A SEMI-PARAMETRIC APPROACH TO ESTIMATING ITEM RESPONSE FUNCTIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Longjuan Liang, M. A., M. A. S.

* * * * *

The Ohio State University

2007

Dissertation Committee:

Professor Michael W. Browne, Adviser

Professor Michael C. Edwards

Professor Steven N. MacEachern

Approved by

_____

Adviser

Graduate Program in Psychology

ABSTRACT

In Item Response Theory (IRT), normal ogive functions or logistic functions are typically used to model the Item Characteristic Curve (ICC). Although the one parameter (1PL), two parameter (2PL) or three parameter (3PL) logistic models have been shown to be useful in a variety of situations, there are cases where these models do not produce a good fit to the data. The Logistic function of a Monotonic Polynomial (L-MP) is a model proposed in this dissertation aiming to improve the model-data fit.

The L-MP model replaces the linear exponent of the 1PL or 2PL model with a monotonic polynomial. It is a general model which includes the 1PL or 2PL model as a special case. A surrogate-based two-stage approach is used to obtain the estimates from the L-MP model.

The L-MP model is illustrated using both simulation studies and two real world examples. Performance of the L-MP model in the simulation studies is evaluated by examining the Root Integrated Mean Square Error (RIMSE) for the item curves and the ability estimates, and also the rank correlations between the estimated and true abilities. The L-MP model is compared with the 2PL model with Marginal Maximum Likelihood (MML) estimates and Joint Maximum Likelihood (JML) estimates. It is also compared with two nonparametric approaches, namely TESTGRAF which uses a kernel smoothing method, and the Nonparametric Bayesian model. Results show that: (1) The L-MP estimation method is able to recover the true values of person

and of item parameters reasonably well. (2) If a standard logistic model holds, the L-MP method can provide very close estimated ICCs to those of the MML method and much better estimated ICCs than those of the JML method. For ability parameters, θ, the L-MP method can provide slightly better estimates than MML and much better estimates than JML in terms of the RIMSE_θ. (3) When the true models are not standard logistic functions, the L-MP model with a higher order polynomial is preferable to the 2PL model. A comparison between TESTGRAF and L-MP shows that generally L-MP and TESTGRAF produce very similar estimated ICCs for most items. TESTGRAF has a slightly smaller RIMSE (in third decimal place) for the estimated item curves, but L-MP model produces better estimates of abilities in terms of RIMSE_θ and rank correlations. The comparison between L-MP and the Nonparametric Bayesian model shows that these two methods produce very similar results. The Nonparametric Bayesian model may yield better estimated ICCs than the L-MP model, but differences are too small to interpret with any certainty. The computational time for the Nonparametric Bayesian program is much longer than for the L-MP program. In summary, our experiments indicate that results from the L-MP model are comparable to the best of those from other approaches considered. This demonstrates that the surrogate ability approach, adapted from TESTGRAF and used in L-MP, yields results that are completely suitable for practical use.

ACKNOWLEDGMENTS

I wish to express my deepest appreciation to my adviser, Dr. Michael Browne, for his advice, patience and support throughout my graduate studies. My gratitude goes to Dr. Michael Edwards for his invaluable suggestions in improving the quality of this dissertation. I also want to thank Dr. Steven MacEachern for all the advice and support that he offered for my dissertation.

I want to thank Guangjian Zhang for his help in every aspect of my graduate studies. I want to thank Kristin Duncan for her kind help in offering the irtNP package, the data and the analysis result from her program. I also want to thank Xiaopeng Li for his support during my completion of this dissertation.

I am particularly indebted to my parents. Without their continuous love and support, I would not have gone this far.

VITA

April 15, 1977	Born - Guangxi, China
1998	B. S. Psychology, Beijing Normal University
2001	M. A. Psychology, Beijing Normal University
2004	M. A. S. Statistics, The Ohio State University
2001-present	Graduate Teaching and Research Associates and Statistical Consultant, The Ohio State University

FIELDS OF STUDY

Major Field: Psychology

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

A half century after its creation, Item Response Theory (IRT) has been widely applied to many areas in educational testing and psychological measurement, such as computer adaptive testing, item banking and test equating, etc. The first model in IRT (the normal ogive model) and associated parameter estimation methods were developed by Lord (1952, 1953a, 1953b). Birnbaum (1957, 1958a, 1958b) made another important development by replacing the normal ogive functions with logistic functions which are more mathematically tractable and easier to deal with. The Item Characteristic Curve (ICC) is a key term in IRT and it appears to be first used by psychometrician Tucker (1946). However, IRT was not widely applied in practice until Lord suggested a parameter estimation method and made a computer program LOGIST available. Development of IRT models was restricted because of the heavy computational load in estimating the parameters, but this difficulty has now been solved with the development of modern computer technologies.

The original work on IRT began with tests of dichotomous items which had only two choices or multiple choice items that were scored as right/wrong. The model was built on the assumption of unidimensionality. Later, models for other types of data were developed. For example, the Nominal Response Model by Bock (1972) and the Multiple Choice Model by Thissen and Steinberg (1984) were intended to be applied

to categorical data without order information such as multiple choice items. There were also many other models developed for polytomous response data such as ordered ratings for essays or responses from Likert-type scales in psychological measurements. Samejima (1969), for example, introduced an important Graded Response Model to analyze polytomous items. The Partial Credit Model from Masters (1982) and the Generalized Partial Credit Model from Muraki (1992) are some variations from this graded response model. Extended from unidimensional IRT and based on the concepts on factor analysis, multidimensional IRT has also been an area that attracts many IRT researchers. This dissertation will be started from the simplest scenario and will propose a new IRT model under the unidimensional assumption for dichotomous response items only. Future work might extend this to polytomous response data or to incorporate multiple dimensions.

In classical test theory, the proportion of correct responses for an item is used as the item difficulty parameter and the biserial correlation between the item score and the total test score is used as the item discrimination parameter. However in IRT, the item difficulty and discrimination parameters and a latent variable at the subject level, usually named "ability", are incorporated into the model and are estimated from data. An ICC is used to plot the probability of a correct response as a function of the unobserved latent trait, subject ability. Goodness of fit of the model to the data is usually assessed afterwards. Graphical procedures and statistical tests are the two common ways to assess the model-data fit. In the graphical procedure, the model-data fit is assessed only by comparing an estimated ICC with a so-called "empirical" ICC. No statistical tests are conducted. This method is straightforward and obvious. Moving from this visual technique for exploring item-fit, some researchers also developed some statistical tests for misfit. For example, Bock (1972) presented a chi-square index that compares the observed and the expected frequencies for each

ability interval. However, this chi-square test, like any other statistical test, is subjected to the criticism that it will reject the model whenever the sample size is large enough. Clearly the statistical results cannot be used solely to determine the adequacy of model-data fit. This dissertation will focus more on the graphical outputs and provide only some descriptive statistics as indices of model-data fit.

After assessing model-data fit, it is very natural to ask what to do about those items that do not fit the data well. One way is to keep the items with the poor fit. But this is obviously not a good choice, since all the inferences and interpretations should be made with much caution. Another way is to discard such items which, of course, is not satisfactory either, since item construction is very expensive. A better solution to these poorly fitted items is to obtain an ICC which has greater flexibility and is not constrained to the currently used family of parametric functions so that it can fit the data better. This is the ultimate goal of this dissertation.

The method proposed in this dissertation can be considered a semi-parametric approach since we still apply a certain functional form to the probability function. In this new method, the original linear exponent in the two parameter logistic (2PL) function is replaced with a monotonic polynomial of an uncertain degree. Thus the functional form for the items is undetermined. In this sense, we call it a semi-parametric approach. Also, the item parameters are the coefficients of the monotonic polynomial. They are used only to define the Item Response Function (IRF) and are not intended to be interpreted in any way. The parameter estimation method for this L-MP model is a surrogate-based two-stage approach. This method is very similar to Ramsay (1991)'s procedure in TESTGRAF except that to avoid tied ranks we use the normalized principal component scores instead of test scores as surrogate ability scores. Newton-Raphson method is then applied to obtain the maximum likelihood estimates of the item parameters. Based on the estimated ICCs, the Bayes expected a posteriori

(EAP) estimates of the latent trait abilities are computed. This model can be considered an extension of the 1PL or 2PL models. Performance of the new model will be evaluated by comparing with the estimates from other parametric techniques (like Joint Maximum Likelihood (JML) estimates from SYSTAT and Marginal Maximum Likelihood (MML) estimates from MULTILOG) and also by comparing with other nonparametric techniques, for example, TESTGRAF from Ramsay (1991) and the Nonparametric Bayesian method from Qin (1998).

Chapter 2 will give a review of IRT including the parametric and nonparametric approaches, the commonly used parameter estimation methods which include JML, MML and the Bayesian method. The filtered polynomial density estimation method will be introduced in Chapter 3. In Chapter 4, I will discuss how the monotonic polynomial described in Chapter 3 is applied to IRT. The computational details of the parameter estimation for the proposed Logistic function of a Monotonic Polynomial (L-MP) model are also given in this Chapter. Performance of the new model is evaluated via simulation studies. Results are presented in Chapter 5. Two real data sets will be tested using the new model and will be presented in Chapter 6. Discussion and concluding remarks will be given in Chapter 7.

CHAPTER 2

REVIEW OF THE ITEM RESPONSE THEORY

The IRT model presented in this dissertation assumes unidimensionality and is appropriate for dichotomous responses. In this chapter, both parametric and nonparametric approaches for estimating the ICCs are reviewed. In the parametric procedure, each item response function is assumed to have an explicit form or a family of mathematical functions. The uncertainty lies in the estimation of the specific parameters in the functional form. For the nonparametric procedure, however, no definite mathematical forms are imposed for the items. In this chapter, I will start by reviewing the two most commonly used mathematical models in the parametric procedures, the normal ogive model and the logistic model. The associated estimation methods for the item parameters and the examinees' abilities in the logistic model are reviewed next. Then we will move to the nonparametric procedures, including a frequentist method from Ramsay (1991) and a Bayesian method from Qin (1998). A brief summary of some other semi-parametric and nonparametric procedures will also be presented.

2.1 The parametric procedures

2.1.1 Mathematical models for ICCs

IRT is a psychometric theory with the basic concepts built on items. For each item in a test, a smooth function is fitted to the data to model the relationship between the observed proportions of the correct responses and the examinees' abilities. Among many other possible functions, the cumulative normal ogive and the cumulative logistic distribution functions are the two most widely used mathematical models in IRT.

2.1.1.1 The normal ogive model

Baker (1992) gave a summary of the justification for using the normal ogive model for the ICC pragmatically and theoretically. From Baker's summary, Richardson (1936), Ferguson (1942) and Finney (1944) justified the use of the normal ogive model on pragmatic grounds. And then Lord & Novick (1968, Chapter 16) provided a theoretical justification.

The normal ogive model has the form of a cumulative normal distribution function. For a two parameter normal ogive model, the IRF for item i is defined as

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz, \tag{2.1.1}$$

where θ is the unobserved latent trait ability, $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly. a_i and b_i are parameters characterizing item i. Parameter b_i is usually called the item difficulty parameter and represents the point on the ability scale where an examinee has a 50 percent of chance of answering item i correctly. This parameter is the location parameter in the normal ogive function.

Parameter a_i is usually called the item discrimination parameter and is the scaling parameter of the normal ogive function. The larger the item discrimination, the steeper the ICC.

2.1.1.2 The logistic model

Birnbaum (1957, 1958a, 1958b, 1968) proposed a 2PL function for the ICCs

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \tag{2.1.2}$$

where $P_i(\theta)$, b_i and a_i have the same meanings as in the normal ogive model in Equation (2.1.1). The constant D is a scaling factor. It has been shown that when $D = 1.7$, the absolute values of $P_i(\theta)$ from the normal ogive model and the logistic model differ by less than .01 for all values of θ (Haley, 1952). For simplicity purpose, when referring to the logistic model in the later section of this dissertation, the scaling parameter D will be omitted.

Since these two models produce very similar result but the logistic function does not involve integration as in the normal ogive function, it is computationally simpler. The logistic model is also more mathematically tractable and thus it has been applied in practice more often than the normal ogive model.

Contrasting with the 2PL model described in Equation (2.1.2), there are also the very commonly used one parameter model (1PL) and the three parameter model (3PL). When the item discrimination parameter a_i remains constant over all items, the model in Equation (2.1.2) becomes a 1PL model. When a so-called *guessing* parameter is introduced to the model where a lower asymptote is added to the ICC, the model becomes a 3PL model. The 1PL model is the same as the model given in Equation (2.1.2) except that the discrimination parameter does not have a subscript.

The 3PL model is written as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}. \tag{2.1.3}$$

2.1.2 Estimation methods

Item parameters ($a_i,\ b_i$ and c_i) and the ability θs are the two sets of parameters in the IRT model that need to be estimated. Depending on the purpose of the test, there are different ways to handle the parameter estimation. Some test issuers only try to get information about the item parameters to define the ICCs and don't care about the performance of the individual test taker. In this case the estimation of θs seems to be unimportant. However, in some tests especially in psychological areas, researchers also want to obtain the information about the specific examinee. The estimates of θ become important in this case. We will briefly review three commonly used methods in estimating parameters of an IRT model in this section: Joint Maximum Likelihood Method (JML) (Birnbaum, 1968), Marginal Maximum Likelihood Method (MML) (Bock & Lieberman, 1970; Bock & Aitkin, 1981) and the Bayesian Method (Swaminathan & Gifford, 1982, 1985, 1986; Patz & Junker, 1999). Other estimation methods like Conditional Maximum Likelihood (CML) will be omitted here because it requires a sufficient statistic which is not available in the 2PL and the 3PL models.

2.1.2.1 Joint Maximum Likelihood estimation(JML)

Birnbaum (1968) proposed a method to estimate the ability parameters and the item parameters jointly for the IRT model. The resulting parameters are called the JML estimates. The log likelihood function used is

$$\log L = \sum_{s=1}^{N}\sum_{i=1}^{n}\big[U_{si}\log P_{si} + (1 - U_{si})\log(1 - P_{si})\big], \tag{2.1.4}$$

where P_{si} is the probability for examinee s ($s = 1,\ 2,\ \ldots,\ N$) of getting item i ($i = 1,\ 2,\ \ldots,\ n$) correct and can be computed as in Equation (2.1.2) or Equation (2.1.3). U_{si} is the response of examinee s on item i which takes only 0 for a failure and 1 for a pass for dichotomous items.

The 1PL, 2PL and 3PL models are unidentified since the scale and the location of θ are not defined. For example, for a 2PL model defined in Equation (2.1.2), if we transform $\theta^* = c\theta + d$, a corresponding transformation of $a^* = \frac{a}{c}$ and $b^* = cb + d$ will result in $a^*(\theta^* - b^*) = a(\theta - b)$. If this indeterminacy is not solved, the procedure might not converge (Baker, 1992). In practice, the problem is usually solved by setting the location and the scale for θ by constraining the mean to be zero and standard deviation to be one. This is accomplished via the standardization, e.g.

$$\theta^* = \frac{\theta - \bar{\hat{\theta}}}{S_{\hat{\theta}}}. \tag{2.1.5}$$

An iterative two-stage method is applied to obtain the estimates. At stage one, starting with some initial values of θ, the item parameters are estimated by maximizing the conditional log likelihood function. At stage two, the item parameters from stage one are treated as known, the maximum likelihood estimates of the abilities are computed. The θs are standardized using Equation (2.1.5) and this ends a cycle. The cycle is repeated until the difference of the likelihood function values between two consecutive cycles is smaller than a predefined precision. Once the procedure converges, the item parameters will be rescaled accordingly. A numerical method such as the Newton-Raphson procedure is usually applied within each cycle to obtain the estimates of item parameters. When the number of items and the number of examinees are large, it could be difficult to handle all items or all ability parameters simultaneously since the dimension of the Hessian matrix will be large. However, due

to the local independence assumption, this procedure can be greatly simplified by estimating item by item and person by person.

This two-stage method to compute the JML estimates is very straightforward. The computational load is also much smaller comparing to the other methods discussed later. This JML procedure, however, is subjected to some criticism due to the following possible problems. First of all, the procedure can not handle examinees with perfect scores or zero scores since under such circumstances, the estimates of ability parameters will tend to be infinity or negative infinity. A more serious concern is the issue of consistency. Neyman & Scott (1948) showed that in the presence of incidental parameters (in an IRT model, this would be the ability parameters), the maximum likelihood estimates of the structural parameters (in an IRT models, this would be the item parameters) need not be consistent. A more obvious explanation to this problem is that in most situations, when the number of parameter remains constant, the larger the sample size, the more information. However, in IRT, when sample size increases, the number of parameters to estimate also increases. This leads to possible inconsistency of the estimated parameters. In practice, many users have moved away from JML to MML, although JML was the dominant estimation method when IRT was first proposed.

In this dissertation, we will propose a general IRT model which includes the 1PL and 2PL models as special cases. We will apply a similar method in estimating the item parameters and the ability parameters from the model jointly. But we will drop the iteration between the cycles and apply a surrogate-based procedure which is similar to the estimation scheme used in TESTGRAF (Ramsay, 1991; Ramsay, 2000). Normalized principal component scores will be used as surrogates for the examinee ability parameters when estimating item parameters. More details of this procedure will be provided in Chapter 4.

2.1.2.2 Marginal Maximum Likelihood estimation(MML)

Bock & Lieberman (1970) provided MML estimators of item parameters for the normal ogive IRF. Bock & Aitkin (1981) improved the procedure by implementing an EM algorithm. The MML method treats ability parameters as nuisance parameters and removes them by integrating over the ability distribution. The resulting likelihood function is thus called a "marginal" likelihood function. A normal distribution for ability is assumed. The likelihood function to be maximized in Bock & Aitkin (1981) procedure is

$$\log L = C + \sum_{l=1}^{m} r_l \log P_l, \tag{2.1.6}$$

where C is a constant and l refers to a response pattern with $\boldsymbol{x}_l = [x_{l1}\ x_{l2}\ \cdots\ x_{ln}]$, m is the total number of response patterns, and the unconditional probability of getting a response pattern $\boldsymbol{x}_l$ is defined as

$$P_l = P(\boldsymbol{x} = \boldsymbol{x}_l) = \int_{-\infty}^{\infty} P(\boldsymbol{x} = \boldsymbol{x}_l|\theta) g(\theta)\ d\theta. \tag{2.1.7}$$

The integration in the above equation is approximated by Gauss-Hermite quadrature, i.e.

$$P(\boldsymbol{x} = \boldsymbol{x}_l) = \sum_{k}^{q} P(\boldsymbol{x} = \boldsymbol{x_l}|X_k) A(X_k), \tag{2.1.8}$$

where X_k is a tabled quadrature point and $A(X_k)$ is the associated weight.

For the EM algorithm, each cycle includes an E step and an M step. In the E step, the primary goal is to compute the "expected frequency", $\overline{r}_{jk}$, of correct responses to item j at each quadrature point k and the "expected sample size", $\overline{N}_k$, at each quadrature point k. For a binary response item, $\overline{N}_k$ is the expected number

of examinees with ability level k and $\overline{r}_{jk}$ is the number of people expected to respond with a "1" at this ability level. In the E step, the item parameters are treated as known. The M step is to obtain the improved item parameters by maximizing the marginal likelihood function using the $\overline{r}_{jk}$ and $\overline{N}_k$ from the E step. The EM cycles are continued until the estimates become stable.

The expected a posteriori (EAP) estimates of abilities can be computed after the MML estimates of item parameters are obtained. From Bayes' theorem, the conditional distribution of θ given $\boldsymbol{x} = \boldsymbol{x}_i$ is

$$g(\theta|\boldsymbol{x}_i) = \frac{P(\boldsymbol{x} = \boldsymbol{x}_i|\theta)g(\theta)}{P(\boldsymbol{x} = \boldsymbol{x}_i)}. \tag{2.1.9}$$

So the conditional expectation of θ given $\boldsymbol{x} = \boldsymbol{x}_i$ is

$$E(\theta|\boldsymbol{x}_i) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) P(\boldsymbol{x} = \boldsymbol{x}_i|\theta)\ d\theta}{\int_{-\infty}^{\infty} g(\theta) P(\boldsymbol{x} = \boldsymbol{x}_i|\theta)\ d\theta}. \tag{2.1.10}$$

One advantage MML has over JML is that it can obtain estimates of the item parameters without estimating the ability parameters and it is theoretically expected to yield more accurate estimates (Lord, 1986). Another advantage is that this procedure has no problems with perfect scores or zero scores. The disadvantage of the method is that the computational load could be heavy. For example, Bock & Lieberman (1970)'s procedure could only handle a test with no more than 12 items. With the improvement of using the EM algorithm (Bock & Aitkin, 1981), however, this is not a problem anymore. Even with hundreds of items and thousands of subjects, MML/EM still converges in relatively little time (less than 1 minute).

2.1.2.3 Bayesian estimation

Both JML and MML are in the frequentist framework. They provide point estimates for the parameters in the model. Standard errors can be obtained based on asymptotic distributions. Therefore the number of items and the number of examinees need to be large. Bayesian estimation in IRT has been an active research area recently. In the frequentist framework, the parameters in the model are considered fixed. However, in the Bayesian framework, the parameters in the model are considered as random variables with their own distributions. There are basically two major methods of estimating parameters in the Bayesian framework. One is focusing on obtaining direct estimations at the mode of the joint posterior distribution (Bayesian Modal Estimation, Swaminathan & Gifford, 1982, 1985, 1986) while the second method is focused on obtaining the whole posterior distributions of the parameters using Markov Chain Monte Carlo (MCMC) (Patz & Junker, 1999). Both methods require the specification of prior distributions for the model parameters. Bayesian practitioners argue that the specification of a prior is not a problem but an advantage in IRT since IRT models has been applied in testing area for a long time and the information of the item parameters and the ability parameters are well collected.

Despite all the seemingly attractive advantages of the Bayesian method over the frequentist method, the Bayesian method usually comes with a heavier computational load and is more time consuming. It is also not easily understood by most test practitioners. Since it is not easy to produce a ready-to-use program for people with limited knowledge in Bayesian analysis, the Bayesian estimates are not used very often in practice. The new model we propose in this dissertation is in the frequentist framework so that more details on the Bayesian estimation methods will not be provided here.

2.2 The nonparametric procedures

The methods discussed in the previous section are all parametric approaches. Functional forms like logistic functions are usually assumed for the ICCs. The goal for the parametric procedure is to estimate the parameters in a functional form. Any parametric approach, however, has the risks of not accounting for features such as nonmonotonicity of the ICCs or other systematic departures of shape from what can be accommodated in existing models (Ramsay, 1991). An additional problem with parametric approaches involving more than one parameter is that the current formulations and estimation procedures might produce parameter estimates with very large sampling covariation (Lord, 1980; Thissen & Wainer, 1982). If an item clearly violates the parametric family, when using any parametric method to fit the item, we will observe a poor fit. When that occurs, we can either retain the items with poor fit or discard them. Neither approach is ideal. Developing some nonparametric models that have greater flexibility on fitting the ICCs would provide an attractive alternative. In this section, I will give a review of the kernel smoothing nonparametric approach from Ramsay (1991) and a nonparametric Bayesian approach from Qin (1998) and we will also mention some other semi-parametric or nonparametric approaches as summarized in Ramsay (1991) and Ramsay & Winsberg (1991).

2.2.1 Kernel Smoothing approach

Ramsay (1991) aimed at providing a "simple to understand, easy to program, non-iterative, very fast and remarkably efficient" technique for IRT. He used a kernel smoothing approach to provide a tool that is useful for modest-sized samples such as college classes.

The goal of the procedure is to estimate the probability of correctly choosing option m for item i (Option Characteristic Curve), which is defined as $P_{im}(\theta) = P[Y_{im} = 1|\theta]$, where Y_{im} has a value of 1 when option m is chosen and a value of 0 when option m is not chosen.

The procedure involves four steps:

1. Rank. Estimate the rank (r_s) for examinee s ($s = 1, 2, \ldots, N$) by some statistic T_s. T_s is usually chosen to be the total test score. For a binary response test, this is the total number of items that are correctly chosen. Ties are broken by randomly reordering within ties.

2. Enumerate. The ranks (r_s) are then replaced by the quantiles (q_s) of a standard normal distribution. These quantiles will be employed as surrogate ability scores in step 4.

3. Sort. The examinees' response patterns are then sorted by their estimated ability rankings.

4. Smooth. The smoothed estimate of $P_{im}(\theta)$ is

$$\hat{P}_{im}(\theta) = \frac{\sum_{s=1}^{N} K\left[\frac{q_s-\theta}{h}\right] Y_{im}^{(s)}}{\sum_{s=1}^{N} K\left[\frac{q_s-\theta}{h}\right]}, \tag{2.2.1}$$

where $K(u)$ is a Kernel function with the property of $K(u) \geq 0$, and $K(u)$ takes its maximum at $u = 0$ and goes to zero as u moves away from 0. The commonly used kernel functions are:

Uniform: $K(u) = 0.5$, $|u| \leq 1$, and 0 otherwise;

Epanechnikov or *Quadratic*: $K(u) = 0.75(1 - u^2)$, $|u| \leq 1$, and 0 otherwise;

Gaussian: $\exp(-u^2/2)$.

To ensure the differentiability of $\hat{P}$, the quadratic or Gaussian kernel is preferred.

This kernel smoothing method can be thought of as a weighted average with weight $\frac{K\left[\frac{q_s-\theta}{h}\right]}{\sum_{s=1}^{N} K\left[\frac{q_s-\theta}{h}\right]}$. Thus only the points that are close to the evaluation points are effectively weighted. h in Equation (2.2.1) is called the smoothing parameter. It is used to control the trade-off between bias and sampling variance. When h is small, the bias will be small since only a few observations very close to θ are effectively weighted. But the sampling variance will be correspondingly large. When h is large, the bias will be large since more observations are effectively weighted but the sampling error will be small. It was suggested that $h = N^{-1/5}$ be used for a Gaussian kernel, where N is the total number of examinees.

Once the estimates of the ICCs or the OCCs are obtained, one can proceed to compute the maximum likelihood estimates of the examinees' proficiency. The estimates of the examinees' proficiency can then be fed back into the process as the basis for ranking and to start the iterative procedure.

One weakness of this procedure is that it uses the test scores to rank the examinees. Test scores can be biased and insufficient for estimating examinees' proficiency, especially when the test is short. Ramsay (2000) pointed out that since the goal of the program is not to estimate the proficiency but merely the proficiency rank order, this limitation doesn't affect the estimate of the ICCs seriously (provided that the test has at least 15 items).

This procedure allows nonmonotonicity of the curve. To comply with the assumption that students with higher ability have higher probability of answering one question correctly, in the semi-parametric approach that we are proposing in this dissertation, we will put a constraint on the estimated ICC to make it monotonic increasing. As to estimation, we will apply a similar procedure to that in Ramsay (1991) but use the normalized first principal component scores instead of test scores as starting points to estimate the ICCs. The advantage of use of principal component

scores over test scores is that fewer ties occur in the ranking procedure. Once the ICCs are obtained, we will return to obtain the ability estimates.

2.2.2 Nonparametric Bayesian procedure

In the Bayesian framework, the unknown parameters are not considered fixed but random variables with their own distributions. Typically, these parameters are considered as coming from a parametric family of distributions. In the nonparametric Bayesian method, this constraint is released. The nonparametric Bayesian method thus actually provides greater support of the prior distributions for the parameters.

Qin (1998) applied a Dirichlet process in the middle stage of the hierarchical model. The Dirichlet process is a means to release the constraint on the prior distributions of the parameters to be from any specific parametric family. If P has a Dirichlet process, then it can be denoted as $P \sim Dirichlet(MP_0)$ where M is a real positive constant and P_0 is the best guess of the probability P. The positive constant can be interpreted as our confidence in the prior P_0. In IRT, this P_0, for example, can be specified as a 2PL function with the form as in Equation (2.1.2). A large value of M implies a strong belief in this prior. With $M \to \infty$, the nonparametric model tends to be the basic parametric model. A small value of M implies no prior information about P.

The general model in Qin (1998), with Dirichlet process prior for the ICCs in the middle stage of the hierarchical model, is given as

$$\begin{aligned}
\theta_s|\mu_\theta,\sigma_\theta^2 &\overset{iid}{\sim} N(\mu_\theta,\sigma_\theta^2)\\
P_i|a_i,b_i,M_i &\overset{iid}{\sim} Dir(M_iP_{0i})\\
z_{si}|P_i &\sim P_i\\
Y_{si}|\theta_s,z_{si} &= \begin{cases} 1 & \text{if } z_{si}\le\theta_s\\ 0 & \text{if } z_{si}>\theta_s \end{cases}\\
P_{0i} &= \frac{1}{1+\exp\left(-a_i(\theta-b_i)\right)}\\
b_i|\mu_b,\sigma_b^2 &\overset{iid}{\sim} N(\mu_b,\sigma_b^2)\\
a_i|v,w &\overset{iid}{\sim} H_{v,w}(v,w),
\end{aligned}$$

where $H_{v,w}$ is a distribution specified as a chi or log normal with parameters v and w.

The Gibbs sampler was used to obtain the posterior distributions of the unknown parameters. Bush and MacEachern (1996)'s algorithm was used to split the sampling of the latent variable into two groups which helped to speed up the mixing over the posterior distribution.

The performance of this nonparametric Bayesian model was compared with a fully Bayesian parametric method via a simulation study. Each model performed best when the data were generated from that model. When the true model was a third model, the nonparametric model outperformed the parametric model in terms of the mean squared error of the ability estimates and the difference between the estimated predictive distribution and the true distribution.

2.2.3 Other quasiparametric or nonparametric procedures

Ramsay (1991), Ramsay & Winsberg (1991) gave a summary of some quasiparametric and nonparametric approaches in estimating the IRFs.

