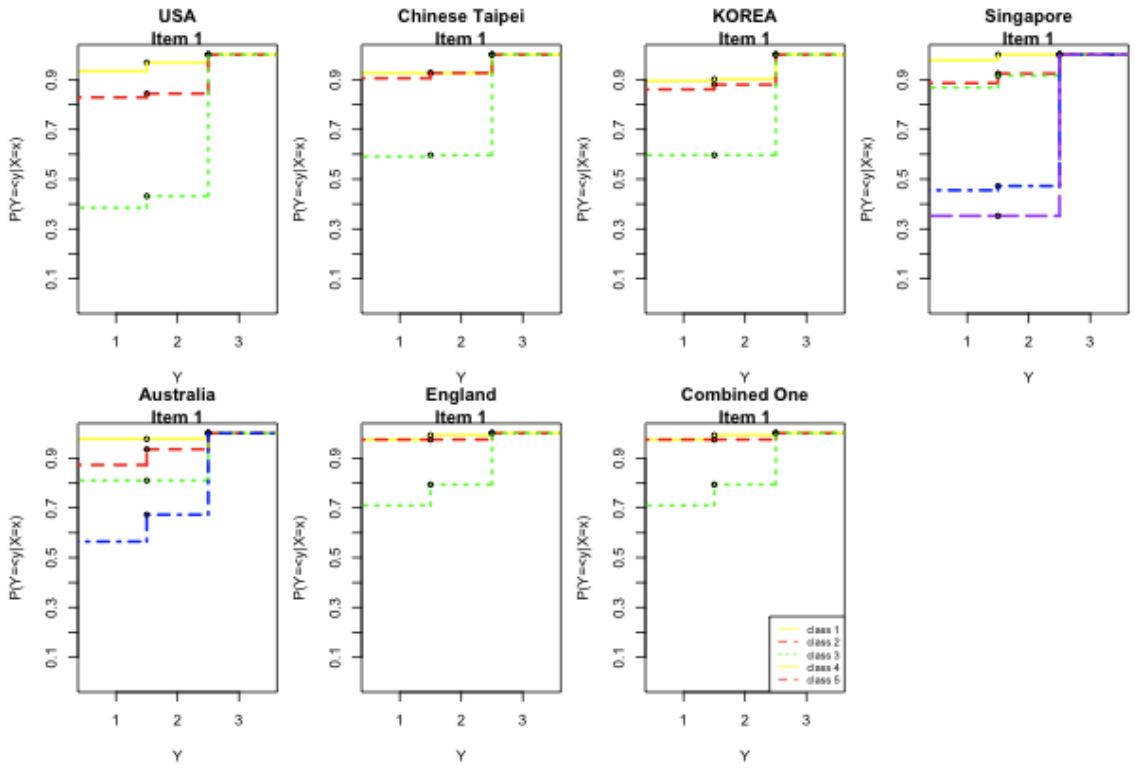

Figure A.1: IRFs of the best-fit models for Item 1

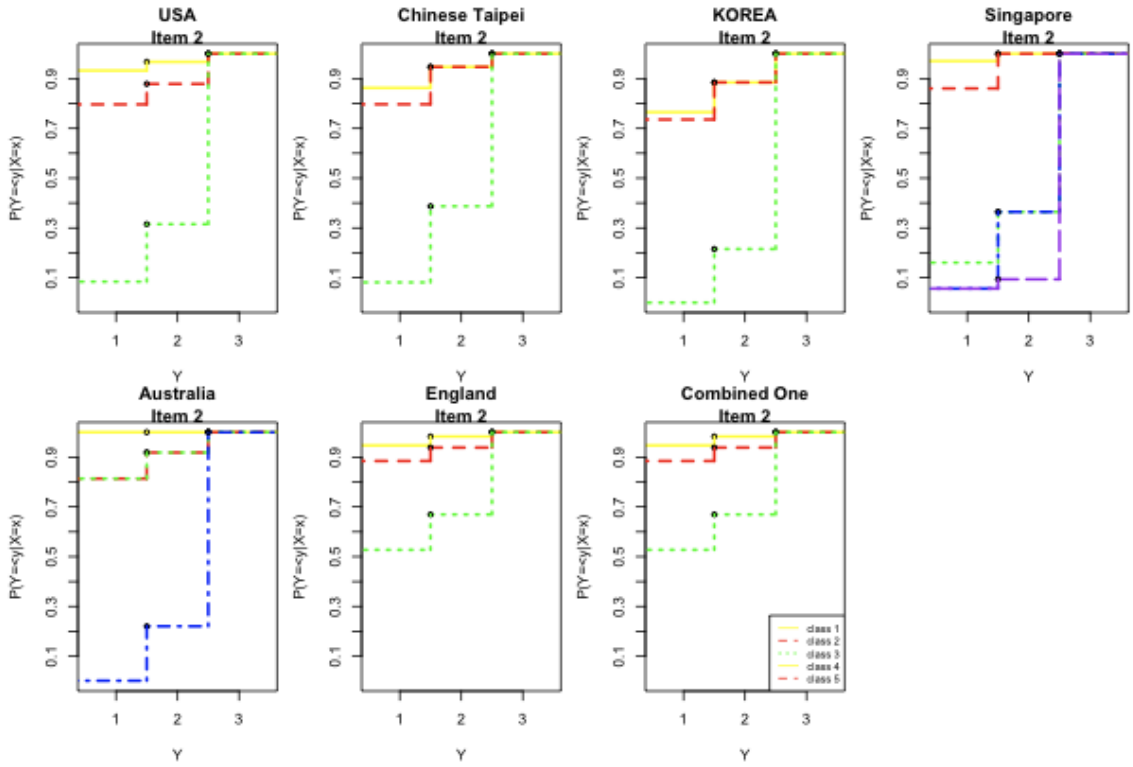

Figure A.2: IRFs of the best-fit models for Item 2

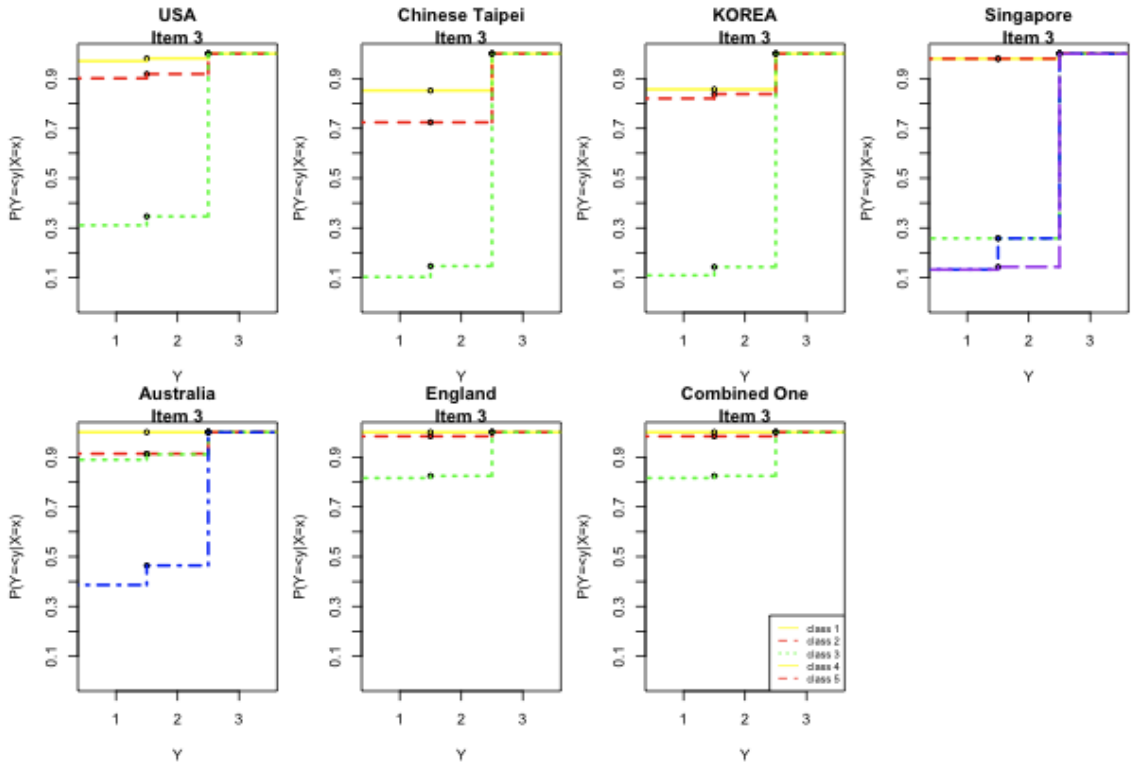

Figure A.3: IRFs of the best-fit models for Item 3

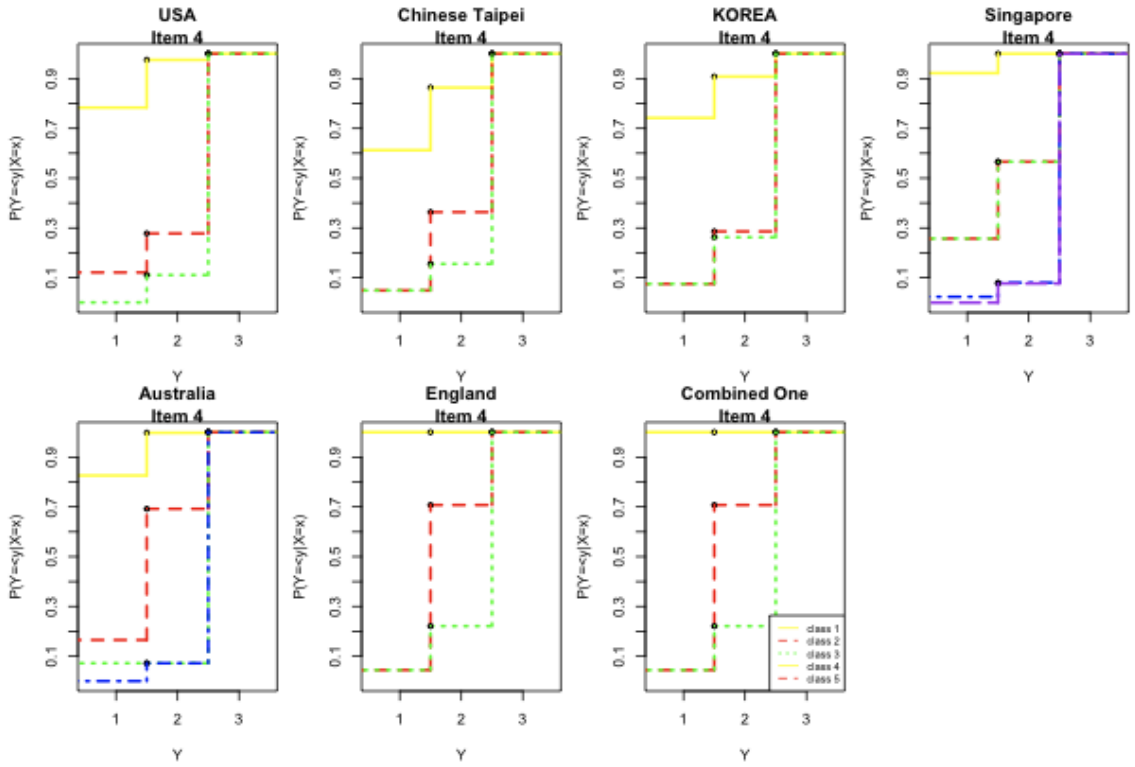

Figure A.4: IRFs of the best-fit models for Item 4

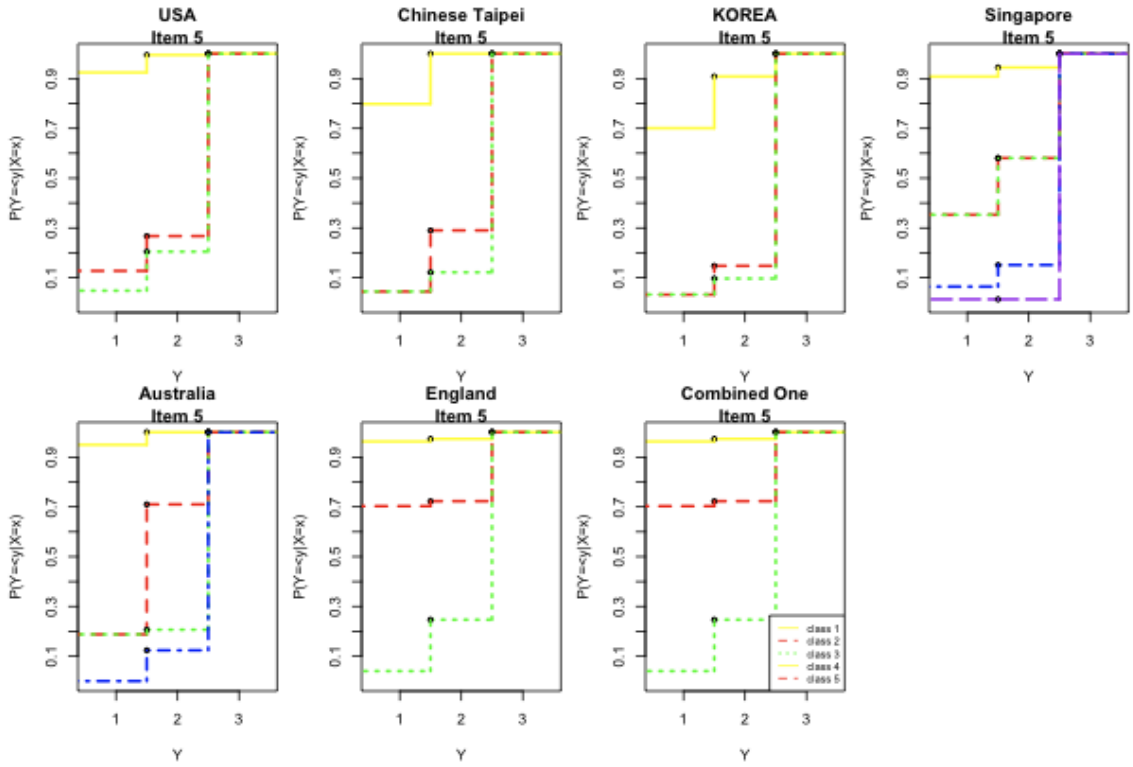

Figure A.5: IRFs of the best-fit models for Item 5

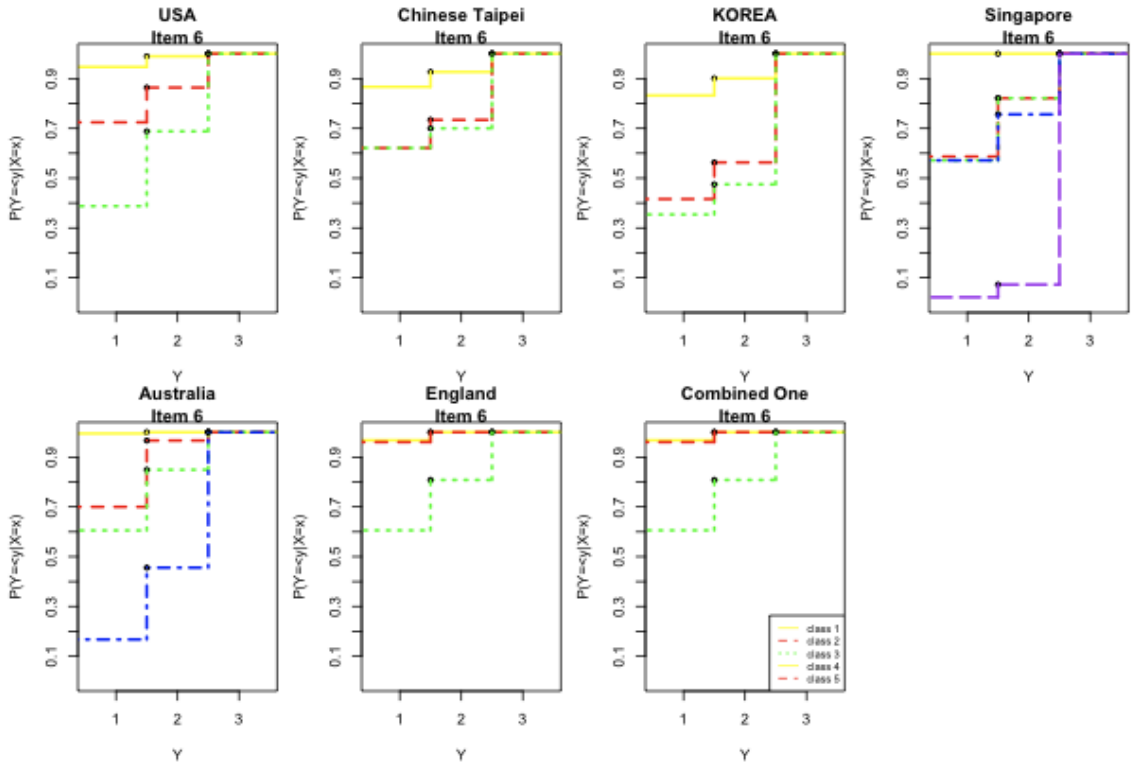

Figure A.6: IRFs of the best-fit models for Item 6

Table A.1: Regression Parameters of M4

	2 class		3 class		4 class		5 class	
γ	γ	SE	γ	SE	γ	SE	γ	SE
Intercept	0.99	(.50)	0.25	(.44)	1.89	(.37)	0.20	(.56)
Gender	-0.79	(.31)	0.19	(.14)	-0.36	(.18)	-0.46	(.23)
