BAYESIAN ANALYSIS OF HIERARCHICAL IRT MODELS:
COMPARING AND COMBINING THE UNIDIMENSIONAL &
MULTI-UNIDIMENSIONAL IRT MODELS

A Dissertation
presented to
the Faculty of the Graduate School
University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree

Doctor of Philosophy

by
YANYAN SHENG

Dr. Steven Osterlind, Dissertation Supervisor

JULY 2005
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled

**BAYESIAN ANALYSIS OF HIERARCHICAL IRT MODELS: COMPARING AND COMBINING THE UNIDIMENSIONAL & MULTIDIMENSIONAL IRT MODELS**

Presented by Yanyan Sheng

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

[Signatures]

1. [Signature]
2. [Signature]
3. [Signature]
4. [Signature]
To Gang, Zhaohui & Angela
ACKNOWLEDGEMENTS

There are many individuals whom I would like to thank for making it possible for me to complete this dissertation. A special note of appreciation goes to Dr. Steven Osterlind and Dr. Chris Wikle. It was an honor working with them both. Dr. Osterlind provided a considerable amount of mentoring throughout my entire doctoral program. He was a source of encouragement, guidance and patience during this process. Dr. Wikle provided valuable advice with the Bayesian analyses and with understanding of the results in the study. He helped with minute technical details as well as the overall model setting up, giving unsparingly of his time and expertise.

Special thanks also go to other members of my doctoral committee, Dr. Jessica Summers, Dr. Alex Wagaindt and Dr. Richard Cox. Their continuing support, together with their time and assistance is very much appreciated.

Appreciation is extended to Dr. Robert Tsutakawa who has provided support and help during my study. His understanding of IRT and sincere advice made Bayesian IRT a not intimidating area to work on.

Finally, I would like to thank my parents, Yongji Sheng and Ming Wang, for helping to make all of my achievements possible and for always believing in me.
# Table of Contents

Acknowledgements ................................................................................................................... ii

Abstract ..................................................................................................................................... v

List of Tables ............................................................................................................................... vi

List of Illustrations .................................................................................................................... vii

Chapter 1. Introduction ............................................................................................... - 1 -

1.1 Statement of the Problem .................................................................................................. - 2 -

1.2 Purpose of the Study .......................................................................................................... - 7 -

1.3 Definition of Terms ........................................................................................................... - 9 -

1.4 Research Questions ........................................................................................................... - 11 -

1.5 Significance of the Study ................................................................................................... - 11 -

1.6 Delimitation of the Study ................................................................................................... - 13 -

1.7 Overview of the Subsequent Chapters ............................................................................ - 13 -

Chapter 2. Review of the Literature ........................................................................ - 15 -

2.1 Unidimensional IRT Models .......................................................................................... - 20 -

2.1.1 Rasch model .............................................................................................................. - 21 -

2.1.2 Two parameter logistic model ................................................................................. - 23 -

2.1.3 Three parameter logistic model .............................................................................. - 23 -

2.1.4 Parameter estimation for the unidimensional models .............................................. - 24 -

2.1.5 Local independence and unidimensionality .............................................................. - 33 -

2.2 Multidimensional IRT Models ..................................................................................... - 35 -

2.2.1 Multi-unidimensional IRT models—a special case of MIRT ................................. - 38 -

2.2.2 Estimation for the multidimensional models .......................................................... - 39 -

2.3 Factor Analysis & MIRT ............................................................................................... - 40 -

2.3.1 Early development in FA ....................................................................................... - 41 -

2.3.2 Comparison of FA and IRT .................................................................................... - 47 -

2.3.3 Bifactor analysis ...................................................................................................... - 50 -

Chapter 3. Methodology ............................................................................................. - 52 -

3.1 Overview of Research Questions ...................................................................................... - 52 -

3.2 Bayesian Unidimensional & Multi-unidimensional IRT Models ............................... - 53 -

3.2.1 Hierarchical U2PNO model .................................................................................. - 53 -

3.2.2 Hierarchical multi-unidimensional 2PNO model .................................................... - 54 -

3.3 Proposed Bayesian IRT models ...................................................................................... - 56 -

3.3.1 Additive model ........................................................................................................ - 56 -
ABSTRACT

As item response theory models gain increased popularity in large scale educational and measurement testing situations, many studies have been conducted on the development and applications of unidimensional and multidimensional models. However, to date, no study has yet looked at models in the IRT framework with an overall ability dimension underlying all test items and several ability dimensions specific for each subtest. This study is to propose such a model and compare it with the conventional IRT models using Bayesian methodology. The results suggest: 1) models with general and specific abilities can be developed, 2) fully Bayesian method is proved to be more accurate and efficient in parameter estimation compared with the usual marginal maximum likelihood method, 3) compared with the conventional IRT models, the proposed model describes the actual data conceivably better. Therefore, the proposed model offers a better way to represent the test situations not realized in existing models. The model specifications for the proposed model also give rise to implications for test developers on test designing. In addition, the proposed IRT model can be applied in other areas, such as intelligence or psychology, among others.
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sum of squared difference between the actual and recovered item parameters ($\alpha_0$, $\alpha_1$, and $\gamma$) using different additive models or TESTFACT under five simulated scenarios.</td>
<td>77</td>
</tr>
<tr>
<td>2. Correlations between the actual person ability parameters ($\theta_0$, $\theta_1$, $\theta_2$) and posterior estimates for the ability parameters ($\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$) using 4 different additive models or TESTFACT under five simulated scenarios.</td>
<td>80</td>
</tr>
<tr>
<td>3. Sum of squared difference between the actual and recovered item parameters ($\alpha$ and $\gamma$) for Models 1 and 2 under seven simulated scenarios.</td>
<td>85</td>
</tr>
<tr>
<td>4. Correlations between the actual person ability parameters ($\theta_0$, $\theta_1$, $\theta_2$) and posterior estimates for the ability parameters ($\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$) for Models 1 and 2 under seven simulated scenarios.</td>
<td>86</td>
</tr>
<tr>
<td>5. Posterior estimates and MC uncertainty for the item parameters with the three proposed models.</td>
<td>89</td>
</tr>
<tr>
<td>6. Approximated marginal densities of the data and Bayesian deviances for the five IRT models.</td>
<td>91</td>
</tr>
<tr>
<td>7. Sum of squared difference between the actual and recovered item parameters using for 3 additive models and 2 hierarchical models under four simulated scenarios.</td>
<td>100</td>
</tr>
<tr>
<td>8. Correlations between the actual person ability parameters ($\theta_0$, $\theta_1$, $\theta_2$) and posterior ability estimates ($\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$) for 3 additive models and 2 hierarchical models in four simulated scenarios.</td>
<td>101</td>
</tr>
</tbody>
</table>
# List of Illustrations

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Plots of the correlations between the actual and recovered item parameters ($\alpha_0$, $\alpha$, and $\gamma$) using 4 different additive models or TESTFACT for the five simulation studies</td>
<td>75</td>
</tr>
<tr>
<td>2.</td>
<td>Plot of the correlations between the actual and recovered item parameters ($\alpha$ and $\gamma$) using the 2 hierarchical models for the seven simulation studies</td>
<td>84</td>
</tr>
<tr>
<td>3.</td>
<td>Plots of extreme tail-area PPP-values for odds ratios with the five IRT models</td>
<td>94</td>
</tr>
<tr>
<td>4.</td>
<td>Graphical illustrations of the two proposed hierarchical IRT models (a) Model 1 (b) Model 2. Circles represent latent traits, and squares represent observed items</td>
<td>98</td>
</tr>
<tr>
<td>5.</td>
<td>Four classes of IRT models: (a) the unidimensional IRT model, (b) the multi-unidimensional IRT model, (c) the proposed additive IRT model, (d) the proposed hierarchical IRT model. Circles represent latent traits, and squares represent observed items</td>
<td>103</td>
</tr>
</tbody>
</table>
CHAPTER 1.

INTRODUCTION

As item response theory (IRT) models gain increased popularity in large scale educational and measurement testing situations, many studies have been conducted that explore their applications in areas such as ability estimation, equating, differential item function (DIF), computerized adaptive testing (CAT), and elsewhere (Lord, 1980; Petersen, Kolen et al., 1989; Kolen & Brennan, 1995). One of the major considerations in measurement theory is to ensure meaningful inferences made from test scores, which requires fit of the empirical test data obtained from test indicators to a theoretical framework. This forms the basis for legitimizing any application of the modern IRT models. However, it is well accepted among measurement theorists that none of the theoretical models fully represents a complex reality (van der Linden & Hambleton, 1997). Instead, the models are only a simplified approximation of the real world.

Therefore, when employing IRT in a testing situation one has to choose a model which provides the most complete description of the data. In this way the test scores are valid interpretations of reality. It has to be pointed out that this point of view, namely, choosing the best fitted model for the test data, is to be differentiated from the Rasch approach to IRT, where test items are believed to be prepared by criteria so they fit into the Rasch model, a special type of IRT model.

Selecting an appropriate IRT model is complex. First, the correct number of latent ability dimensions must be identified a priori. Usually, this is assumed to be just one. Understandably, unidimensional models require that the test is essentially unidimensional in its underlying structure. Any violation of this assumption would
result in inadequacy of the model in describing the data and hence unreliable estimation of the examinee’s ability. Therefore, the correct specification of the number of the latent dimensions is directly tied to the construct validity of a test.

In classical test theory (CTT), factor analysis (FA; Spearman, 1904; Thurstone, 1938) is considered as a conventional method for evaluating dimensionality and much work has been done on extending the method to the test settings where binary data are the major concern (e.g., McDonald, 1981; Muthén, 1978; Bock & Aitkin, 1981). Nevertheless, in spite of the numerous efforts on assessing the dimensionality, especially the unidimensionality assumption, FA does not seem to be a satisfactory solution. As a matter of fact, no known methods provide clear criteria for determining it (Hattie, 1985).

Researches have shown that FA and IRT are similar to certain extent (e.g., Reckase, 1997; McDonald, 1985) and their parameters can be easily converted from one to the other. However, FA is a data reduction method with a focus on the number of factors whereas IRT is a model-based theory modeling the interaction between the examinee’s response and individual item (Reise et al., 1993). Thus, IRT models are more appropriate when the examinee’s ability level has to be estimated once the dimensionality is decided. In the field of factor analysis or latent structure analysis, the multiple-factor model is considered as an extension of the one-factor model emerged in the beginning of the 20th century. Another extension of the one-factor model is the bifactor model (Holzinger & Swineford, 1937) where each test item loads on a general factor and a group or specific factor. The idea of including both general factors and specific factors in one model can be extended to the framework of item response theory.

1.1 Statement of the Problem
Much research has been conducted on the development and application of both unidimensional IRT models where one ability dimension is assumed (e.g., Bock & Aitkin, 1981; Mislevy, 1985; Tsutakawa & Lin, 1986; Patz & Junker, 1999), and in multidimensional IRT models where multiple ability dimensions are involved in one test (e.g., Hoijtink & Molenaar, 1997; Béguin & Glas, 2001; Lee, 1995). However, to date, no study has yet looked at models in the IRT framework with an overall ability dimension underlying all test items and several ability dimensions specific for each subtest. This circumstance is comparable to a bifactor model with a general factor and specific/group factors. Before the problem is further detailed, it is essential to briefly overview the IRT models and how they are related to models in the factor-analytic framework.