These quasiparametric approaches are similar in that the probability functions are represented as a linear combination of some basis functions $\phi_1(\theta)$, $\phi_2(\theta)$, ..., $\phi_q(\theta)$,

$$P_i(\theta) = \sum_{q=0}^{Q} a_{qi}\phi_q(\theta). \tag{2.2.2}$$

Levine (1984, 1985) and Drasgow, Levine, Williams, McLaughlin, and Candell (1989) developed basis functions by computing the Q orthonormal principal functions. Ramsay & Winsberg (1991) used monotone spline basis functions for their quasiparametric approach and calculated Maximum Marginal Likelihood estimates for the item parameters.

Ramsay (1991) and Ramsay & Winsberg (1991) also summarized the work by Samejima (1977, 1979, 1984) in a series of unpublished reports where she developed a nonparametric approach to estimate the ICCs. To use that technique, examinees' responses to an independent test with known ICCs must be available.

CHAPTER 3

FILTERED POLYNOMIAL DENSITY ESTIMATION

Elphinstone (1985) proposed a unidimensional nonparametric approach to estimate an unknown distribution function $F(x)$ on the basis of a sample data from the distribution. This method was later called a filtered polynomial density estimation method by Sinnott (1997) when she extended this work to multivariate settings. Heinzmann (2005) has reviewed and evaluated the work of Elphinstone and Sinnott and provided a computer program. Since the filtered polynomial method is the foundation of the model that we are proposing, I will give a review of this approach in this chapter. Section 3.1 is a general introduction to the method, Section 3.2 will give a review on how to construct a positive polynomial which is essential in this filtered polynomial density estimation method, Section 3.3 will present a recurrence relation for evaluating the positive polynomial and Section 3.4 will describe the discrepancy function used to estimate the parameters and the model selection criteria.

3.1 Introduction

Let F be an unknown, continuous, one-dimensional distribution function, with derivative f. The goal is to estimate the density function, f, of the unknown distribution based on a given sample data X from this distribution. Elphinstone (1985)

used a known target distribution, H, and a monotonic polynomial[1], m, to approximate F. The cumulative distribution function (CDF) is estimated as

$$\hat{F} = H(m(x)), \tag{3.1.1}$$

and the density function is estimated as

$$f = \hat{F}' = [H(m(x))]' = h(m(x)) \cdot m'(x). \tag{3.1.2}$$

Here H is referred to as a filter, or a target distribution. It could be any known distribution, but usually is chosen to be a normal, exponential, Gamma or Beta distribution. Since any monotonic transformation may be approximated to an arbitrary degree by a monotonic polynomial of sufficiently high order, many continuous non-defective distributions may be approximated to arbitrary closeness by this method (Elphinstone, 1985).

Since the target distribution H could be chosen as any known distribution, no fixed functional form was specified for the unknown distribution F. However, this technique is not a pure nonparametric technique. The pure nonparametric approach usually makes no assumptions on the structure of the unknown distribution. But this filtered polynomial density approximation method does need an assumption of differentiability although it is a very weak assumption. Thus this method was considered semi-parametric by Elphinstone.

There are usually two classes of nonparametric density estimation techniques. One focuses on local construction. The estimate of $\hat{f}$ is based on the pieces of information provided by observations in a small neighborhood. The Kernel smoothing estimation method is one that falls into this category. Filtered polynomial density approximation

[1]In this dissertation, a monotonic polynomial refers to a monotonic increasing polynomial.

is a method in the other class which focuses on all information simultaneously and searches for a smooth function that maximizes the likelihood of the observed data. Elphinstone (1985), Sinnott (1997) and Heinzmann (2005) all provided mathematical justifications of using this filtered polynomial method in estimating density functions.

3.2 Construction of a positive polynomial

In Equation (3.1.1), $m(x)$ is a monotonic polynomial of the observed data. A necessary and sufficient condition for a continuous and differentiable polynomial to be monotonic increasing is that its derivative $m'(x)$, be positive.

Let the monotonic polynomial $m(x)$ be defined as

$$m_{2k+1}(x) = \sum_{i=0}^{2k+1} b_i x^i, \tag{3.2.1}$$

and its derivative $m'(x)$ be defined as

$$m'_k(x) = \sum_{i=0}^{2k} a_i x^i, \tag{3.2.2}$$

where a_i's and b_i's are all real numbers but must satisfy certain conditions (discussed below) to ensure that $m(x)$ is a monotonic increasing polynomial and $m'(x)$ is a positive polynomial.

Elphinstone (1985) showed that the conditions for a polynomial to be positive are that it must be of even order, its roots are complex and the coefficient multiplying the highest-order term is positive. These conditions lead to the following representation

$$m'_k(x|\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k1}, \gamma_{k2}) = \begin{cases} \gamma \prod_{j=1}^{k} [x - (\gamma_{j1} + i\gamma_{j2})][x - (\gamma_{j1} - i\gamma_{j2})] & k > 0 \\ \gamma & k = 0 \end{cases},$$

(3.2.3)

where γ is a real positive number, γ_{j1} and γ_{j2} are real numbers. $\gamma_{j1} \pm i\gamma_{j2}$ $(j = 1, 2, \ldots, k)$ are the complex roots of the polynomial. The roots will be real when γ_{j2} is zero.

Equation (3.2.3) can also be written as

$$m'_k(x|\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k1}, \gamma_{k2}) = \begin{cases} \gamma \prod_{j=1}^{k} (x^2 - 2\gamma_{j1}x + \gamma_{j1}^2 + \gamma_{j2}^2) & k > 0 \\ \gamma & k = 0 \end{cases} . \quad (3.2.4)$$

In computation, we will start from $k = 0$ and search through all k levels to decide what value of k is sufficient to yield acceptable fit. At step j, it is convenient to start with the values of γ_1 and γ_2 from step $j - 1$ (i.e. $\gamma_{j-1,1}$ and $\gamma_{j-1,2}$) and set $\gamma_{j1} = \gamma_{j2} = 0$. This requires that the functional form can be written in a recurrent form such that when $\gamma_{k1} = \gamma_{k2} = 0, m'_{k-1}(x|\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k-1,1}, \gamma_{k-1,2}) = m'_k(x|\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k,1}, \gamma_{k,2})$.

Equation (3.2.4) doesn't have this property. Elphinstone (1985) provided a reparametrization that can be written in a recurrent form.

Let $z_j = \gamma_{j1} + i\gamma_{j2}$ and $\overline{z}_j = \gamma_{j1} - i\gamma_{j2}$. When $k > 0$, a little algebra applied to Equation (3.2.3) gives,

$$\begin{aligned} m'_k(x|\gamma, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{k,1}, \gamma_{k,2}) &= \gamma \prod_{j=1}^{k} (x - z_j)(x - \overline{z}_j) \\ &= \gamma \prod_{j=1}^{k} \left[z_j \overline{z}_j (\tfrac{x}{\overline{z}_j} - 1)(\tfrac{x}{z_j} - 1) \right] \\ &= \gamma \prod_{j=1}^{k} [z_j \overline{z}_j] \prod_{j=1}^{k} \left[(\tfrac{x}{\overline{z}_j} - 1)(\tfrac{x}{z_j} - 1) \right] \\ &= \lambda \prod_{j=1}^{k} [(w_j x - 1)(\overline{w}_j x - 1)] , \end{aligned}$$

where

$\lambda = \gamma \prod_{j=1}^{k} [z_j \overline{z}_j] > 0$ (Since $\gamma > 0$ and $z_j \overline{z}_j = \gamma_{j1}^2 + \gamma_{j2}^2 > 0$),

$w_j = \frac{1}{\overline{z}_j} = \alpha_j + i\beta_j$ and $\overline{w}_j = \frac{1}{z_j} = \alpha_j - i\beta_j$, with

$\alpha_j = \frac{\gamma_{j1}}{\gamma_{j1}^2 + \gamma_{j2}^2}$ and $\beta_j = \frac{\gamma_{j2}}{\gamma_{j1}^2 + \gamma_{j2}^2}$.

Further expansion shows that a positive polynomial can be written as

$$m_k'(x|\lambda, \alpha_1, \beta_1, \ldots, \alpha_k, \beta_k) = \begin{cases} \lambda \prod_{j=1}^{k} (1 - 2\alpha_j x + (\alpha_j^2 + \beta_j) x^2) & k > 0 \\ \lambda & k = 0 \end{cases}, \quad (3.2.5)$$

where $\lambda > 0$ and $\beta_j > 0, j = 1, 2, \ldots, k$.

Note that with this representation, when α_k and β_k are zero, m_k' will be a positive polynomial with degree of $2k - 2$ and we have a nested structure.

It is not difficult to obtain the relationship between the coefficients of a positive polynomial with those of the corresponding monotonic polynomial. A positive polynomial with order $2k$ is specified as

$$m_k'(x|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = a_{k,0} + a_{k,1} x + a_{k,2} x^2 + \ldots + a_{k,2k} x^{2k}, \quad (3.2.6)$$

and the corresponding monotonic polynomial is given by

$$\begin{aligned} m_{2k+1}(x|\xi, \lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \xi + \textstyle\int_0^x m_k'(t|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) dt \\ &= \xi + a_{k,0} x + \tfrac{a_{k,1}}{2} x^2 + \tfrac{a_{k,2}}{3} x^3 + \cdots + \tfrac{a_{k,2k}}{2k+1} x^{2k+1} \\ &= b_{k,0} + b_{k,1} x + b_{k,2} x^2 + b_{k,3} x^3 + \cdots + b_{k,2k+1} x^{2k+1}, \end{aligned} \quad (3.2.7)$$

where $\boldsymbol{\alpha}^2 = [\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_k]$ and $\boldsymbol{\beta} = [\beta_1,\ \beta_2,\ \ldots,\ \beta_k]$.

Clearly, we have $b_{k,0} = \xi$, $b_{k,j} = \frac{a_{k,j-1}}{j}$, $j = 1, 2, \ldots, 2k + 1$.

[2]In this dissertation, bold Greek letter represents a vector and bold capital letter represents a matrix

3.3 Recurrence of the nonnegative polynomial

It could be shown from Equation (3.2.5) that the positive polynomial has a nested structure and can be represented in a recurrent form. This recurrent functional relationship can be represented in a matrix form (Browne, 1997).

Let $f_k(x) = 1 - 2\alpha_k x + \phi_k x^2$, where $\phi_k = \alpha_k^2 + \beta_k$, we then have

$m'_k(x) = f_k(x) \times m'_{k-1}(x)$.

For example,

$m'_0(x) = \lambda$,

$m'_1(x) = f_1(x) \times m'_0(x) = f_1(x) \times \lambda$,

$m'_2(x) = f_2(x) \times m'_1(x) = f_2(x) \times f_1(x) \times \lambda$,

$\vdots$

$m'_k(x) = f_k(x) \times m'_{k-1}(x) = f_k(x) \times \cdots \times f_2(x) \times f_1(x) \times \lambda$.

In the step from $k = 0$ to $k = 1$,

$m'_1(x) = \lambda f_1(x) = \lambda(1 - 2\alpha_1 x + \phi_1 x^2) = a_{10} + a_{11}x + a_{12}x^2$,

the coefficients are

$a_{10} = \lambda$

$a_{11} = -2\alpha_1\lambda$

$a_{12} = \phi_1\lambda$.

In matrix form, this can be written as

$$\begin{bmatrix} a_{10} \\ a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ -2\alpha_1 \\ \phi_1 \end{bmatrix} \lambda.$$

In the step from $k = 1$ to $k = 2$,

$$
\begin{aligned}
m_2'(x) &= f_2(x) \times m_1'(x) \\
&= (1 - 2\alpha_2 x + \phi_2 x^2)(a_{10} + a_{11}x + a_{12}x^2) \\
&= a_{20} + a_{21}x + a_{22}x^2 + a_{23}x^3 + a_{24}x^4,
\end{aligned}
$$

where

$$
\begin{aligned}
a_{20} &= a_{10} \\
a_{21} &= (-2\alpha_2)a_{10} + a_{11} \\
a_{22} &= \phi_2 a_{10} + (-2\alpha_2)a_{11} + a_{12} \\
a_{23} &= \phi_2 a_{11} + (-2\alpha_2)a_{12} \\
a_{24} &= \phi_2 a_{12} \quad .
\end{aligned}
$$

In matrix form, this can be written as

$$
\begin{bmatrix} a_{20} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{24} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2\alpha_2 & 1 & 0 \\ \phi_2 & -2\alpha_2 & 1 \\ 0 & \phi_2 & -2\alpha_2 \\ 0 & 0 & \phi_2 \end{bmatrix} \begin{bmatrix} a_{10} \\ a_{11} \\ a_{12} \end{bmatrix}.
$$

In the step from $k = 2$ to $k = 3$,

$$
\begin{aligned}
m_3'(x) &= f_3(x) \times m_2'(x) \\
&= (1 - 2\alpha_3 x + \phi_3 x^2)(a_{20} + a_{21}x + a_{22}x^2 + a_{23}x^3 + a_{24}x^4) \\
&= a_{30} + a_{31}x + a_{32}x^2 + a_{33}x^3 + a_{34}x^4 + a_{35}x^5 + a_{36}x^6,
\end{aligned}
$$

where

$$
\begin{aligned}
a_{30} &= a_{20} \\
a_{31} &= (-2\alpha_3)a_{20} + a_{21} \\
a_{32} &= \phi_3 a_{20} + (-2\alpha_3)a_{21} + a_{22} \\
a_{33} &= \phi_3 a_{21} + (-2\alpha_3)a_{22} + a_{23} \\
a_{34} &= \phi_3 a_{22} + (-2\alpha_3)a_{23} + a_{24} \\
a_{35} &= \phi_3 a_{23} + (-2\alpha_3)a_{24} \\
a_{36} &= \phi_3 a_{24} \quad .
\end{aligned}
$$

In matrix form, this can be written as

$$
\begin{bmatrix} a_{30} \\ a_{31} \\ a_{32} \\ a_{33} \\ a_{34} \\ a_{35} \\ a_{36} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2\alpha_3 & 1 & 0 & 0 & 0 \\ \phi_3 & -2\alpha_3 & 1 & 0 & 0 \\ 0 & \phi_3 & -2\alpha_3 & 1 & 0 \\ 0 & 0 & \phi_3 & -2\alpha_3 & 1 \\ 0 & 0 & 0 & \phi_3 & -2\alpha_3 \\ 0 & 0 & 0 & 0 & \phi_3 \end{bmatrix} \begin{bmatrix} a_{20} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{24} \end{bmatrix},
$$

$\vdots$

$\vdots$

In summary, let vector $\boldsymbol{a}_k$ represent the coefficients of the positive polynomial with length of $2k+1$,

$$
\boldsymbol{a}_k = \begin{bmatrix} a_{k0} \\ a_{k1} \\ \vdots \\ a_{k,2k} \end{bmatrix},
$$

and $\boldsymbol{T}_k$ represent a $(2k+1) \times (2k-1)$ matrix with the typical form as

$$\boldsymbol{T}_k = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -2\alpha_k & 1 & 0 & \cdots & 0 & 0 & 0 \\ \phi_k & -2\alpha_k & 1 & \cdots & 0 & 0 & 0 \\ 0 & \phi_k & -2\alpha_k & \cdots & 0 & 0 & 0 \\ 0 & 0 & \phi_k & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \phi_k & -2\alpha_k & 1 \\ 0 & 0 & 0 & \cdots & 0 & \phi_k & -2\alpha_k \\ 0 & 0 & 0 & \cdots & 0 & 0 & \phi_k \end{bmatrix},$$

when $k = 0$, we have $m_0'(x|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = a_{00} = \lambda$, and we can easily show that

$\boldsymbol{a}_0 = [a_{00}] = \lambda$,

$\boldsymbol{a}_1 = \boldsymbol{T}_1\boldsymbol{a}_0 = \boldsymbol{T}_1\lambda$,

$\boldsymbol{a}_2 = \boldsymbol{T}_2\boldsymbol{a}_1 = \boldsymbol{T}_2\boldsymbol{T}_1\lambda$,

$\boldsymbol{a}_3 = \boldsymbol{T}_3\boldsymbol{a}_2 = \boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\lambda$,

$\vdots$

$\boldsymbol{a}_k = \boldsymbol{T}_k\boldsymbol{a}_{k-1} = \boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda$.

This greatly simplifies the computational load since when calculating the derivatives, only the corresponding $\boldsymbol{T}$ matrix will be involved, for example,

$\frac{d\boldsymbol{a}_3}{d\lambda} = \boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1$,

$\frac{d\boldsymbol{a}_4}{d\alpha_2} = \boldsymbol{T}_4\boldsymbol{T}_3\frac{d\boldsymbol{T}_2}{d\alpha_2}\boldsymbol{T}_1\lambda$,

and $\frac{d^2\boldsymbol{a}_4}{d\beta_2 d\beta_3} = \boldsymbol{T}_4\frac{d\boldsymbol{T}_3}{d\beta_3}\frac{\boldsymbol{T}_2}{d\beta_2}\boldsymbol{T}_1\lambda$.

Computing time can be saved by performing the multiplication backwards, i.e. $\boldsymbol{T}_4\{\boldsymbol{T}_3[\boldsymbol{T}_2(\boldsymbol{T}_1\lambda)]\}$.

3.4 Density estimation and model selection

It has been shown by Elphinstone (1985) that for any given $\epsilon > 0$ and any continuous distribution F, one can find a monotonic polynomial $m(x)$ such that $|F - H(m(x))| < \epsilon$. Two ways could be used to search for the monotonic polynomial $m(x)$. One relies on minimizing the distance between F and $H(m(x))$. Since F is unknown, a natural candidate would be the empirical distribution function $F_n(x)$. The other way works on the density function itself and aims at maximizing the likelihood of the observed data.

In the distance-minimizing class, there are three commonly used measures of the distance, the Kolmogorov-Smirnov distance, the Cramer-von Mises distance and the Anderson-Darling distance. If we use $\hat{F}_k(x)$ to represent the estimate of $H(m(x))$, the Kolmogorov-Smirnov distance between $F_n(x)$ and $\hat{F}_k(x)$ is given by

$$K_{kn} = \sup_{-\infty<x<\infty} \left| F_n(x) - \hat{F}_k(x) \right|. \tag{3.4.1}$$

The class of weighted Cramer-von Mises distances is defined as

$$C_{kn} = n \int \left[F_n(x) - \hat{F}_k(x) \right]^2 w(x)\, d\hat{F}_k(x), \tag{3.4.2}$$

where $w(x)$ is weight. If $w(x) = 1$, this distance is called the Cramer-von Mises Distance. If $w(x) = \frac{1}{\hat{F}_k(x)(1-\hat{F}_k(x))}$, the distance is called the Anderson-Darling distance.

We shall be primarily concerned with obtaining the maximum likelihood estimates by minimizing the negative log likelihood function. Let $\hat{f}_k(x)$ be the density function derived from $\hat{F}_k(x)$, i.e.

$$\hat{f}_k(x) = h(m_{2k+1}(x))m_k'(x).$$

The maximum likelihood estimates will be obtained by minimizing the negative log likelihood function

$$F = -\log \prod_{j=1}^{n} \hat{f}_k(x_j) = -\sum_{j=1}^{n} \log \hat{f}_k(x_j). \tag{3.4.3}$$

When the values of the observed data x are too large, an overflow problem might arise during the computation especially when we are working with high degree polynomials. To avoid this problem, Elphinstone (1985) suggested applying a linear transformation to the observed data, for example, using the standardized data in the analysis. After the estimates of parameters are obtained, they can be transformed back to the original scale.

The optimal estimates for the monotonic polynomial will be searched for at each stage of k. Usually we start from $k = 0$ where the positive polynomial is a real positive constant ($m_0'(x) = \lambda$) and the corresponding monotonic polynomial is a linear function. Then the search goes to $k = 1$ and so on. One can find a monotonic polynomial, $m(x)$, such that $H(m(x))$ approximates the unknown distribution F to an arbitrary degree of accuracy. However, a model might have excellent fit because there are too many unnecessary parameters in the model that are absorbing some random errors. A model like this would have low generalizability to other datasets. A trade-off between the goodness of fit and the model complexity should be considered. Some criterion should be used to stop the fitting at a certain k value to prevent over-fitting.

Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two criteria that can be used in this filtered polynomial density estimation to decide on the k value for the monotonic polynomial.

The AIC is a very commonly used criterion in model selection developed by Akaike (1973). It is obtained by minimizing the Kullback-Leibler distance and is defined as

$$AIC = -2 \log ML_k + 2p_k, \tag{3.4.4}$$

where $\log ML_k$ is the log likelihood function evaluated at the maximum likelihood estimates for the fitted model at stage k and p_k is the number of parameters in the model. This criterion includes two terms. The first term is twice the negative value of the maximized log likelihood and can be considered as a measure of lack of fit. It is directly related to the maximum likelihood estimates. The AIC attempts to maximize the expected maximum likelihood for the whole data space. Its second term is twice the number of parameters and can be considered as a measure of model complexity. Thus models with more parameters will be penalized by the second term. This would help avoid choosing some overly parameterized model. The model with the minimum AIC value is chosen as the best model.

Schwarz (1978) suggested a criterion which tends to the logarithm of the Bayes factor when sample size increases to infinity. Minus twice the Schwarz criterion is often called the Bayesian Information Criterion (BIC). Thus BIC can be viewed as a large sample approximation to the logarithm of the Bayes factor but it doesn't require the prior density as in calculating the Bayes factor. BIC is defined as

$$BIC = -2 \log ML_k + p_k \log N, \tag{3.4.5}$$

where N is the number of observations. This criterion also includes two terms. The first term is also twice the negative value of the natural logarithm of the maximized likelihood function for model k. This term can be viewed as a measure of the badness of fit. The second term is a measure of model complexity using the number of free

parameters in the model and the number of independent observations. A model with a minimum BIC among a set of competing models should be selected.

Note that the first term of BIC is the same as the first term of AIC but the second term of BIC depends on the sample size as well as the number of parameters. As sample size increases, AIC would favor complex model since the second term is a constant and it will be dominated by the first term where complex model will have smaller values. BIC has some control over the sample size by having the $\log N$ in second term. Compared with AIC, BIC favors a less complex model with fewer parameters when the sample size is large.

CHAPTER 4

LOGISTIC FUNCTION OF A MONOTONIC POLYNOMIAL (L-MP)

The idea of applying a monotonic polynomial to the IRF comes from the similarity between an ICC and a cumulative distribution. Although strictly speaking ICCs are not cumulative distributions, they do have the mathematical properties of cumulative distributions. The ICCs are all non-decreasing and the dependent variable in the ICC is a probability ranging from 0 to 1. These similarities initiate the idea of applying Elphinstone (1985)'s procedure to the IRF and using a monotonic polynomial transformation on the abilities to estimate the IRF. This procedure has the possibility of increasing the model-data fit, especially when the true ICC doesn't follow a logistic curve. Only the binary response data under the unidimensional assumption are considered in this dissertation. This work could, however, be extended to non-binary response data or to the three parameter logistic model.

An introduction to the proposed L-MP model will be provided in Section 4.1. Details of the parameter estimation procedure will be given in Section 4.2. Model selection criteria will be discussed in Section 4.3.

4.1 Model specification

Let us revisit the 1PL and 2PL functions,

$$P_i(\theta) = \frac{1}{1+e^{-a(\theta-b_i)}}$$
$$P_i(\theta) = \frac{1}{1+e^{-a_i(\theta-b_i)}}.$$

The exponents are both linear for the above equations. In the L-MP model, this linear exponent will be replaced by a monotonic polynomial $m_i(\theta)$, and the new equation will be

$$P_i(\theta) = \frac{1}{1+e^{-m_i(\theta)}}, \tag{4.1.1}$$

where $m_i(\theta)$ is the monotonic polynomial represented as

$$m_i(\theta) = \xi_i + b_{1i}\theta + b_{2i}\theta^2 + \cdots + b_{2k+1,i}\theta^{2k+1}, \tag{4.1.2}$$

and its corresponding positive polynomial is

$$m'(\theta) = b_{1i} + 2b_{2i}\theta + \cdots + (2k+1)b_{2k+1,i}\theta^{2k}$$

$$= a_{0i} + a_{1i}\theta + \cdots + a_{2k,i}\theta^{2k}. \tag{4.1.3}$$

The relationship between the coefficients of the monotonic polynomial and the corresponding positive polynomial is

$$b_{j,i} = \frac{a_{j-1,i}}{j},\ j = 1, 2, \ldots, 2k+1. \tag{4.1.4}$$

Clearly, the 1PL and 2PL models are special cases of this more general model. When $k = 0$, $m_i(\theta) = \xi_i + b_{1i}\theta$ will be equivalent to the exponent in the 2PL model. A further constraint of $b_{1i} = 1$ results in a 1PL model.

The construction of the positive polynomial has already been discussed in Section 3.2. After some reparametrization, the positive polynomial is defined as

$$m'_k(\theta|\lambda,\alpha_1,\beta_1,\ldots,\alpha_k,\beta_k) = \begin{cases} \lambda\prod_{j=1}^{k}\left[1-2\alpha_j\theta+(\alpha_j^2+\beta_j)\theta^2\right] & k>0 \\ \lambda & k=0 \end{cases} \tag{4.1.5}$$

with $\lambda > 0$. The above equation can be represented using matrix notation (Browne, 1997) as

$$m'_k(\theta|\lambda,\boldsymbol{\alpha},\boldsymbol{\beta}) = \begin{cases} (\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda)'\boldsymbol{\theta} & k>0 \\ \lambda & k=0 \end{cases}, \tag{4.1.6}$$

where $\boldsymbol{\theta} = [1\ \theta\ \theta^2\ \cdots\ \theta^{2k}]'$ and $\boldsymbol{T}_k$ is a $(2k+1)\times(2k-1)$ matrix with parameters (α_k,β_k). For example, $\boldsymbol{T}_3$ is a 7×5 matrix with parameters (α_3,β_3), i.e.

$$\boldsymbol{T}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2\alpha_3 & 1 & 0 & 0 & 0 \\ \phi_3 & -2\alpha_3 & 1 & 0 & 0 \\ 0 & \phi_3 & -2\alpha_3 & 1 & 0 \\ 0 & 0 & \phi_3 & -2\alpha_3 & 1 \\ 0 & 0 & 0 & \phi_3 & -2\alpha_3 \\ 0 & 0 & 0 & 0 & \phi_3 \end{bmatrix},$$

where $\phi_3 = \alpha_3^2+\beta_3$ $(\beta_3 > 0)$.

There are two sets of parameters to be estimated in this model, item parameters and examinees' abilities. For each item i, there are $(2k_i+2)$ elements in the coefficient vector $\boldsymbol{b}$. These coefficients are functions of $[\xi,\lambda,\boldsymbol{\alpha},\boldsymbol{\beta}]$. Thus for each item i, there are $(2k_i+2)$ real parameters to be estimated, i.e. $[\xi_i,\lambda_i,\alpha_{1i},\ldots,\alpha_{ki},\beta_{1i},\ldots,\beta_{ki}]$ with constraints $\lambda_i > 0$ and $\beta_{ji} > 0, j = 1,2,\ldots,k$. For a test with N examinees, there will be N ability parameters to estimate, i.e. $[\theta_1,\theta_2,\ldots,\theta_N]$. For a test of n items

applied on N examinees there will be a total of $\sum_{i=1}^{n}(2k_i+2)+N$ parameters to be estimated.

In IRT models, both examinees' abilities and item parameters are to be estimated. This leads to an identification problem. For example, in the 2PL model,

$$P = \frac{1}{1+\exp -a(\theta-b)},$$

suppose that a linear transformation is imposed on the ability

$\theta^* = c\theta + d$,

and a corresponding transformation is imposed on item parameters

$a^* = \frac{a}{c}$ and $b^* = cb + d$,

we will have $P^*(\theta) = P(\theta)$ and the ICCs will be the same for two sets of parameters.

In general, if any monotonic transformation is applied on θ,

$$\tau = g(\theta),$$

then a curve $p^*(\tau)$ where $p^* = p \circ g^{-1}$ will generate the same curve on the original metric since $p^*(\tau) = p[g^{-1}(\tau)] = p[g^{-1}(g(\theta))] = p(\theta)$.

In the JML estimation method, this problem is usually addressed by fixing the location and scale of θ by standardizing it to a standard normal distribution after each estimation cycle. In the MML estimation method, a prior density function of ability can be thought of as solving this lack of identifiability problem to some extent. In the L-MP model, we will solve this problem by imposing a standard normal distribution on the surrogate abilities and using a surrogate-based two-stage estimation procedure to estimate the ICCs and abilities.

4.2 Parameter estimation

There are two sets of parameters in the model. If we use $\boldsymbol{\gamma}$ to denote the whole parameter vector, $\boldsymbol{\gamma}_1$ to represent the item parameters and $\boldsymbol{\gamma}_2$ to be the abilities, we

have

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma_1} \\ \boldsymbol{\gamma_2} \end{bmatrix},$$

where $\boldsymbol{\gamma_1} = [\xi_i, \lambda_i, \alpha_{1i}, \beta_{1i}, \ldots, \alpha_{ki}, \beta_{ki}]', i = 1, 2, \ldots, n$, and $\boldsymbol{\gamma_2} = [\theta_1, \theta_2, \ldots, \theta_N]'$.

Let $U_i = 1$ represents a correct response for item i and $U_i = 0$ for an incorrect response. The probability of a response, U_i, can be expressed as

$$\begin{aligned} P_i(U_i|\theta) &= P_i(U_i = 1|\theta)P_i(U_i = 0|\theta) \\ &= P_i^{U_i} Q_i^{1-U_i}, \end{aligned}$$

where $Q_i = 1 - P_i$.