Korea	-0.03	(.54)	0.35	(.45)	-0.64	(.40)	0.78	(.23)
Singapore	-1.19	(.55)	-0.59	(.43)	-1.00	(.36)	0.26	(.52)
USA	-2.35	(.57)	-0.47	(.40)	-3.31	(.39)	-2.74	(.49)
Australia	-1.80	(.59)	0.01	(.41)	-3.52	(.54)	-1.86	(.58)
England	-2.33	(.67)	-0.25	(.42)	-2.87	(.40)	-3.12	(.83)

Table A.2: Regression Parameter of M5

	Class 2		Class 3		Class 4		Class 5	
γ	γ	SE	γ	SE	γ	SE	γ	SE
Intercept	0.98	(0.70)	0.17	(0.66)	1.76	(0.54)	0.27	(0.71)
Gender	0.83	(0.83)	0.34	(0.74)	-0.10	(0.60)	-0.81	(0.84)
Korea	0.18	(0.85)	0.36	(0.79)	-0.14	(0.68)	0.69	(0.78)
Singapore	-0.83	(0.81)	-0.28	(0.73)	-0.60	(0.59)	0.49	(0.70)
U.S.	-2.41	(0.78)	-0.31	(0.65)	-3.14	(0.58)	-3.41	(0.91)
Australia	-1.80	(0.81)	-0.02	(0.67)	-3.30	(0.73)	-2.52	(0.95)
England	-2.43	(0.87)	-0.27	(0.67)	-2.74	(0.60)	-3.33	(1.19)