Item response theory was introduced in the 1950s as an alternative to classical test theory (CTT; Lord & Novick, 1968). It originated from psychophysics and latent structure analysis in sociology instead of test theory. IRT models are stochastic models for responses to individual test items by individual examinee. The probabilities of these responses are defined as a function of separate parameters for the item and the person to represent item characteristics and the examinee’s ability or attribute level. When the probability of a response is represented as a function of the person parameter assuming the item parameters known, the function is known as an item response function (IRF) in IRT, or sometimes called item characteristic curve (ICC; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). IRT response models differ from response models in psychophysics in that all parameters are latent; they differ from factor analysis or latent structure analysis in that the person parameters do not represent a finite set of latent classes but are real valued.

Early work in IRT dealt with binary responses, assuming that all test items are measuring one ability in common, but extensions to other item formats or response
processes have followed soon. One of these extensions is on models for items that measure multiple abilities or require response processes with different cognitive components (Davey, Oshima, & Lee, 1996; Samejima, 1974). In general, those modeling a single ability dimension are referred to as the unidimensional IRT (UIRT) models whereas those for multiple abilities are the multidimensional IRT (MIRT) models, which are more complex and hence less restrictive than the unidimensional models. In addition, MIRT can be considered as a special case of factor analysis (Reckase, 1997). The MIRT models in the IRT literature postulate that each item measures more than one underlying ability or the response to each item involves multiple cognitive processes. In the situations where an overall test consists of several subsets, each measuring slightly different trait, a more specific model (such as what is described in Lee, 1995) has to be emphasized. Since each subtest is unidimensional (i.e., the items in each subtest are designed to measure one thing in common), this special multidimensional type can be referred to as the multi-unidimensional IRT models.

An analogy between factor analysis and IRT can be drawn to illustrate the differences between the above three classes of IRT models, namely, the UIRT models, the MIRT models and the multi-unidimensional IRT models. Generally, as one can see, the UIRT models are appropriate in the situation when only one factor is extracted from the test items whereas MIRT models have to be adopted when more than one factor are found to be significant. In an exploratory factor analysis, the factor solution is usually visually or analytically rotated. Often the rotation scheme is devised to approximate simple structure (McDonald, 1985) so that the factor loadings are split into two groups, the elements of one tending to zero and those of the other tending toward unity. Hence, each item has a unity loading on one factor and 0 loadings on other factors. To put it in other words, the test involves multiple abilities and each test item measures only one of them.
This is probably more common in large scale testing situation than the more general case of MIRT where each item measures more than one ability as described earlier. The models specific for this type of situations are multi-unidimensional IRT models, which can be viewed as an extension of the UIRT models or a special class of the MIRT models.

In the testing setting, more often a test consists of several subtests with each focusing on one specific ability so that the items in a particular subtest are designed to measure one ability in common. One can fit a unidimensional model or a multidimensional model based on different assumptions. However, it is natural to hypothesize that the model with both general and specific abilities would be more efficient and describe the test data more adequately. Furthermore, it would be interesting to investigate how the model compares with the bifactor model, with which it shares similar ideas.

The UIRT models make simple but strong assumptions about the relationships between item responses and the latent ability. One of them is unidimensionality, that is, each test item is designed to measure some facet of the same underlying ability or so called unified latent trait. This is closely related to the concept of local independence, for it is said that a data set is unidimensional when item responses are not correlated based on a single latent trait (McDonald, 1981). It is necessary that a test intending to measure one certain ability should not be affected by other ability dimensions, especially when only the overall test scores are reported and used for assessment. Although it is well accepted that the real-world test data will never be strictly unidimensional, it is important to ensure “essential” unidimensionality so that the resulting parameter estimates (ability estimates) derived from application of the UIRT model will be reliable and consistent.

This is directly related to the confidence one may have in the construct validity for useful, meaningful and appropriate interpretations of the test’s scores.
Conceptually, we know that the UIRT models are appropriate when there is one dominant factor and the MIRT models should be adopted when examinee ability estimates are affected by the presence of smaller specific factors. However, in many applications, due to the lack of existence of reliable statistical tests for unidimensionality (see Hattie, 1985 for an overview), it is not easy to choose one model over the other. For instance, suppose an English test consists of three subtests, listening, reading and writing. One can assume all items are measuring a unified English ability dimension and fit a unidimensional model. However, on the other hand, the test items are designed to measure specifically listening, reading or writing, then a multi-unidimensional model might seem to be more appropriate. The key difference between the two models lies in whether a single composite score is sufficient or if sub-scores have to be reported for illustrating the examinee’s proficiency on answering all test items. In order to produce reliable and valid scores, one has to decide which model to adopt, for “(two sorts of items) unless highly correlated, the meanings of scores based on (their) such a composite are questionable” (McNemar, 1946 p.298).

The lack of a well-developed statistical test for assessing the latent dimensions or specifically the unidimensionality assumption creates problems as to which IRT model (e.g., the unidimensional or the multi-unidimensional model) is to adopt when carrying out ability estimation. It is said that the multi-unidimensional IRT model is an extension of the unidimensional model. Similarly, the model with both general and specific abilities is another extension to it, only in a slightly different direction. Thus, its comparison with the unidimensional model can provide information on whether the model assumption, i.e., unidimensionality, is satisfied.

Meanwhile, the critical difference between the unidimensional model and the multi-unidimensional model makes model comparison necessary as to which model one
can adopt to produce reliable test scores. In certain circumstances, one may want to report both composite score and sub-scores. To this end, one may 1) implement a multi-unidimensional model to obtain the sub-scores and then average the estimated sub-scores to get the composite score, or 2) implement both the unidimensional and the multi-unidimensional models. However, averaging may give rise to biased composite score and two separate implementations overlooks the relationships between the latent abilities and could be time consuming. It is then reasonable to have the overall ability as well as specific ability dimensions in the model, so that both composite score and sub-scores can be obtained with one single implementation.

1.2 Purpose of the Study

The purpose of the study is to propose new IRT models under the Bayesian framework so that both general ability and specific ability dimensions can be estimated with one implementation. A Markov chain Monte Carlo (MCMC) algorithm known as Gibbs sampling is adopted to implement the model. MCMC is powerful for complicated models where the probabilities or expectations are intractable by analytical methods or other numerical approaches. Its methods have been influential in modern Bayesian analyses where they are used to summarize the posterior distributions that arise in the context of the Bayesian prior-posterior framework (e.g., Tanner & Wong, 1987; Gelfand & Smith, 1990; Chib & Greenberg, 1995; Carlin & Louis, 2000; Gelman et al., 2004). MCMC methods have proved useful in practically all aspects of Bayesian inference, such as parameter estimation and model comparisons. A key reason for the widespread interest in the MCMC method is that they are extremely general and flexible and hence can be used to sample univariate and multivariate distributions when other methods (for example, the classical maximum likelihood methods) either fail or are difficult to implement. In addition, with MCMC, it is straightforward to construct one more Markov
chains whose limiting invariant distribution is the desired target distribution (Gelman et al., 2004). One of the simplest Markov chain Monte Carlo algorithms is Gibbs sampling (Casella & George, 1992). The method is straightforward to implement when each full conditional distribution is a known distribution that is easy to sample. To illustrate the Gibbs sampling procedure for the proposed model, a subset of College Basic Academic Subjects Examination (CBASE)-English subject data is used.

CBASE is a criterion-referenced achievement examination adopted by over 140 colleges and universities across the country to evaluate knowledge and skills in four subject areas: English, math, science and social studies of sophomore-level students (usually after they complete the core curriculum). It has 180 total academic items, with 41 for English, 56 for mathematics, 41 for science and 42 for social studies. During the administration of the test, nine different forms were used, coded as forms LF, LG, LH, till LO. These forms were the same in their major constructs and item number in the four academic areas. In addition, each subject area is further organized into levels of increasing specificity by two to three clusters1, e.g., English test consists of two clusters, writing and reading/literature.

Conceptually, the proposed IRT model shares similarities with the bifactor model (Holzinger & Swineford, 1937) in the factor analytic framework. Since bifactor analysis is implemented in TESTFACT program, the parameter estimates from the proposed MCMC procedures can be further compared with the estimates from TESTFACT to illustrate the efficiency of the procedures. Moreover, the proposed model is more complicated than the UIRT model or the multi-unidimensional IRT model and thus is supposed to provide better fit to the actual testing data. Model comparisons are thus carried out using Bayes factors, Bayesian deviance (Spiegelhalter et al., 1998) and a

---

1 It has to be noted that clusters are used to refer to subtests in the four subject areas in CBASE.
Bayesian predictive approach, i.e., posterior predictive model checks (Sinharay & Stern, 2003). Since the adequacy of a model is directly related to model assumptions, when comparing a unidimensional model with the proposed model or a multi-unidimensional model, one is automatically testing whether the unidimensionality assumption holds for actual test data. The Bayesian model comparison techniques provide an alternative method of checking model assumptions.

1.3 Definition of Terms

For the purpose of the study, some of the major terms are defined as follows:

*Item response theory (IRT)* - It is a modern test theory that describes the interaction between item characteristics and person abilities. Because the ability is not manifested directly, it is also sometimes referred to as latent trait theory.

*Unidimensional IRT* - An examinee’s response to a specific test item is determined by a latent mental trait of the examinee. The underlying ability is assumed to vary continuously along a single dimension so that the examinees can be arrayed on this dimension from lowest to highest.

*Multidimensional IRT* - This IRT is more general than the unidimensional IRT. It deals with the situation of complexity in psychological measurement when multiple latent traits affect the examinee’s performance on a given item.

*Multi-unidimensional IRT* - It is a special case of multidimensional IRT and it applies when an overall test consists of several subtests, each measuring a different latent trait. The test is multidimensional whereas each subtest is unidimensional.
Marginal maximum likelihood – It is one of the popular techniques for estimating parameters. Marginal maximum likelihood treats the person parameters as nuisance parameters, assumes they are random effects sampled from some larger continuous distribution, and removes them from model to estimate item parameters.

Bayesian Inference - It is a branch of mathematical probability theory that allows one to model uncertainty about the world and outcomes of interest by combining common-sense (prior) knowledge and observational evidence (likelihood).

Empirical Bayes – In Bayesian inference, the posterior distribution requires both the likelihood and prior specifications of the parameters. When these priors are estimated from the data, the procedure is referred to as empirical Bayes.

Markov chain Monte Carlo – Markov chain Monte Carlo simulation techniques provide a fully Bayesian methodology that allows one to estimate item parameters and person abilities at the same time while incorporating uncertainty of the item estimates in calculations of uncertainty about abilities for persons.

Gibbs sampling – It is one of the simplest Markov chain Monte Carlo algorithms. Gibbs sampling is applicable when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. The algorithm is to generate a sample from the distribution of each variable in turn, conditional on the current values of the other variables.

College Basic Academic Subjects Examination (CBASE) – It’s a criterion-referenced examination assessing sophomore-level students’ knowledge and skills in English, mathematics, science and social studies.
1.4 Research Questions

The overall research question is whether or not a new IRT-based model for undimensionality can be developed. Under this general topic, the specific research questions to be answered in this study are:

(1) How does the proposed IRT model, which incorporates a general ability as well as several specific ability dimensions, perform when implementing it to various simulated situations as well as to the CBASE English data.

(2) How does the proposed model compare with the UIRT or the multi-unidimensional IRT model as far as the CBASE English data are concerned.

(3) How does the proposed MCMC procedure compare with the bifactor analysis implemented in TESTFACT.