Let $\boldsymbol{u}_s = [u_{s1}, u_{s2}, \ldots, u_{sn}]$, $s = 1, 2, \ldots, N$, be the response vector of the examinee s on n items, and $[\theta_1, \theta_2, \ldots, \theta_N]$ be the vector of abilities for the N examinees. The likelihood function for the N examinees on n items is

$$L(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \prod_{s=1}^{N} \prod_{i=1}^{n} P_{si}^{U_{si}} Q_{si}^{1-U_{si}}, \tag{4.2.1}$$

where $P_{si} = P_i(\theta_s)$ is the probability of getting the ith item correct by examinee s. The estimates of the item parameters and the latent abilities are obtained by minimizing the negative logarithm likelihood function,

$$F = -\sum_{s=1}^{N} \sum_{i=1}^{n} \left[U_{si} \log P_{si} + (1 - U_{si}) \log (1 - P_{si}) \right]. \tag{4.2.2}$$

The user will need to specify the value of k which determines the degree of the monotonic polynomial. The procedure will start from $k = 0$, and then $k = 1$, ..., until the k value specified by the user. To avoid the lack of identification problem, a surrogate-based two-stage estimation method similar to Ramsay (1991)'s procedure will be applied to estimate the item parameters and the abilities. In Ramsay's procedure, test scores are used as the ranking basis to obtain the quantiles of a standard

normal distribution. The most obvious problem with test scores is that ties occur very frequently especially for a short test with many examinees. In TESTGRAF, ranks are randomly assigned to the tied test scores. To avoid this problem, the first principal component scores will be used as the ranking basis. The first principal component scores retain the information of rankings in examinees' abilities yet reduce the chances of tied ranks. The ranks are then transformed to the quantiles of a standard normal distribution q_i. By fixing the distribution of the surrogate abilities as a standard normal distribution, the identification problem is avoided. This step also helps to avoid the possibility of the overflow problem with high degree polynomials. The parameters in the IRF are then obtained by minimizing the negative log likelihood function as defined in Equation (4.2.2).

A Newton-Raphson algorithm will be applied at stage one to obtain the estimates of item parameters. Details of the first and second derivatives of the negative log likelihood function with respect to each item parameter are given in the next section. Once the estimates of the item parameters have been obtained, we will move on to stage two to estimate the abilities.

4.2.1 Estimation of item parameters

A Newton-Raphson method is applied at stage one to obtain the item parameters. The estimates of $\boldsymbol{\gamma_1}$ at the $(j+1)$th iteration is obtained by

$$\hat{\boldsymbol{\gamma}}_{\mathbf{1}}{}^{(j+1)} = \hat{\boldsymbol{\gamma}}_{\mathbf{1}}{}^{(j)} - \alpha \boldsymbol{H_j}^{-1} \boldsymbol{g_j}, \tag{4.2.3}$$

where $\boldsymbol{g_j}$ is the gradient of the negative log likelihood function in Equation (4.2.2) with respect to $\boldsymbol{\gamma}_1$ evaluated at the jth iteration, e.g. $\frac{dF}{d\boldsymbol{\gamma}_1}$, and it will be a $[\sum_{i=1}^{n}(2k_i + 2)] \times 1$ vector. $\boldsymbol{H}_j$ is the Hessian matrix, i.e. $\frac{d^2F}{d\boldsymbol{\gamma}_1 d\boldsymbol{\gamma}_1'}$, and will be a $[\sum_{i=1}^{n}(2k_i + 2)] \times$

$[\sum_{i=1}^{n}(2k_i+2)]$ matrix. The details of elements in the Hessian matrix will be given below. α is the step size and it is usually set to be between 0 and 1. This helps to adjust the change of the parameters by stepping back when the increment is too large to ensure that the function value is less than that from the previous step. We will start from $\alpha = 1$ and will reduce the α by half until we reach $F(\hat{\gamma}_{j+1}) < F(\hat{\gamma}_j)$.

The first derivative of the negative log likelihood function with respect to any parameter π is

$$\begin{aligned}\frac{dF}{d\pi} = \frac{d}{d\pi}(-\log L) &= \frac{d}{d\pi}\left(-\sum_{s=1}^{N}\sum_{i=1}^{n}\left[U_{si}\log P_{si} + (1-U_{si})\log\left(1-P_{si}\right)\right]\right) \\ &= -\sum_s\sum_i\left(\frac{U_{si}-P_{si}}{P_{si}(1-P_{si})}\frac{dP_{si}}{d\pi}\right),\end{aligned}$$

and we have

$\frac{dP_{si}}{d\pi} = \frac{d}{d\pi}(\frac{1}{1+e^{-m_i(\theta_s)}}) = P_{si}(1-P_{si})\frac{dm_i(\theta_s)}{d\pi}$.

Thus we have a general formula

$$\frac{dF}{d\pi} = -\sum_s\sum_i(U_{si}-P_{si})\frac{dm_i(\theta_s)}{d\pi}, \tag{4.2.4}$$

where $m_i(\theta_s)$ represents the monotonic polynomial.

Note the summation can be dropped depending on what parameter we are taking derivative with respect to. If the parameter π is related to a specific item, then $\sum_i$ can be dropped because all other terms will be zero.

With the local independence assumption, the Hessian matrix of the item parameters will be a block diagonal matrix. Thus we can work item by item to avoid a huge dimensional Hessian matrix. In the following derivation, the derivatives are calculated for a specific item. For simplicity purpose, the subscript for the item will be dropped.

The gradient in Equation (4.2.3) for a specific item is

$$\boldsymbol{g} = \begin{bmatrix} \frac{dF}{d\xi} \\ \frac{dF}{d\lambda} \\ \frac{dF}{d\alpha_1} \\ \vdots \\ \frac{dF}{d\alpha_k} \\ \frac{dF}{d\beta_1} \\ \vdots \\ \frac{dF}{d\beta_k} \end{bmatrix}.$$

Suppose a typical element in $\boldsymbol{\gamma}_1$ is η_l, from the result in Equation (4.2.4), the typical element in $\boldsymbol{g}$ will be

$$\frac{dF}{d\eta_l} = -\sum_s (U_s - P_s)\frac{dm}{d\eta_l}, \tag{4.2.5}$$

where m is a monotonic polynomial function with coefficients $(b_1, \ldots, b_{2k+1})$, b_j is a function of a_{j-1} and thus is also a function of the item parameters $[\xi, \lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}]$.

The chain rule is used to obtain $\frac{dm}{d\eta_l}$ in general:
$\frac{dm}{d\eta_l} = \sum_{j=0}^{2k} \frac{dm}{da_j}\frac{da_j}{d\eta_l}$.

Applying the chain rule again,

$$\begin{aligned} \frac{dm}{da_i} &= \sum_{j=1}^{2k+1} \frac{dm}{db_j}\frac{db_j}{da_i} \\ &= \frac{dm}{db_1}\frac{db_1}{da_i} + \frac{dm}{db_2}\frac{db_2}{da_i} + \cdots + \frac{dm}{db_{2k+1}}\frac{db_{2k+1}}{da_i} \\ &= \frac{dm}{db_{i+1}}\frac{db_{i+1}}{da_i} \\ &= \frac{1}{i+1}\theta^{i+1}, \ i = 0, 1, 2, \ldots, 2k. \end{aligned}$$

We know that a_0, a_1, …, a_{2k} are all functions of $(\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\frac{da_j}{d\eta_l}$, $j = 0, 1, 2, \ldots, 2k$, are the elements of $\frac{d\boldsymbol{a}_k}{d\eta_l}$, where $\boldsymbol{a}_k$ can be computed as

$$\boldsymbol{a}_k = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{2k-1} \\ a_{2k} \end{bmatrix} = \boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda.$$

So we have

$$\begin{aligned} \frac{dm}{d\eta_l} &= \begin{bmatrix} \frac{dm}{da_0} & \frac{dm}{da_1} & \frac{dm}{da_2} & \cdots & \frac{dm}{da_{2k}} \end{bmatrix} \begin{bmatrix} \frac{da_0}{d\eta_l} \\ \frac{da_1}{d\eta_l} \\ \frac{da_2}{d\eta_l} \\ \vdots \\ \frac{da_{2k}}{d\eta_l} \end{bmatrix} \\ &= \begin{bmatrix} \theta & \frac{1}{2}\theta^2 & \frac{1}{3}\theta^3 & \cdots & \frac{1}{2k+1}\theta^{2k+1} \end{bmatrix} \frac{d\boldsymbol{a}_k}{d\eta_l}, \end{aligned}$$

and

$$\frac{d\boldsymbol{a}_k}{d\lambda} = \frac{d(\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda)}{d\lambda} = \boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1.$$

For $\frac{d\boldsymbol{a}_k}{d\alpha_j}$ and $\frac{d\boldsymbol{a}_k}{d\beta_j}$, $j = 1, 2, \ldots, k$,

$$\frac{d\boldsymbol{a}_k}{d\alpha_j} = \frac{d(\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda)}{d\alpha_j} = \boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\frac{d\boldsymbol{T}_j}{d\alpha_j}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda,$$

$$\frac{d\boldsymbol{a}_k}{d\beta_j} = \frac{d(\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda)}{d\beta_j} = \boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\frac{d\boldsymbol{T}_j}{d\beta_j}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda.$$

So it is easy to obtain

$$\frac{dm}{d\xi} = 1, \tag{4.2.6}$$

$$\frac{dm}{d\lambda} = \begin{bmatrix} \theta & \frac{1}{2}\theta^2 & \frac{1}{3}\theta^3 & \cdots & \frac{1}{2k+1}\theta^{2k+1} \end{bmatrix} (\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\boldsymbol{T}_2\boldsymbol{T}_1), \tag{4.2.7}$$

$$\frac{dm}{d\alpha_j} = \begin{bmatrix} \theta & \frac{1}{2}\theta^2 & \frac{1}{3}\theta^3 & \cdots & \frac{1}{2k+1}\theta^{2k+1} \end{bmatrix} (\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\frac{d\boldsymbol{T}_j}{d\alpha_j}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda), \tag{4.2.8}$$

$$\frac{dm}{d\beta_j} = \begin{bmatrix} \theta & \frac{1}{2}\theta^2 & \frac{1}{3}\theta^3 & \cdots & \frac{1}{2k+1}\theta^{2k+1} \end{bmatrix} (\boldsymbol{T}_k\boldsymbol{T}_{k-1}\cdots\frac{d\boldsymbol{T}_j}{d\beta_j}\cdots\boldsymbol{T}_2\boldsymbol{T}_1\lambda), \tag{4.2.9}$$

for $j = 1, 2, \ldots, k$.

Now the elements in the gradient can be obtained by substituting Equations (4.2.6), (4.2.7), (4.2.8) and (4.2.9) into Equation (4.2.5).

The Hessian matrix in Equation (4.2.3) for a specific item is

$$\boldsymbol{H} = \begin{bmatrix} \frac{d^2F}{d\xi d\xi} & \frac{d^2F}{d\xi d\lambda} & \frac{d^2F}{d\xi d\alpha_1} & \cdots & \frac{d^2F}{d\xi d\alpha_k} & \frac{d^2F}{d\xi d\beta_1} & \cdots & \frac{d^2F}{d\xi d\beta_k} \\ \frac{d^2F}{d\lambda d\xi} & \frac{d^2F}{d\lambda d\lambda} & \frac{d^2F}{d\lambda d\alpha_1} & \cdots & \frac{d^2F}{d\lambda d\alpha_k} & \frac{d^2F}{d\lambda d\beta_1} & \cdots & \frac{d^2F}{d\lambda d\beta_k} \\ \frac{d^2F}{d\alpha_1 d\xi} & \frac{d^2F}{d\alpha_1 d\lambda} & \frac{d^2F}{d\alpha_1 d\alpha_1} & \cdots & \frac{d^2F}{d\alpha_1 d\alpha_k} & \frac{d^2F}{d\alpha_1 d\beta_1} & \cdots & \frac{d^2F}{d\alpha_1 d\beta_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{d^2F}{d\alpha_k d\xi} & \frac{d^2F}{d\alpha_k d\lambda} & \frac{d^2F}{d\alpha_k d\alpha_1} & \cdots & \frac{d^2F}{d\alpha_k d\alpha_k} & \frac{d^2F}{d\alpha_k d\beta_1} & \cdots & \frac{d^2F}{d\alpha_k d\beta_k} \\ \frac{d^2F}{d\beta_1 d\xi} & \frac{d^2F}{d\beta_1 d\lambda} & \frac{d^2F}{d\beta_1 d\alpha_1} & \cdots & \frac{d^2F}{d\beta_1 d\alpha_k} & \frac{d^2F}{d\beta_1 d\beta_1} & \cdots & \frac{d^2F}{d\beta_1 d\beta_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{d^2F}{d\beta_k d\xi} & \frac{d^2F}{d\beta_k d\lambda} & \frac{d^2F}{d\beta_k d\alpha_1} & \cdots & \frac{d^2F}{d\beta_k d\alpha_k} & \frac{d^2F}{d\beta_k d\beta_1} & \cdots & \frac{d^2F}{d\beta_k d\beta_k} \end{bmatrix}.$$

Let η_i and η_j be two typical elements in the item parameter vector $\boldsymbol{\gamma}_1$. The monotonic polynomial m will be a function of η_i and η_j. The typical element in the Hessian matrix will be $\frac{d^2F}{d\eta_i d\eta_j}$. Using the results in Equation (4.2.4), it could be computed as

$$\frac{d^2F}{d\eta_i d\eta_j} = \sum_s \left[PQ \frac{dm}{d\eta_i} \frac{dm}{d\eta_j} - (u_s - P_s) \frac{d^2m}{d\eta_i d\eta_j} \right], \tag{4.2.10}$$

where $\frac{dm}{d\eta_i}$ $(\eta_i = \xi, \lambda, \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k)$ can be obtained through Equations (4.2.6), (4.2.7), (4.2.8) and (4.2.9) and $\frac{d^2m}{d\eta_i d\eta_j}$ can be computed as

$$\frac{d^2m}{d\eta_i d\eta_j} = \frac{d}{d\eta_i} \left(\frac{dm}{d\boldsymbol{a}_k}' \frac{d\boldsymbol{a}_k}{d\eta_l} \right) = \frac{dm}{d\boldsymbol{a}_k'} \frac{d^2\boldsymbol{a}_k}{d\eta_i d\eta_j}, \tag{4.2.11}$$

since the item parameters and the abilities are assumed to be uncorrelated. Note that there are some special elements in this expression, i.e. $\frac{d^2 m}{d\xi d\eta_l} = 0$.

The term $\frac{d^2 \boldsymbol{a}_k}{d\eta_i d\eta_j}$ is the second derivatives of $\boldsymbol{a}_k$ with respect to the parameters η_i and η_j. For example, if $\eta_i = \alpha_2$, $\eta_j = \beta_2$,

$\frac{d^2 \boldsymbol{a}_k}{d\eta_i d\eta_j} = \frac{d}{d\alpha_2}(\frac{d\boldsymbol{a}_k}{d\beta_2}) = \boldsymbol{T}_k \boldsymbol{T}_{k-1} \cdots \frac{d^2 \boldsymbol{T}_2}{d\alpha_2 d\beta_2} \boldsymbol{T}_1 \lambda$,

or if $\eta_i = \alpha_3$, $\eta_j = \beta_2$,

$\frac{d^2 \boldsymbol{a}_k}{d\eta_i d\eta_j} = \frac{d}{d\alpha_3}(\frac{d\boldsymbol{a}_k}{d\beta_2}) = \boldsymbol{T}_k \boldsymbol{T}_{k-1} \cdots \frac{d\boldsymbol{T}_3}{d\alpha_3} \frac{d\boldsymbol{T}_2}{d\beta_2} \boldsymbol{T}_1 \lambda$.

Thus the values of the second derivatives of the monotonic polynomial with respect to the item parameters as in Equation (4.2.11) can be substituted into Equation (4.2.10) to obtain the elements in the Hessian matrix.

The Hessian matrix will be replaced by the negative information matrix if the Hessian matrix is not positive definite. The information matrix is the negative of the expectation of the Hessian matrix. Since the response U is a Bernoulli distribution with expectation P, the second term in Equation (4.2.10) can be dropped when computing the information matrix.

The first and second derivatives have been checked by comparing with the numerical derivatives $f' = \frac{f(x+\epsilon)-f(x-\epsilon)}{2\epsilon}$. When $\epsilon = 1.0 \times 10^{-4}$, it was found that the difference between the numerical derivative and the analytical derivative was always less than 10e-3.

A modification of the Newton-Raphson method (Jennrich & Sampson, 1968), based on stepwise linear regression techniques, will be used to handle the constrained boundaries of parameters λ and $\boldsymbol{\beta}$ ($\lambda > 0$ and $\boldsymbol{\beta} > 0$).

The convergence criterion for the Newton-Raphson procedure in Equation (4.2.3) in this stage is set to be that the maximum element of the gradient be less than 10e-4.

4.2.2 Estimation of the abilities

The estimates of item parameters from stage one will be treated as known at stage two to estimate the abilities. To help prevent obtaining unreasonable ability estimates, a standard normal prior distribution will be used. If the prior contains useful information and if the estimates are substantially different from the mean of the prior distribution, shrinkage will occur and will restrain estimates from unreasonable values.

Bayes expected a posteriori (EAP) estimates as discussed in Bock & Aitkin (1981) will be used to estimate abilities. From Bayes theorem, the distribution of θ given the response patterns and the item parameters is

$$g(\theta|\boldsymbol{U}_s, \boldsymbol{\gamma}_1) = \frac{P(\boldsymbol{U} = \boldsymbol{U}_s|\theta, \boldsymbol{\gamma}_1)g(\theta)}{P(\boldsymbol{U} = \boldsymbol{U}_s)}, \tag{4.2.12}$$

where

$$P(\boldsymbol{U} = \boldsymbol{U}_s) = \textstyle\int_{-\infty}^{\infty} P(\boldsymbol{U} = \boldsymbol{U}_s|\theta)g(\theta)\ d\theta.$$

So the conditional expectation of θ given $\boldsymbol{U}_s$ and $\boldsymbol{\gamma}_1$ is

$$E(\theta|\boldsymbol{U}_s, \boldsymbol{\gamma_1}) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) P(\boldsymbol{U} = \boldsymbol{U}_s|\theta, \boldsymbol{\gamma_1})\ d\theta}{\int_{-\infty}^{\infty} g(\theta) P(\boldsymbol{U} = \boldsymbol{U}_s|\theta)\ d\theta}. \tag{4.2.13}$$

Since the above equation involves integrals, we use Gauss-Hermite numerical approximations:

$$E(\theta|\boldsymbol{U}_s, \boldsymbol{\gamma_1}) = \frac{\sum_k^q X_k L(X_k) A(X_k)}{\sum_k^q L(X_k) A(X_k)}, \tag{4.2.14}$$

where X_k is the evaluation point, $A(X_k)$ is the probability density from a standard normal distribution at X_k and $L(X_k)$ is the likelihood function value evaluated at X_k.

When the true ability parameters are from a standard normal distribution, it was found from the simulation studies that adding a prior to estimate the abilities gives substantially better estimates than the pure maximum likelihood estimates. The reason is that the assumption of a normal prior for the abilities involved in the EAP estimates retains the normalization imposed on the surrogate abilities and thus prevents unreasonable ability estimates. Another advantage is that since the pure ML estimates can't handle zero or perfect scores, a proportion of examinees will need to be deleted from the analysis. These cases with zero or perfect scores carry useful information, however, in estimating the item curves. With the EAP estimates, zero or perfect scores are not a problem and all cases can be included in the analysis.

4.3 Model Selection Criteria

Three types of model selection criteria will be used for this L-MP model. Two of them are the AIC and BIC as described in Section 3.4. Since the model with $k = j$ is nested within the model with $k = j + 1$, a likelihood ratio test can be used to test if the additional parameters in the general model significantly improve the goodness of fit. If M_j is used to represent the L-MP model with $k = j$ and M_{j+1} is used to represent the model with $k = j + 1$, when the model goes from $k = j$ to $k = j + 1$, two additional parameters are introduced to the model. The statistic $-2(\log M_j - \log M_{j+1})$ will be a χ^2 distribution with 2 degrees of freedom, i.e.

$$-2(\log M_j - \log M_{j+1}) \ \sim \ \chi^2_2. \tag{4.3.1}$$

This could be used to test the null hypothesis: H_0: $\alpha_{j+1} = 0$, $\beta_{j+1} = 0$. A significant result implies that additional parameters in model M_{j+1} help to improve the model-data fit significantly.

Because H_0 implies that β_{j+1} is on the boundary of zero, standard regularity conditions for the derivation of asymptotic results are violated. This implies that the χ^2 test statistic might not have an asymptotic chi-square distribution. It may still be used, however, to give an indication of equivalent goodness-of-fit for $k = j$ and $k = j + 1$. The effectiveness of these three criteria will be evaluated through the simulation studies in Chapter 5.

CHAPTER 5

SIMULATIONS: PERFORMANCE OF THE L-MP MODEL

The L-MP model proposed in this dissertation is a general model which includes the 1PL and 2PL as special cases. The 1PL and 2PL models will capture most of the data features if the true model is a logistic function. However, if the data are not from such a logistic function, these methods will not be sufficient in modeling the data. The L-MP item response function is more flexible because of the extra parameters in the exponent. Correspondingly it should be able to capture the data characteristics better. The simulations run in this chapter aim at illustrating that this general model can produce similar results to other currently used programs like MULTILOG (MML/EM) (Thissen, Chen & Bock, 2003) and SYSTAT TESTATLOG (version 10.2) (JML) for the 2PL model, and can produce a better fit to the data when the true model is not a logistic function. We will also make comparisons of the L-MP model with the other nonparametric programs, such as TESTGRAF (Ramsay, 1991) and the irtNP package in R (Duncan & MacEachern, in press) for the Nonparametric Bayesian model.

5.1 Simulation design

Several factors could affect the estimates including: the number of items, the number of examinees, the degree of the polynomial and the dispersion of item difficulty. These factors will be considered jointly or partially in different simulation studies.

Four models were considered. Three of them differed in the degree, $2k+1$, of the polynomials as described in Equation (4.1.1) and (4.1.2)

$$P_i(\theta) = \frac{1}{1+e^{-m_i(\theta)}},$$

where $m_i(\theta)$ is the monotonic polynomial represented as

$$m_i(\theta) = \xi_i + b_{1i}\theta + b_{2i}\theta^2 + \cdots + b_{2k+1,i}\theta^{2k+1}.$$

Specifically, the models corresponding to different k values ($k=0$, $k=1$, $k=2$) are

$$\begin{aligned} P_i(\theta) &= \frac{1}{1+e^{-(\xi_i+b_{1i}\theta)}} && \text{for } k=0, \\ P_i(\theta) &= \frac{1}{1+e^{-(\xi_i+b_{1i}\theta+b_{2i}\theta^2+b_{3i}\theta^3)}} && \text{for } k=1, \\ \text{and} \quad P_i(\theta) &= \frac{1}{1+e^{-(\xi_i+b_{1i}\theta+b_{2i}\theta^2+b_{3i}\theta^3+b_{4i}\theta^4+b_{5i}\theta^5)}} && \text{for } k=2. \end{aligned}$$

When $k=0$, the L-MP model is equivalent to the 2PL model. The data were generated using the 2PL function with *difficulty* and *discrimination* parameters. To avoid the situation where a test is composed of items with similar difficulty levels, for example, a too easy test or a too hard test, such that examinees' abilities outside this range are hard to estimate, two ways were used to draw the difficulty parameter. One way was to fix the difficulty parameters, b, for a set of items as equally spaced on the range of -2.5 to 2.5. The other way was to make random draws from a truncated normal distribution from -2.5 to 2.5. To generate data sets that are practically meaningful, items with discrimination parameters that were too small or too large were avoided. In the current simulation study, the discrimination parameters, a, were

randomly drawn from a uniform distribution within the range of $[1.1, 1.8]$. This model was used to illustrate that the L-MP model with its associated estimation method could adequately recover parameters from the simpler parametric models. In summary, when $k = 0$, the parameters were drawn from

$a \sim$ unif$[1.1, 1.8]$,

$b \sim$ equally spaced on $[-2.5, 2.5]$ or from $N(0, 1)$ truncated at ± 2.5.

For the models corresponding to $k = 1$ or $k = 2$, it is more difficult to interpret the coefficients of the monotonic polynomial. After some initial experiments to determine values for the item parameters that would generate commonly seen IRFs, the item parameters were chosen to be randomly drawn from uniform distributions within certain ranges. Specifically, $\xi \sim$unif$[-1, 1]$, $\lambda \sim$unif$[0.3, 2.5]$, $\boldsymbol{\alpha} \sim$unif$[-1, 1]$ and $\boldsymbol{\beta} \sim$unif$[0, 1]$. These two models were used to investigate to what extent the true parameter values could be recovered for this L-MP model. The effectiveness of using AIC, BIC or LRT as model selection criterion was also investigated.

A fourth model was used to evaluate how the L-MP model performed compared to other nonparametric procedures. This model should not have a simple logistic functional form and should be different from the L-MP model with higher order polynomials. It is created as the CDF of a mixture of two normal distributions defined as $p_1 N_1(m_1, s_1) + p_2 N_2(m_2, s_2)$, where $p_1 \sim$unif$[0.3, 0.7]$, $p_2 = 1 - p_1$, $m_1 \sim N(-1.5, 0.1)$, $s_1 \sim N(1, 0.1)$, $m_2 \sim N(1.0, 0.1)$ and $s_2 \sim N(0.4, 0.1)$.

The examinees' abilities for all four models were generated from a truncated standard normal distribution, specifically, $\theta \sim N(0, 1)$ truncated at ± 3.

To investigate how the model performs with different number of examinees, two levels of sample size (300 or 2000) were used. Test length is another factor to be considered. We tend to believe that short tests cannot provide enough information for the estimation of abilities. The estimation of item curves is based on the estimated

abilities. When the test is too short, the estimates of abilities will be biased and the estimated ICCs will also be in doubt. We do not expect this procedure to provide reliable estimates for the ICCs and the abilities for a very short test.

To illustrate this point of view and to see how the estimation procedure performs under different test length, a simple illustrative example with a combination of test length (10 items, 20 items or 100 items) and the number of examinees (300 or 2000) was conducted with only one dataset in each condition. The precision of the estimated ICCs and the estimated abilities will be investigated.

To generate the 0/1 response for each examinee on each item, a probability was computed by the four models considered, using the generated item parameters and abilities. The probability was then compared with a random number from a uniform distribution of $[0, 1]$. If the probability was greater than the random number, a 1 was assigned and 0 otherwise. The functions that were used to generate data were all written in R language (version 2.4.1).

Evaluation of the performance of the models is from two perspectives: the closeness of the estimated ICC to the true ICC and the precision of the estimated abilities. The Root Integrated Mean Square Error (RIMSE) (Ramsay & Winsberg, 1991; Ramsay, 1991) will be used as the measure of the closeness of two ICCs. The RIMSE is defined as

$$\text{RIMSE} = \left[\frac{\sum_{k=1}^{q} (\hat{P}(\theta_k) - P(\theta_k))^2 \phi(\theta_k)}{\sum_{k=1}^{q} \phi(\theta_k)} \right]^{\frac{1}{2}}, \tag{5.1.1}$$

where θ_k's $(k = 1, 2, \ldots, q)$ are the evaluation points and are equally spaced on the range of abilities. In this simulation study, 801 evaluation points ranging from -4 to 4 were usually used. But to accommodate the default settings in TESTGRAF, 51 points on $[-2.5, 2.5]$ were used for small samples $(N = 300)$ and 51 points on $[-3, 3]$

were used for large samples ($N = 2000$) when comparing the performance of L-MP with TESTGRAF. In Equation (5.1.1), $\phi(\theta)$ is a weight which allows the measure to be most sensitive to departures for values of θ that are most commonly observed. This weight $\phi(\theta)$ is chosen to be the density of a standard normal distribution.

The idea of the RIMSE for the estimated ICCs can also be applied to the estimated abilities. The θs in the middle of the standard normal distribution received larger weights since they are more frequently observed. The closeness of the estimated and the true values is measured by RIMSE_θ defined as

$$\text{RIMSE}_\theta = \left[\frac{\sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2 \phi(\theta_i)}{\sum_{i=1}^{N} \phi(\theta_i)} \right]^{\frac{1}{2}}, \tag{5.1.2}$$

where N is the number of examinees.

Ramsay (1991, p614) concluded that due to the lack of identifiability, in the context of item analysis, a test cannot yield anything more than rank order information about examinees. From this point of view, the rank correlations between the ability estimates $\hat{\theta}$ and θ are also informative. Spearman rank correlations were computed and were used as alternative indices of how closely the estimated θs were to the true values.

5.2 Results

This section presents results regarding the performance of the L-MP model. I have chosen three primary indicators of performance: (1) Is the estimation method able to recover the true parameters for models with different k values? (2) Are the L-MP estimates comparable to the estimates from other currently used programs like MULTILOG (version 7.0.2327.3) yielding MML/EM estimates and TESTATLOG procedure in SYSTAT (version 11.00.01) yielding JML estimates when the true model

is a 2PL model? (3) Are the L-MP estimates comparable to the other nonparametric procedures like TESTGRAF and the Nonparametric Bayesian method when the true model is a different model? The simulation studies in this chapter are aimed at answering these questions. Before addressing these central questions, preliminary experiments were conducted to explore how the test length affects the performance of the L-MP model and to decide what number of items would be used for the main simulations.

5.2.1 Effect of test length

The data for this preliminary study were generated from the L-MP model,

$$P_i(\theta) = \frac{1}{1+e^{-m_i(\theta)}},$$

where $m_i(\theta)$ is the monotonic polynomial represented as

$$m_i(\theta) = \xi_i + b_{1i}\theta + b_{2i}\theta^2 + \cdots + b_{2k+1,i}\theta^{2k+1}$$

with three different levels as $k = 0$, $k = 1$ and $k = 2$. We investigated a short test of 10 items, a modest-sized test with 20 items and a long test with 100 items. The number of examinees was 300 or 2000. This led to a $3 \times 3 \times 2$ design. Since this was only a preliminary study to help decide the number of items to use in the main simulation studies, only one dataset was generated for each condition. The data were then fitted to the L-MP model with true k level.