Gen-Korea	-0.13	(1.12)	-0.07	(0.97)	-0.79	(0.84)	0.43	(1.05)
Gen-Singapore	-0.31	(1.13)	-0.52	(0.89)	-0.55	(0.72)	-0.28	(0.95)
Gen-U.S.	0.35	(1.08)	-0.29	(0.77)	-0.16	(0.74)	-1.23	(1.22)
Gen-Australia	-0.45	(1.44)	0.03	(0.80)	-0.27	(0.97)	1.19	(1.21)
Gen-England	0.89	(1.22)	0.06	(0.80)	-0.11	(0.77)	0.32	(1.78)

Table A.3: Chinese Taipei

Model	N class	N parm	loglikelihood	AIC	BIC
OLCA	T=2	25	-1299.35	2648.71	2740.11
OLCA	T=3	38	-1280.13	2636.27	2775.19
OLCA	T=4	51	-1273.69	2649.39	2835.84
OLCA	T=5	64	-1263.33	2654.66	2888.64
OLCR	T=3	40	-1279.49	2638.99	2785.23
OLCR	T=4	54	-1271.76	2651.52	2848.94
OLCR	T=5	68	-1263.81	2663.63	2912.24
GPCM	nogender	18	-1302.01	2640.01	2705.82
GPCM	gender	18	-1301.77	2641.55	2711.02

Table A.4: Korea