1.5 Significance of the Study

The significance of this study lies in the fact that a new IRT model can offer several advantages to measurement validation not now realized in existing models. Specifically, the advantages of the current approach are three-fold. First, the proposed model incorporates one general ability for the overall test and several specific abilities for the subtests. Thus, both levels of the ability dimensions can be estimated with one implementation. The model can be regarded as the combination of the UIRT and the multi-unidimensional IRT models. This way, one does not have to rely on a well-developed index for checking the unidimensionality assumption to choose between the two models. In addition, neither does one have to wonder if a single composite score is sufficient or if sub-scores are to be reported.
Secondly, estimating parameters in the fully Bayesian framework allows incorporating the dependencies among variables and sources of uncertainty. One prerequisite for the application of IRT models to testing problems is efficient statistical procedures for parameter estimation and testing the goodness of fit of the model for actual data sets (van der Linden, 1999). For years, the standard methodology for parameter estimation has been focusing on first estimating item parameters using the EM algorithm (Bock & Aitkin, 1981), treating person parameters as missing data, and then using the estimated item parameters and taking them as known when making inference regarding the latent abilities. Much research has been conducted using this empirical Bayes procedure (e.g., Mislevy, 1985; Tsutakawa & Lin, 1986) and the associated marginal maximum likelihood (MML) is considered to be more accurate, hence less biased, than other ML estimating methods (Lord, 1986). Nonetheless, when the model gets more complex, integrating out ability parameters is not straightforward. Moreover, and most importantly, ML or EM fail to take into account the uncertainty about item parameters when making inference on examinees and may seriously underestimate the uncertainty in abilities (Tsutakawa & Soltys, 1988; Tsutakawa & Johnson, 1990). With current enhanced computational technology and the emergence of Markov chain Monte Carlo (MCMC) simulation techniques (e.g., Chib & Greenberg, 1995), the methodology has rapidly moved from an empirical Bayes to a fully Bayesian approach. The fully Bayesian approach allows one to estimate item parameters and examinee abilities at the same time while incorporating uncertainty of the item estimates in calculations of uncertainty about abilities for examinees (e.g., Albert, 1992; Patz & Junker, 1999; Béguin & Glas, 2001). Thus, the proposed MCMC procedure is more efficient in accurately estimating the parameters.

Lastly, the Bayesian model comparisons, especially those between the UIRT and the multidimensional models are comparable to the $\chi^2$ test in factor analysis. Hence, the
adequacy of the unidimensional model would suggest satisfaction of the model assumption. Moreover, the comparisons are confirmatory in nature, which has been noticeably lacking in the IRT literature (Segall, 2002).

1.6 Delimitation of the Study

The delimitations of the current study are as follows:

- In the study, we only consider the normal ogive IRT models and assume items are characterized by two parameters, i.e., item discrimination and item difficulty.
- When implementing the IRT models, we limit our focus on form LP of the CBASE data.
- Since CBASE consists of four subject matters, only English test is used in the study. It is then anticipated that other tests might perform differently.
- With the CBASE English data, a convenience sample is used instead of being randomly selected. Therefore, the findings from the study may not generalize to other samples.

1.7 Overview of the Subsequent Chapters

The subsequent chapters are organized as follows. Chapter 2 reviews the related literature on the IRT models, estimation procedures and the relationship between factor analysis and item response theory. Chapter 3 describes the hierarchical parameterizations for the unidimensional and multi-unidimensional models and presents proposed models incorporating one general ability and several specific abilities. This chapter also describes the data used in the analysis and the specific procedures adopted. Chapter 4 presents the results as well as some discussions with the simulation
studies and real data. And finally, the summary of findings, implication of the study and direction for future research are concluded in Chapter 5.
CHAPTER 2.

REVIEW OF THE LITERATURE

The review of literature starts with the early development in item response theory. Three main sections are included in this chapter. The first section focuses on the unidimensional IRT model, including the estimation procedure and model assumptions. The second section describes the multidimensional IRT model, with the multi-unidimensional model as the special case. As mentioned in the previous chapter, factor analysis is similar to IRT in certain ways. Thus, Section 3 reviews factor analysis and how it is compared with IRT. Studies of bi-factor analysis are also reviewed in the later part of this section.

Modern item response theory can be traced back to Binet-Simon intelligence scale (Binet & Simon, 1908) and Binet’s global conceptualization of intelligence (Binet & Simon, 1905), which first appeared at the beginning of the 20th century. An important feature of Binet’s conception is that both items and examinees are placed on a common scale. That is, items and examinees are referenced to mental age. More specifically, cognitive tasks were scaled for mental age from the empirical performance of children at various ages. In addition, examinees were scaled for mental age by their relative success in solving the age-referenced tasks. Due to the age-calibrations of the tasks, examinees may be compared even if they do not receive the same items. This concept also appears in IRT and is one of its advantages over the classical test theory (CTT). Moreover, like the Binet scale, IRT does not use the number-correct as a statistic for estimating a person’s position on the latent continuum. Although there are differences between the Binet scale and IRT,
it is fair to say that modern IRT owes more to the concept of Binet’s clinical assessment of intelligence in children than to CTT pioneered by Spearman.

In 1925, Thurstone took a step further. He actually offered a solution to the problem of how best to place the Binet-Simon items on an age-referenced scale by plotting the proportion of children in succeeding age cross-sections on successive Binet tasks. Persons and items were placed on a common scale by using the normal distribution to scale item-solving probabilities. According to Thurstone (1925, p. 436), “Each test question is located at a point on the scale so chosen that the percentage of the distribution to the right of that point is equal to the percentage of right answers to the test question for children.” The population of examinees, of course, could be designated by z-scores. Thurstone (1925, p. 449) illustrated a graph of the resulting scaling of items on mental ability. These transformations resulted in a common scaling of persons and items that is similar to that given by the normal ogive IRT model that was developed decades later. Thurstone’s solution bears certain basic features in common with IRT. Both posit a response model where the probability of success on a particular item is a function of a continuous variable measuring an attribute of the examinee (i.e., the person’s ability or proficiency, as described in much of the IRT literature). Meanwhile, both share the same objective, namely, to represent the item locations and the examinee attributes on the scale of quantitative variable.

The early development of IRT models in educational and psychological measurement was, in a large sense, under the influence of other fields. As early as 1860, to represent qualitative response probabilities in psychophysics, Gustav Fechner used the integrated normal curve,
\[ P(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-t^2 / 2)dt, \]  

where \( y \) is a normal deviate. When \( y = a(x-b) \), the function is referred to as the normal ogive IRT model, where \( a \) is the item slope or discrimination power and \( b \) is the item location or difficulty level. Later, in 1952, to solve the problem in toxicology of natural mortality of the controls, David Finney advanced the methods of probit analysis to correct for the effect of natural causes so that the observed proportion of death among the treated insects is assumed to be the sum of the corresponding proportion \( C \) in the controls plus the expected proportion \( P(x) \) of treated insects dying at dosage level \( x \):

\[ P^* = C[1 - P(x)] + P(x) = C + (1 - C)P(x). \]  

This problem shares much in common with the guessing correctly on multiple-choice items. Hence, equation 2.2 was adopted in IRT to account for the guessing effects on item analysis and test scoring (the model later came to be known as the three-parameter IRT model). A further important development that shaped the IRT literature was the introduction of the logistic item response function as an alternative to the normal ogive model in bioassay applications by Fisher and Yates in 1938. The logit form offers two computational advantages over the normal ogive form in that fitting the logistic response model simplifies the solution to the maximum likelihood equations and that the iteration procedure converges faster with the logistic model than the normal ogive response model.

IRT is regarded as having two distinct origins, Georg Rasch (1960) and Frederic Lord (Lord, 1953; Lord & Novick, 1968), who pioneered the formation and development of the IRT, or more specifically, UIRT models.
An important theoretical and practical contribution in the history of IRT can be summarized in the following timetable (Summary based on Hambleton & Swaminathan, 1985, and Bock, 1997):

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1905</td>
<td>Binet-Simon intelligence scale</td>
</tr>
<tr>
<td>1925</td>
<td>Thurstone was the first to plot performance levels against an independent variable and use the plots in the test development.</td>
</tr>
<tr>
<td>1936</td>
<td>Richardson derived relationship between IRT model parameters and classical item parameter, which provided an initial way for obtaining IRT parameter estimates.</td>
</tr>
<tr>
<td>1943-44</td>
<td>Lawley produced some new procedure for parameter estimation.</td>
</tr>
<tr>
<td>1952</td>
<td>Lord described the two-parameter normal ogive model, derived model parameter estimates, and considered applications of the model.</td>
</tr>
<tr>
<td>1957-58</td>
<td>Birnbaum substituted the more tractable logistic models for the normal ogive models, and developed the statistical foundation for these new model.</td>
</tr>
<tr>
<td>1960</td>
<td>Rasch developed three item response models and described them in his book, <em>Probabilistic Model for Some Intelligence and Attainment Tests</em>. His work influenced Wright in the United States and psychologists such as Andersen and Fischer in Europe.</td>
</tr>
<tr>
<td>1967</td>
<td>Wright was the leader and catalyst for most of the Rasch model research in United States through the 1970s. His presentation at the ETS Invitational Conference on Testing Problems served as a major stimulus for work in IRT, especially with the Rasch model. Later, his highly successful AERA Rasch model Training programs contributed substantially to the understanding of the Rasch model by many researchers.</td>
</tr>
</tbody>
</table>
| 1967    | Lord and Novick provided five chapters on the theory of latent traits (four
of the chapters were prepared by Birnbaum). The authors’ endorsement of IRT stimulated a considerable amount of research.

1967 Wright and Panchapakesan described parameter estimation methods for the Rasch model and the computer program BICAL, which utilized the procedures described in the paper. BICAL was of immense importance because it facilitated applications of the Rasch model.

1972 Bock contributed several important new ideas about parameter estimation.

1974 Lord described his new parameter estimation methods, which were utilized in a computer program called LOGIST.

1975 Fischer described his extensive research program with linear logistic models.

1976 Lord et al. made available LOGIST, a computer program for carrying out parameter estimation with logistic test models. LOGIST was one of the two most commonly used programs (the other is BICAL).

1977 Baker provided a comprehensive review of parameter estimation methods.

1977 Researchers such as Bashaw, Lord, Marco, Rentz, Urry, and Wright in the Journal of Educational Measurement special issue of IRT applications described many important measurement breakthroughs.

1979 Wright and Stone in Best Test Design described the theory underlying the Rasch model, and many promising applications.

1979 Lord in Applications of Item Response Theory to Practical Testing Problems provided an up-to-date review of theoretical developments and applications of the three-parameter model.

1982 Lord and his staff at ETS made available the second edition of LOGIST. This updated computer program was faster, somewhat easier to set up, and had more additional worthwhile output than the 1976 edition of the program.
2.1 Unidimensional IRT Models

As is known, IRT models the interaction between persons and individual test items. Lord (1980) expressed the goal of IRT as follows:

We need to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinee to any item, even if similar examinees have never taken similar test before. (p.11)

Under IRT, or specifically UIRT, an examinee's response to a specific test item or question is determined by an unobserved or latent mental trait of the examinee. Each of these underlying traits, or what is usually called abilities, is assumed to vary continuously along a single dimension usually denoted $\theta$ so that the examinees responding to the test items can be arrayed on $\theta$ from lowest to highest. The position of person $i$ on the continuum $\theta$, denoted $\theta_i$, is usually referred to as the person's ability or proficiency. Intuitively, we expect the probability of a correct response to each item to increase monotonically as $\theta_i$ increases. In terms of binary scored test items, i.e., items on which responses are designated either correct or incorrect, the UIRT models express the probability of a correct response to a test item as a function of $\theta$ given one or more parameters of the item.