The RIMSEs for the item curves are reported in Table 5.1. The first observation from Table 5.1 is that estimates are more accurate with a larger sample size. The general trend is that a longer test has smaller RIMSE than a shorter test and a more complex model has larger RIMSE than a simpler model, e.g. the RIMSE tends to be larger for $k = 2$ than for $k = 1$ etc. There are a few exceptions for the smaller

		$n = 10$	$n = 20$	$n = 100$
$N = 300$	$k = 0$	0.0452	0.0307	0.0352
	$k = 1$	0.0722	0.0600	0.0506
	$k = 2$	0.0766	0.0571	0.0538
$N = 2000$	$k = 0$	0.0428	0.0257	0.0121
	$k = 1$	0.0480	0.0305	0.0165
	$k = 2$	0.0654	0.0356	0.0214

Table 5.1: Effect of test length: RIMSE for estimated ICCs.
N is the number of examinees, n is the number of items, and k defines the highest degree of the monotonic polynomial (which is $2k + 1$). For example, $k = 2$ represents a monotonic polynomial of fifth order. Each cell is based on one dataset.

		$n = 10$	$n = 20$	$n = 100$
$N = 300$	$k = 0$	0.4723	0.3438	0.1839
	$k = 1$	0.4032	0.2886	0.1608
	$k = 2$	0.3487	0.2977	0.1461
$N = 2000$	$k = 0$	0.4385	0.3469	0.1690
	$k = 1$	0.3781	0.2713	0.1340
	$k = 2$	0.3685	0.2568	0.1199

Table 5.2: Effect of test length: RIMSE_θ for estimated abilities.
N is the number of examinees, n is the number of items, and k defines the highest degree of the monotonic polynomial (which is $2k + 1$). For example, $k = 2$ represents a monotonic polynomial of fifth order. Each cell is based on one dataset.

sample size. For example, when the true model is a 2PL model ($k = 0$), the test with 100 items has slightly larger RIMSE than the test with 20 items. But with only one sample in each condition, the small differences (in third decimal place) should be disregarded.

The RIMSE_θs are reported in Table 5.2 and the rank correlations between the true abilities and the estimated abilities are reported in Table 5.3. The precision of the ability estimate $\hat{\theta}$ is consistently improved when the test gets longer. The RIMSE_θ

		$n = 10$	$n = 20$	$n = 100$
$N = 300$	$k = 0$	0.8408	0.9268	0.9819
	$k = 1$	0.9018	0.9548	0.9887
	$k = 2$	0.9243	0.9467	0.9898
$N = 2000$	$k = 0$	0.8722	0.9285	0.9844
	$k = 1$	0.9156	0.9601	0.9905
	$k = 2$	0.9243	0.9633	0.9921

Table 5.3: Effect of test length: rank correlations for abilities $\rho(\hat{\theta}, \theta)$.
N is the number of examinees, n is the number of items, and k defines the highest degree of the monotonic polynomial (which is $2k + 1$). For example, $k = 2$ represents a monotonic polynomial of fifth order. Each cell is based on one dataset.

is much less for a test with 100 items than for a test with smaller number of items. Similar conclusions are found for rank correlations between the estimated and the true abilities. The rank correlations are almost perfect for a test with 100 items.

Based on these preliminary results, the test length does play a role in the estimations, particularly when estimating the abilities. We would expect some estimation problems for a very short test, especially when we try to fit the data to a model with high degree of monotonic polynomial. Although the model is able to provide estimates for a test with 10 items, the estimates might not be very stable. The RIMSEs for both item curves and abilities are also large in this case. That being said, a test of 20 items is considered not too long and not too short and thus seems to be good for our simulation study. If the L-MP model works well for a test with 20 items, it generally can produce better estimates for a test with more items. This does not imply one could not use the L-MP model with fewer than 20 items. However, more research needs to be done to evaluate the performance of the model under that situation.

5.2.2 Simulation 1

The purpose of this simulation is to examine to what extent the L-MP model and its associated estimation method can recover the true parameters. The models that were used to generate the data are

$$P_i(\theta) = \frac{1}{1+e^{-m_i(\theta)}},$$

where $m_i(\theta)$ is the monotonic polynomial represented as

$$m_i(\theta) = \xi_i + b_{1i}\theta + b_{2i}\theta^2 + \cdots + b_{2k+1,i}\theta^{2k+1}.$$

Three different k values ($k = 0$, $k = 1$, and $k = 2$) were used. Theoretically, k could be set to any value, but in practice, a high value of k leads to a model with too many parameters in the IRF and limits the generalizability of the estimated ICC. The upper limit employed, $k = 2$, represents a monotonic polynomial of fifth order which should be sufficient for most circumstances. Two sample sizes ($N = 300$ and $N = 2000$) were investigated. This led to a 3×2 design. Each cell contained 100 datasets.

When the monotonic polynomial is linear ($k = 0$), the data will be generated from the 2PL model. The difficulty parameter was randomly drawn from a standard normal distribution truncated at ± 2.5. The data were fitted using $k = 0$ and $k = 1$. Figure 5.1 shows the ICCs for the first four items in a sample of 300 examinees when the true model is the 2PL model. Figure 5.3 shows the ICCs for the same four items in a larger sample of 2000 examinees. The k values selected by AIC, BIC or LRT are shown in the upper left corner of the figure for each item. Figure 5.2 and Figure 5.4 provide a different way of presenting the estimated ICCs from Figure 5.1 and Figure 5.3. In Figures 5.2 and 5.4, deviations from the true probabilities are plotted against the abilities. Observing these figures we find that the estimated ICCs with $k = 0$

are very close to the true ICCs for the majority of the four items. There are some departures for item 4 for examinees with low abilities. The estimates are improved a little with a larger sample (Figure 5.3 and Figure 5.4). Fitting curves using an unnecessary higher degree polynomial results in worse estimated curves especially when the sample size is small. Generally speaking, the precision of the estimated ICCs improves when sample size increases (from 300 to 2000).

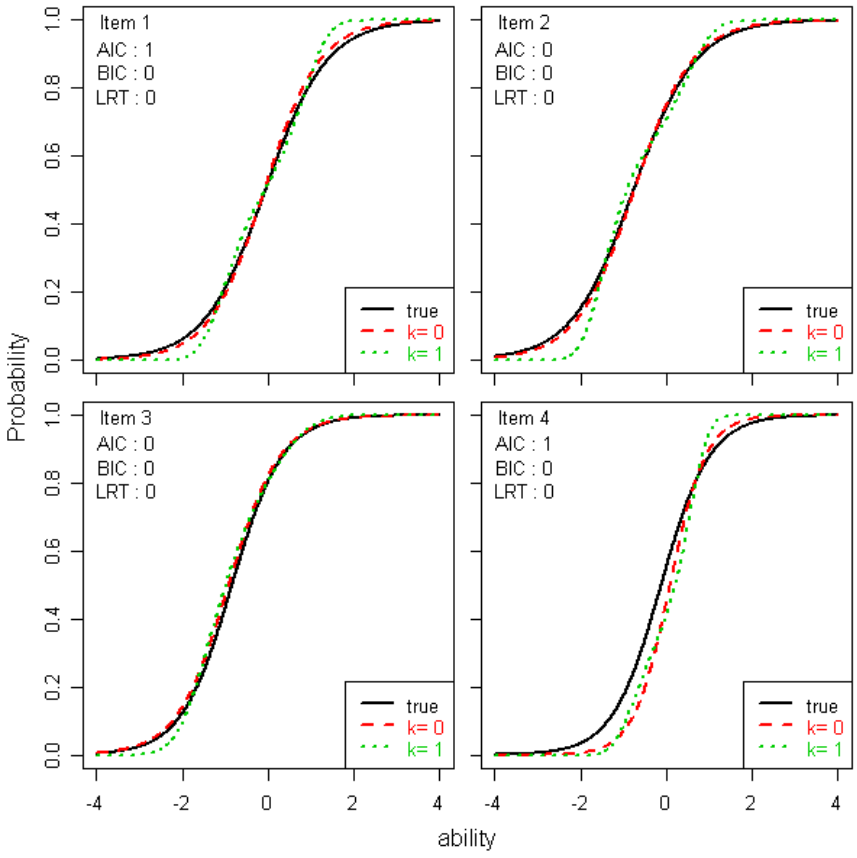

Figure 5.1: Estimated ICCs for some selected items in a typical data set ($k = 0, N = 300$). The upper left corner shows the k values selected by AIC, BIC and LRT.

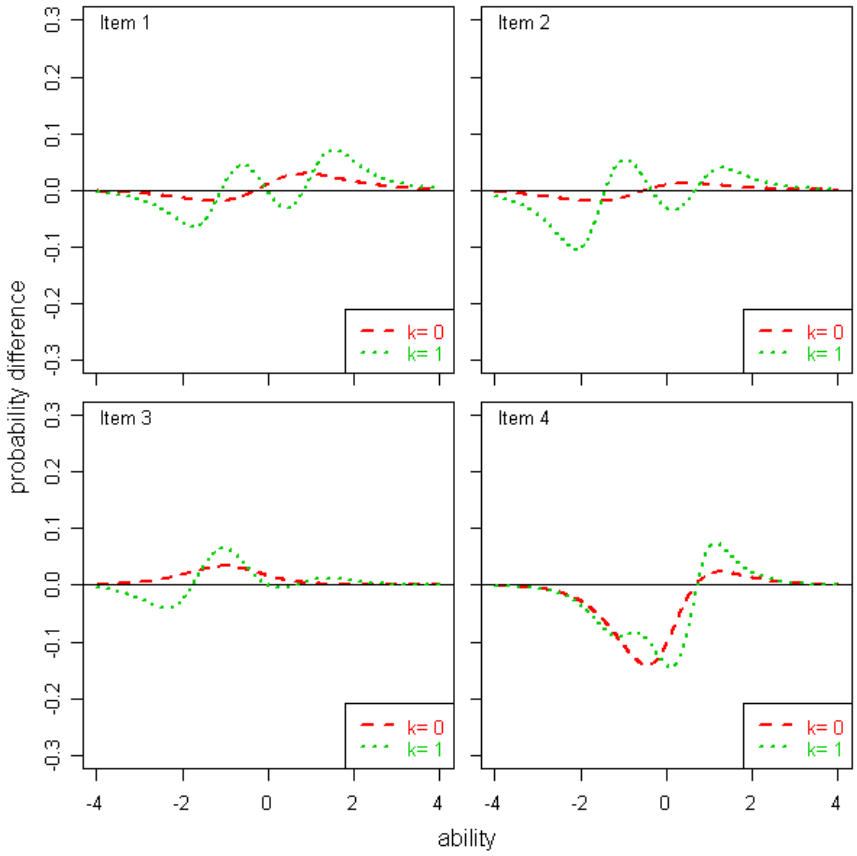

Figure 5.2: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 0, N = 300$).

Figure 5.5 and Figure 5.7 are the ICCs for some selected items in a typical data set with 300 examinees or 2000 examinees when the true monotonic polynomial is cubic (with $k = 1$). Figure 5.6 and Figure 5.8 are the corresponding plots for the probability difference. The data were fitted using $k = 0$, $k = 1$ and $k = 2$. The estimated curves with linear monotonic polynomials seem to be inadequate for the majority of the four items. The estimates under $k = 1$ and $k = 2$ are much closer to each other and to the true curve. The estimated ICCs are also improved with larger

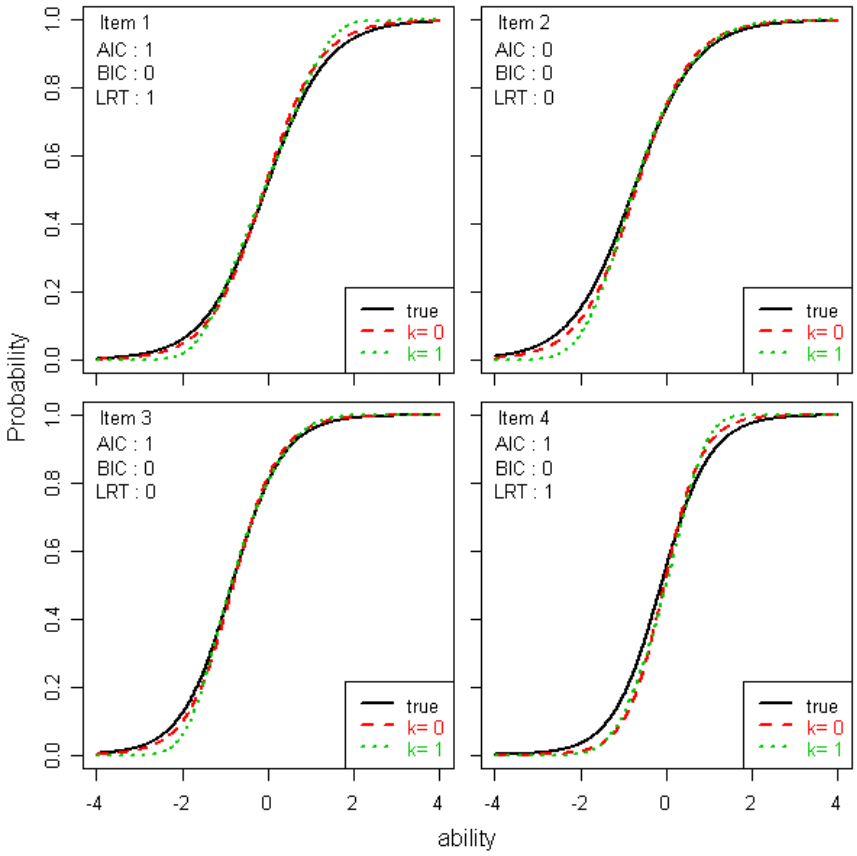

Figure 5.3: Estimated ICCs for some selected items in a typical data set ($k = 0, N = 2000$). The upper left corner shows the k values selected by AIC, BIC and LRT.

sample size. For example, the estimated ICC for item 11 gets much closer to the true curve for the sample of size 2000 especially for examinees with medium ability level. However, not all three criteria choose the correct model. For the plotted items, AIC and LRT are relatively consistent and prefer same models while BIC favors simpler models especially when the sample size is small.

Figure 5.9 and Figure 5.11 are the ICCs for some selected items in typical datasets, one with 300 examinees and another with 2000 examinees when the true monotonic

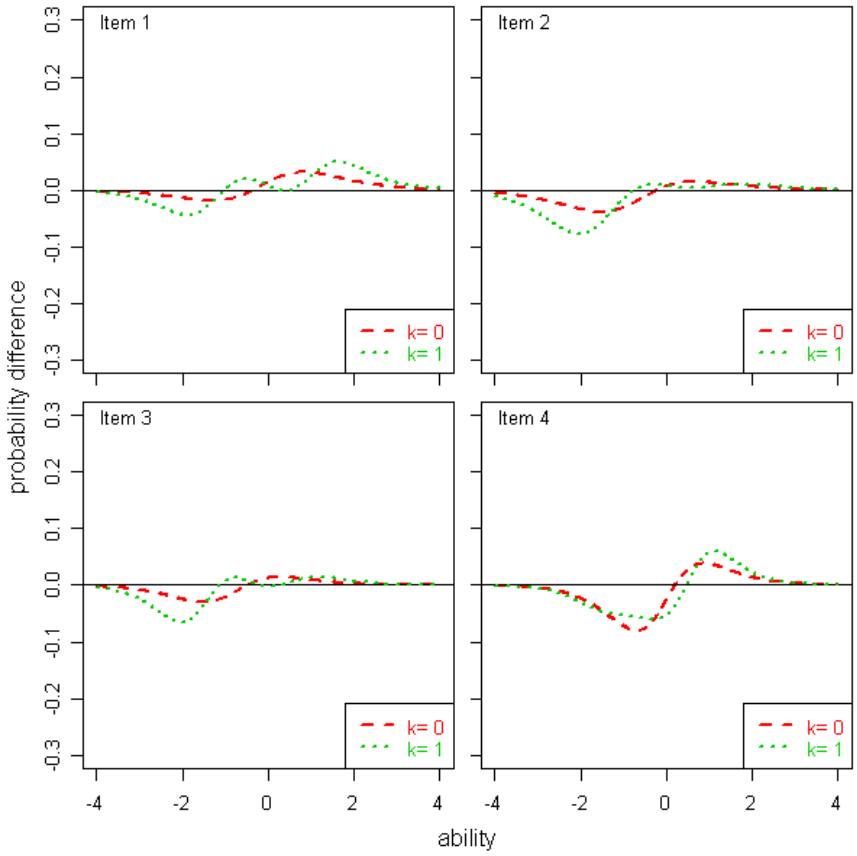

Figure 5.4: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 0, N = 2000$).

polynomial is of fifth order ($k = 2$). Figure 5.10 and Figure 5.12 are the corresponding plots for the probability difference. The data were fitted using $k = 0$, $k = 1$, $k = 2$ and $k = 3$. It can be observed that for most items the IRF with linear polynomial are not close to the true ICCs. The estimated ICCs with $k = 1$, $k = 2$ and $k = 3$ are relatively close to each other and to the true curves. The improvement from $k = 1$ to $k = 2$ or higher k values is not big enough to justify the additional parameters added to the model. AIC, BIC and LRT favor different models. Again, AIC and LRT are

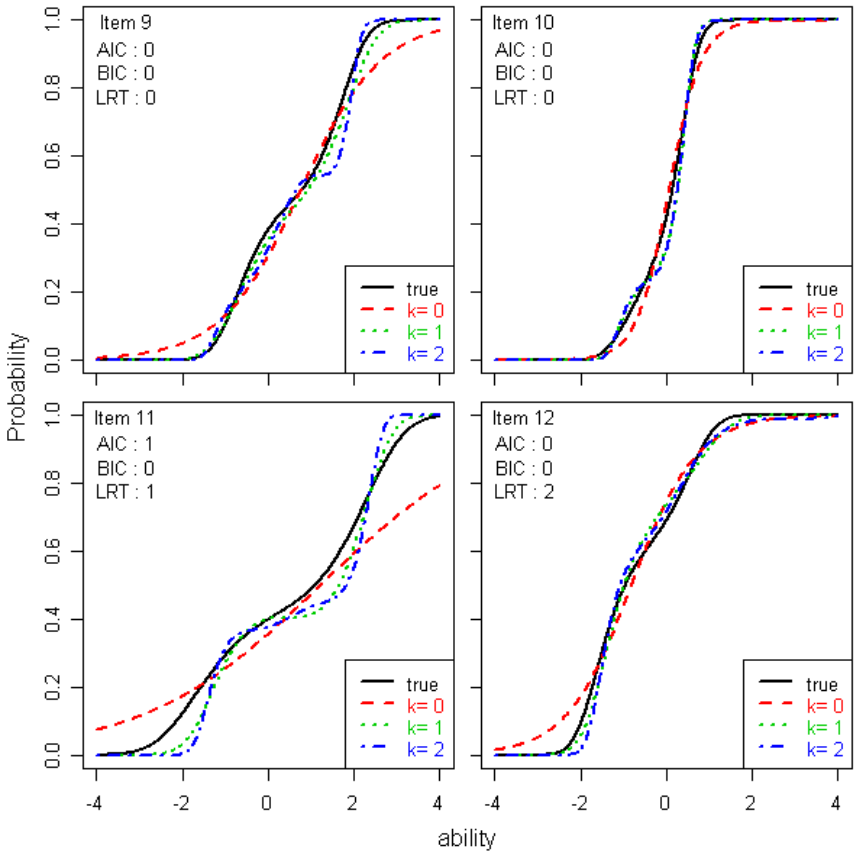

Figure 5.5: Estimated ICCs for some selected items in a typical dataset ($k = 1, N = 300$). The upper left corner shows the k values selected by AIC, BIC and LRT.

relatively consistent and BIC favors simpler models in general. Not all indices choose the true models for all items again. It appears that neither AIC, BIC nor LRT are good model selection criteria for this L-MP model since they are not able to choose the correct models. However, based on the observation of the discrepancies of the estimated ICCs to the true ICCs, the models selected by these indices are adequate enough for approximating the true curves.

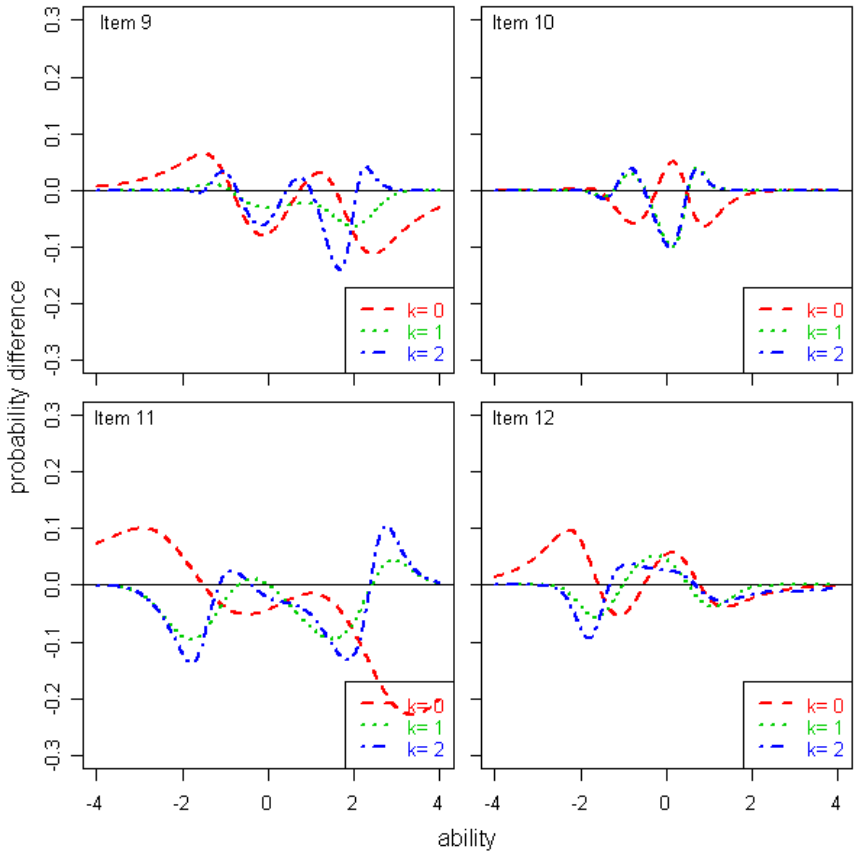

Figure 5.6: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 1, N = 300$).

Table 5.4 to Table 5.9 give summaries of the frequencies of indices AIC, BIC and LRT choosing the true model under different sample sizes. When the true model is a 2PL model ($k = 0$), the percentages of AIC, BIC and LRT choosing the true model are high for the smaller sample. For the larger sample, although all three criteria tend to choose the true model, BIC has larger odds of choosing the true model. When the true model is a nonstandard logistic function and the sample size is small, all model selection indices favor the simpler model, especially BIC. The most extreme case is

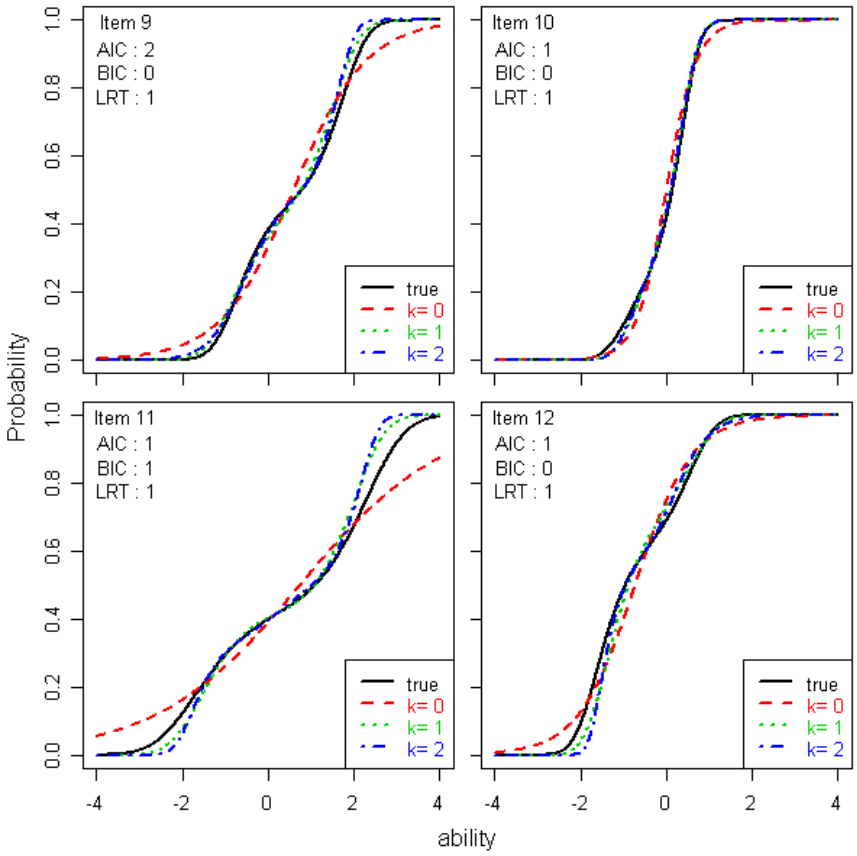

Figure 5.7: Estimated ICCs for some selected items in a typical dataset ($k = 1, N = 2000$). The upper left corner shows the k values selected by AIC, BIC and LRT.

in Table 5.5, when the true k is 2, BIC never chooses true model! When sample size increases, the percentages of choosing the true models increase under same conditions for all model selection criteria. But still, the results are not very positive in terms of the percentages of true models being selected.

It is very natural to question the usefulness of these model selection criteria. We perceive this from a different perspective. We have already observed from Figure 5.5 to Figure 5.12 that the difference of the estimated curves between $k = 0$ and $k = 1$

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	80.95	34.65	22.10
selected k	$k=1$	19.05	52.85	62.50
	$k=2$		12.50	13.75
	$k=3$			1.65

Table 5.4: Frequency table for AIC selected items with $N = 300$.
Each row represents the k values selected by AIC. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	99.35	83.90	72.05
selected k	$k=1$	0.65	16.10	27.95
	$k=2$		0.00	0.00
	$k=3$			0.00

Table 5.5: Frequency table for BIC selected items with $N = 300$.
Each row represents the k values selected by BIC. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	91.75	50.55	36.35
selected k	$k=1$	8.25	43.95	56.75
	$k=2$		5.50	5.50
	$k=3$			1.40

Table 5.6: Frequency table for LRT selected items with $N = 300$.
Each row represents the k values selected by LRT. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	56.85	2.35	0.35
selected k	$k=1$	43.15	62.10	51.45
	$k=2$		35.55	20.70
	$k=3$			27.50

Table 5.7: Frequency table for AIC selected items with $N = 2000$.
Each row represents the k values selected by AIC. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	98.50	22.40	10.60
selected k	$k=1$	1.50	75.75	85.90
	$k=2$		1.85	3.25
	$k=3$			0.25

Table 5.8: Frequency table for BIC selected items with $N = 2000$.
Each row represents the k values selected by BIC. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

		true k values		
		$k=0$	$k=1$	$k=2$
	$k=0$	74.70	4.20	0.85
selected k	$k=1$	25.30	74.50	60.45
	$k=2$		21.30	16.75
	$k=3$			21.95

Table 5.9: Frequency table for LRT selected items with $N = 2000$.
Each row represents the k values selected by LRT. Each column represents the true k values. 100 datasets are used for each true k value. The number in each cell is a percentage. Each column should sum to 100.