Model	N class	N parm	loglikelihood	AIC	BIC
LCA	T=2	25	-1446.84	2943.68	3036.77
LCA	T=3	38	-1408.39	2892.79	3034.29
LCA	T=4	51	-1402.52	2907.05	3096.96
LCA	T=5	64	-1390.67	2909.34	3147.65
LCR	T=3	40	-1406.31	2892.63	3041.57
LCR	T=4	54	-1393.19	2894.39	3095.47
LCR	T=5	68	-1387.92	2911.84	3165.05
GPCM	nogender	18	-1429.06	2894.12	2961.15
GPCM	gender	18	-1428.79	2895.58	2966.33

Table A.5: Singapore

Model	N class	N parm	loglikelihood	AIC	BIC
LCA	T=2	25	-1529.40	3108.81	3203.56
LCA	T=3	38	-1495.70	3067.40	3211.42
LCA	T=4	51	-1476.97	3055.94	3249.23
LCA	T=5	64	-1462.09	3052.18	3294.74
LCR	T=3	40	-1493.78	3067.57	3219.16
LCR	T=4	54	-1476.65	3061.30	3265.95
LCR	T=5	68	-1454.69	3045.38	3303.10
GPCM	nogender	18	-1512.74	3061.49	3129.71
GPCM	gender	18	-1510.96	3059.92	3131.93