For dichotomously (0/1) scored items, the IRT models generally have two main variants. One is Gaussian or so-called "normal ogive" models and the other consists of logistic models. Suppose a $k$-item (multiple choice item) test is measuring a single unified ability $\theta$, which means that the abilities for answering $k$ items are highly correlated, then the probability of person $i$ obtaining correct response for item $j$ is defined as follows:

$$P^* = C[1 - P(x)] + P(x) = C + (1 - C)P(x).$$ (2.3)
where $\alpha_j$ and $\gamma_j$ are item parameters for the $j$-th item, $\theta_i$ is the $i$-th examinee ability parameter, and $F$ is either the logistic cdf, i.e., $F(x) = 1/(1+exp(-x))$ or the standard normal cdf $F(x) = \Phi(x)$. In practice, it often does not make much difference which variation one adopts.

On the other hand, as far as item parameters are concerned, there are three common classes of parametric models in the UIRT literature, namely, the Rasch (one-parameter or 1P) model, the two-parameter (2P) model and the three-parameter (3P) model.

**2.1.1 Rasch model**

The Rasch model (Rasch, 1960), usually referred to as the one parameter logistic model (1PL), assumes that the logit of the item response function is a linear function of $\theta$ and that the slopes of these linear function are equal across all items.

$$P(y_j) = \frac{1}{1 + \exp\{-\alpha(\theta - \beta_j)\}}, \quad j = 1, ..., k$$  \hspace{1cm} (2.4)

The intercepts $\gamma_j = \alpha \beta_j$ are constructed so that the parameter $\beta_j$ can be interpreted as the difficulty of the item. That is, items with large values of $\beta_j$ have lower proportion of examinees correctly endorsing them. The discrimination parameter $\alpha$ can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individual’s abilities is allowed to vary. Common values for the discrimination are $\alpha = 1$ and $\alpha = 1.7$. The latter is usually adopted so that the IRF is similar to the normal ogive model.

Rasch models assume that items differ only in difficulty levels. In other words, the IRFs can only differ in the intercepts. If plotting them, the IRFs do not intersect. This property
is referred to as *invariant item ordering* (Sijtsma & Junker, 1996), which is related to another property of the Rasch model called *specific objectivity* (Rasch, 1960). The invariant property says that comparisons between two examinees (items) are independent of the items (examinees) used to measure them. More specifically, if the probabilities $P(y_{ij})$ and $P(y_{i'j})$ are known for two examinees $i$ and $i'$, then the difference between their abilities is:

$$\theta_i - \theta_{i'} = \log \frac{P(y_{ij})}{P(y_{i'j})},$$

(2.5)

which is independent of the characteristic of item $j$ chosen for comparison. Likewise, if the probabilities $P(y_{ij})$ and $P(y_{ij'})$ are known for two items $j$ and $j'$, then the difference between the two item difficulties is:

$$\beta_i - \beta_{i'} = \log \frac{P(y_{ij})}{P(y_{ij'})},$$

(2.6)

which is independent of the examinee $i$ chosen for comparison. The Rasch model is shown to be the only IRT model that has this property and hence is considered as specifically objective. Another attractive feature of the Rasch model is that the raw score $\sum_{j=1}^{k} y_{ij}$ is a minimal sufficient statistic for the person ability parameter $\theta_i$. In fact, the Rasch model is the only possible IRT model for which there exists a one-dimensional minimal sufficient statistic for the ability parameter (Andersen, 1977).

The Rasch model is a relative simple but peculiar model with nice properties. To proponents of Rasch IRT modeling, who believe test items are prepared by criteria so they fit into the Rasch model, it is a separate IRT model differentiated from two- or three-parameter models. However, for those who acknowledge that varying IRT models
are tried out to determine which provides best description to the test data, it does not fit all test data. The model has to be expanded so as to provide better fit.

### 2.1.2 Two parameter logistic model

In many situations the assumption that items differ only in difficulties is too restrictive. Birnbaum (1968) introduced a model called the two-parameter logistic (2PL) UIRT, or U2PL model which generalizes the Rasch model by allowing the slopes to vary so that

\[
P(y_j) = \frac{1}{1 + \exp\{-\alpha_j(\theta - \beta_j)\}}, \quad j = 1, \ldots, k. \tag{2.7}
\]

The slope parameter \( \alpha_j \), also known as the discrimination of the item, is a measure of how much information an item provides about the latent ability \( \theta \). As \( \alpha_j \to \infty \), the IRF approaches a step function with a jump at \( \beta_j \) (this type of items are also referred to as Guttman items). Compared with the Rasch model, the U2PL model is not specifically objective. The differences between the logits of the response functions do not yield independent comparisons of person abilities under the U2PL model. In addition, the U2PL model does not have a simple sufficient statistic for the ability parameters, unless the discrimination parameters are fixed and known.

### 2.1.3 Three parameter logistic model

The response functions approaches zero as \( \theta \to -\infty \) for both the Rasch and U2PL models as can be seen from equations 2.4 and 2.7. However, for multiple choice items, it’s possible for examinees to guess correctly. Under this situation, the assumption

\[
\lim_{\theta \to -\infty} P(y_j) = 0
\]

is not reasonable for the cognitive process the model is trying to describe. For this reason, Birnbaum (1968) generalized the U2PL model to allow the IRF to have a
lower asymptote above zero. The generalization is defined as three-parameter logistic (3PL) UIRT, or U3PL model.

\[ P(y_j) = c_j + \frac{1-c_j}{1 + \exp\{-\alpha_j(\theta - \beta_j)\}}, \quad j = 1,\ldots,k, \tag{2.8} \]

where \( c_j \) is the guessing parameter. The U3PL model assumes that the examinee knows the correct answer of item \( j \) with probability as specified in 2.7 or guesses the item correctly with probability \( c_j \).

The three UIRT models described previously are the logistic variant of the IRT models. If replacing \textit{logit} with \( \Phi \), the IRFs are defined as one-parameter normal ogive UIRT (U1PNO), two-parameter normal ogive UIRT (U2PNO) and three-parameter normal ogive UIRT (U3PNO) models. With the parametric models specified, fast and efficient estimation procedures are needed before the models could be applied to the educational or psychological problems.

2.1.4 Parameter estimation for the unidimensional models

Although IRT models appeared in the 50s, their application to practical test problem remained limited due to the lack of a well-defined estimation procedure. As early as 1925, R. A. Fisher introduced the Newton-Raphson iterative method to solve the nonlinear likelihood equations involved in the estimation of dosage response models in bioassay. This method, referred to \textit{Fisher scoring} in the statistics literature, uses the second derivatives of the log-likelihood function and has a significant role in IRT maximum likelihood (ML) procedures for estimating item and person parameters. However, the procedure couldn’t be readily applied to estimate the item and person
parameters jointly, either because of model under-identification or because of bias introduced by treating a random component in the model as fixed.

Some basic estimation techniques for IRT include joint maximum likelihood (Birnbaum, 1969), conditional maximum likelihood (Molenaar, 1995), marginal maximum likelihood (Bock & Aitkin, 1981), and Bayesian estimation with Markov chain Monte Carlo. It has to be noted that basic estimation techniques for UIRT models rely heavily on the assumptions that the individuals taking the tests are independent of one another and that items behave in the same way for all individuals, and that the item responses of a given individual are independent given that individual’s ability level \( \theta_i \). Under the assumption of conditional independence (or local independence) the joint probability of the item response vector \( y_i \), conditional on \( \theta_i \) is

\[
L(\theta_i \mid y_i, \xi) = \Pr(y_i \mid \theta_i, \xi) = \prod_{j=1}^{k} \Pr(y_{ij} \mid \theta_i, \xi_j),
\]

(2.9)

where \( \xi_j \) is the vector of all item parameters for item \( j \). For example, the likelihood for ability \( \theta \) under the U2PL model, where \( \xi_j = (\alpha_j, \beta_j)' \), is:

\[
L(\theta_i \mid y_i, \xi) = \frac{\exp\{\theta \sum_j y_{ij} \alpha_j - \sum_j y_{ij} \alpha_j \beta_j\}}{\prod_j (1 + \exp\{\alpha_j (\theta_i - \beta_j)\})},
\]

(2.10)

The following sections describe the four basic methods for the estimation of item response models.

2.1.4.1 Joint maximum likelihood

The joint maximum likelihood (JML) estimation procedure treats both item parameters (e.g. \( \alpha_j, \beta_j \)) and person abilities \( \theta_i \) as unknown, but fixes model parameters. Under
the JML procedure the \( n \times k \) item responses are essentially treated as the observational units in the analysis. The JML procedure estimates the item parameters (\( \xi \)) and examinee abilities \( \theta \) by maximizing \( L(\xi, \theta | y) = \prod_{j=1}^{k} L(\theta_j | y_j, \xi_j) \) with respect to \( \xi \) and \( \theta \) simultaneously.

The model is not identified, however, which means there is no unique solution to the maximization equation. Thus, constraints have to be placed on the model parameters to ensure the existence of a unique solution. For two parameter models like the U2PL model, two constraints are necessary, i.e., a location constraint and a scale constraint. The location constraint can be made by constraining either a single ability or difficulty to some fixed number, or by constraining the average ability or difficulty to some fixed number (typically zero). The scale constraint can be made by forcing the product of discrimination parameters to one (i.e., \( \prod_j \alpha_j = 1 \)). However, even with constraints, the maximization equation cannot be solved analytically unless some numerical method is used. In addition, JML estimates suffer from a problem of being inconsistent (Andersen, 1970). This is caused by the fact that a limited number of parameters of interest (item characteristics \( \xi \)) are to be estimated in the presence of many nuisance parameters (person abilities \( \theta \)). Eliminating the nuisance parameters gives the solution for this problem with the use of conditional or the marginal maximum likelihood method.

2.1.4.2 Conditional maximum likelihood

Andersen (1970) proposed conditional maximum likelihood (CML) as an alternative technique for ML estimation method for the Rasch model. His method conditions on raw scores, \( \sum_j y_{ij} = s_i \), which is a sufficient statistic for the person ability \( \theta \):
\[ P(y_j | \theta, \xi, s_i) = \frac{\exp\left\{-\sum_{j} y_{ij} \beta_j \right\}}{\sum_{x \in \{y_{ij} = s_i\}} \exp\left\{-\sum_{j} x_{ij} \beta_j \right\}}. \]

The formula does not depend on the value of \( \theta \). The CML estimates the item parameters by maximizing the conditional likelihood \( L(\xi | y, s) = \prod_i P(y_i | \xi, s_i) \).

Although it has been showed that CML estimates for item parameters are consistent (Andersen, 1970), there could be some loss of information when CML estimation is applied in that by using the conditional likelihood to estimate the item parameters, the marginal distribution of the raw score, which possibly contains some information on the item parameter, is neglected. Moreover, the CML method works only when there is a simple sufficient statistic like the raw scores \( s_i \) for the Rasch model. This condition cannot be satisfied by the more complicated UIRT models, such as the U2PL model, which do not have simple sufficient statistics.

### 2.1.4.3 Marginal maximum likelihood

Similar to CML, marginal maximum likelihood (MML) takes a different approach to removing the nuisance (ability) parameters from the model. Unlike JML estimation, which treats each of the \( n \times k \) item responses as separate observational units, the MML technique treats only the \( n \) individuals as the observational units by assuming that the ability parameters are random effects sampled from some larger continuous distribution, denoted \( F(\theta) \). Integrating the random ability effects out of the individual likelihoods defined in 2.9 results in the marginal probability of observing the item response vector \( y_i \):

\[ P(y_i | \xi) = \int_{\theta} L(\theta | y_i, \xi)dF(\theta). \]  \hspace{1cm} (2.11)
Taking the product of the probabilities in 2.11 over examinee $i$ produces the marginal likelihood of the item parameter $\xi$

$$L(\xi \mid y) = \prod_i P(y_i \mid \xi),$$

which is maximized with respect to the item parameters $\xi$ to derive the MML estimates. Numerical methods are necessary to maximize the likelihood and numerical integration techniques are required to approximate the integral in 2.11 as well. Similar to the JML estimation procedure, constraints have to be placed on the location and scale parameters to identify the model. In effect, the constraints can be either placed on the mean and standard deviation of the ability distribution $F(.)$ or on the item parameters. As described earlier, the ability distribution is a part of the IRT model so a correct specification is needed. Otherwise, the resulting estimates of the item parameters can be useless. Further, in MML the loss in efficiency of the parameter estimation due to the joint estimation of item parameters with the parameters of the ability distribution is not clear.