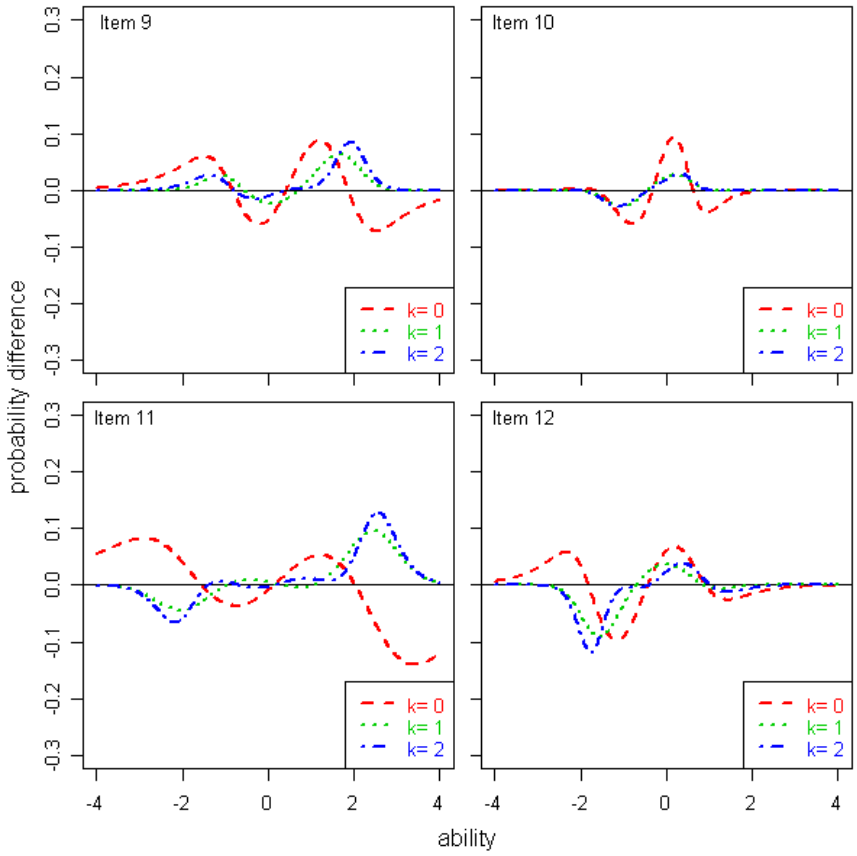

Figure 5.8: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 1, N = 2000$).

or above is obvious, however, the difference of the estimated curves between $k = 1$ and $k = 2$ or higher is only small. Meanwhile, when k goes one point higher, two additional parameters are introduced to the model for each item. With smaller sample size, it is not worthwhile to achieve the little improvement at the cost of unnecessary parameters in the model. In such case, it is better to use an "incorrect" model with few parameters than a "correct" model with many parameters. The AIC is also not intended to select the "correct" model. Thus we argue that we are not concerned

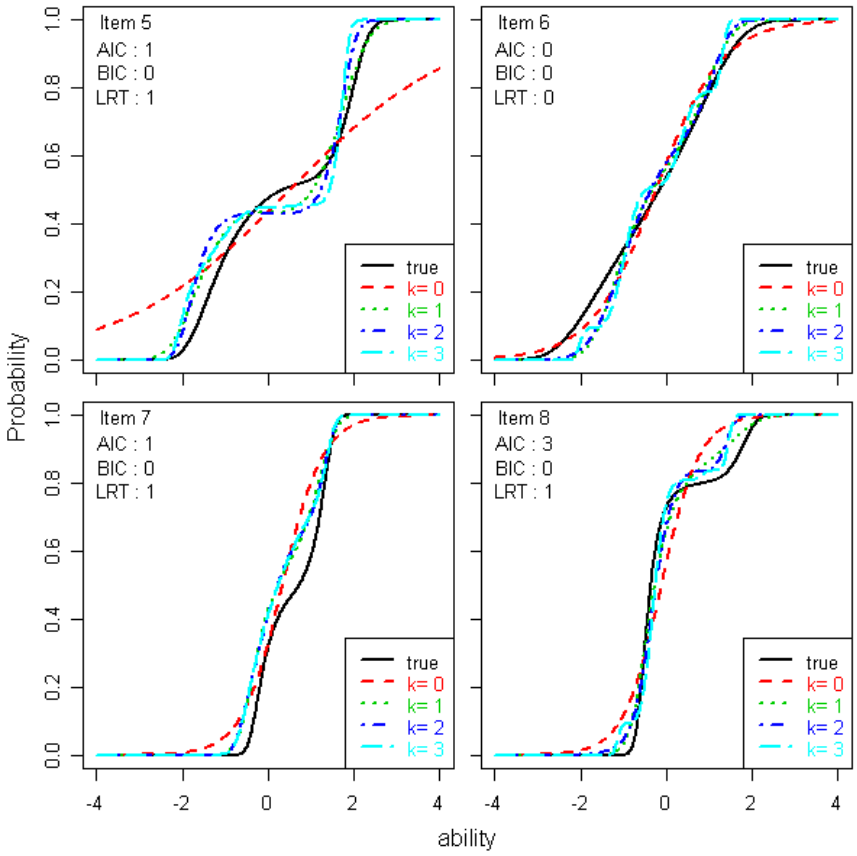

Figure 5.9: Estimated ICCs for some selected items in a typical dataset ($k = 2, N = 300$). The upper left corner shows the k values selected by AIC, BIC and LRT.

about the percentages of the true model being chosen, but the models selected by these criteria actually having better estimates in terms of RIMSE for both the ICCs and the abilities, or having higher rank correlations between the estimated and the true ability parameters. In this sense, AIC, BIC and LRT might still be useful criteria for this L-MP model.

Table 5.10 and Table 5.11 present the RIMSE for the estimated ICCs. It is natural to think that the estimated curves fitted to the true k values will have the smallest

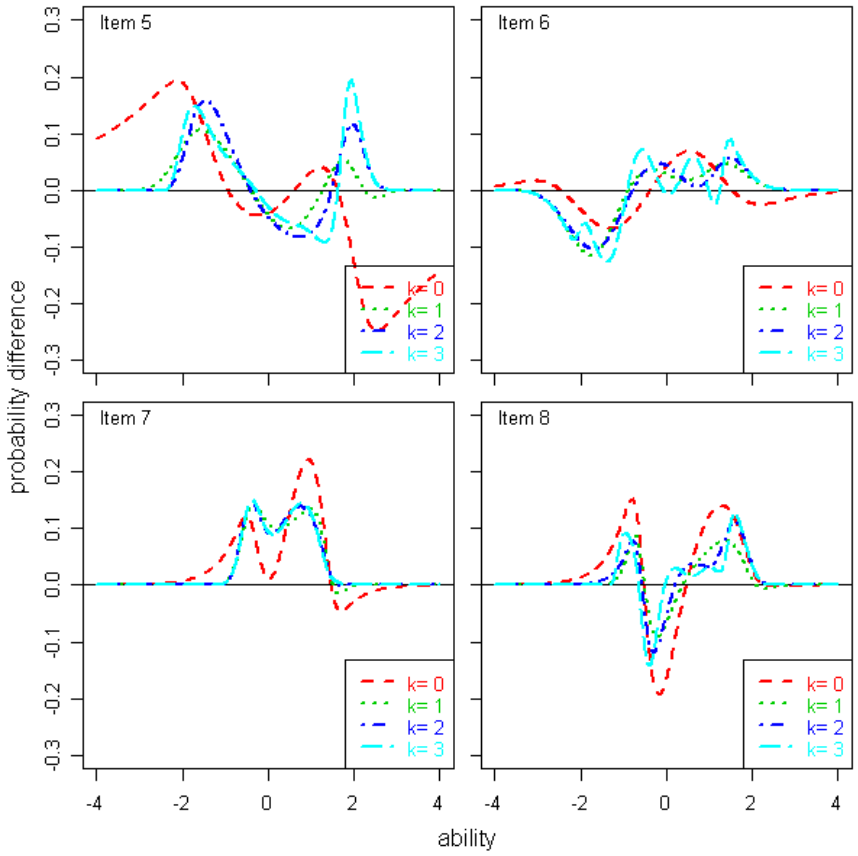

Figure 5.10: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 2, N = 300$).

RIMSE. However, this conjecture does not always hold for all samples and models with different k values. For example, for both sample sizes, when the true model is $k = 2$, the model with $k = 1$ actually produces the smallest RIMSE for estimated ICCs. This supports our argument of not investigating the percentage of true models being selected but the better RIMSE for selected models. As can be expected, the RIMSE increases as the model becomes more complicated (k value gets larger). And also, not surprisingly, larger sample size helps to improve the estimates. In practice,

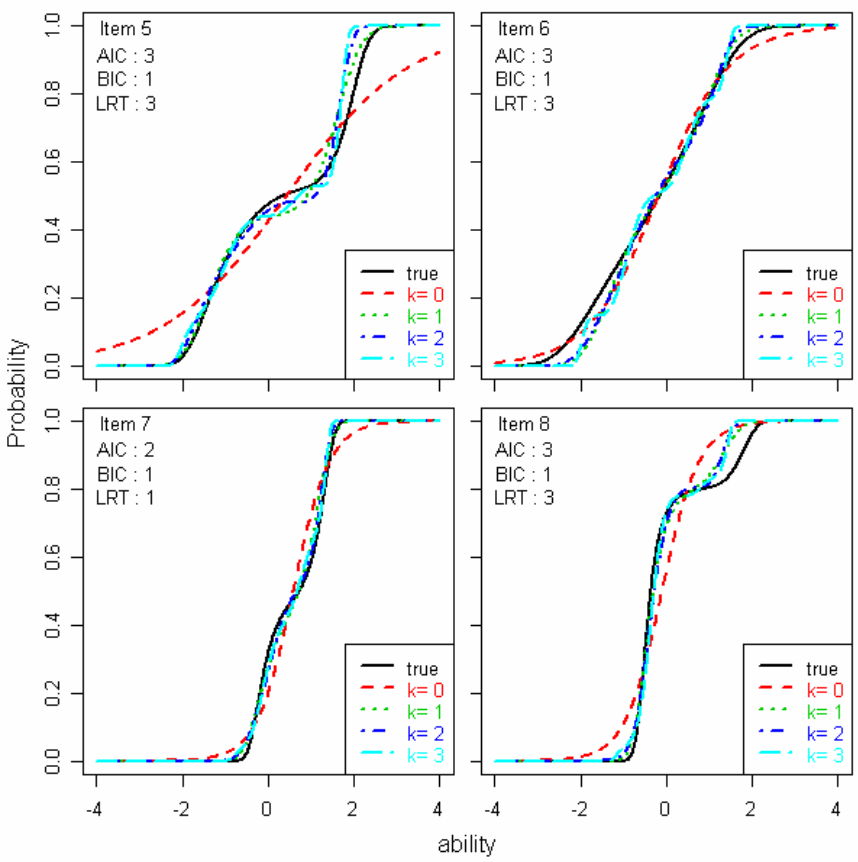

Figure 5.11: Estimated ICCs for some selected items in a typical dataset ($k = 2, N = 2000$). The upper left corner shows the k values selected by AIC, BIC and LRT.

no true value of k is known, AIC, BIC and LRT are the criteria that are used to help us select the models. Based on this simulation result, the overall trend is that models selected by BIC produce the smallest RIMSE when the true model is the 2PL model ($k = 0$) and the AIC selected models have smallest RIMSE when the true models are not the standard logistic curves. This is because BIC favors simpler model when sample size is large (even the smaller sample size 300 in this simulation can be considered as "large" when using BIC). Thus when the true model is a simpler model,

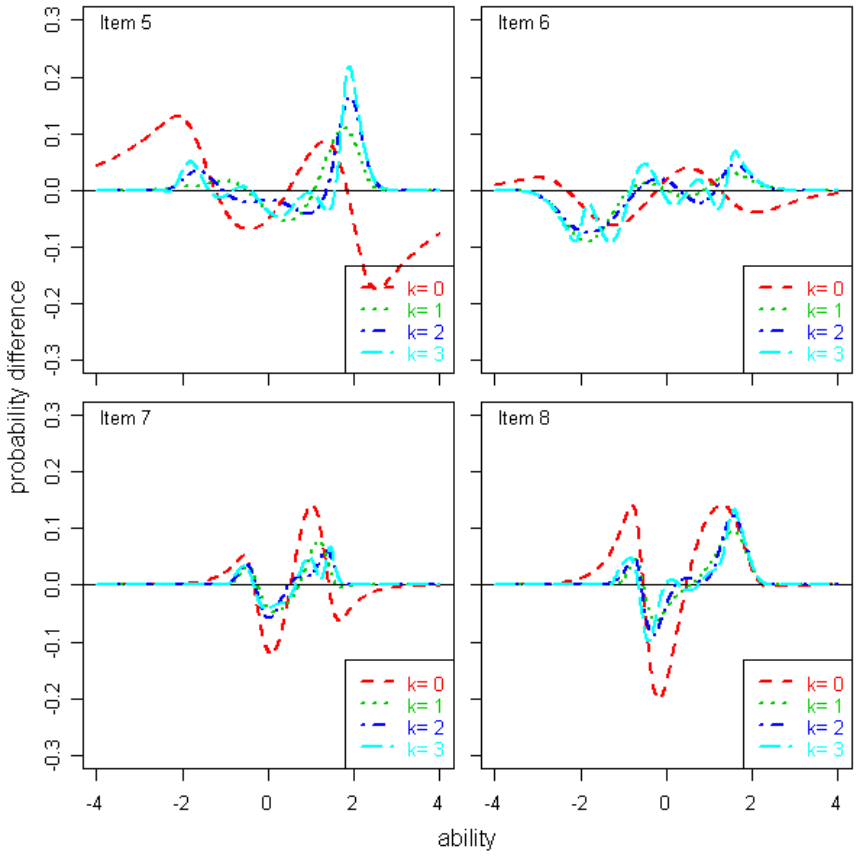

Figure 5.12: Deviations of the estimated probabilities from the true probabilities for some selected items in a typical dataset ($k = 2, N = 2000$).

AIC has bigger chance of choosing a more complex model which apparently just adds some error to the model. But when the true model is actually a more complex model, BIC will still select the simpler model which is not sufficient in approximating the true curve. Overall speaking, although differences exist, they are too small to have any practical significance.

Table 5.12 to Table 5.15 give information on the precision of the estimates of abilities in terms of RIMSE_θ and rank correlations. Again, the smallest RIMSE_θ

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.0431 (0.0057)	0.0575 (0.0062)	0.0598 (0.0068)
BIC	0.0391 (0.0058)	0.0622 (0.0067)	0.0674 (0.0063)
LRT	0.0412 (0.0058)	0.0589 (0.0062)	0.0612 (0.0067)
$k = 0$	0.0388 (0.0058)	0.0646 (0.0062)	0.0754 (0.0060)
$k = 1$	0.0476 (0.0055)	0.0533 (0.0062)	0.0557 (0.0067)
$k = 2$		0.0591 (0.0056)	0.0610 (0.0064)
$k = 3$			0.0641 (0.0061)

Table 5.10: RIMSE for estimated ICCs for various true k values ($N = 300$).
Each cell is based on 100 datasets. A mean RIMSE is calculated across items for each dataset. The number presented in the table is the average of the mean RIMSE across 100 datasets. The number in parentheses is the standard deviation of the 100 averaged RIMSEs. The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial.

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.0257 (0.0025)	0.0324 (0.0032)	0.0358 (0.0035)
BIC	0.0240 (0.0027)	0.0333 (0.0034)	0.0349 (0.0038)
LRT	0.0252 (0.0027)	0.0322 (0.0033)	0.0357 (0.0036)
$k = 0$	0.0238 (0.0027)	0.0541 (0.0048)	0.0660 (0.0048)
$k = 1$	0.0262 (0.0024)	0.0309 (0.0033)	0.0334 (0.0037)
$k = 2$		0.0326 (0.0032)	0.0347 (0.0036)
$k = 3$			0.0368 (0.0034)

Table 5.11: RIMSE for estimated ICCs for various true k values ($N = 2000$).
Each cell is based on 100 datasets. A mean RIMSE is calculated across items for each dataset. The number presented in the table is the average of the mean RIMSE across 100 datasets. The number in parentheses is the standard deviation of the 100 averaged RIMSEs. The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial.

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.3570 (0.0168)	0.3075 (0.0227)	0.2817 (0.0202)
BIC	0.3531 (0.0165)	0.3110 (0.0227)	0.2889 (0.0204)
LRT	0.3553 (0.0166)	0.3089 (0.0228)	0.2834 (0.0204)
$k = 0$	0.3528 (0.0164)	0.3120 (0.0220)	0.2944 (0.0206)
$k = 1$	0.3600 (0.0174)	0.3036 (0.0227)	0.2779 (0.0196)
$k = 2$		0.3068 (0.0226)	0.2813 (0.0202)
$k = 3$			0.2815 (0.0199)

Table 5.12: RIMSE_θ for various true k values ($N = 300$).
The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

does not always occur when fitting to true k values. When the true model is $k = 2$, the models fitted to $k = 1$ actually have the smallest RIMSE_θ and the highest rank correlations. When comparing the different model selection criteria in terms of the ability estimates, these two tables show the same trend as in Table 5.10 and Table 5.11. When the sample size is small and the true model is $k = 0$, BIC selected models have smallest RIMSE_θ and highest rank correlations for estimated abilities. When the true models are not standard logistic functions, AIC selected models have best ability estimates. But again, these differences are too small to be interpreted with any practical meanings.

The sample size has a larger effect on estimating the ICCs than estimating the abilities. The estimates of abilities do improve in terms of both RIMSE_θ and rank correlations with a larger sample size, but this improvement is not large. An easy explanation on this is that large number of examinees provides more information that could be used to estimate the ICCs and thus results better estimated curves. The

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.3468 (0.0109)	0.2920 (0.0158)	0.2682 (0.0139)
BIC	0.3457 (0.0109)	0.2931 (0.0162)	0.2680 (0.0142)
LRT	0.3467 (0.0109)	0.2919 (0.0158)	0.2680 (0.0140)
$k = 0$	0.3455 (0.0108)	0.3064 (0.0166)	0.2890 (0.0157)
$k = 1$	0.3466 (0.0110)	0.2908 (0.0160)	0.2666 (0.0142)
$k = 2$		0.2921 (0.0158)	0.2677 (0.0141)
$k = 3$			0.2687 (0.0140)

Table 5.13: RIMSE_θ for various true k values ($N = 2000$).
The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.9229 (0.0101)	0.9473 (0.0096)	0.9566 (0.0081)
BIC	0.9241 (0.0100)	0.9458 (0.0099)	0.9544 (0.0086)
LRT	0.9234 (0.0100)	0.9470 (0.0097)	0.9561 (0.0082)
$k = 0$	0.9242 (0.0100)	0.9450 (0.0097)	0.9522 (0.0090)
$k = 1$	0.9217 (0.0103)	0.9485 (0.0096)	0.9576 (0.0079)
$k = 2$		0.9475 (0.0096)	0.9566 (0.0081)
$k = 3$			0.9564 (0.0080)

Table 5.14: Rank correlations for abilities for various true k values ($N = 300$).
The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

	true model		
	$k = 0$	$k = 1$	$k = 2$
AIC	0.9271 (0.0109)	0.9517 (0.0067)	0.9595 (0.0070)
BIC	0.9275 (0.0109)	0.9513 (0.0068)	0.9595 (0.0059)
LRT	0.9272 (0.0109)	0.9517 (0.0067)	0.9595 (0.0059)
$k = 0$	0.9276 (0.0108)	0.9465 (0.0073)	0.9528 (0.0059)
$k = 1$	0.9271 (0.0110)	0.9519 (0.0067)	0.9598 (0.0059)
$k = 2$		0.9517 (0.0067)	0.9596 (0.0059)
$k = 3$			0.9594 (0.0059)

Table 5.15: Rank correlations for abilities for various true k values ($N = 2000$). The rows represent the different criteria used to select the model. For example, $k = 0$ means that the items in a test are all chosen to have linear polynomial. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

estimates of abilities are obtained based on the estimated curves. Thus increased sample size has an indirect effect on estimating abilities.

In practice, there is no way to know what the true model is, and there is no definite conclusion of what selection criterion works consistently better than the other. Observations on Table 5.10 to Table 5.15 find that the differences of the RIMSE and the rank correlations for the models selected via AIC, BIC and LRT are not very large. The choice of the criterion in practice is not very critical in this sense.

5.2.3 Simulation 2

As we have mentioned before, the proposed L-MP model is a general model which includes 2PL as a special case. We are also interested in comparing the surrogate based estimates from the L-MP model with those from the currently used methods, particularly the MML estimates and the JML estimates. The data were generated using the standard 2PL model. As described in Section 5.1, two ways were used to generate the difficulty parameters. One way was to let the difficulty parameter

equally spaced on the range of $[-2.5, 2.5]$, the other way was to randomly draw from a standard normal distribution truncated at ± 2.5. Two sample sizes (300 and 2000) were considered. Although investigators seldom use JML estimates, the estimation method used for this L-MP model can be viewed as a version of the JML truncated after one iteration. We are interested to see how the estimates perform compared to the JML estimates and if consistency is still an issue. Thus the results will be compared with other software programs with different estimation methods: MULTILOG which produces MML/EM estimates and SYSTAT TESTATLOG which produces JML estimates.

Figure 5.13 and Figure 5.15 show plots for the first four items with estimates from L-MP, MML and JML in a typical dataset when the difficulty parameter is equally spaced, under two different sample sizes 300 and 2000. Figure 5.14 and Figure 5.16 are the corresponding plots for the probability differences. In general, the estimated item curves for the L-MP model and MML estimates for the 2PL model are much better than the JML estimates for the 2PL model. Sample size has a larger effect on the estimates from the L-MP model and the MML estimates. For example, the estimated ICCs for item 2 and item 3 are almost identical to the true curves when the number of examinees increase from 300 to 2000 (with equally spaced difficulty parameters). However, increasing sample size has no obvious effect in improving the estimated ICCs using JML. This suggests that the estimates of JML might not be consistent. Although JML uses a standardization on the abilities after each cycle of the iteration between the item parameters and abilities to solve the identification problem, it doesn't seem like it eliminates the possibility of inconsistency.

Figure 5.17 and Figure 5.19 produce plots for 300 examinees but with difficulty parameters randomly drawn from a standard normal distribution truncated at ± 2.5. Figure 5.18 and Figure 5.20 are the corresponding plots for the deviations of each

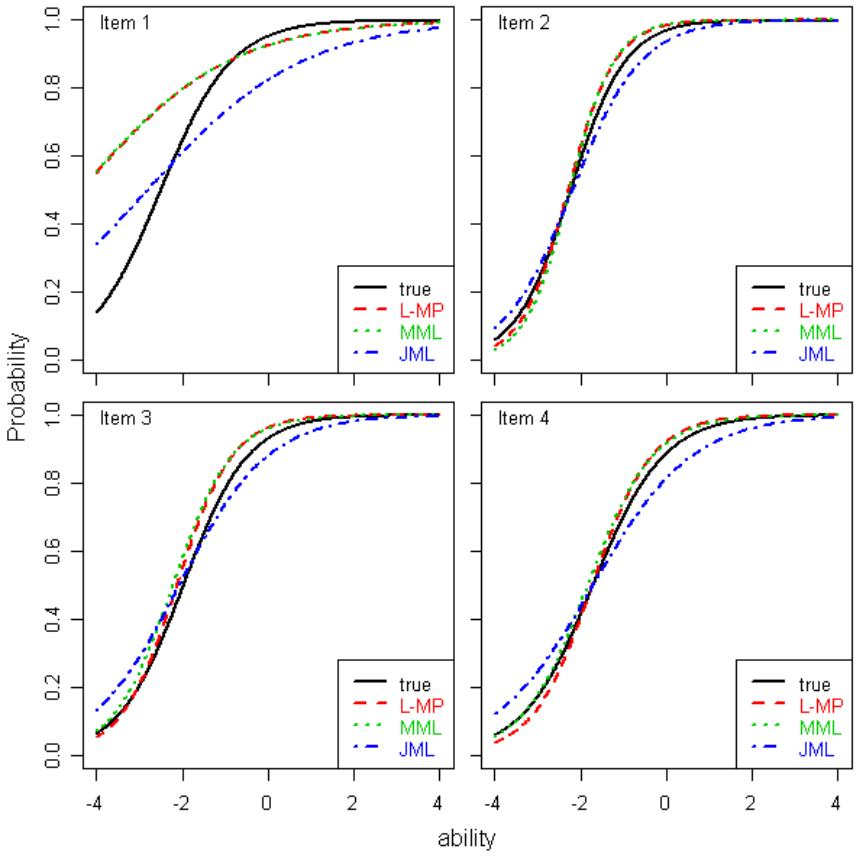

Figure 5.13: Comparisons of estimated ICCs among L-MP, MML and JML ($N = 300$, and equally spaced *difficulty* parameters).

estimated curve to the true curves. These two graphs show very similar results as Figure 5.13 and Figure 5.15. The estimates from L-MP and MML are very close to each other and to the true curve, while the estimates of JML are further away from the true values. The increasing of sample size helps to improve the estimates from L-MP and MML but does not help to improve the estimates from JML.

To check the performance of the estimates of abilities, Figure 5.21 to Figure 5.24 provide comparisons of L-MP with MML and JML in terms of RIMSE_θ and the rank

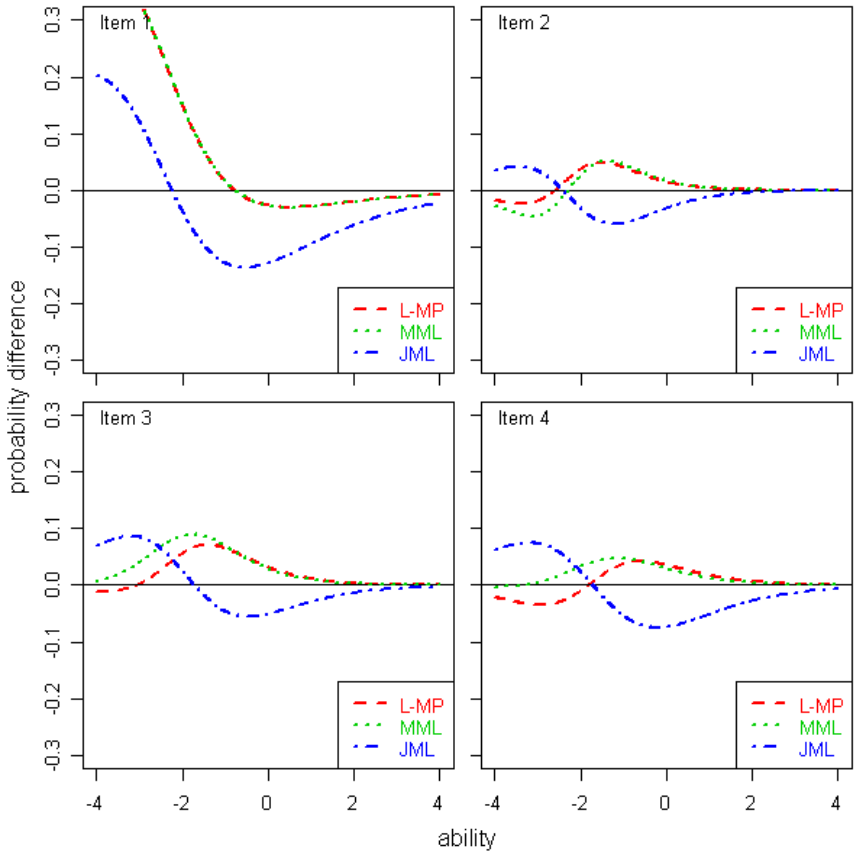

Figure 5.14: Comparisons of estimated ICCs among L-MP, MML and JML in probability difference ($N = 300$, and equally spaced *difficulty* parameters).

correlations. In each figure, the top two plots are for the comparisons of the L-MP with MML. The bottom two plots are for the comparisons of L-MP with JML. The RIMSE_θ and the rank correlations from L-MP are plotted against those from MML, and also those from JML. Each point in a plot represents the RIMSE_θ or the rank correlation for the estimated abilities in one dataset. There are 100 datasets under each condition and thus 100 points in each plot. If the two procedures being compared perform equally well, the RIMSE_θ or the rank correlations should be scattered around

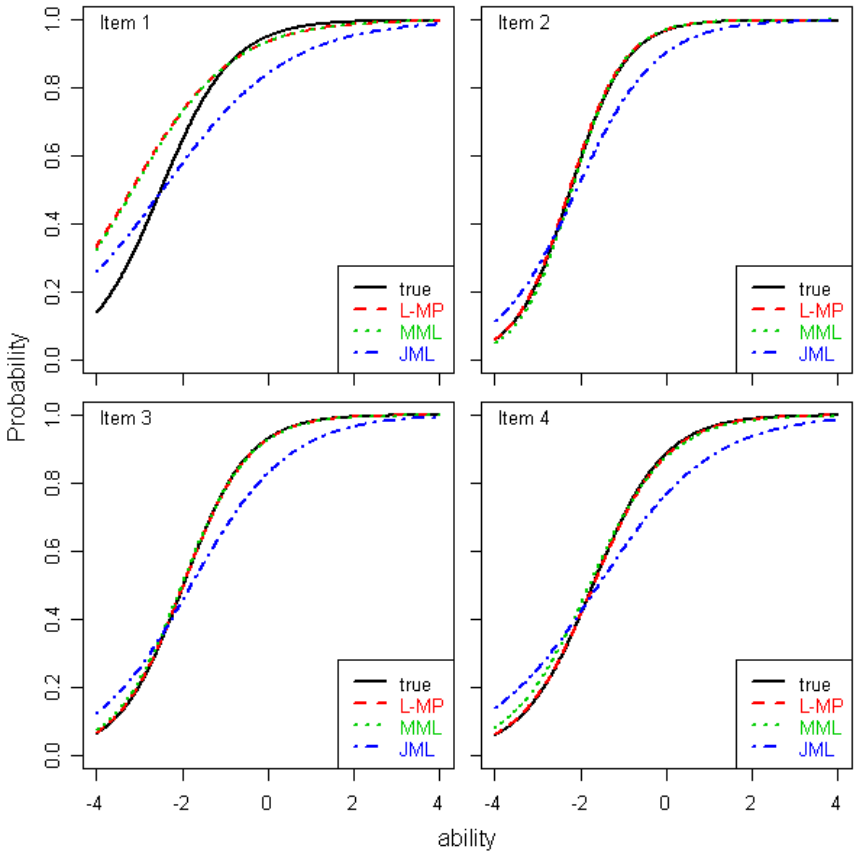

Figure 5.15: Comparisons of estimated ICCs among L-MP, MML and JML ($N = 2000$, and equally spaced *difficulty* parameters).

the $y = x$ line. If all points are sitting above or below the $y = x$ line, then one method outperforms the other. For example, when comparing L-MP with MML under the condition of 300 examinees and equally spaced difficulty parameters, the upper left plot in Figure 5.21 compares the RIMSE_θ for ability estimates and the upper right plot compares the rank correlations. Clearly, the L-MP estimates are slightly better than the MML estimates in terms of the RIMSE_θ since most points are sitting above the $y = x$ line. However, when it comes to the rank correlations, all the points are very

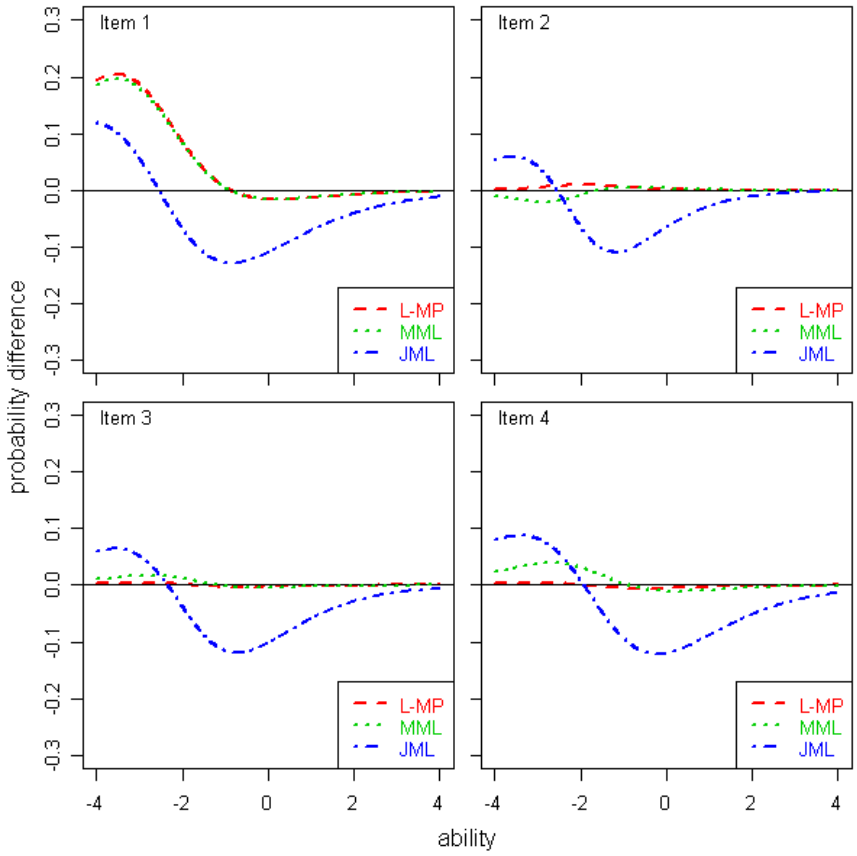

Figure 5.16: Comparisons of estimated ICCs among L-MP, MML and JML in probability difference ($N = 2000$, and equally spaced *difficulty* parameters).

close to the $y = x$ line which means these two methods produce similar estimates of the rankings to the abilities. Similar trends are found for samples with 2000 examinees or with normally distributed difficulty parameters. When comparing the L-MP estimates with the JML estimates, the L-MP estimates produce much smaller RIMSE_θ for abilities than the JML estimates for both sample sizes and different difficulty parameter distributions. When it comes to the rank correlations, the L-MP estimates and JML estimates are very close with the equally spaced difficulty

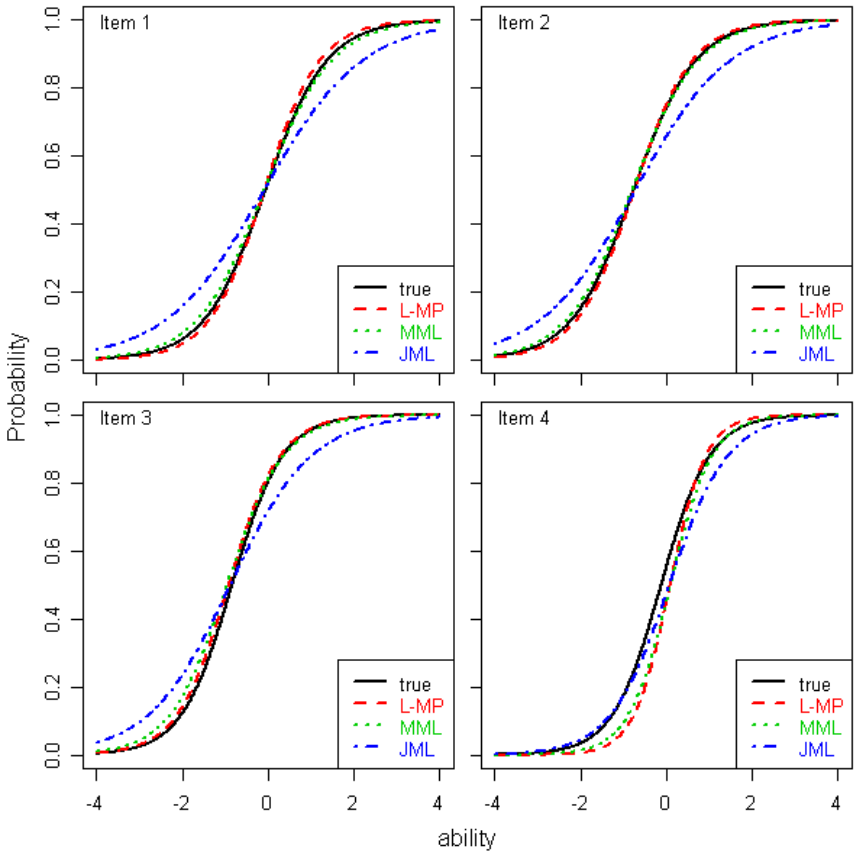

Figure 5.17: Comparisons of estimated ICCs among L-MP, MML and JML ($N = 300$, and normally distributed *difficulty* parameters).

parameter for both sample sizes. But the L-MP estimates are consistently better than the JML estimates when the difficulty parameter is from a truncated normal distribution.