Table A.6: USA

Model	N class	N parm	log-like	AIC	BIC
LCA	T=2	25	-2041.63	4133.26	4240.03
LCA	T=3	38	-1999.52	4075.04	4237.34
LCA	T=4	51	-1990.37	4082.75	4300.57
LCA	T=5	64	-1985.67	4099.34	4372.69
LCR	T=3	40	-1999.32	4078.65	4249.49
LCR	T=4	54	-1989.15	4086.31	4316.94
LCR	T=5	68	-1980.12	4096.24	4386.67
GPCM	nogender	18	-2022.93	4081.86	4158.74
GPCM	gender	18	-2022.91	4083.82	4164.97

Table A.7: Australia

Model	N class	N parm	log-like	AIC	BIC
LCA	T=2	25	-1156.62	2363.24	2455.16
LCA	T=3	38	-1123.05	2322.10	2461.81
LCA	T=4	51	-1107.26	2316.52	2504.04
LCA	T=5	64	-1105.12	2338.25	2573.56
LCR	T=3	40	-1121.34	2322.69	2469.76
LCR	T=4	54	-1103.15	2314.30	2512.85
LCR	T=5	68	-1096.015	2328.03	2578.05
GPCM	nogender	18	-1139.20	2314.41	2380.59
GPCM	gender	18	-1139.06	2316.12	2385.98

Table A.8: England

Model	N class	N parm	loglikelihood	AIC	BIC
LCA	T=2	25	-1191.69	2433.39	2525.30
LCA	T=3	38	-1158.51	2393.03	2532.75
LCA	T=4	51	-1152.03	2406.06	2593.57
LCA	T=5	64	-1150.84	2429.69	2665.01
LCR	T=3	40	-1158.28	2396.56	2543.63
LCR	T=4	54	-1151.29	2410.59	2609.13
LCR	T=5	68	-1148.21	2432.42	2682.44
GPCM	nogender	18	-1194.80	2425.61	2491.79
GPCM	gender	18	-1006.71	2427.58	2497.43

Table A.9: Parameter values of 2 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.9354	0.9353	0.9718	0.7364	0.8570	0.9392
	2	0.6753	0.5531	0.7066	0.0454	0.0690	0.5975
2	1	0.0299	0.0340	0.0110	0.1967	0.0873	0.0488
	2	0.0266	0.1359	0.0240	0.1312	0.1355	0.1963
3	1	0.0347	0.0307	0.0172	0.0669	0.0557	0.0120
	2	0.2981	0.3110	0.2694	0.8234	0.7956	0.2062
π	1	0.5361					
	2	0.4638					

Table A.10: Bias of 2 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0010	-0.0008	0.0011	-0.0039	0.0000	0.0015
	2	-0.0049	-0.0027	-0.0040	-0.0032	0.0025	0.0015
2	1	0.0005	-0.0012	-0.0005	0.0022	-0.0010	-0.0013
	2	-0.0009	-0.0013	-0.0003	0.0003	-0.0035	-0.0033
3	1	-0.0015	0.0020	-0.0007	0.0016	0.0010	-0.0002
	2	0.0058	0.0040	0.0042	0.0029	0.0009	0.0018
π	1	0.0038					
	2	-0.0038					

Table A.11: MSE of 2 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0002	0.0003	0.0001	0.0006	0.0007	0.0002
	2	0.0007	0.0011	0.0011	0.0003	0.0005	0.0011
2	1	0.0001	0.0002	0.0001	0.0005	0.0004	0.0002
	2	0.0001	0.0006	0.0001	0.0006	0.0005	0.0007
3	1	0.0001	0.0001	0.0001	0.0003	0.0003	0.0001
	2	0.0007	0.0011	0.0010	0.0008	0.0009	0.0008
π	1	0.0007					
	2	0.0007					

Table A.12: Population values of 3 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.9344	0.9323	0.9704	0.7834	0.9246	0.9468
	2	0.8287	0.7964	0.9010	0.1213	0.1274	0.7243
	3	0.3853	0.0842	0.3105	0.0007	0.0477	0.3876
2	1	0.0335	0.0349	0.0109	0.1917	0.0705	0.0430
	2	0.0152	0.0823	0.0176	0.1564	0.1398	0.1396
	3	0.0466	0.2308	0.0356	0.1099	0.1571	0.3005
3	1	0.0320	0.0328	0.0188	0.0248	0.0049	0.0103
	2	0.1561	0.1213	0.0814	0.7223	0.7327	0.1361
	3	0.5681	0.6850	0.6539	0.8894	0.7952	0.3119
π	1	0.4710					
	2	0.3851					
	3	0.1437					

Table A.13: Bias of 3 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	-0.0001	-0.0014	0.0019	-0.0097	-0.0074	-0.0001
	2	0.0016	0.0030	0.0031	-0.0144	-0.0004	-0.0075
	3	0.0029	0.0072	0.0044	0.0041	0.0023	0.0105
2	1	0.0003	0.0003	0.0002	0.0053	-0.0011	-0.0006
	2	-0.0017	-0.0027	0.0003	0.0034	0.0015	0.0037
	3	0.0028	0.0004	0.0006	0.0037	-0.0070	-0.0058
3	1	-0.0001	0.0011	-0.0020	0.0044	0.0085	0.0007
	2	0.0000	-0.0003	-0.0034	0.0110	-0.0011	0.0037
	3	-0.0057	-0.0076	-0.0050	-0.0078	0.0048	-0.0047
π	1	0.0096					
	2	-0.0180					
	3	0.0084					