### 2.1.4.4 Bayes modal estimation

The Bayesian method for estimating IRT model parameters is similar to the MML technique described previously. However, in addition to assuming a distribution for ability parameters, Bayesian analysis also places a prior distribution for each of the model parameters. In addition, with Bayesian technique, it is possible to simultaneously estimate posterior estimates for both item characteristics and person traits. Early Bayesian procedures focused on Bayes marginal modal solutions.

As early as 1968, Birnbaum proposed a two-stage procedure of estimation that assumes first-stage sampling of item responses from each respondent and second-stage sampling of respondents from the population. This procedure makes possible Bayes estimation of
the examinee’s ability level at the first stage, and ML estimation or Bayes estimation of item parameters and latent distributions at the second stage. Under such circumstances, the Bayesian estimate of ability, $\theta$, can be either the mode of the posterior ability distribution, the Bayes modal or maximum a posteriori (MAP) estimate or the mean of the posterior distribution, the Bayes estimate or the expected a posteriori (EAP) estimate. This procedure could be cumbersome if the number of items is large, which directly affects the size of the information matrix. The computational difficulty was solved after Dempster, Laird & Ruben (1977) proposed a new method, the so-called expectation-maximization (EM) algorithm of ML estimation of two-stage models. This expectation-maximization procedure only requires the first derivatives of the log-likelihood and is numerically robust whereas easy to implement.

It was Bock & Aitkin (1981) who formulated the standard estimation procedure in IRT using the EM method with MML estimation for a two-stage sampling scheme, treating person parameters as missing data, and then using the estimated item parameters and taking them as known when making inference regarding the latent abilities. In light of this, BILOG program was developed to carry out the estimation by placing stochastic constraints on the parameter estimates so there is a normal prior distribution for thresholds, a log normal prior for slope and a beta prior for lower asymptotes (Mislevy & Bock, 1983). Much research has been conducted using the empirical Bayes procedure (e.g., Mislevy, 1985; Tsutakawa & Lin, 1986) and the associated MML is considered to be more accurate, hence less biased, than other ML estimating methods (Lord, 1986). Nonetheless, when the model gets more complex, integrating out ability parameters is not straightforward. Moreover, and most importantly, ML or EM procedures fail to take into account the uncertainty about item parameters when making inference on
examinees and may seriously underestimate the uncertainty in abilities (Tsutakawa & Soltys, 1988; Tsutakawa & Johnson, 1990).

2.1.4.5 BILOG

BILOG (Mislevy & Bock, 1983) is a computer program designed to estimate unidimensional IRT model parameters using MML together with the Bayes modal solution described by Mislevy (1986). The population density \( F \) for \( \theta \) in equation 2.11 is approximated by a step function with jumps at finite number of points, which are usually referred to as quadrature points. By using the finite number of points, the program assumes that the only values \( \theta \) can take are those represented by the quadrature points. Thus, the posterior distribution of \( \theta \) can be obtained using Bayes theorem from the examinee’s responses, the item parameters and \( F \). With this, the expected value of the log likelihood can be calculated and maximized with respect to item parameters. However, the item parameters and \( F \) are not readily known. To solve the problem, the program adopts the EM algorithm to iteratively recompute expected value and \( F \) with updated item parameters estimates.

Additionally, BILOG fits in a formal Bayesian framework by incorporating external information, i.e., assuming the normal prior distribution for item difficulties, the log-normal for the discrimination parameters and the beta distribution for guessing under the 3PL model. The prior distributions may be either specified by the user or partially estimated from the data. The latter is called floating priors and is the default in BILOG. Floating priors signifies that all item parameters shrink toward the mean, which is estimated from the data instead of being arbitrarily prespecified (Mislevy, 1986). As is the case with Bayesian analysis, as sample size increases, the influence of prior distributions decreases. So does the shrinkage toward the mean. For small samples,
particularly when all possible item-person interactions are obtained, the statistical procedures based on large sample MML theory may not be accurate, thus BILOG is less useful.

### 2.1.4.6 Fully Bayesian approach with Markov chain Monte Carlo

With current enhanced computational technology and the emergence of Markov chain Monte Carlo (MCMC) simulation techniques (e.g., Chib & Greenberg, 1995), the methodology has rapidly developed from the empirical Bayes to a fully Bayesian approach. MCMC is a promising estimation methodology whose appeal lies in its ease of implementation. The rationale behind using MCMC methods is given by Metropolis & Ulam (1949) and Hastings (1970). The fully Bayesian approach allows one to estimate item parameters and examinee abilities at the same time while incorporating uncertainty of the item estimates in calculations of uncertainty about abilities for examinees. Wollack et al. (2002) compared the effectiveness of MCMC with MML estimates by recovering the model parameters and found no marked difference between the two methods, suggesting MCMC an alternative to MML estimation when MML is not available. Albert (1992) was the first to apply an MCMC algorithm known as Gibbs sampling (Gelfand & Smith, 1990; Casella & George, 1992) method to IRT problems. Specifically, Albert (1992) applied the data augmentation idea of Tanner & Wong (1987) to the U2PNO model. Other similar attempts on Bayesian estimation for UIRT models can be found for the U2PL model (Patz & Junker, 1999), the U3PNO model (Béguin & Glas, 2001), the U3PL model, etc.

The Gibbs sampling procedure, as described by Gelfand & Smith (1990), proceeds as follows. Suppose one is interested in simulating samples from the \( k \)-parameter posterior distribution \( \pi(\Theta) = \pi(\Theta_1, \ldots, \Theta_k) \) of the random vector \( \Theta = (\Theta_1, \ldots, \Theta_k) \) and it is
difficult to sample from the joint posterior because of the complexity of the model.

However, it is easy to simulate samples from the full conditional distributions

\[ \pi(\Theta_i \mid \Theta_j, j \neq i, j = 1, \ldots, k), \ i = 1, \ldots, k. \]

Then given arbitrary starting values

\[ \Theta_i^{(0)}, \Theta_2^{(0)}, \ldots, \Theta_k^{(0)}, \]

one can draw

\[ \Theta_1^{(1)} \text{ from } \pi(\Theta_1 \mid \Theta_2^{(0)}, \ldots, \Theta_k^{(0)}), \]
\[ \Theta_2^{(1)} \text{ from } \pi(\Theta_2 \mid \Theta_1^{(1)}, \Theta_3^{(0)}, \ldots, \Theta_k^{(0)}), \]
\[ \vdots \]
\[ \Theta_k^{(1)} \text{ from } \pi(\Theta_k \mid \Theta_1^{(1)}, \ldots, \Theta_{k-1}^{(1)}). \]

This completes a cycle from the starting values \( \Theta^{(0)} = (\Theta_1^{(0)}, \ldots, \Theta_k^{(0)}) \) to a new sample \( \Theta^{(1)} = (\Theta_1^{(1)}, \ldots, \Theta_k^{(1)}) \). Iteratively generating random variables from each of the full conditional distributions in turn produces a sequence \( \Theta^{(0)}, \Theta^{(1)}, \ldots, \Theta^{(m)}. \) This procedure is a realization of Markov chain, with transition kernel from \( \Theta^{(t)} \) to \( \Theta^{(t+1)} \) given by

\[ P(\Theta^{(t+1)} \mid \Theta^{(t)}) = \prod_{i=1}^{k} \pi(\Theta_i^{(t+1)} \mid \Theta_j^{(t)}, i < l, \Theta_j^{(t+1)}, i < l). \]

With the essential properties that the joint distribution of \( \Theta^{(t)} = (\Theta_1^{(t)}, \ldots, \Theta_k^{(t)}) \) converges geometrically to

\[ \pi(\Theta) = \pi(\Theta_1, \ldots, \Theta_k) \text{ as } t \to \infty \]

and that for any initial distribution, \( f_t = \frac{1}{t} \sum_{i=1}^{t} f(\Theta^{(i)}) \)

asymptotically converges to \( Ef(\Theta) = \int f(\Theta)\pi(\Theta)d\Theta \) as \( t \to \infty \), the joint distribution of \( \Theta \) can be approximated by the empirical distribution of the \( M \) values \( (\Theta_1^{(t)}, \ldots, \Theta_k^{(t)}), t=m+1, \ldots, m+M, \) where \( m \) is large enough so that the Gibbs sampler converges. Hence, the proposed marginal density estimate for \( \Theta_i, l = 1, \ldots, k \) can be approximated by
\[ \frac{1}{M} \sum_{i=m+1}^{m+M} \pi(\Theta_j | \Theta_1^{(i)}, \ldots, \Theta^{(i)}_{j-1}, \Theta^{(i)}_{j+1}, \ldots, \Theta^{(i)}_k). \]

The posterior mean and variance of \( \Theta \) can be approximated by 
\( \Theta = \frac{1}{M} \sum_{i=m+1}^{m+M} \Theta^{(i)} \)
and 
\( \frac{1}{M} \sum_{i=m+1}^{m+M} (\Theta^{(i)})^2 - \Theta^2 \) respectively.

### 2.1.5 Local independence and unidimensionality

The cornerstone of IRT is the assumption of local independence (LI), namely, statistical independence of responses to test items given item parameters and examinee abilities. Or equivalently, the joint distribution of item responses is equal to the marginal distributions (Lord & Novick, 1968, p.361). This principle was first introduced by Lazarsfeld (1950), who postulated that if there exists a suitable function 
\[ P(y_i = 1 | \Psi_i) = P(\Psi_i) \]
for the probability of correct response, given a value \( \Psi_i \) for the latent variable, the conditional probability of the responses \( y = (y_1, y_2, \ldots, y_n) \) can be expressed as the product
\[ P(y | \Psi) = \prod_{i=1}^{n} P(\Psi_i)^{y_i} (1 - P(\Psi_i))^{1-y_i}. \]  

(2.12)

In the context of maximum likelihood estimation, this is called the likelihood of \( \Psi \).

Another essential assumption which shares similar idea is the assumption of unidimensionality, more specifically, the ability parameter vary on only one dimension (Lord & Novick, 1968). In other words, each test item is designed to measure some facet of the same underlying ability or so called unified latent trait. Theoretically, LI includes but goes beyond unidimensionality, although Lord (1980) pointed out, "local independence follows automatically from unidimensionality. It is not an additional assumption" (p.19). The unidimensionality assumption posits that a test intending to
measure one certain ability should not be affected by other abilities, especially when
only the overall test scores are reported and used as an assessment criterion for various
levels of abilities. Therefore, by adopting the above models in parameter estimation and
item scoring, one is confident that a single composite score for all \( k \) items is sufficient.
This is directly related to the construct validity on useful, meaningful and appropriate
interpretations of the test’s scores. This point is further supported by research. For
example, Walker & Berebtas (2000) fit a UIRT model for data known to be
multidimensional and found that error of measurement increases, which resulted in
incorrect inferences about an examinee’s proficiency in a given test. Furthermore, as
more and more attention has been paid to IRT, unidimensionality of the latent trait is
becoming more important and there is a need for well-developed statistical tests (Lord,
1980).