The above observations are summarized in Table 5.16 to Table 5.18. Table 5.16 shows the RIMSE for the estimated item curves with L-MP, MML and JML under different conditions based on 100 datasets in each condition. The results show that the estimates of L-MP and MML are very close to each other with MML having

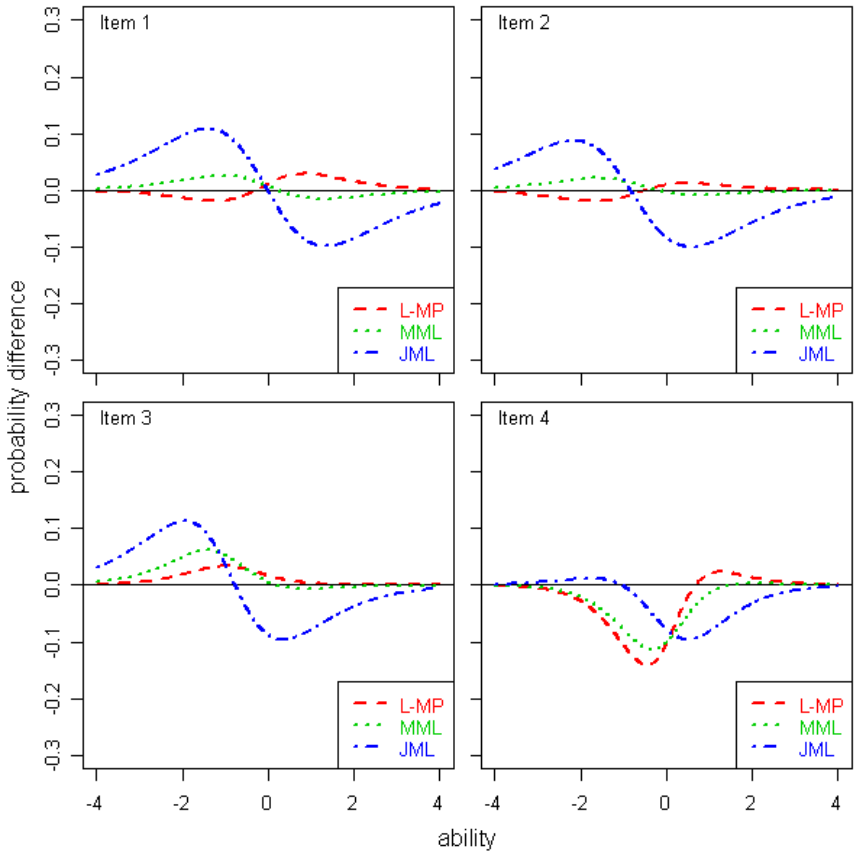

Figure 5.18: Comparisons of estimated ICCs among L-MP, MML and JML in probability difference ($N = 300$, and normally distributed *difficulty* parameters).

slightly smaller RIMSE for the estimated curves but the difference is very small (in the third decimal place) and can be ignored. The estimates from both L-MP and MML are consistently better than JML. An increase in sample size noticeably improves the estimates from L-MP and MML but makes no notable difference for the JML estimates. As to the distribution of the difficulty parameter, MML and JML work better with equally spaced difficulty parameters while L-MP works slightly better under normally distributed difficulty parameters.

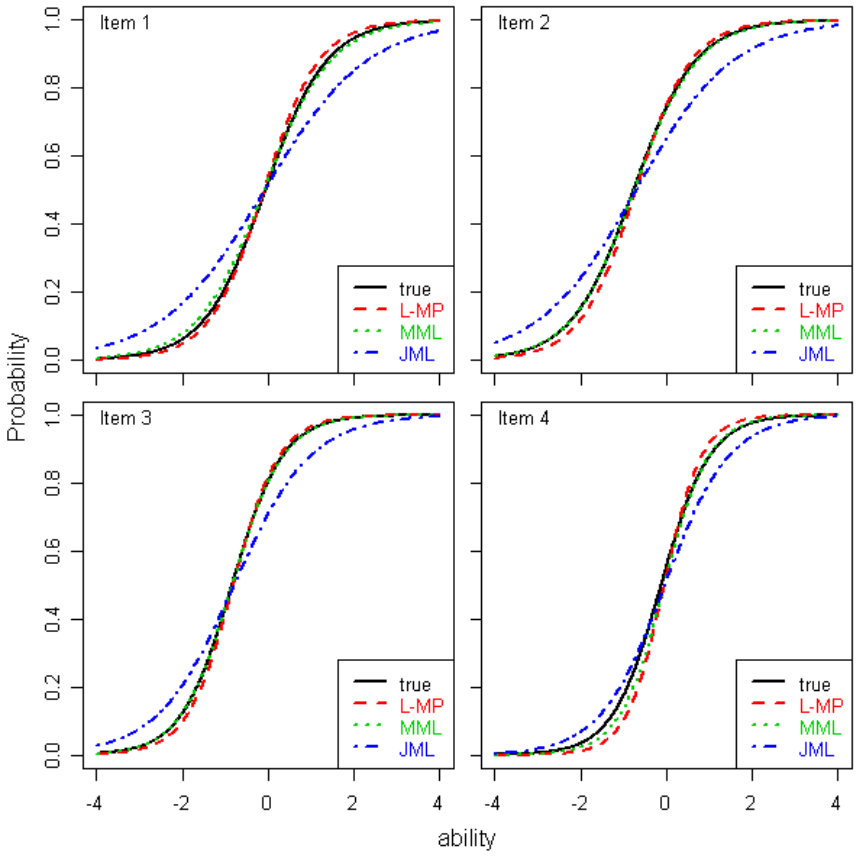

Figure 5.19: Comparisons of estimated ICCs among L-MP, MML and JML ($N = 2000$, and normally distributed *difficulty* parameters).

Table 5.17 and Table 5.18 provide the comparisons of L-MP, MML and JML on the estimates of abilities. The increase in sample size has no significant effect on the estimates of abilities. The RIMSE_θ and the rank correlations both get better for all three estimation methods when the difficulty parameters are randomly drawn from a truncated normal distribution than when they are equally spaced. When comparing the performance of these three methods on the estimates of abilities, L-MP gives

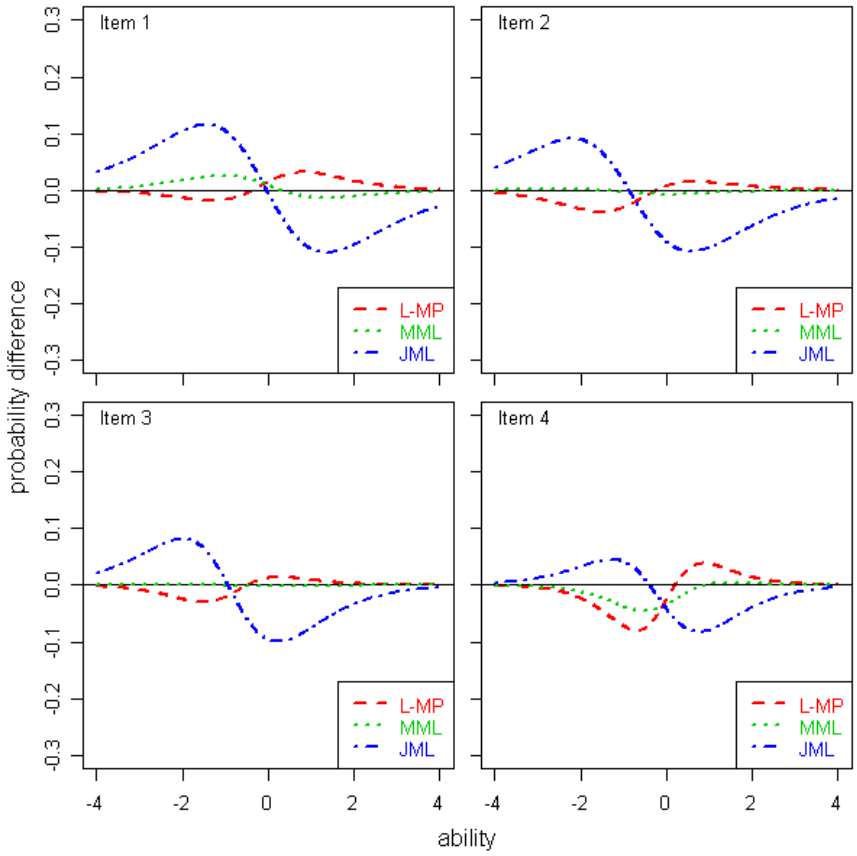

Figure 5.20: Comparisons of estimated ICCs among L-MP, MML and JML in probability difference ($N = 2000$, and normally distributed *difficulty* parameters).

smallest RIMSE_θ among all three methods. However, when it comes to the rank correlations, the results are very close for all three methods.

5.2.4 Simulation 3

It has been illustrated that when the model does not follow a standard logistic function, fitting the data to a 2PL model is not sufficient for most items. In this simulation, we are interested in comparing the performance of the L-MP model with

		L-MP	MML	JML
$N = 300$	ES	0.0392 (0.0011)	0.0326 (0.0009)	0.0752 (0.0009)
	ND	0.0388 (0.0012)	0.0332 (0.0011)	0.0781 (0.0011)
$N = 2000$	ES	0.0246 (0.0004)	0.0125 (0.0003)	0.0730 (0.0003)
	ND	0.0238 (0.0005)	0.0135 (0.0005)	0.0761 (0.0006)

Table 5.16: Comparisons of RIMSE for estimated ICCs among L-MP, MML and JML. ES (Equally Spaced) and ND (Normally Distributed) are two ways to generate the *difficulty* parameter. Each cell is based on 100 datasets. A mean RIMSE is calculated across items for each dataset. The number presented in the table is the average of the mean RIMSE across 100 datasets. The number in parentheses is the standard deviation of the 100 averaged RIMSEs.

		L-MP	MML	JML
$N = 300$	ES	0.3867 (0.0162)	0.3957 (0.0172)	0.4244 (0.0214)
	ND	0.3528 (0.0164)	0.3582 (0.0178)	0.3961 (0.0205)
$N = 2000$	ES	0.3819 (0.0091)	0.3887 (0.0094)	0.4140 (0.0109)
	ND	0.3455 (0.0108)	0.3501 (0.0114)	0.3865 (0.0120)

Table 5.17: Comparisons of RIMSE_θ among L-MP, MML and JML. ES (Equally Spaced) and ND (Normally Distributed) are two ways to generate the *difficulty* parameter. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

		L-MP	MML	JML
$N = 300$	ES	0.9022 (0.0119)	0.9028 (0.0119)	0.9014 (0.0124)
	ND	0.9042 (0.0100)	0.9042 (0.0100)	0.9041 (0.0101)
$N = 2000$	ES	0.9057 (0.0059)	0.9065 (0.0058)	0.9057 (0.0059)
	ND	0.9276 (0.0060)	0.9278 (0.0059)	0.9242 (0.0058)

Table 5.18: Comparisons of rank correlations $\rho(\theta, \hat{\theta})$ among L-MP, MML and JML. ES (Equally Spaced) and ND (Normally Distributed) are two ways to generate the *difficulty* parameter. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

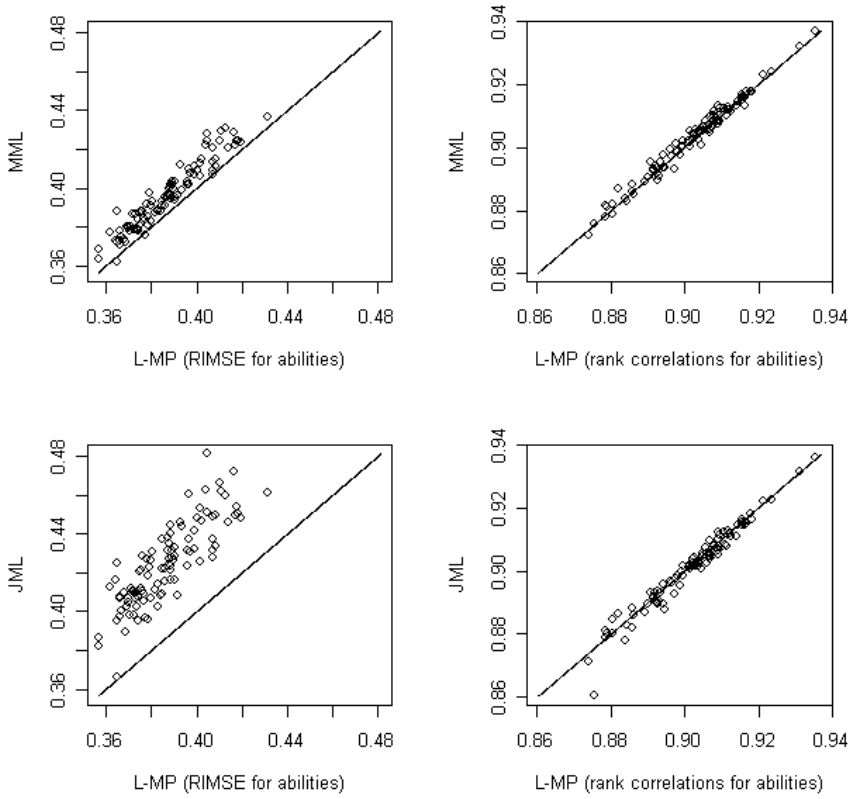

Figure 5.21: Comparisons of $\hat{\theta}$s among L-MP, MML and JML ($N = 300$, and equally spaced *difficulty* parameters).

two other nonparametric techniques. The model used to generate the data is a mixed normal distribution with parameters described as in Section 5.1. Two sample sizes were considered ($N = 300$, $N = 2000$) and 100 datasets were generated for each sample size. The results were compared to TESTGRAF and the Nonparametric Bayesian method separately.

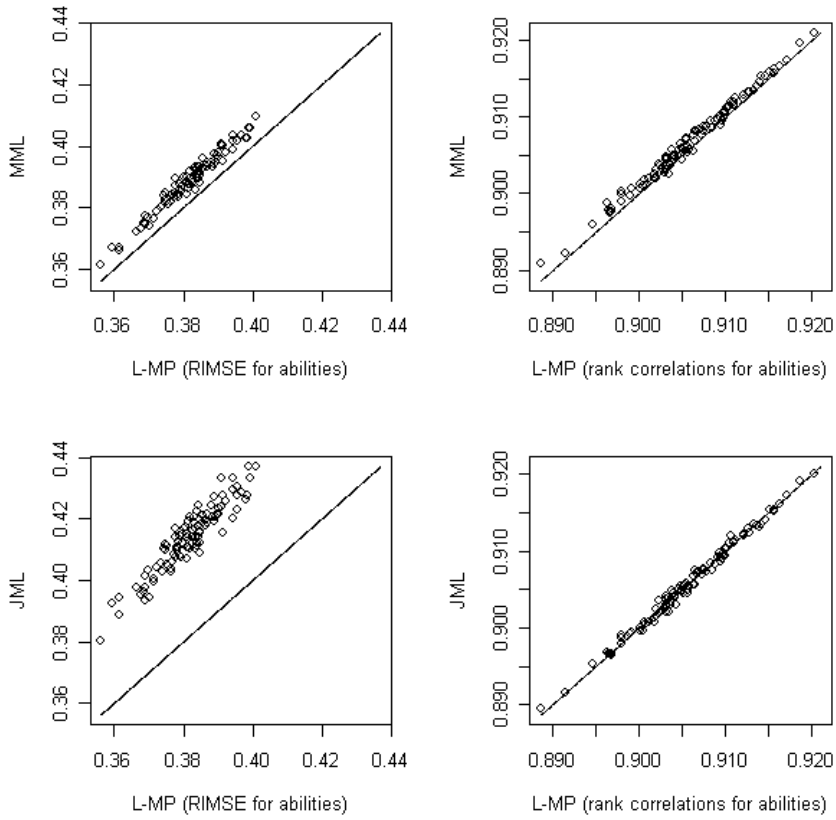

Figure 5.22: Comparisons of $\hat{\theta}$s among L-MP, MML and JML ($N = 2000$, and equally spaced *difficulty* parameters).

5.2.4.1 L-MP and TESTGRAF

The data were analyzed using TESTGRAF with default parameters. For datasets with 300 examinees, the default smoothing parameter is 0.35, and the default evaluating points are 51 points between -2.5 to 2.5. For datasets with 2000 examinees, the default smoothing parameter is 0.24 and the default evaluating points are 51 points between -3.0 and 3.0. However, for most datasets, the program broke down by issuing a warning message indicating that the default smoothing parameter needed to be

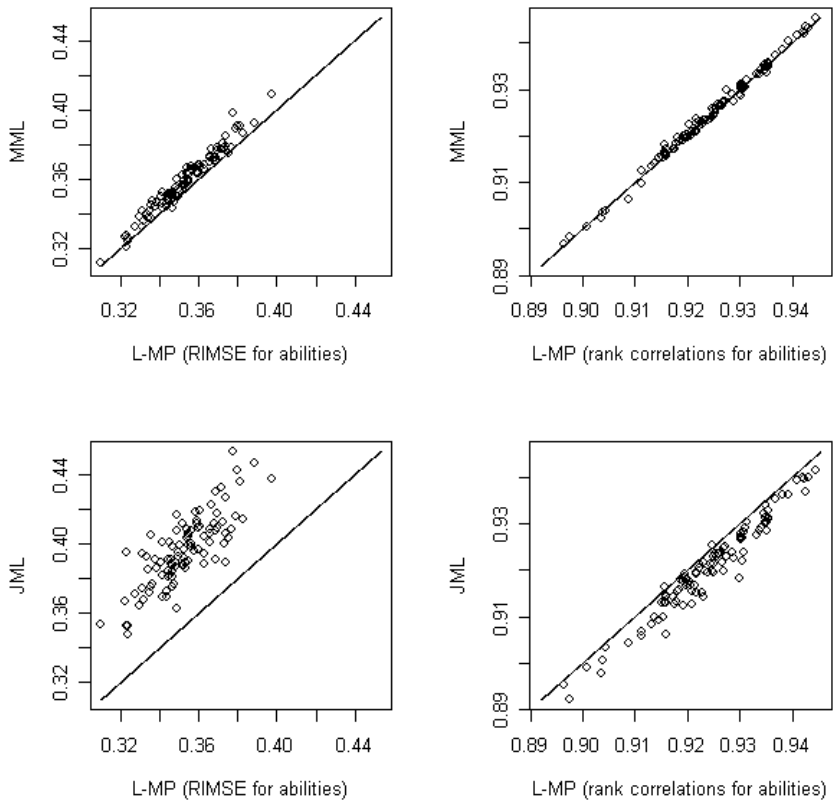

Figure 5.23: Comparisons of $\hat{\theta}$s among L-MP, MML and JML ($N = 300$, and normally distributed *difficulty* parameters).

increased. The smoothing parameter used for samples with size 2000 was increased to 0.40.

The data were fitted up to $k = 4$ (ninth order polynomial) using the L-MP program. Figure 5.25 and Figure 5.27 plot the ICCs for some selected items in a typical dataset estimated by L-MP and TESTFRAF with 300 examinees or 2000 examinees. Figure 5.26 and Figure 5.28 are the corresponding plots for the deviations of probabilities.

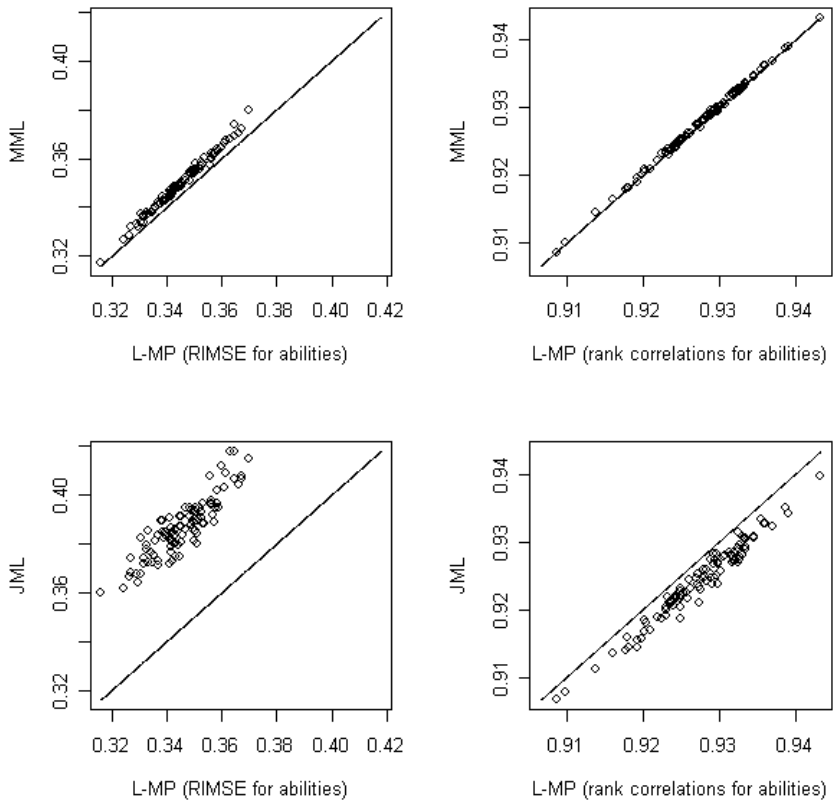

Figure 5.24: Comparisons of $\hat{\theta}$s among L-MP, MML and JML ($N = 2000$, and normally distributed *difficulty* parameters).

Overall, the estimated ICCs from TESTGRAF and L-MP with $k \geq 1$ are similar. For the displayed four items in Figure 5.25, the L-MP with $k \geq 1$ estimates seem to be slightly closer to the true curve. There are some deviations for the estimated ICCs from TESTGRAF, especially for examinees with low abilities or high abilities. When sample size increases to 2000, the estimates from TESTGRAF improved greatly and the estimated ICCs get much closer to the true curves. One important feature of TESTGRAF is that it doesn't put a constraint of monotonicity on the estimated ICCs which means that the estimated ICCs could decrease as ability increased. The

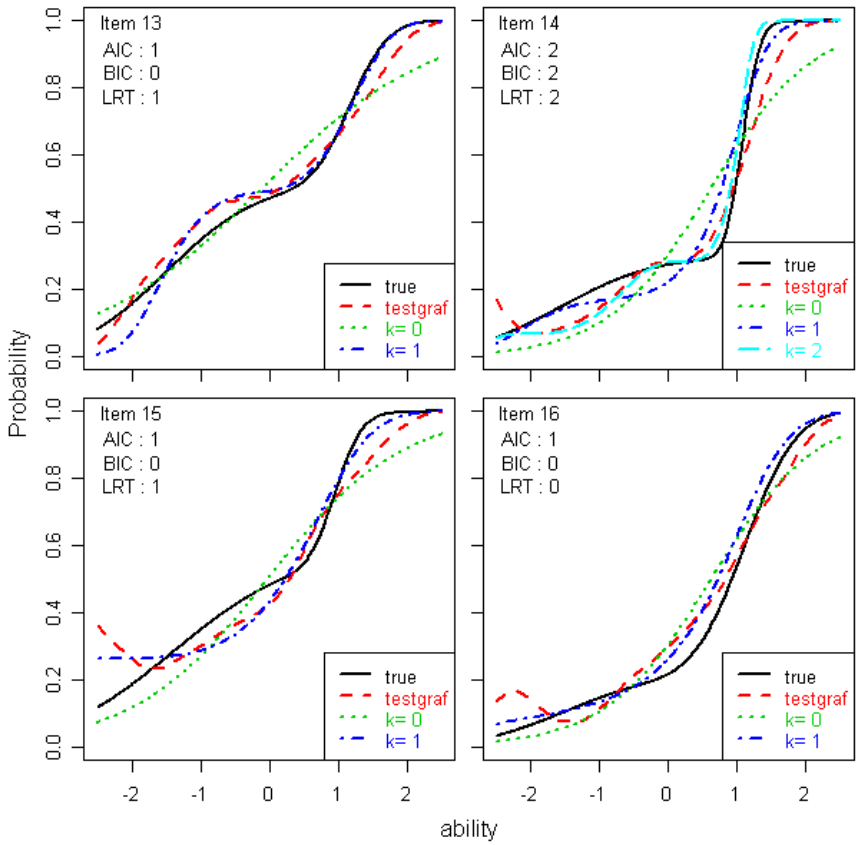

Figure 5.25: Comparisons of the estimated ICCs between L-MP and TESTGRAF ($N = 300$).

nonmonotonicity in the ICCs could be an arguable feature. For any well functioned item, it is reasonable to assume that examinees with higher abilities will have higher probability of endorsing the item and thus the ICC should be nondecreasing. However, the feature of allowing the ICC to be nonmonotonic can sometimes work as a diagnostic tool in deciding if something unexpected happens for a particular item. From this typical dataset, it seems that TESTGRAF can easily pick up any nonmonotonicity in the responses even when the models are truly nondecreasing. This can be seen from item 14, item 15 and item 16 in Figure 5.25. These estimated ICCs from

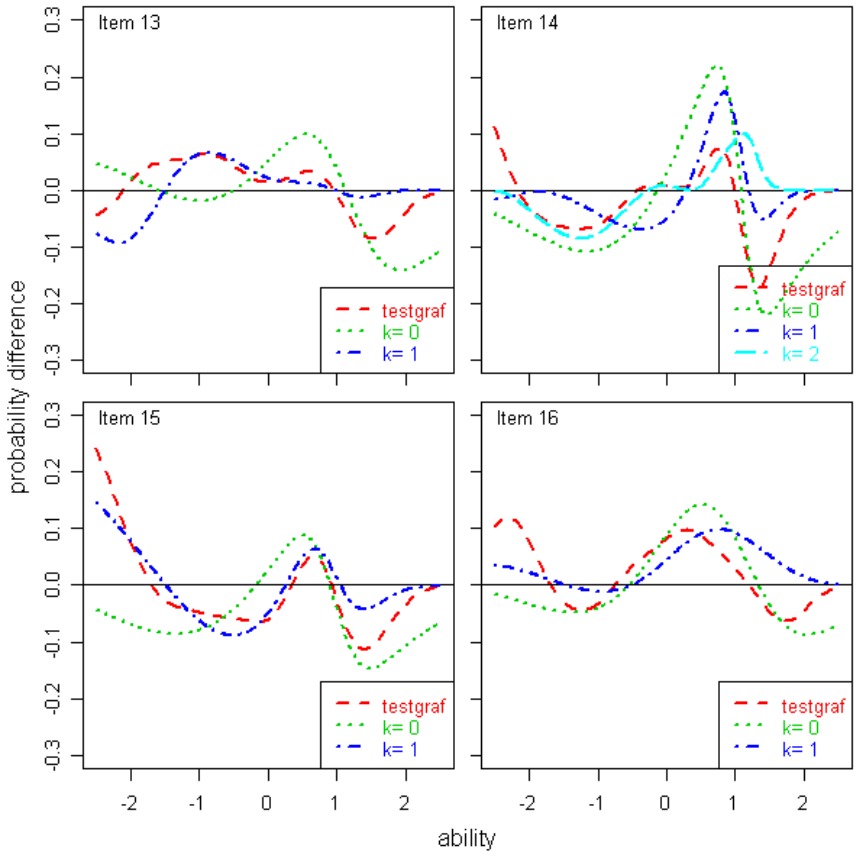

Figure 5.26: Comparisons of the estimated ICCs between L-MP and TESTGRAF in probability difference ($N = 300$).