Table A.14: MSE of 3 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0002	0.0003	0.0002	0.0011	0.0009	0.0003
	2	0.0016	0.0021	0.0009	0.0016	0.0018	0.0020
	3	0.0050	0.0050	0.0088	0.0001	0.0009	0.0045
2	1	0.0001	0.0002	0.0001	0.0007	0.0004	0.0002
	2	0.0001	0.0006	0.0001	0.0009	0.0009	0.0013
	3	0.0008	0.0043	0.0006	0.0020	0.0016	0.0031
3	1	0.0001	0.0001	0.0001	0.0006	0.0004	0.0001
	2	0.0016	0.0013	0.0008	0.0024	0.0016	0.0010
	3	0.0040	0.0085	0.0083	0.0022	0.0022	0.0030
π	1	0.0011					
	2	0.0018					
	3	0.0007					

Table A.15: Population parameter values of 4 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.9340	0.9635	0.9995	0.8721	0.9858	0.9767
	2	0.9191	0.8406	0.8971	0.4934	0.7029	0.8696
	3	0.8210	0.7870	0.8971	0.1065	0.0934	0.7094
	4	0.3741	0.0834	0.3017	0.0000	0.0306	0.3728
2	1	0.0354	0.0365	0.0005	0.1222	0.0000	0.0121
	2	0.0252	0.0355	0.0312	0.3478	0.2829	0.1192
	3	0.0156	0.0891	0.0170	0.1402	0.1002	0.1404
	4	0.0479	0.2297	0.0364	0.0957	0.1630	0.3075
3	1	0.0306	0.0000	0.0000	0.0057	0.0142	0.0112
	2	0.0556	0.1239	0.0717	0.1588	0.0142	0.0112
	3	0.1634	0.1239	0.0859	0.7533	0.8064	0.1502
	4	0.5780	0.6869	0.6620	0.9043	0.8064	0.3197
π	1	0.3301					
	2	0.1842					
	3	0.3475					
	4	0.1379					

Table A.16: Bias of 4 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0014	-0.0166	-0.0148	-0.0254	-0.0209	-0.0068
	2	-0.0344	-0.0104	0.0209	-0.1543	-0.2814	-0.0514
	3	-0.0167	-0.0389	-0.0204	-0.0328	0.0092	-0.0241
	4	-0.0163	-0.0064	-0.0294	0.0028	-0.0048	-0.0121
2	1	-0.0029	0.0011	0.0057	0.0153	0.0290	0.0101
	2	0.0008	0.0110	-0.0044	0.0390	0.0097	0.0186
	3	0.0009	0.0086	-0.0019	-0.0440	-0.0046	0.0025
	4	-0.0041	0.0049	0.0033	-0.0025	-0.0127	0.0124
3	1	0.0015	0.0155	0.0091	0.0101	-0.0081	-0.0033
	2	0.0335	-0.0006	-0.0164	0.1153	0.2717	0.0329
	3	0.0158	0.0304	0.0223	0.0769	-0.0047	0.0215
	4	0.0204	0.0014	0.0261	-0.0003	0.0176	-0.0003
π	1	0.0669					
	2	-0.0076					
	3	-0.0541					
	4	-0.0051					

Table A.17: MSE of 4 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0004	0.0007	0.0004	0.0039	0.0014	0.0004
	2	0.0028	0.0025	0.0017	0.0516	0.1162	0.0075
	3	0.0034	0.0046	0.0018	0.0029	0.0026	0.0076
	4	0.0051	0.0044	0.0086	0.0001	0.0005	0.0053
2	1	0.0002	0.0003	0.0001	0.0031	0.0016	0.0004
	2	0.0007	0.0018	0.0006	0.0331	0.0255	0.0055
	3	0.0002	0.0028	0.0003	0.0053	0.0028	0.0034
	4	0.0006	0.0056	0.0007	0.0017	0.0021	0.0050
3	1	0.0002	0.0004	0.0002	0.0005	0.0002	0.0001
	2	0.0026	0.0016	0.0014	0.0593	0.1198	0.0026
	3	0.0031	0.0026	0.0017	0.0106	0.0025	0.0034
	4	0.0049	0.0085	0.0087	0.0018	0.0025	0.0040
π	1	0.0068					
	2	0.0031					
	3	0.0066					
	4	0.0004					

Table A.18: Parameter Values of 5 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.9331	0.9586	1.0000	0.8739	0.9847	0.9893
	2	0.9290	0.8687	0.9107	0.5791	0.7768	0.8536
	3	0.9166	0.8328	0.9107	0.1151	0.1207	0.7458
	4	0.4321	0.4610	0.8045	0.0792	0.1207	0.5782
	5	0.3855	0.0871	0.1850	0.0000	0.0241	0.3612
2	1	0.0320	0.0414	0.0000	0.1006	0.0068	0.0000
	2	0.0362	0.0206	0.0282	0.3953	0.2147	0.1357
	3	0.0118	0.0565	0.0000	0.1585	0.1463	0.1119
	4	0.0326	0.2706	0.1061	0.0689	0.0696	0.2795
	5	0.0474	0.1743	0.0000	0.1154	0.1661	0.2795
3	1	0.0349	0.0000	0.0000	0.0256	0.0085	0.0107
	2	0.0349	0.1107	0.0611	0.0256	0.0085	0.0107
	3	0.0716	0.1107	0.0893	0.7264	0.7331	0.1423
	4	0.5354	0.2684	0.0893	0.8519	0.8098	0.1423
	5	0.5672	0.7387	0.8150	0.8846	0.8098	0.3593
π	1	0.3140					
	2	0.1698					
	3	0.3013					
	4	0.1066					
	5	0.1082					