On the other hand, in other instances where distinct multiple abilities (i.e., more than
one \( \theta \)) are involved in producing the manifest responses for an item, more general
models, i.e., multidimensional IRT (MIRT) models have to be adopted, for a single
composite score illustrating one ability will be more likely affected by some other
abilities involved. Thus, a critical problem involves checking the unidimensionality
assumption. Over the past several decades, numerous indices for assessing this
assumption for a pool of binary items have been proposed. Hattie (1985) identified 87
such indices and provided rationale and discussion of the strengths and weaknesses of
each. In particular, he categorized the indices into several groups: indices based on
answer patterns, reliability, correcting for the number of items, principle components,
factor analysis, nonlinear factor analysis, and some other approaches. After reviewing
extensively the literature and discussing the indices of dimensionality, he concluded that
“there are no known satisfactory indices. None of the attempts to investigate
unidimensionality have provided clear decision criteria for determining it” (Hattie, 1985, p.158).

2.2 Multidimensional IRT Models

The UIRT model is useful when tests are designed to measure only one ability. While this ability may be explained by one latent trait or a specific combination of traits, each item on the test measures only one feature of an examinee that can be represented by $\theta$. However, psychological processes have consistently been found to be more complex than they first appear, and an increasing number of educational measurements assess an examinee on more than one trait factor (Reckase, 1997). For instance, a math test composed of short answer items could potentially be evaluating students on their knowledge of arithmetic operations as well as their ability to verbally express their problem solving strategies. If only one score is reported, that is, if the score is only expressed in terms of mathematical understanding, an inaccurate interpretation of the examinee’s ability could arise. It is possible that an examinee has a strong grasp of arithmetic operations but fails in writing technique. If the test is interpreted solely in terms of mathematical understanding, he or she may be unfairly judged to lack algebraic reasoning. Research has shown that when the test known to be multidimensional is modeled using a unidimensional model, error of measurement increases and incorrect inferences about an examinee’s proficiency in a given subject may be made (Walker & Berevtas, 2000). It is hence necessary to distinguish between the multiple latent traits that affect a person’s performance on a given item. With regard to this, multidimensional IRT (MIRT) is a methodology that shows promise for dealing with this form of complexity in educational and psychological measurement (Reckase, 1997).
MIRT is a methodology that shows promise for dealing with the situation of complexity in psychological measurement when multiple latent traits affect the examinee’s performance on a given item (Reckase, 1997). The MIRT model allows separate inferences to be made about an examinee for each distinct ability dimension being measured by introducing person trait and item discrimination parameters for each skill measured by a test question (Ackerman, 1993). A distinction is frequently made between compensatory and noncompensatory MIRT models. Suppose a test consists of \( k \) multiple choice items, each measuring \( m \) ability dimensions, \( \theta_1, \ldots, \theta_m \). Let \( y_{ij} \) represent a matrix of \( n \) examinees’ responses to \( k \) dichotomous items, so that \( y_{ij} \) is defined as

\[
y_{ij} = \begin{cases} 1, & \text{if person } i \text{ answers item } j \text{ correctly} \\ 0, & \text{if person } i \text{ answers item } j \text{ incorrectly} \end{cases}, \ i = 1, \ldots, n, \ j = 1, \ldots, k
\]

Reckase (1985) derived a multivariate extension of the U2PL model with the form

\[
P(y_{ij} = 1 | \theta, \alpha_j, \gamma_j) = \frac{1}{1 + \exp\{-\sum_{v=1}^{m} \alpha_v \theta_v - \gamma_j\}}.
\]

This model was labeled as a two-parameter logistic compensatory MIRT (M2PL) model. When the link function is \( \Phi \) instead of logit, the model is called two-parameter normal orgive compensatory MIRT (M2PNO) model. Rather than being classified by one latent trait, \( \theta_i \), as in the UIRT, examinees are represented by an ability vector in the multidimensional model: \( \theta_i = (\theta_{i1}, \ldots, \theta_{im}) \), where \( m \) is the number of dimensions being measured by a specific item. Similar to those in the unidimensional model, the ability levels are usually assumed to be normally distributed. The discrimination parameters for an item measuring multiple dimensions are also represented in vector form:

\( \alpha_j = (\alpha_{j1}, \ldots, \alpha_{jm}) \), where \( j \) represents the item number and \( m \) indicates the dimension to
which the discrimination value applies. The higher a particular $\alpha_{vj}$, the more important that dimension becomes in determining an examinee’s success on item $j$. $\gamma_j$ is a scalar parameter determining the location in the latent space where the item is maximally informative. It is related to the difficulty parameter in the UIRT model through a function that includes the vector of discrimination parameters and $\gamma_j$. Due to the additive nature of the ability parameters, an examinee with a low ability on one dimension is able to compensate by having a higher ability on another, thus increasing his or her probability of correctly responding to the item.

The two-parameter logistic noncompensatory MIRT model (Whitely, 1980) is defined as

$$P(y_{ij} = 1 | \theta_i, \alpha_j, \beta_j) = \frac{1}{\prod_{j=1}^{k} [1 + \exp\{-(\alpha_{vj} - \beta_{vj})\}]},$$

where $y_{ij}$, $\theta_i$, and $\alpha_{vj}$ are defined as above. $\beta_{vj}$ is the difficulty parameter for item $j$ on dimension $k$. Different from the compensatory model, the noncompensatory model specifies a separate difficulty parameter for each ability dimension. With this model, for fixed values of the exponents, the probability of success decreases with an increase in the number of dimensions, $m$.

As can be inferred from equations 2.13 and 2.14, in compensatory MIRT models, the latent abilities interact with each other so that a lack in one ability dimension can be offset by an increase in other ability dimensions. On the contrary, in noncompensatory MIRT models, a lack in one ability cannot be compensated through an increase in others. Given this major distinction, the former may be more appropriate for items having disjunctive component processes whereas the latter may be appropriate for items that have conjunctive items that have conjunctive component processes (Mavis, 1999).
However, noncompensatory models exhibit greater estimation challenge due to dimension-specific difficulty parameters. Therefore, they are practically disadvantageous to compensatory models (Knol & Berger, 1988).

2.2.1 Multi-unidimensional IRT models—a special case of MIRT

Sometimes a test involves multiple abilities but each item measures only one of them. This is probably more common in large scale testing situations than the more general case where each item measures multiple traits as described earlier. More often a test consists of several subtests with each focusing on one specific ability and the items in a particular subtest are designed to measure one ability in common. The IRT models particularly appropriate in this situation are defined as multi-unidimensional IRT models to differentiate them from the usual MIRT models in the literature. Suppose an English test consists of three subtests, listening, reading and writing. Each subtest measures a slightly different trait in English. The overall exam is multidimensional while each subtest is actually unidimensional. In this situation, the vector of discrimination parameters $\mathbf{a}_j = (\alpha_{ij}, ..., \alpha_{mj})$, as specified in the compensatory MIRT model in equation 2.13, is simplified to $\mathbf{a}_j = (0, ..., 0, \alpha_{vj}, 0, ..., 0)$. Alternatively, suppose a $k$-item test consisting of $m$-subtests, each containing $k_v$ multiple choice items which measure one ability dimension. With a logit link, the probability of person $i$ obtaining correct response for item $j$ of the $v$-th subtest can be defined as follows:

$$P(y_{iv} = 1 | \theta_v, \alpha_{vj}, \gamma_{vj}) = \log \frac{\exp \{ \alpha_{vj} \theta_v - \gamma_{vj} \}}{1 + \exp \{ - \alpha_{vj} \theta_v + \gamma_{vj} \}} = \frac{1}{1 + \exp \{ - \alpha_{vj} \theta_v + \gamma_{vj} \}},$$

(2.15)

where $\alpha_{vj}$ and $\theta_v$ are scalar parameters representing the item discrimination and the examinee ability in the $v$-th ability dimension, and $\gamma_{vj}$ is a scalar parameter indicating the
location in that dimension where the item provides maximum information. As described in details later, the discrimination parameters \( \alpha_j = (\alpha_{j1},...,\alpha_{jm}) \) in 2.13 can be interpreted as similar to factor loadings in factor analysis. The model reduces to the multi-unidimensional model if a rotation is performed so that each item loads on one factor only. Therefore, the discrimination parameters have a specific pattern of \( \alpha_j = (0,...,0,\alpha_y,0,...,0) \). Hence, with each item loads on one factor only, the model can be regarded as some special type of the MIRT models.

### 2.2.2 Estimation for the multidimensional models

The estimation procedures for the MIRT model are relatively challenging because sufficient statistics do not exist for the parameters and that the person and item parameters have to be estimated simultaneously. One basic type of procedures for estimating parameters in compensatory MIRT models involves full information procedures which work directly with the observed response vectors of examinees and the corresponding computer program, TESTFACT (Wilson et al., 1984), has been devised to estimate multidimensional item parameters.

TESTFACT, developed by Wilson, Woods, and Gibbons (1984), is a computer program designed to perform a non-linear, exploratory full information factor analysis on dichotomous item responses. As an exploratory program, prior restrictions on item parameters are not specified. However, the number of latent traits hypothesized to underlie the test construct must be specified. The model, in essence, predicts the dimensional structure of individual items based upon the number of traits, defined a priori, that contribute to their responses (McDonald, 1999). This is done using marginal maximum likelihood (MML) estimation in combination with the EM algorithm (Bock & Aitkin, 1981). The algorithm considers examinees to be a random sample from the
population and assumes their latent trait levels to come from a normal distribution with mean zero and standard deviation of one (Knol & Berger, 1988). This procedure is an iterative one, in which the expected number of examinees at each ability level is first computed along with the expected number of people answering each item correctly. Then, using MML estimation equations, based on the logistic or normal ogive MIRT model (e.g., the model as in 2.13), item parameter estimates are obtained to maximize the likelihood given the observed item responses. The item parameters are then used to re-estimate the expected frequencies, which are then placed once again in the MML equations, and so on. Once the expected frequencies converge with the known response pattern, the final item parameters are found using a Newton-Gauss procedure (Embretson & Reise, 2000).

Due to certain limitations with MML and the empirical Bayes estimates, as illustrated earlier, researchers have explored MCMC techniques for estimating MIRT parameters. Hoijtink & Molenaar (1997) illustrated a Gibbs sampling procedure for a two dimensional latent structure. Their focus was actually on the class of nonparametric multidimensional IRT models. Moreover, Béguin & Glas (2001) generalized Albert’s (1992) approach and proposed using Gibbs sampling for estimating parameters for multidimensional normal ogive models, particularly the 3-parameter models. Lee (1995) also extended Albert’s procedure and formulated MCMC algorithms for multi-unidimensional IRT models (namely, 2-parameter normal ogive and logistic models). Other studies (e.g., Bolt & Lall, 2003; Segall, 2002) further illustrated and supported the effectiveness of the MCMC algorithms, especially for the multidimensional compensatory models.

2.3 Factor Analysis & MIRT
IRT, especially MIRT is, to a certain extent, closely related to factor analysis (FA) in spite of the different focuses of the two approaches (Reckase, 1997). More specifically, as one can see, the UIRT model is appropriate when only one factor is extracted from all the test items whereas MIRT models have to be adopted when more than one factor is found to be significant. With the latter, it should be noted that each item measures multiple abilities, i.e., each test item has loadings on all factors extracted (the loadings can be either zero or not). In an exploratory factor analysis, the factor solution is usually visually or analytically rotated. Often the rotation scheme is devised to approximate simple structure (McDonald, 1985) so that the factor loadings are split into two groups, the elements of one tending to zero and those of the other tending toward unity. Hence, each item has a unity loading on one factor and 0 loadings on other factors. To put it in other words, the test involves multiple abilities and each item measures only one of them.