TESTGRAF are actually giving some false information on the monotonicity for these items. While on the other hand, if the items truly have some problems, TESTGRAF will be able to capture them by having nonmonotonicity in the estimated ICCs where the L-MP model will only produce a monotone function but probably with poor fit. The nonmonotonicity could go away in TESTGRAF if the smoothing parameter is increased.

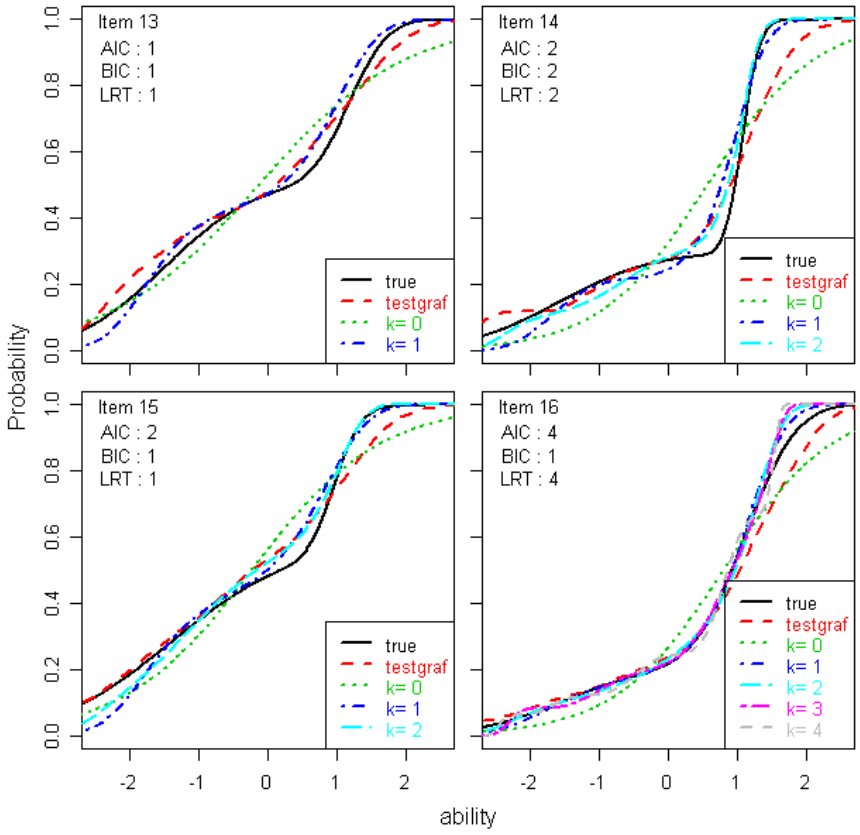

Figure 5.27: Comparisons of the estimated ICCs between L-MP and TESTGRAF ($N = 2000$).

Table 5.19 - Table 5.21 present the summaries for this simulation based on 100 datasets for each sample size. Table 5.19 shows the RIMSE for the estimated ICCs for TESTGRAF and the L-MP model with different selection criteria. With both sample sizes, TESTGRAF has a smaller RIMSE for the estimated ICCs than the L-MP model. But the difference is very small especially with sample size 2000. From Table 5.20 and Table 5.21 where estimates of abilities are evaluated, L-MP model has a smaller RIMSE_θ and higher rank correlations for estimated abilities no matter what model selection criterion was used.

	TESTGRAF	L-MP(AIC)	L-MP(BIC)	L-MP(LRT)
$N = 300$	0.0565 (0.0048)	0.0657 (0.0065)	0.0754 (0.0067)	0.0678 (0.0066)
$N = 2000$	0.0405 (0.0033)	0.0417 (0.0030)	0.0423 (0.0035)	0.0417 (0.0032)

Table 5.19: Comparisons of RIMSE for estimated ICCs between TESTGRAF and L-MP. The last three columns represent the L-MP models using three different criteria. Each cell is based on 100 datasets. A mean RIMSE is calculated across items for each dataset. The number presented in the table is the average of the mean RIMSE across 100 datasets. The number in parentheses is the standard deviation of the 100 averaged RIMSEs.

	TESTGRAF	L-MP(AIC)	L-MP(BIC)	L-MP(LRT)
$N = 300$	0.6257 (0.0353)	0.4735 (0.0223)	0.4727 (0.0215)	0.4733 (0.0224)
$N = 2000$	0.6264 (0.0186)	0.4624 (0.0115)	0.4611 (0.0113)	0.4622 (0.0115)

Table 5.20: Comparisons of RIMSE_θ between TESTGRAF and L-MP. The last three columns represent the L-MP models using three different criteria. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

	TESTGRAF	L-MP(AIC)	L-MP(BIC)	L-MP(LRT)
$N = 300$	0.7626 (0.0500)	0.8280 (0.0274)	0.8291 (0.0266)	0.8284 (0.0274)
$N = 2000$	0.7686 (0.0366)	0.8342 (0.0122)	0.8348 (0.0123)	0.8341 (0.0122)

Table 5.21: Comparisons of rank correlations for abilities $\rho(\theta, \hat{\theta})$ between TESTGRAF and L-MP. The last three columns represent the L-MP models using three different criteria. Each cell is based on 100 datasets. The number in parentheses is the standard deviation of the 100 replications.

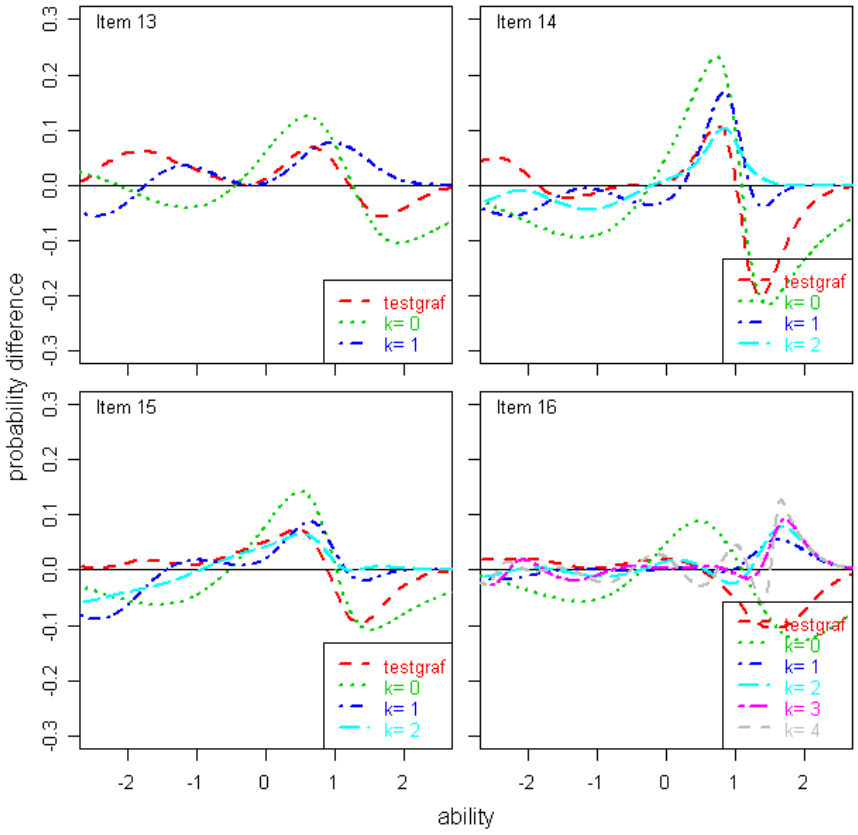

Figure 5.28: Comparisons of the estimated ICCs between L-MP and TESTGRAF in probability difference ($N = 2000$).

This information is also shown in Figure 5.29 and Figure 5.30. When looking at RIMSE$_\theta$, almost all points lie above the $y = x$ line. And for rank correlations, most of the points lie below $y = x$ line. This means that the L-MP has smaller RIMSE$_\theta$ and higher rank correlations than TESTGRAF program in this simulation.

However, for simplicity purpose, the above comparisons were only made when TESTGRAF used same smoothing parameters for all datasets. Ramsay (1991) suggested that a smoothing parameter of $h = N^{-\frac{1}{5}}$ will produce good results. According to TESTGRAF manual, the default parameter in TESTGRAF is set to be $1.1N^{-\frac{1}{5}}$.

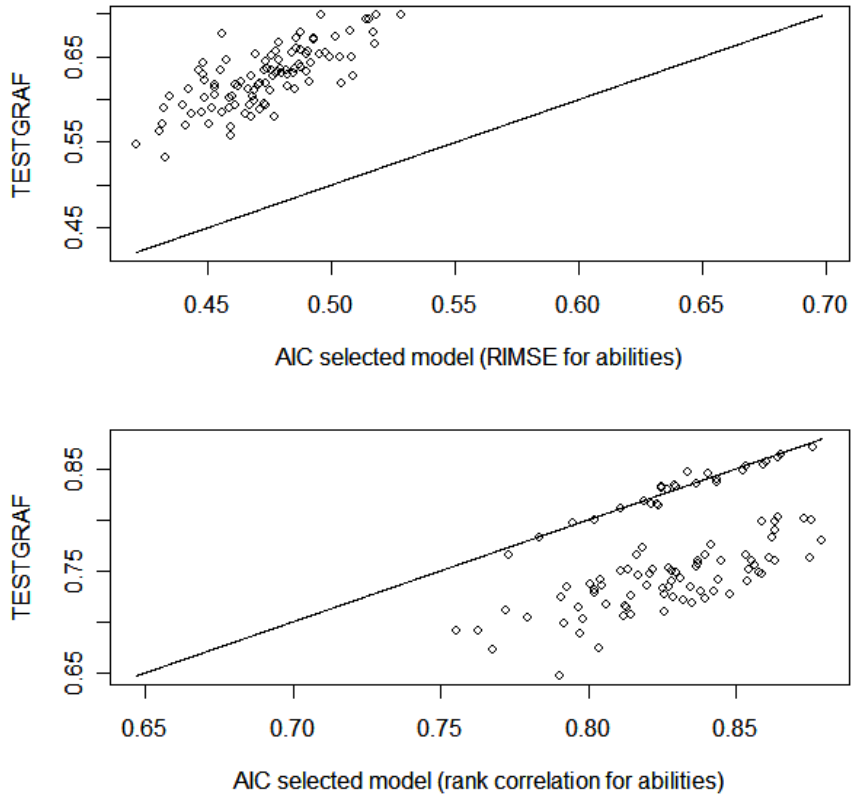

Figure 5.29: Comparisons of $\hat{\theta}$s between TESTGRAF and L-MP ($N = 300$).

A small experiment was run to see how the comparisons between TESTGRAF and L-MP change with various smoothing parameters. The different levels of smoothing parameter were chosen to be around the values actually used. For simplicity purpose, this little experiment was only run with two datasets, one for each sample size. The results are shown in Table 5.22 and Table 5.23.

It can be seen in Table 5.22 and Table 5.23, for this particular dataset, the smallest RIMSE value for the estimated ICCs occurs at the default h value for $N = 300$. This is also true when the estimates of abilities are considered. When sample size

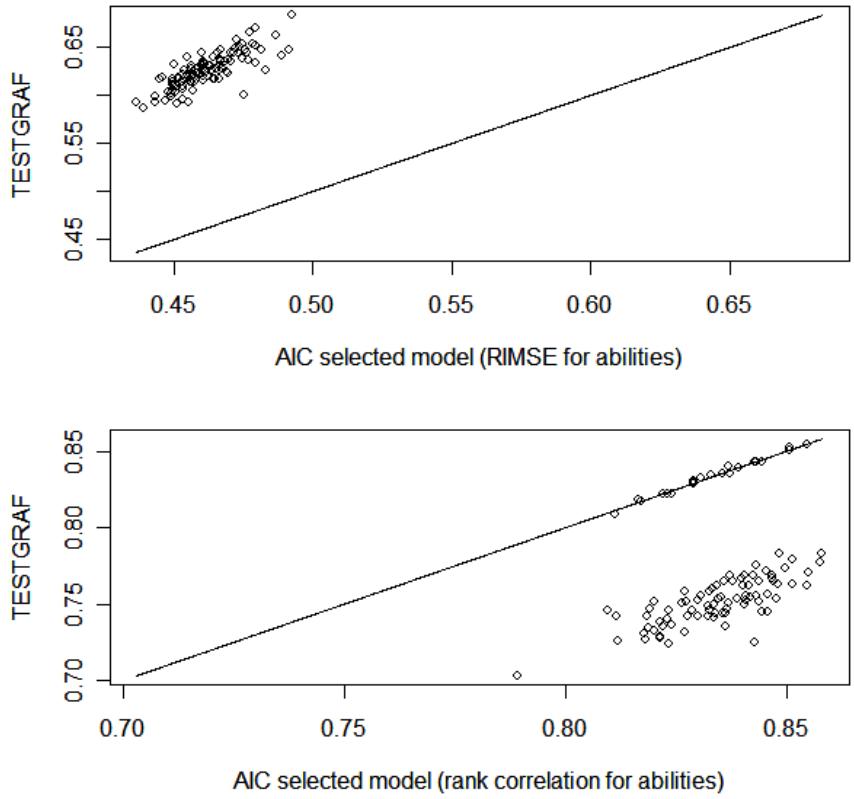

Figure 5.30: Comparisons of estimated $\hat{\theta}$s between TESTGRAF and L-MP ($N = 2000$).

increases to 2000, the smallest RIMSE for the estimated ICCs from TESTGRAF occurs at $h = 0.36$ which is less than the value we used in our simulations. However, TESTGRAF will break down using $h = 0.36$ for many datasets. Changing the smoothing parameters in TESTGRAF might be able to improve the RIMSE for estimated ICCs slightly, but it does not appear to affect the comparisons we have made between TESTGRAF and the L-MP model.

TESTGRAF has a strong influence on the L-MP model, from the idea of producing a more flexible ICC than the 1PL or 2PL model to the estimation method. There are

	L-MP			TESTGRAF		
	AIC	BIC	LRT	$h = .30$	$h = .35^*$	$h = .40$
RIMSE(ICC)	0.0628	0.0760	0.0667	0.0580	0.0567	0.0570
RIMSE(θ)	0.4780	0.4739	0.4801	0.6160	0.6160	0.6580
rank corr	0.8367	0.8419	0.8347	0.8300	0.8300	0.8359

Table 5.22: Comparisons between L-MP and TESTGRAF with various smoothing parameters ($N = 300$). The h value with * is the default smoothing parameter TESTGRAF used.

	L-MP			TESTGRAF		
	AIC	BIC	LRT	$h = .36$	$h = .40^*$	$h = .44$
RIMSE(ICC)	0.0357	0.0378	0.0351	0.0338	0.0363	0.0394
RIMSE(θ)	0.4509	0.4491	0.4502	0.5864	0.6026	0.6195
rank corr	0.8330	0.8351	0.8334	0.8334	0.8340	0.8338

Table 5.23: Comparisons between L-MP and TESTGRAF with various smoothing parameters ($N = 2000$). The h value with * is the actual smoothing parameter used in the simulations.

some major differences, however, between these two methods. These differences are listed in Table 5.24.

5.2.4.2 L-MP and Nonparametric Bayesian

As described in Section 2.2.2, Qin (1998) handled the item response data in a nonparametric, fully Bayesian manner. The approach uses a Dirichlet process to release the constraint of the prior distribution and is preferred when the true ICC deviates from the parametric family. It will be interesting to also compare the performance of the L-MP model with this nonparametric Bayesian procedure.

	TESTGRAF	L-MP
surrogate ability	normalized test score	normalized 1st principal component scores
item curves	Option Characteristic Curve	Item Characteristic Curve
estimation of curves	kernel smoothing	ML estimates
estimation of abilities	ML estimates	EAP estimates

Table 5.24: Major differences between L-MP and TESTGRAF

The program for the nonparametric Bayesian model is a R package *irtNP* (Duncan & MacEachern, in press). Due to the long executing time of the R package, only one sample with $N = 300$ was used for comparison purpose.

The plots of the last four items in a simulated dataset with $N = 300$ are shown in Figure 5.31. The result from the nonparametric Bayesian model is based on 2500 updates with the first 1000 iterations as the burn-in period. The thinning interval was chosen to be 10.

When the true curve is the CDF for a mixture of two normal distributions, the nonparametric Bayesian model is also able to pick up some nonstandard logistic characteristics in the curve. Generally speaking, the nonparametric Bayesian and the AIC selected L-MP model produce curves with similar shape. Differences usually occur at both tails. Considering ability scale in the plot is from -4 to 4, the number of examinees in the tails are really small and thus the differences between two curves on the tails are not considered as a problem. Although not strictly following the true curve, the estimates from the nonparametric Bayes are reasonably close to the true curve especially for medium abilities. With larger number of examinees and larger number of items, the estimates might be improved.

The RIMSE for the estimated ICCs, the RIMSE_θ and the rank correlation for the abilities for this particular dataset are shown in Table 5.25. The comparison

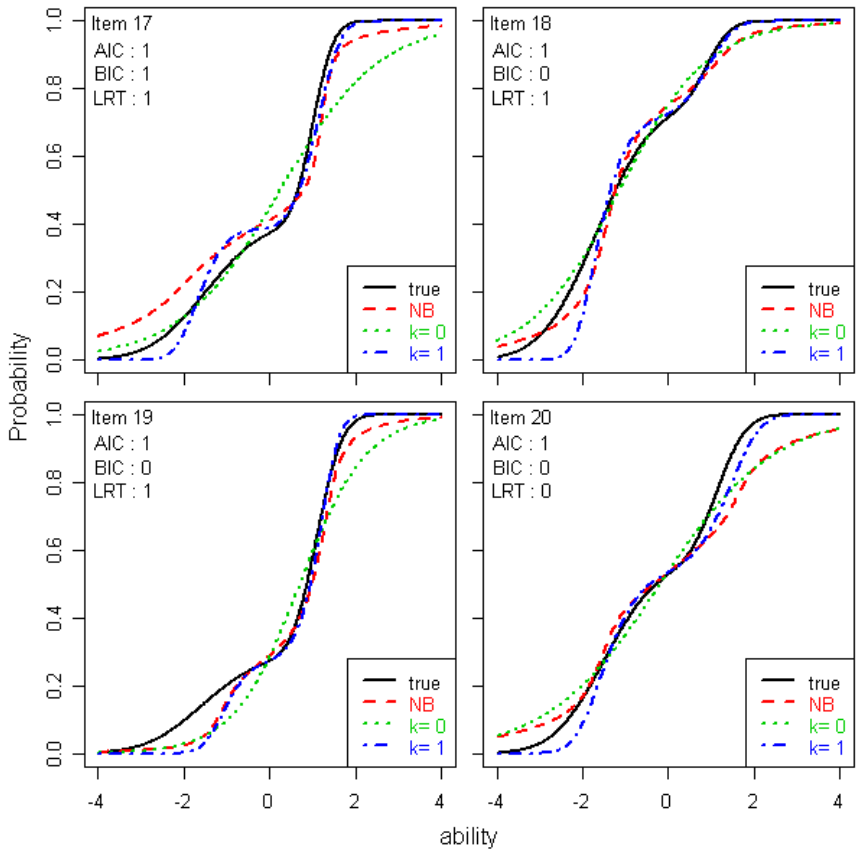

Figure 5.31: Comparisons of estimated ICCs between L-MP and Nonparametric Bayesian Models ($N = 300$). "NB" refers to the Nonparametric Bayesian Model.

shows that for this particular dataset, the Nonparametric Bayesian model has smaller RIMSE than the L-MP model. But in terms of the estimated abilities, these two models produce very similar results.

It would be good if the comparisons were based on more datasets or datasets with larger sample size. However, the irtNP package for the nonparametric Bayesian model is very time consuming and it is unrealistic to run too many datasets especially with large sample size. For example, for the current dataset with 300 examinees and

	AIC	BIC	LRT	NB
RIMSE(ICC)	0.0610	0.0689	0.0627	0.0587
RIMSE(θ)	0.4743	0.4801	0.4728	0.4747
rank corr	0.7833	0.7832	0.7854	0.7881

Table 5.25: Comparisons between the L-MP and the Nonparametric Bayesian models ($N = 300$). "NB" refers to the Nonparametric Bayesian model. The comparison is based on one dataset.

20 items to have 2500 updates and 1000 burn-in points, it took 1.78 days to run on a computer with P4 2.66GHz CPU and 512M RAM, but it only took 31.50 seconds for the L-MP program fitted to $k = 4$ (ninth order polynomial). The fact that irtNP program is written in R, which is known to be much slower than FORTRAN, and that it uses MCMC are the reasons for the slow speed.

5.3 Discussion

Some questions were raised from the results of the simulation study. From the comparisons between L-MP and TESTGRAF in Section 5.2.4, we have seen that for some items, TESTGRAF produced some estimated ICCs with nonmonotonicity. It will be interesting to see what would happen if we release the monotonic constraint on the ICCs. One other question is related to the statement made in Section 3.1 for the filtered polynomial density estimation which claims that with increased degree of monotonic polynomial, many continuous non-defective distributions may be approximated to arbitrary closeness. One difference between the L-MP model and the filtered polynomial density estimation method is that in the L-MP model, the independent variable, θ, is an unobserved latent trait and needs to be estimated from the model. It will be interesting to investigate if this statement for the filtered polynomial density estimation is still true for the L-MP model with this difference.

5.3.1 Logistic function with unconstrained polynomial

To model the data using a logistic function with an unconstrained polynomial, we used the estimates from the L-MP at each k stage as starting points, fit the data to an IRF with same degree of ordinary polynomial. For example, when $k = 2$, the fitted model is

$$P_i(\theta) = \frac{1}{1+e^{-m_i(\theta)}},$$

where $m_i(\theta)$ is an ordinary polynomial defined as

$$m_i(\theta) = b_{i0} + b_{i1}\theta + b_{i2}\theta^2 + \cdots + b_{i,5}\theta^5.$$

The same data as in Figure 5.25 was reanalyzed using an IRF with an unconstrained polynomial. To avoid too many curves on the same plot and make the comparisons easier, we only plot the ICCs from the L-MP model with unconstrained polynomial and monotonic polynomial at $k = 1$ and the ICC from TESTGRAF. The same four items are plotted as an illustration in Figure 5.32.

Based on the observation of all items in the dataset, the estimated ICCs with monotonic polynomial and with unconstrained polynomial are very close for most items, for example, item 13, item 14 and item 16. The estimated ICCs with monotonic polynomial or with ordinary polynomial for item 15 are very close for most ability levels. The small differences occur for examinees with extreme abilities. For items that are non-monotonic by TESTGRAF, for example, item 15 in Figure 5.32, the ICC with unconstrained polynomial is very close to that from TESTGRAF. Releasing the monotonic constraint is provided as an additional facility in the L-MP program. A warning message will be generated if the item has a negative principal component loading. Releasing the constraint in such cases would be a good way to check if there is any problem with the item. It is always worthy to investigate why that

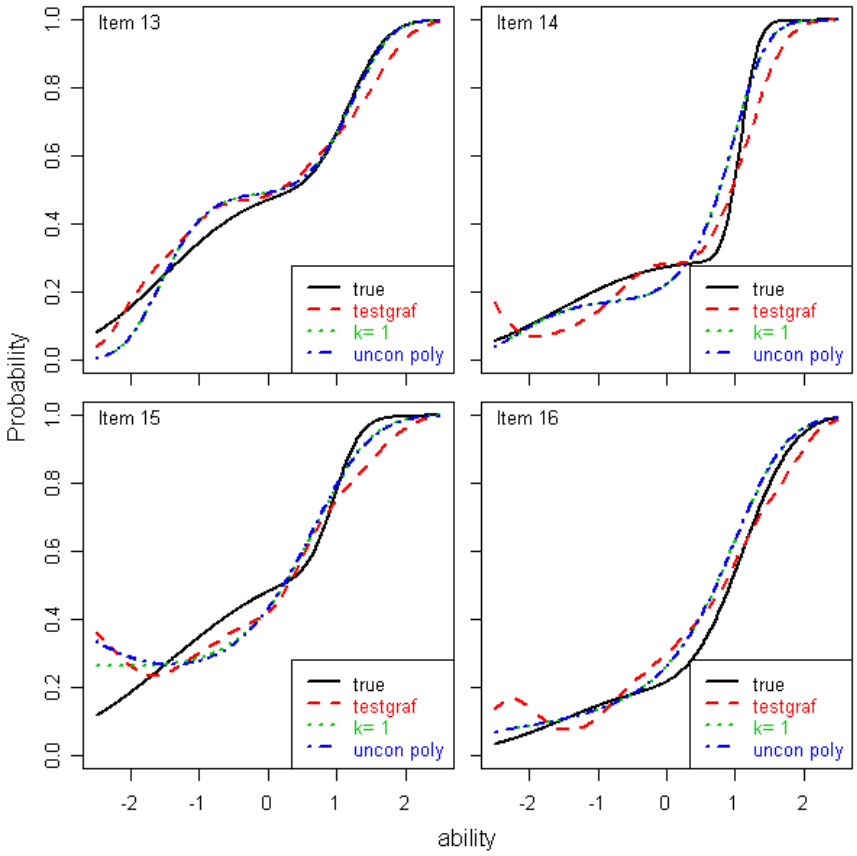

Figure 5.32: Comparisons of estimated ICCs among TESTGRAF, L-MP ($k = 1$) and IRFs with unconstrained polynomial ($N = 300$).

non-monotonicity occurs. However, we should be cautious when interpreting such non-monotonic IRFs. As was mentioned in Section 5.2.4, non-monotonicity could occur because the estimation process is too sensitive to random departures from monotonicity. The problem would be more severe if the non-monotonicity occurs in the middle instead of the tails. If the probabilities of endorsing an item consistently decrease as the abilities increase, we can make the conclusion that something has gone wrong with the item with much more confidence.

5.3.2 The estimation errors

By checking Figure 5.25 and Figure 5.27, although the curves are significantly different from the standard logistic function, it can be seen that the differences between the estimated ICCs with higher k values are very small. It looks as if increasing the order of the monotonic polynomial has no effect in moving the estimated curves to the true curve. The estimated ICCs are closer to the true curve for the data with 2000 examinees than the data with only 300 examinees. These reflect the two causes for the deviations of the estimated ICCs from the true curves. One of them is the abilities. In the filtered polynomial density estimation procedure, the method is applied to a sample of observed data. In IRT, however, the abilities are unobserved and we used surrogate abilities to estimate the item curves. The surrogate abilities and the true abilities are not equal. A second cause is the sampling errors. If using the true simulated abilities to estimate the ICCs and if we have large enough sample, these two sources of error are supposed to be eliminated. This is illustrated in Figure 5.33.

The best k value for the estimated ICCs to be close to the true curve varies from item to item. For example, in Figure 5.33, when $k = 1$, the estimated ICC for item 4 is almost identical to the true curve. For item 17 and item 20, when k gets to 2, the estimated ICCs are almost identical to the true curves. Increasing the k values for these items actually introduce some deviations for the lower tail. The estimated ICC for item 8 requires $k = 5$ to follow along the true curve almost exactly. These plots illustrate that when the two sources of error are eliminated by using the true simulated abilities instead of the surrogates and the large enough sample ($N = 10000$), increasing the k value will result in an estimated ICC close enough to the true curve.