Table A.19: Bias of 5 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0062	-0.0098	-0.0140	-0.0029	-0.0146	-0.0103
	2	-0.0206	0.0038	0.0119	-0.1535	-0.1598	-0.0409
	3	-0.0294	-0.0104	-0.0093	-0.0134	0.0054	-0.0050
	4	0.0209	-0.0347	-0.0681	-0.0148	-0.0398	-0.0517
	5	-0.0388	-0.0311	0.0080	0.0026	-0.0053	-0.0016
2	1	0.0011	-0.0039	0.0058	0.0127	0.0148	0.0130
	2	0.0006	0.0056	-0.0081	0.0518	0.0501	0.0137
	3	0.0009	-0.0001	0.0040	-0.0300	-0.0289	0.0127
	4	0.0114	0.0407	-0.0021	0.0215	0.0363	-0.0002
	5	-0.0060	-0.0111	0.0165	-0.0231	-0.0251	-0.0115
3	1	-0.0073	0.0137	0.0082	-0.0099	-0.0003	-0.0028
	2	0.0200	-0.0094	-0.0038	0.1017	0.1096	0.0272
	3	0.0285	0.0105	0.0053	0.0432	0.0232	-0.0077
	4	-0.0322	-0.0060	0.0703	-0.0068	0.0036	0.0519
	5	0.0448	0.0421	-0.0245	0.0212	0.0310	0.0131
π	1	0.0475					
	2	-0.0163					
	3	-0.0200					
	4	-0.0041					
	5	-0.0061					

Table A.20: MSE in 5 class OLCA in Study 1

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0004	0.0005	0.0003	0.0038	0.0012	0.0004
	2	0.0020	0.0018	0.0009	0.0494	0.0554	0.0051
	3	0.0031	0.0023	0.0009	0.0022	0.0023	0.0024
	4	0.0160	0.0285	0.0181	0.0025	0.0039	0.0134
	5	0.0092	0.0063	0.0202	0.0001	0.0004	0.0089
2	1	0.0003	0.0003	0.0001	0.0034	0.0009	0.0004
	2	0.0012	0.0011	0.0006	0.0251	0.0194	0.0045
	3	0.0002	0.0018	0.0001	0.0050	0.0039	0.0026
	4	0.0024	0.0314	0.0065	0.0042	0.0049	0.0156
	5	0.0011	0.0069	0.0012	0.0027	0.0033	0.0075
3	1	0.0003	0.0004	0.0002	0.0006	0.0002	0.0001
	2	0.0010	0.0013	0.0006	0.0287	0.0344	0.0020
	3	0.0030	0.0013	0.0008	0.0074	0.0032	0.0013
	4	0.0144	0.0188	0.0177	0.0037	0.0024	0.0067
	5	0.0096	0.0121	0.0210	0.0026	0.0038	0.0084
π	1	0.0045					
	2	0.0019					
	3	0.0026					
	4	0.0021					
	5	0.0004					

Table A.21: Bias 2 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	-0.0008	0.0002	0.0013	-0.0001	-0.0031	0.0005
	2	-0.0001	0.0005	0.0054	0.0009	-0.0003	0.0087
2	1	0.0013	-0.0005	-0.0004	0.0041	-0.0003	-0.0019
	2	-0.0013	0.0003	-0.0000	-0.0016	0.0043	-0.0022
3	1	-0.0005	0.0003	-0.0009	-0.0040	0.0034	0.0014
	2	0.0014	-0.0009	-0.0054	0.0007	-0.0040	-0.0066
π	1	0.0026					
	2	-0.0026					

Table A.22: MSE 2 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0003	0.0003	0.0001	0.0009	0.0006	0.0002
	2	0.0010	0.0011	0.0010	0.0003	0.0004	0.0009
2	1	0.0001	0.0001	0.0000	0.0007	0.0003	0.0001
	2	0.0001	0.0005	0.0001	0.0006	0.0006	0.0007
3	1	0.0001	0.0001	0.0001	0.0005	0.0004	0.0001
	2	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007
π	1	0.0006					
	2	0.0006					

Table A.23: Bias of 3 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	-0.0111	-0.0063	0.0081	0.0137	-0.1230	-0.0115
	2	-0.0393	0.0403	0.0161	-0.1440	0.0339	0.0654
	3	0.0538	0.0572	0.0418	0.0250	-0.0490	0.0849
2	1	0.0129	-0.0032	-0.0014	-0.0249	0.0233	-0.0047
	2	-0.0013	-0.0022	0.0111	0.0284	0.0310	-0.0248
	3	0.0223	-0.0455	0.0658	-0.0085	-0.0254	-0.0241
3	1	-0.0018	0.0095	-0.0067	0.0111	0.0997	0.0162
	2	0.0406	-0.0381	-0.0273	0.1157	-0.0650	-0.0406
	3	-0.0761	-0.0117	-0.1076	-0.0165	0.0745	-0.0608
π	1	0.0703					
	2	-0.1212					
	3	0.0509					

Table A.24: MSE of 3 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0026	0.0032	0.0009	0.0087	0.0094	0.0020
	2	0.0104	0.0137	0.0085	0.0135	0.0175	0.0179
	3	0.0448	0.0460	0.0623	0.0007	0.0055	0.0376
2	1	0.0014	0.0012	0.0004	0.0035	0.0038	0.0015
	2	0.0010	0.0049	0.0014	0.0061	0.0081	0.0097
	3	0.0051	0.0247	0.0066	0.0134	0.0157	0.0340
3	1	0.0013	0.0014	0.0006	0.0047	0.0049	0.0006
	2	0.0097	0.0087	0.0066	0.0182	0.0214	0.0072
	3	0.0410	0.0592	0.0649	0.0142	0.0191	0.0301
π	1	0.0089					
	2	0.0121					
	3	0.0045					