2.3.1 Early development in FA

Spearman, who pioneered the use of the pattern of correlations between a set of measures to determine the number of abilities, proposed his famous \( g \) theory (1927), in which he argued that there was one general ability \( (g) \) running through all cognitive abilities. He was the first to try to build a statistical theory for intelligence tests. Each test consists of a general factor \( g \) and a specific factor \( s \). With the evidence that many mental tests, even when they are different, have positive correlations with each other, he argued that the general ability \( (g) \) was more meaningful than any specific facet of intelligence \( (s) \). This theory, although known as the Two-factor theory of intelligence, is in effect a single factor theory. However, one might argue, that, in mental tests, many factors play a role in affecting the final test score. In that case \( g \) must be a composite of more than one underlying factor. The question arises: how, at the same time, the correlation matrix
could still be uni-factorial. According to Spearman the solution for this problem could be as follows:

There are, however, certain special cases where g does admit of resolution into a plurality of sub-factors ... say, for example, ability and zeal. If in all tests the respective influences of these two always remained in any constant ratio, then both could quite well enter into g together; for the tetrad equation would still be able to hold. (Spearman, 1927, p. 93).

If Spearman’s theory were generalized to all of the scientific situations, it would mean that no matter what set of measures one might obtain for a particular situation on a specific entity, the intercorrelations among the variables could all be accounted for by a single vector of factor loadings. This hypothesis would be highly restrictive and could not be expected to be appropriate for all phenomena in various situations. In response to this, Thurstone (1931), believing that different tests reflect different content, disagreed that all mental activities could be explained by a single g. Instead, he emphasized on group factors. Tests with the same unique content are supposed to be more related than tests with different content. For example, reading comprehension shares more similarity with verbal analogies than with algebra. He was among the first to propose and demonstrate that there are numerous ways in which a person can be intelligent. Thurstone's Multiple-factors theory identified seven primary mental abilities, namely, Verbal Comprehension, Word Fluency, Number Facility, Spatial Visualization, Associative Memory, Perceptual Speed and Reasoning. He subsequently applied FA to study multiple abilities and found results conflict with Spearman’s theory and concluded (Thurstone, 1938):

So far in our work we have not found the general factor of Spearman, but our methods do not preclude it. The presence of a general factor could be indicated by a large part of the communality of each test that remains unaccounted for by the common factors identified in a simple structure. So far we have found no conclusive evidence for a general factor. (p. 7)
His exclusive use of orthogonal rotation was criticized by researchers. Some of them re-analyzed Thurstone’s data using various combinations of oblique and orthogonal rotation, and in most studies the variance of test scores is evenly distributed between $g$ and different combinations of group factors. Therefore, Thurstone’s emphasis on group factors in explaining individual differences in intelligence may be justified, as long as they are considered to be equal with regard to $g$.

Spearman and Thurstone’s theories on intelligence provided foundational work for the development of FA. From the rich and long history of FA, Reckase (1997) identified several researchers as having special contribution to the factor analytic foundations for MIRT, namely, Horst, Christofferson, Muthén, McDonald and Bock & Aitkin.

Horst was one of the early researchers in the field of FA who laid foundations for MIRT. In his work on FA, he suggested using and thus reproducing the raw data matrix instead of the correlation matrix, as he illustrated (1965):

It should be observed at the outset that most treatments of factor analysis do not, however, begin with a consideration of the $x$ matrix (the matrix of observed scores) as such and the determination of the $\mu$ matrix (the matrix of true scores). These treatments usually begin with correlation matrices derived from the $x$ matrix. This approach has led to much misunderstanding because the analyses applied to the correlation matrix sometimes imply that there is more information in the correlation matrix than in the data matrix $x$. This can never be the case. For this reason, as well as others, it is much better to focus attention first on the data matrix, as such, in considering the problems and techniques of factor analysis. (p. 96)

In his work on the FA of binary data matrices, as summarized in *Factor Analysis of Data Matrices* (Horst, 1965), he summarized issues related to the characteristics of the observed variables, i.e., the correlations among objective tests are affected by the variations of item difficulties, or what he called “item preferences” (p.514). The problem was considered to be that the correlation coefficient between two binary items could never be as high as one when the items have different difficulty levels and hence must be
misrepresenting the exact relationship. Some researchers proposed using tetrachoric correlations to remove the effects of item difficulty on correlation. However, Horst argued against this solution on standardizing binary variables in that:

We cannot resort to tetrachoric correlations for item intercorrelations of unequal preference (difficulty) value when we know that this phenomenon is an essential characteristic of the attributes which we are measuring. We know that for all psychological test scores which are obtained from 0-1 scoring of the constituent items, we are dealing with unweighted sums of binary variables. Therefore, the properties of the binary variables are embedded in all test measures which are sums of binary measures. (Horst, 1965, p. 515-516)

Rather, he suggested working on the segregated data matrix that is not influenced by the variation of item difficulties, or in his term “eliminating or partialing out that part of the dimensionality which is due to dispersion of item preference from the data matrix” (p.516). This idea is conceptually similar to conditionally modeling the data, i.e., estimating the item difficulty parameters first and then using the estimates to model the data. Horst’s work shares much similarity with the current MIRT conceptions. However, focusing on the factors and modeling the actual responses, he did not estimate item parameters or model probabilities of correct responses (Reckase, 1997).

Christofferson (1975) took a step further. He actually produced a probabilistic model. Suppose $y_{i}^{*}$ is the observed response to item $i$. It takes a value of 1 if the corresponding continuous latent variable $y_{i}$ is greater than the item threshold $h_{i}$ and 0 otherwise. That is,

$$y_{i}^{*} = \begin{cases} 1, & \text{if } y_{i} \geq h_{i} \\ 0, & \text{if } y_{i} < h_{i} \end{cases}$$

Then with a normal ogive model, the marginal distribution of a single $y_{i}^{*}$ is

$$P(y_{i}^{*} = 1) = \int_{h_{i}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} dx = 0.5 + \frac{1}{2} \Phi \left( \frac{x - h_{i}}{\sqrt{2\sigma^2}} \right) \bigg|_{h_{i}}^{\infty}$$

(2.16)
The threshold parameters, which were defined as the normal deviates that specified the area under the normal curve equal to the proportion of incorrect responses to the items, are essentially the same as the difficulty parameters in a MIRT model (Reckase, 1997). Christofferson formulated the model, i.e., \( P = P^* + \epsilon \), which “expresses the observed proportions \( P \) in terms of the threshold levels \( h_i \), \( i = 1, 2, \ldots, M \), the loadings \( \Lambda \), the factor correlation \( \Phi \) and a random component \( \epsilon \)” (p.8) and further presented a generalized least squares (GLS) method to estimate the factor loadings and item thresholds. As can be seen from the probabilistic model in 2.16, Christofferson’s approach differs from the current MIRT mainly in that he was modeling the hypothetical continuous item trait rather than the probability of correct response that the probability was not modeled conditional on the item and person parameters, but rather it was modeled as population statistics (Reckase, 1997).

Based on Christofferson (1975)’s work, Muthén (1978) extended the model by presenting it for the observed proportions of correct responses in \( m \)-dimensional vector \( \mathbf{p} \). That is, \( \mathbf{p} = f(\theta) + \epsilon \), where \( \theta = (\theta_1, \theta_2) \) is an \( m \)-dimensional vector with \( \theta_1 \) denoting the vector of thresholds and \( \theta_2 \) as the vector of elements below the diagonal of the matrix of population tetrachoric correlations. In this way, the probabilities of correct responses were modeled with regard to the item threshold. However, they were not conditional on the person parameters.

Both Christofferson and Muthén used a normal ogive model to present probabilistic models very similar to the MIRT conceptualization. What still needed is the conditional probability of correct response to an item as a function of the vector of person parameters, or ability space \( \theta \).
The nonlinear FA approach taken by McDonald (1967) is most similar to MIRT. However, his focus was still more on estimating the factors instead of understanding the interaction between item and person characteristics. In addition, he used polynomial models that are mathematically intractable and can result in values beyond the permissible range of 0 to 1. It was more recent that he explicitly connected his early work with IRT. McDonald (1985) pointed out that for a binary variable, the probability of correct response as a function of a factor was actually the same as the regression curve on that factor and that when the regression curve was allowed to be nonlinear, the problem of variation of item difficulties in the analysis of binary data could be dealt with easily. In his description of the solution to the factor analysis of binary data, he clearly emphasized the importance of local independence as a basic principle, which is a key assumption in IRT. Further, he actually made the relationship between FA and IRT clear in indicating that FA is a special case of IRT, as he states, “The view taken here is that common factor analysis is a special case of latent trait theory (IRT), based on the principle of local independence, but one in which for convenience only the weak, zero-partial-correlation version of the principle is typically tested” (p.203). McDonald also illustrated the relationship of the regression function to the conditional probability of correct response as \( p_i | \Psi = E(y_i | \Psi) \), where \( \Psi \) is the vector of latent characterizations and \( p_i \) is the probability of correct response to item \( i \).

Bock & Aitkin (1981) defined a normal ogive model for a multidimensional ability space that included item characteristics, namely, item difficulty and discrimination, labeled by them as intercept and slope, which are essential in both FA and IRT. In working on the solution for FA, they used a two-dimensional extension of the 2PNO model along with a compensatory MIRT type parameterization of the item parameters. The model is showed as
\[ P(y_j = 1 | \theta_{i1}, \theta_{i2}) = \frac{1}{2\pi^{1/2}} \int_{-z_j(\theta_j)}^\infty \exp\{-\frac{t^2}{2}\} dt, \quad (2.17) \]

where \( z_j(\theta_j) = d_j + a_{ii}\theta_{i1} + a_{ij}\theta_{i2} \). The \( a, d \) and \( \theta \) are item discrimination, item difficulty and person ability parameters respectively, which are the same as the parameterization in MIRT models. Although the model in 2.17 as presented by Bock & Aitkin (1981) was in effect an MIRT model, their focus was still on defining factors instead of emphasizing the interaction of persons and items in that the item parameters were interpreted in the FA sense for labeling the factors. Moreover, they did not point out that the item parameters were descriptive measures of the interaction between items and persons.

2.3.2 Comparison of FA and IRT

FA and IRT share much in common. The latent trait in IRT carries the same meaning as the term common factor in FA (McDonald, 1985). As illustrated in the previous section, the statistical formulation of FA and IRT to the analysis of binary data matrices of item responses is actually identical. The item discrimination parameter \( a \) in IRT models (e.g., as in equation 2.17) is analogous to the factor loading \( \Lambda \) in linear FA models, for it represents the relationship between the latent variable \( \theta \) and item responses. More specifically, the more strongly responses on an item are related to the latent ability \( \theta \), the larger the corresponding discrimination power or factor loading. The difficulty parameter \( d \), defined as the point on the ability \( \theta \) scale at which the probability is 50% that the item response is greater than the threshold, is essentially the item threshold. In working on the nonlinear FA solution for binary item responses, McDonald (1985) actually presented the formulas for converting between item parameters for FA and 2PNO IRT specifications, i.e.,
\[
\lambda_j = \frac{a_j}{\sqrt{1-\lambda_j}} \quad \text{and} \quad h_j = \frac{d_j}{\sqrt{1+a^2}}
\]

where \(a_j\) and \(d_j\) are item discrimination and difficulty in the IRT models; \(h_j\) and \(\lambda_j\) are factor loading and item threshold in the FA models. Takane & Leeuw (1987) further presented a formal proof of the equivalence between IRT and FA models and concluded that they are “two alternative formulations of a same model” (p.397). They derived a general form for the relations between the parameters in FA and IRT as,

\[
a_j = \frac{\lambda_j}{\sqrt{q_j}} \quad \text{and} \quad d_j = \frac{h_j}{\sqrt{q_j}}
\]

where \(q_j\) is the \(j\)-th diagonal element of \(\Sigma\), the variance/covariance matrix for the random component \(\varepsilon\). It has to be noted that \(q_j\) has to be set to an arbitrary value, for the variance cannot be estimated due to the lack of relevant information in the data. Additionally, another key similarity between FA and IRT is in the partial measurement invariance analyses in relation to the function of invariant and noninvariant items.