However, the ideal scenario described above of using the true abilities and the large enough sample size are not practical in real life. But if the surrogates are good

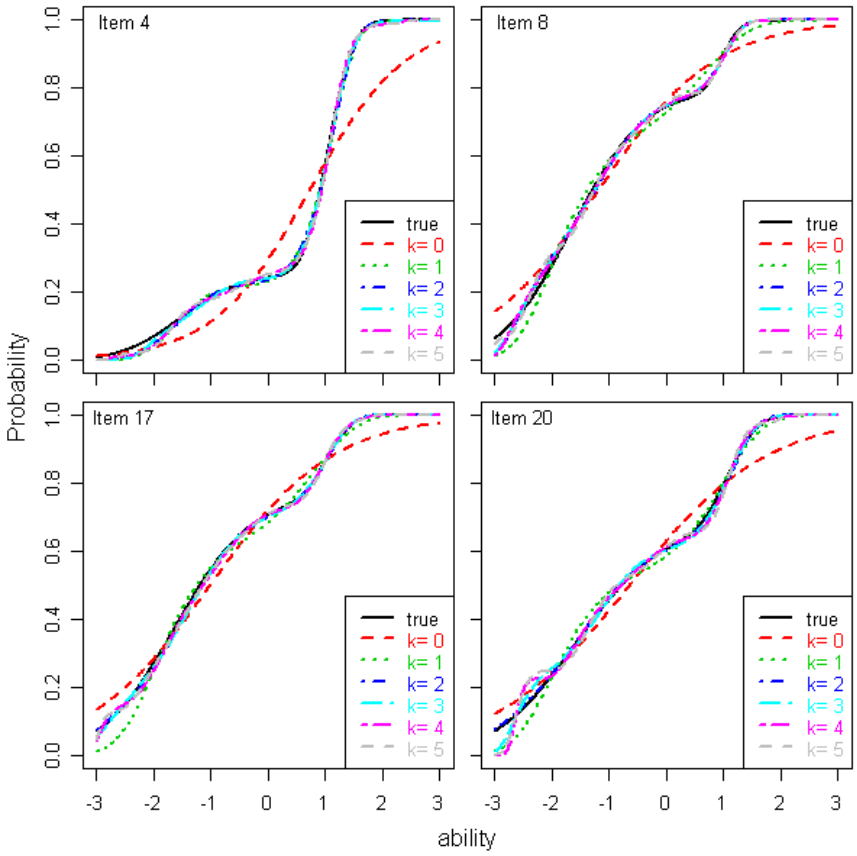

Figure 5.33: Estimated ICCs using simulated abilities for $N = 10000$ and $n = 20$. N is the number of examinees and n is the number of items.

estimates of the abilities, this type of error could be minimized. The better surrogates could be obtained by having a longer test which usually provides more information about the examinee's proficiency. An illustrative example of a simulated test of 100 items on 5000 examinees is shown in Figure 5.34.

No true abilities were used to obtain the estimated ICCs for this dataset. Instead, the normalized principal component scores were used as the surrogates of abilities. For most items, the plots reveal that although not exactly, the estimated ICCs are fairly close to the true curves. This demonstrates that using a longer test, $n = 100$

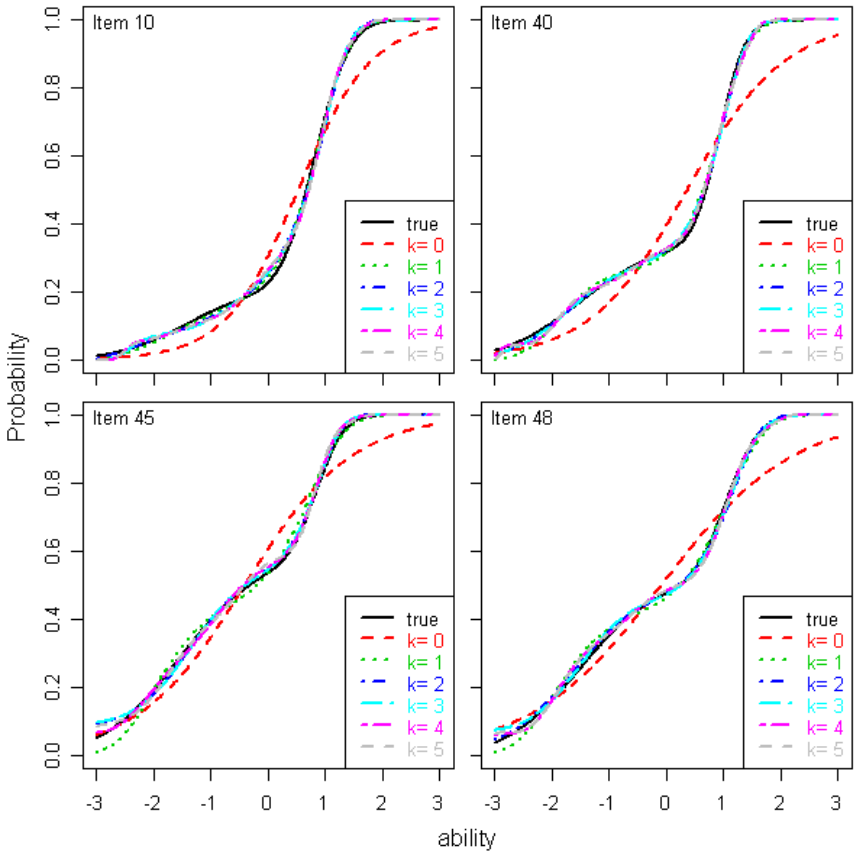

Figure 5.34: Estimated ICCs using surrogate abilities for $N = 5000$ and $n = 100$. N is the number of examinees and n is the number of items.

in this example, will help to improve the goodness of fit of the estimated ICCs.

In summary, the advantage of using the L-MP model is clear. If the true models are the standard logistic functions, the L-MP estimates can provide very close estimated ICCs to those from MML and much better estimated ICCs than those from JML. In terms of the abilities, L-MP model produce slightly better estimates than MML and much better estimates than JML. This suggests that the iteration procedure used in JML procedure actually produce worse estimates. When the true

models are not standard logistic functions, the ordinary 1PL or 2PL models are not able to pick up the additional features. But the L-MP model is capable of doing this. The comparison between TESTGRAF and L-MP shows that generally L-MP and TESTGRAF produce very similar estimated ICCs for most items. TESTGRAF has slightly better RIMSEs for the estimated item curves, but L-MP model produces better estimates of abilities in terms of RIMSE_θ and rank correlations. The comparison between L-MP and the Nonparametric Bayesian model shows that these two methods produce very similar results. The Nonparametric Bayesian model may yield better estimated ICCs than the L-MP model, but differences are too small to interpret with any certainty. The computational time for the Nonparametric Bayesian program is much longer than for the L-MP program. In summary, Our experiments indicate that results from the L-MP model are comparable to the best of those from other approaches considered. This demonstrates that the surrogate ability approach, adapted from TESTGRAF and used in L-MP, yields results that are completely suitable for practical use.

CHAPTER 6

APPLICATIONS

The L-MP model has been shown to be a useful model through simulation studies in Chapter 5. In this chapter, we will apply this model to two real world examples. One is the data example that comes with the TESTGRAF program (Ramsay, 1991). It consists of students' responses on a general psy101 class. The other example is the sample data used in the irtNP package (Duncan & MacEachern, in press) for the nonparametric Bayesian model. It consists of students' responses on an elementary statistics class exam. The two datasets were analyzed using the L-MP program and were compared with TESTGRAF and the Nonparametric Bayesian model separately.

6.1 Psychology 101 data example

This illustration used a dataset from an examination given to 379 students in an introductory course in psychology in the Christmas period of 1989 at University of McGill (TESTGRAF manual, 2000). The examination included 100 multiple choice questions, each had four options with one correct. The data are used as an example in the TESTGRAF program package and can be downloaded from Ramsay's website: http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html.

Since L-MP only deals with binary response data, the data in the original file were recoded. A "1" was assigned to an item if a correct response was made otherwise a "0" was assigned. Missing responses were treated as wrong response in this analysis.

TESTGRAF provides Option Characteristic Curves (OCCs) instead of ICCs. To make the results more comparable, we used the recoded data for both programs. No student got zero score or perfect score in this test. The lowest score among these 379 students was 24 and the highest was 91. The data were analyzed with both TESTGRAF and L-MP. The smoothing parameter in the TESTGRAF was chosen to be the default value of 0.33. The number of evaluation points was set to be 51 points between -3 to 3. When fitting with the L-MP model, the data were fitted up to $k = 5$ (11th order polynomial).

Since no true values of the parameters are known, we investigated the goodness of fit using some empirical points. In order to construct the empirical points, the abilities were divided into 14 intervals. The left most interval was $\theta \leq -3$, and the right most interval was $\theta > 3$. The other 12 intervals were between -3 to 3, with 0.5 apart from each other. The percentage of individuals getting the item correct in the interval was used as the empirical probability for that interval. The plots for some selected items in the test are shown in Figure 6.1. We did not plot the items if the model selection criterion chose the standard logistic function. What was plot is part of the items that the standard logistic functions were not able to fit the data and TESTGRAF or L-MP model with higher degree polynomial were preferable. For all the plotted items, it is obvious that the standard normal logistic function (LMP-k0) are not following the empirical points. But the TESTGRAF or the AIC selected L-MP curves are showing the trend more clearly. For most of these items, the estimates from TESTGRAF and L-MP model are very close except that for some items where TESTGRAF produces curves with nonmonotonicity, for example, item 13 and item 22. If we release the monotonic constraint, the L-MP program tends to produce similar results as TESTGRAF. Since usually we assume abilities follow a standard normal distribution, more examinees fall in the middle of the ability intervals. Thus

any nonmonotonicity in the middle of the curve shows a more severe problem. The nonmonotonicity in the ICC tails could be due to random errors. In that sense, item 96 should definitely call the attention of the test developer since students with higher abilities actually have lower probability to answer the question right. It could either be the item is mis-typed or the item is ambiguous. The feature of allowing nonmonotonicity in this case provides very clear diagnostic information. If releasing the monotonic constraint, the L-MP model also produces an item curve that shows a decreasing trend. With the monotonic constraint, the fitted line by the L-MP model is basically a straight line which should catch the attention of the test developer too.

Figure 6.2 plots the estimates of abilities from the L-MP model with different selection criteria against the estimates from TESTGRAF. If two approaches give consistent solution, the points will scatter around the $y = x$ line. From the three graphs in Figure 6.2, regardless of what model selection criterion was used, the estimates from two methods are generally consistent for the majority of examinees from -1.8 to 1.8. Outside that range, it seems that L-MP model tends to give lower estimates than TESTGRAF for examinees with high ability levels and higher estimates than TESTGRAF for examinees with low ability levels.

6.2 An elementary statistics test example

The dataset used in this section comes from students' responses on a test for an elementary statistics class at The Ohio State University. The test has 32 items with 28 of them are multiple choice questions. The test was administered to 258 undergraduate students. Since the current L-MP model only works with binary data, we discarded the 4 constructed response items and retained the 28 multiple choice items. The data is also used as the sample data for the irtNP package.

The data was analyzed using the L-MP model fitted to as high as 11th order of the monotonic polynomial ($k = 5$). The estimated curve heights from the Nonparametric Bayesian model with 2PL function as the prior were provided by the author of the irtNP package. The estimates are based on 20,000 updates with a burn-in period of 5000 points and a thinning interval of 10. There are 801 evaluating points equally spaced between -4 to 4. The plots for some selected items are shown in Figure 6.3.

The comparison between the L-MP model and the Nonparametric Bayesian model finds that the estimates from these two models agree well for most of the items, especially when the data can be approximated using the standard logistic function. For example, for item 4 and item 5 in Figure 6.3, the estimated curves from these two methods are almost identical. There are some items for which the standard logistic functions are not sufficient, for example, item 2, item 11, item 12, item 14, item 18, item 23 and item 24 in the figure. Among these items, the main trend of the estimated ICCs from these two models agrees reasonably well with slight deviations in the tails. For example, the AIC selects the $k = 2$ model for item 18. The estimated ICC from the L-MP model differs from the one from Nonparametric Bayesian in both tails. For this particular item, the one from the L-MP model seems agreeing with the empirical points on the tails better.

Figure 6.4 plots the estimates of abilities from the L-MP model with different selection criteria against the estimates from the Nonparametric Bayesian model. If two approaches give consistent solution, the points will scatter around the $y = x$ line. From the three graphs in Figure 6.4, regardless of what model selection criterion was used, the estimates from these two methods agree very well.

Overall speaking, the L-MP model and the Nonparametric Bayesian model agree with each other for most of the items. For the items that are truly following a standard logistic curve, both models are able to produce estimates close to the standard

2PL function. When the data are not following the standard logistic functions, the 2PL model will fit the data poorly, while the L-MP model and the Nonparametric Bayesian model are usually able to produce estimated curves that capture the data characteristic better.

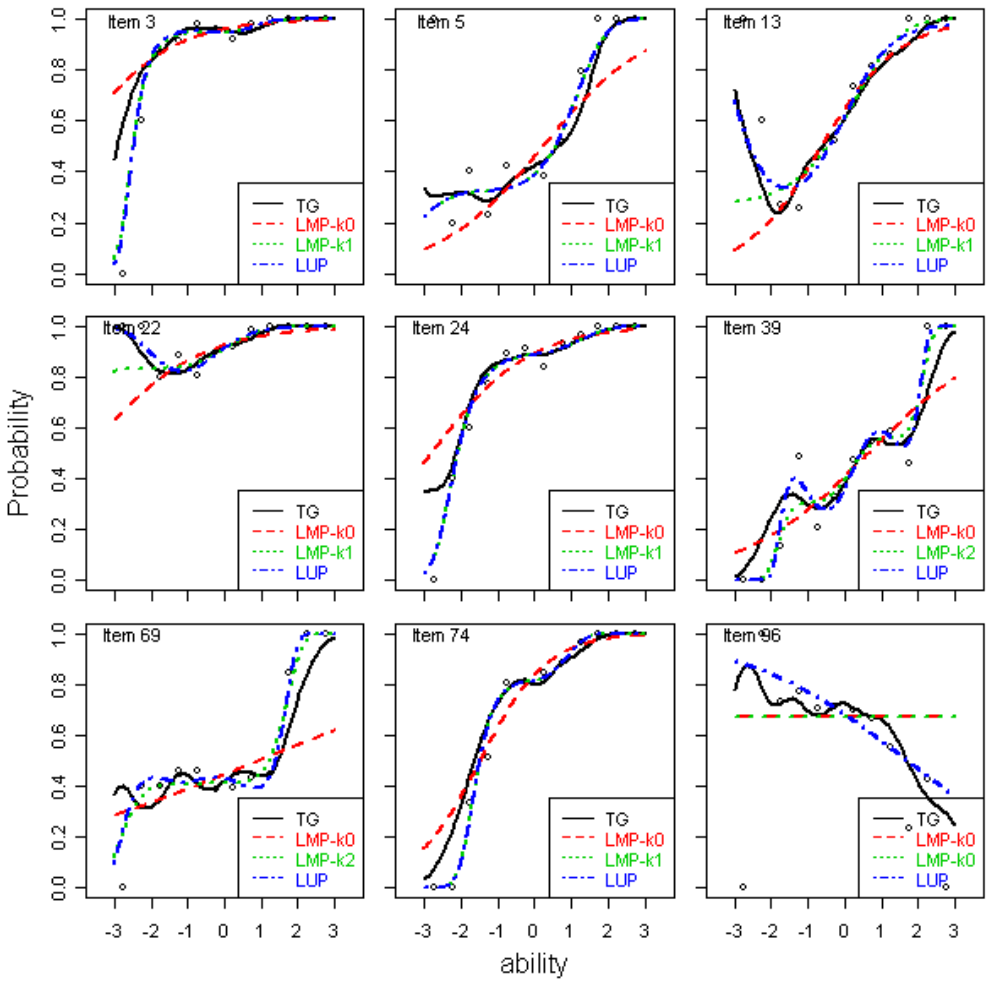

Figure 6.1: Estimated ICCs for some selected items from TESTGRAF and L-MP (psychology 101 data). "TG" represents TESTGRAF; the red curve LMP-k0 is the L-MP model with standard logistic curve; the green curve is the L-MP model selected by AIC criterion; LUP represents the logistic function with unconstrained polynomial.

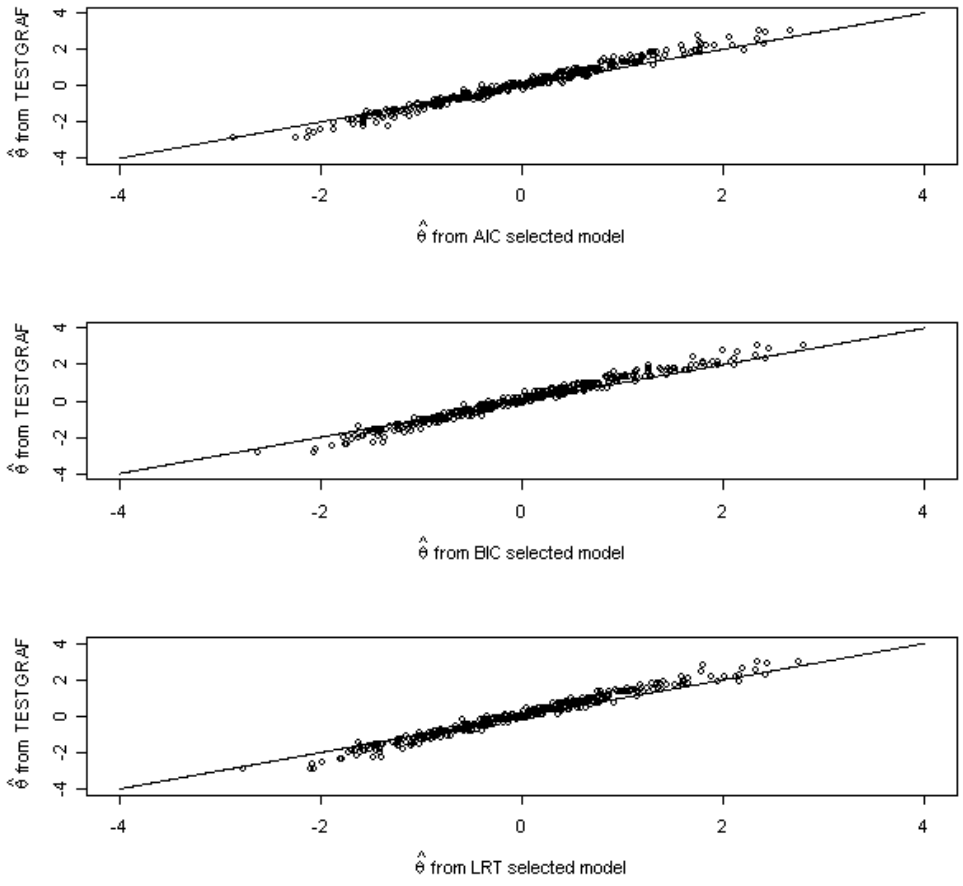

Figure 6.2: Comparisons of $\hat{\theta}$s from TESTGRAF and L-MP (psychology 101 data).

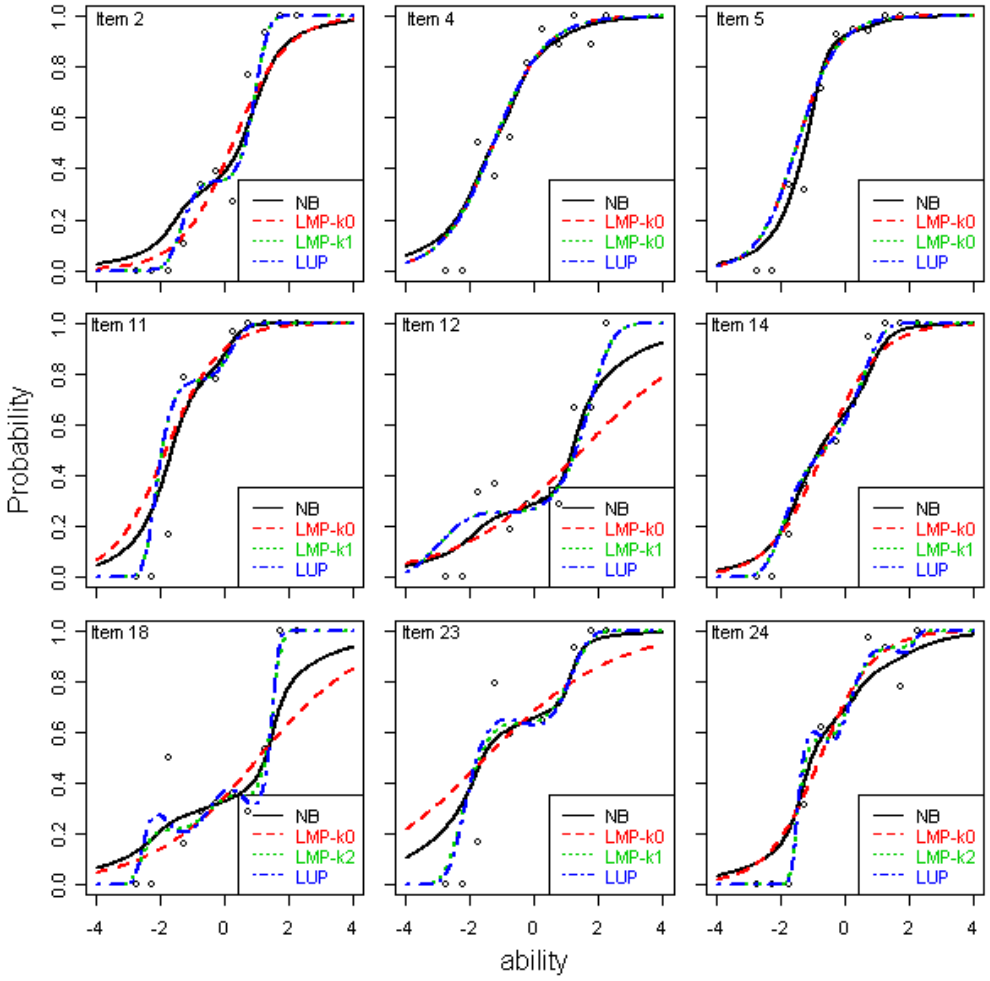

Figure 6.3: Estimated ICCs for some selected items from Nonparametric Bayes and L-MP (elementary statistics exam data).“NB” represents Nonparametric Bayesian model; the red curve LMP-k0 is the L-MP model with standard logistic curve; the green curve is the L-MP model selected by AIC criterion; LUP represents the logistic function with unconstrained polynomial.

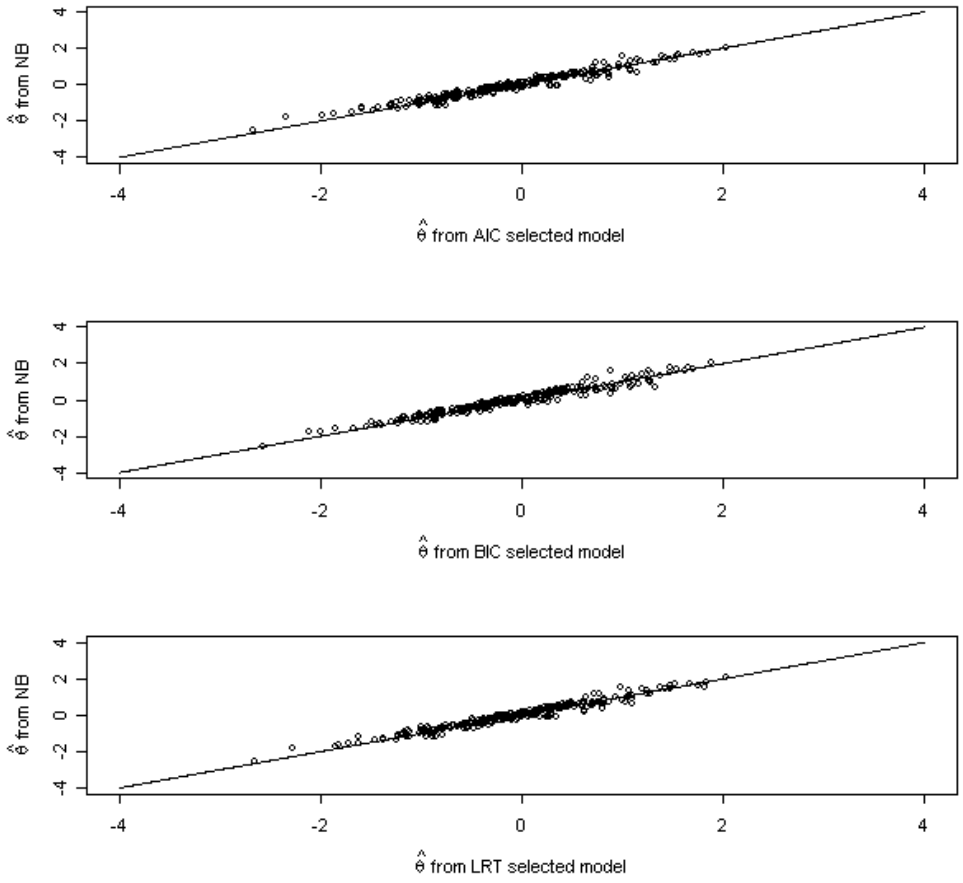

Figure 6.4: Comparisons of $\hat{\theta}$s from Nonparametric Bayesian model and L-MP model (elementary statistics exam data).

CHAPTER 7

DISCUSSION AND FUTURE DIRECTIONS

The 1PL, 2PL and 3PL models have been the dominant IRFs for binary response data since IRT was developed. Although they have been shown to be useful models, they can not handle all data. The L-MP model is a more general model which includes the 1PL or 2PL as special cases. The idea of having a more flexible model with a higher degree polynomial helps to improve the model-data fit when the ordinary 1PL or 2PL model is not adequate.

A surrogate-based two-stage procedure is used to estimate the parameters in the L-MP model. At stage one, it starts with normalized first principal component scores as ability surrogates and then estimates the item curves by minimizing the negative log likelihood function. At stage two, the EAP estimates of abilities are obtained. This procedure can also be viewed as a modified version of the JML estimation method truncated after the first iteration. The estimation procedure is simple and straightforward. The estimates for the L-MP using this estimation technique have been shown to be of reasonably accuracy. When the true model is a 2PL model, the estimates from L-MP are very close to MML, and much better than JML. When the true model is not the standard logistic function, the 2PL model is not adequate for many items. The L-MP function works better to capture those characteristics. The comparisons between L-MP and the other nonparametric procedures like TESTGRAF (Ramsay, 1991) and the irtNP (Duncan & MacEachern, in press) when the true model

is a mixed normal distribution also show that it can provide similar estimates to those nonparametric procedures in terms of the RIMSE for item curves, RIMSE_θ and the rank correlations for abilities.

This surrogate-based two-stage estimation method for the L-MP model tries to estimate the item parameters and the ability parameters jointly, but no iterations between item parameters and abilities are made. The use of the first principal component scores instead of test scores as ranking basis as in TESTGRAF greatly reduces the chance of tied ranks. Although the estimates of item parameters might not converge to exactly the true values since surrogates are only approximations to the person parameters, simulation results show that this procedure yields very good approximations when the sample size increases.

Using the normalized principal component scores as surrogates of the abilities at stage one of the estimation procedure makes it greatly dependent on normally distributed abilities. For the simulation studies in Chapter 5, abilities are all drawn from standard normal distributions truncated at ± 3. Although this is an assumption that is used very frequently in educational testing, it could happen that for some tests, another distribution for the latent trait is preferable. Woods (2005) proposed a spline-based method to estimate the ability distributions.

Only binary response data have been considered here. However, this method can be extended to other settings, for example, for the 3PL model or for ordered response data. For multiple choice items, future work can consider providing option characteristic curve for each incorrect option.

REFERENCES

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In B. N. Petrox and F. Caski (Eds.), Second international symposium on information theory, pp. 267-281.

Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York, Marcel Dekker, Inc.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems*. Series Report 58-16, Project No. 7755-23. USAF school of aviation medicine, Randolph air force base, Texas.

Birnbaum, A. (1958a). *On the estimation of mental ability*. Series Report No. 15. Project No. 7755-23, USAF school of aviation medicine, Randolph air force base, Texas.

Birnbaum, A. (1958b). *Further consideration of efficiency in tests of a mental ability*. Technical Report No. 17. Project No. 7755-23, USAF school of aviation medicine, Randolph air force base, Texas.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, and M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley. pp.399-402

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D. and Aiktin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Browne, M. W. (1997). *Notes on monotonic polynomials*. Unpublished manuscript, Department of Psychology, The Ohio State University.

Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block design. *Biometrika*, *83*: 275-285.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., &Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*, 285-299.

Duncan, K. A. and MacEachern, S. N. (in press). Nonparametric Bayesian modeling for item response. *Statistical modeling: an international journal*.

Elphinstone, C. D. (1985). *A method of distribution and density estimation*. Unpublished dissertation, University of South Africa.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, *7*, 19-29.

Finney, D. J. (1944). The application of probit analysis to the results on mental tests. *Psychometrika*, *19*, 31-39.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error*. (Technical Report No. 15). Stanford, CA: Standford University, Applied Mathematics and Statistics Laboratory.

Heinzmann, D. (2005). *Computational aspects of filtered polynomial density estimation*. Unpublished Master Thesis, The Ohio State University and Swiss Federal Institute of Technology.

Jennrich, R. I. and Sampson, P. F. (1968). Application of stepwise regression to non-linear estimation. *Technometrics*, *10*, 63-72.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84-5). Champaign, IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V. (1985). The trait in latent trait theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory/ computerized adaptive testing conference* (pp. 41-65). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, *No. 7*.

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, *18*, 57-75.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517-548.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N. J.: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157-162.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement. 17*, 351-363.

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16(1)*, 1-32.

Patz, R. and Junker, B. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response theory. *Journal of Educational and Behavioral Statistics*. *24*, 146-178.

Qin, L. (1998) *Nonparametric Bayesian models for item response data*. Doctoral Dissertation, The Ohio State University.

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

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*, TESTGRAF program manual.

Ramsay, J. O. & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, *56*: 365-379.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, *1*, 33-49.

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

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, *42*, 161-191.

Samejima, F. (1979). *A new family of models for the multiple choice item*. (Research Report No. 79-4). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1984). *A plausibility functions of Iowa Vocabulary Test items estimated by the simple sum procedure of the conditional P.D.F. approach*. (Research Report No. 84-1). Knoxville: University of Tennessee, Department of Psychology.

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

Sinnott, L. T. (1997). *Filtered polynomial density approximations and their application to discriminant analysis*. Thesis, The Ohio State University.

Swaminathan, H. and Gifford, J. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*. *7*, 175-191.

Swaminathan, H. and Gifford, J. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349-364.

Swaminathan, H. and Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589-601.

SYSTAT 10.2 Statistics II, 2002, SYSTAT Software Inc.

Thissen, D., Chen, W. and Bock, D. (2003) *Multilog for windows*. Version: 7.0.2327.3. Scientific Software International, Inc.

Thissen, D. and Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Thissen, D. and Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*, 397-412.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*, 1-13.

Woods, C. M. (2005). Item Response Theory with Estimation of The Latent Population Distribution Using Spline-Based Densities. *Psychometrika*, *71*, 281-301.