Table A.25: Bias of 4 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	-0.0001	-0.0112	-0.0133	-0.0166	-0.0291	-0.0060
	2	-0.0245	-0.0088	0.0159	-0.1520	-0.2326	-0.0439
	3	-0.0181	-0.0395	-0.0275	-0.0303	0.0067	-0.0223
	4	-0.0189	-0.0196	-0.0136	0.0046	-0.0084	-0.0068
2	1	0.0024	-0.0028	0.0048	0.0058	0.0303	0.0080
	2	-0.0049	0.0238	-0.0057	0.0455	-0.0061	0.0130
	3	-0.0007	0.0024	0.0022	-0.0267	0.0016	0.0065
	4	0.0005	0.0049	0.0060	-0.0043	-0.0118	0.0023
3	1	-0.0023	0.0140	0.0084	0.0108	-0.0012	-0.0019
	2	0.0294	-0.0150	-0.0101	0.1065	0.2387	0.0309
	3	0.0188	0.0371	0.0253	0.0570	-0.0083	0.0158
	4	0.0183	0.0147	0.0076	-0.0003	0.0203	0.0045
π	1	0.0632					
	2	-0.0102					
	3	-0.0446					
	4	-0.0083					

Table A.26: MSE of 4 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0003	0.0005	0.0003	0.0048	0.0020	0.0003
	2	0.0022	0.0036	0.0014	0.0498	0.1056	0.0072
	3	0.0044	0.0095	0.0037	0.0028	0.0025	0.0032
	4	0.0073	0.0040	0.0094	0.0001	0.0005	0.0053
2	1	0.0002	0.0003	0.0001	0.0039	0.0018	0.0003
	2	0.0006	0.0027	0.0007	0.0304	0.0198	0.0050
	3	0.0002	0.0022	0.0002	0.0034	0.0029	0.0024
	4	0.0009	0.0045	0.0010	0.0017	0.0019	0.0062
3	1	0.0002	0.0004	0.0002	0.0005	0.0003	0.0001
	2	0.0026	0.0022	0.0011	0.0602	0.1091	0.0024
	3	0.0038	0.0051	0.0035	0.0069	0.0027	0.0027
	4	0.0080	0.0052	0.0096	0.0018	0.0020	0.0050
π	1	0.0070					
	2	0.0030					
	3	0.0059					
	4	0.0006					

Table A.27: Bias of 5 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0082	-0.0108	-0.0127	0.0152	-0.0221	-0.0102
	2	-0.0197	0.0045	0.0201	-0.1860	-0.1314	-0.0170
	3	-0.0465	-0.0119	-0.0013	0.0034	0.0193	0.0022
	4	0.1289	0.0060	-0.0480	-0.0162	-0.0261	-0.0121
	5	-0.0461	-0.0257	0.0196	0.0043	-0.0025	-0.0066
2	6	0.0000	-0.0037	0.0059	-0.0014	0.0233	0.0143
	7	-0.0022	0.0076	-0.0110	0.0798	0.0107	0.0003
	8	0.0043	0.0037	0.0064	-0.0312	0.0010	0.0070
	9	-0.0108	-0.0151	-0.0505	0.0374	0.0239	-0.0297
	10	0.0010	0.0012	0.0241	-0.0252	-0.0326	-0.0044
3	11	-0.0082	0.0145	0.0069	-0.0138	-0.0013	-0.0041
	12	0.0219	-0.0121	-0.0091	0.1062	0.1207	0.0167
	13	0.0421	0.0083	-0.0051	0.0277	-0.0203	-0.0092
	14	-0.1181	0.0091	0.0985	-0.0212	0.0021	0.0418
	15	0.0451	0.0245	-0.0438	0.0209	0.0352	0.0109
π	1	0.0467					
	2	-0.0220					
	3	-0.0355					
	4	0.0154					
	5	-0.0045					

Table A.28: MSE of 5 class OLCA in Study 2

Category	Class	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0.0005	0.0006	0.0003	0.0054	0.0013	0.0003
	2	0.0018	0.0018	0.0014	0.0669	0.0493	0.0056
	3	0.0049	0.0030	0.0010	0.0040	0.0046	0.0040
	4	0.0451	0.0426	0.0203	0.0024	0.0032	0.0123
	5	0.0133	0.0048	0.0162	0.0001	0.0006	0.0080
2	1	0.0003	0.0004	0.0001	0.0048	0.0012	0.0004
	2	0.0009	0.0020	0.0005	0.0448	0.0205	0.0059
	3	0.0003	0.0029	0.0002	0.0045	0.0034	0.0033
	4	0.0013	0.0244	0.0055	0.0058	0.0038	0.0158
	5	0.0013	0.0080	0.0016	0.0031	0.0036	0.0077
3	1	0.0003	0.0004	0.0001	0.0004	0.0002	0.0001
	2	0.0015	0.0014	0.0008	0.0366	0.0397	0.0013
	3	0.0044	0.0012	0.0011	0.0078	0.0054	0.0013
	4	0.0419	0.0202	0.0293	0.0041	0.0027	0.0060
	5	0.0134	0.0099	0.0208	0.0027	0.0041	0.0094
π	1	0.0048					
	2	0.0024					
	3	0.0042					
	4	0.0019					
	5	0.0002					