On the other hand, they are not exactly the same. FA, especially exploratory FA is basically a data reduction technique whereas IRT focuses on accurately modeling the interaction between persons and items. Reckase (1997) pointed out that understanding this interaction might be hampered by using too few dimensions. Therefore, MIRT is not a data reduction technique. Rather, it is for “determining stable features of both persons and items that influence responses to test items” (p.30). In addition, FA focuses on the correlation or variance/covariance matrices while ignores other item characteristics, such as the mean or standard deviation (SD) of the variables. MIRT, however, does not use standardized variables and instead uses the mean and SD of the items as represented
by the item difficulty and discrimination. A third difference lies in how the goodness of fit test is considered. FA looks at a global measure of fit in accurately reproducing a variance/covariance matrix for a group of examinees. On the contrary, with IRT, people are more concerned with a single item that is not modeled well. As Reise et al. (1993) stated, “Typically, (for IRT) fit is assessed at the item level by a statistic that tests the congruence between the proportion of item responses in a particular category predicted from an IRF and the proportion of responses in a particular category observed in the data.” (p.558). And lastly, in IRT, efforts have been taken to use the same latent space across tests and examinees to keep a common scale for all analyses. This does not receive much attention in FA approach.

In deriving the relationship between IRT and FA, Takane & Leeuw (1987) illustrated them as two techniques for marginalizing the person parameters, as they stated:

The only crucial difference is where the marginalization is performed. In the IRT formulation dichotomization of \( \tilde{y} \) (continuous latent vector corresponding to the observed responses) is done conditionally on \( u \) and then the marginalization is performed. In the FA tradition, the marginalization is undertaken on continuous \( \tilde{y} \), followed by the dichotomization. An advantage of the IRT formulation is that the dichotomization is relatively straightforward (it can be done separately for each \( \tilde{y}_j \) given \( u \) due to the local independence assumption). ...However, this integration (due to marginalization) usually involves numerical integration, which may be quite time consuming. ... In the FA formulation the marginalization is rather trivial, but the dichotomization is extremely difficult. ...Whereas the IRT formulation uses the maximum likelihood estimation based on the full joint probabilities of response patterns (Bock & Aitkin, 1981; Bock & Lieberman, 1970), the FA approach typically uses a generalized least squares (GLS) estimation based on the first and second order marginal probabilities (Christofferson, 1975; Muthén, 1978). (p.397)

Although IRT focuses on an individual examinee’s score whereas FA focuses on how observations (responses) distributed in the population of examinees, the MML estimation proposed to deal with inconsistent estimators in IRT brought the two dichotomization approaches together. In effect, TESTFACT, the program for the full-information FA as presented in Bock et al. (1988) can be used for either FA or IRT.
2.3.3 Bifactor analysis

Spearman’s fundamental theory on general factor was not appropriate in various situations. In order to explain departures from one common factor, Holzinger & Swineford (1937) proposed the bifactor model by extending the Spearman (1904) one-factor model for intelligence test to include the specific or group factors. This was even before Thurstone’s development of the multiple-factor model. The bifactor model applies to the educational setting when the achievement tests contain more than one subject matter (for example, an English test containing a reading, a writing and a listening section). Such tests are often scored for general English factor, but the multiple content areas induce specific factors.

In the bifactor model, each item is associated with two nonzero factors, a common factor and an uncorrelated group or specific factor. Due to this feature, Gibbons & Hedeker (1992) showed that MML estimation of the bifactor model requires quadratures in only two dimensions, regardless of the number of subtests or content areas. Hence, the conditional dependence problem could be solved in a more computationally practical way. Since the bifactor model accounts for departures from conditional independence or LI of responses to groups of items that depend on a common latent ability, comparing the maximum likelihood of the bifactor solution with that of a one-factor solution also provides a statistical test for violation of the LI assumption. Analysis based on the bifactor model is also included in TESTFACT (Wilson et al., 1984).

In the 50s, another trend of study worth noting shares in common with the bifactor model. Following Thurstone’s procedure of oblique simple structure, Schmid and Leiman (1957) presented a technique for depicting the hierarchical structure of a group of variables and their factors. In this strategy, one first identifies clusters of items and
rotate axes through those clusters; next the correlations between those (oblique) factors is computed, and that correlation matrix of oblique factors is further factor-analyzed to yield a set of orthogonal factors that divide the variability in the items into that due to shared or common variance (secondary factors), and unique variance due to the clusters of similar variables (items) in the analysis (primary factors). The Schmid and Leiman (1957) method was associated with the Structural Equation Modeling (SEM) framework and implemented in Mplus (Muthén & Muthén, 2001) for the second-order factor model.

In summary, with fewer parameters, the UIRT models are much simpler and easier to implement. However, as a subset of the MIRT models, they require more stringent assumptions, which are easily violated in the actual testing situations. IRT is closely related to factor analysis to certain extent. Although with different focuses, their parameters can be converted from one to the other. UIRT models are comparable to the one-factor model whereas MIRT models are comparable to the multiple-factor model. With one common factor and several specific factors, the bifactor model is found to have advantages over the one-factor model. Moreover, the bifactor model is a model in FA framework that focuses on data reduction and factor solutions instead of the interaction of person ability and item characteristics. Thus, as an alternative to the UIRT model, the current study proposes IRT models in a Bayesian hierarchical framework so that they incorporate one general ability and several specific abilities for all test items. In this way, both general and specific abilities can be estimated simultaneously.
CHAPTER 3.

 METHODOLOGY

This chapter is organized with four main sections. Section one restates the research questions and Section two briefly reviews the unidimensional and multi-unidimensional IRT models, in particular, a Bayesian 2 parameter normal ogive unidimensional (U2PNO) model and a 2 parameter normal ogive multi-unidimensional model in Bayesian hierarchical framework. In the third section, the proposed Bayesian IRT models with one general ability and several specific abilities are presented so that each item is related to two latent dimensions directly or indirectly depending on the additive or hierarchical relationship between the general ability and specific abilities. Finally, the statistical procedures, together with the sample used in the analyses are described in the fourth section.

3.1 Overview of Research Questions

The major purpose of this study is to propose IRT models incorporating one general ability and several specific ability dimensions in the Bayesian framework. Gibbs sampling procedure is adopted for implementation of the model. The specific research questions related to the performance of the model are as follows:

(1) How does the proposed IRT model perform when implementing it to various simulated situations as well as to the CBASE English data.

(2) How does the proposed model compare with the unidimensional IRT (UIRT) or the multi-unidimensional IRT model as far as the CBASE English data are concerned.
How does the proposed model, which is considered using Bayesian approach, compared with the bifactor analysis implemented in TESTFACT.

The proposed IRT model, together with the unidimensional and multi-unidimensional models are described under the Bayesian hierarchical framework in the subsequent sections.

3.2 Bayesian Unidimensional & Multi-unidimensional IRT Models

Bayesian approaches require the specification of hierarchical models based on prior distributions for model parameters and hence random samples are simulated from the posterior distribution through simulated Markov chain procedures (Gelman et al., 2004). The Bayesian model specifications of a 2 parameter normal ogive unidimensional (U2PNO) model and a 2 parameter normal ogive multi-unidimensional model are reviewed in this section. It has to be noted that the IRT models involved in the study are exclusively two parameter normal ogive models.

3.2.1 Hierarchical U2PNO model

For the UIRT model, or more specifically the U2PNO model, the study follows Albert (1992)'s procedure. Suppose a $k$-item (multiple choice item) test, designed for measuring one single latent ability, has been administered to $n$ subjects, and let $y_{ij}$ denote the score for the $i$-th examinee on the $j$-th item, where $i=1,2,...,n$ and $j=1,2,...,k$. The binary response $y_{ij}$ takes value 1 for a correct response and 0 for an incorrect response. So, $y_{ij}$ can be assumed to be independent Bernoulli random variables with probabilities $p_{ij} = \text{Prob}(y_{ij}=1)$. With a probit link, the U2PNO model is:

$$p_{ij} = P(y_{ij} = 1) = \Phi(\alpha_j \theta_i - \gamma_j),$$

(3.1)
where, as is defined in the previous chapter, $\alpha_j$ and $\gamma_j$ are item parameters for the $j$-th item, $\theta_i$ is the $i$-th examinee ability parameter, and $\Phi(\cdot)$ is the standard normal CDF.

Let $\xi_j = (\alpha_j, \gamma_j)$ denote the vector of parameters for the $j$-th item, let $\theta = (\theta_1, \ldots, \theta_n)$ denote the vector of ability parameters and $\xi = (\xi_1, \ldots, \xi_k)$ the vector of all item parameters. The likelihood function can be written as:

$$f(y \mid \theta, \xi) = \prod_{i=1}^{n} \prod_{j=1}^{k} p^{y_{ij}} p^{(1 - p_{ij})^{1-y_{ij}}}, \quad (3.2)$$

where $p_{ij}$ is as defined in equation 3.1. Assume $\theta_1, \ldots, \theta_n$ are iid samples from $N(0,1)$, and assume the item discriminating parameter $\alpha_j$ is positive while $p(\gamma_j) \propto 1$. Thus, after introducing latent continuous random variables $Z_{ij}$ so that $Z_{ij} \sim N(\alpha_j \theta_i - \gamma_j, 1)$ and

$$y_{ij} = \begin{cases} 1, & \text{if } Z_{ij} > 0 \\ 0, & \text{if } Z_{ij} \leq 0 \end{cases}$$

the joint posterior distribution of $(\theta, \xi, Z)$ is:

$$p(\theta, \xi \mid y) \propto f(y \mid Z) p(Z \mid \theta, \xi) p(\theta) p(\xi). \quad (3.3)$$

Since one is unable to obtain the normalizing constants for equation 3.3, the Gibbs sampling procedure is adopted to simulate random samples of $\theta$ and $\xi$ from the respective posterior distributions and hence inferences can be made accordingly.

3.2.2 Hierarchical multi-unidimensional 2PNO model

Consider a $K$-item test consists of $m$ subtests, each containing $k_v$ multiple choice items, where $v = 1, 2, \ldots, m$. Let $y_{vij}$ denote the $i$-th examinee’s response on the $j$-th item of the $v$-th subtest, where $i=1,2,\ldots,n$ and $j=1,2,\ldots,k_v$. The binary response $y_{vij}$ has probability $p_{vij} = \text{Prob}(y_{vij}=1)$. With a probit link, the 2 parameter multi-unidimensional model is defined as follows:
\[ p_{vij} = P(y_{vij} = 1) = \Phi(\alpha_j \theta_{vi} - \gamma_{vj}), \quad (3.4) \]

where \( \alpha_j \) and \( \gamma_{vj} \) are item parameters for the \( j \)-th item of the \( v \)-th subtest, and \( \theta_{vi} \) is the \( i \)-th examinee ability parameter corresponding to the \( v \)-th subtest. The likelihood function is then:

\[ f(y | \theta, \xi) = \prod_{v=1}^{n} \prod_{i=1}^{k_v} \prod_{j=1}^{k_v} p_{vij}^{y_{vij}} (1 - p_{vij})^{1-y_{vij}}. \quad (3.5) \]

Denote each examinee’s abilities for all items as \( \theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{im}) \), vectors of \( m \) ability parameters. \( \theta_1, \ldots, \theta_n \) are assumed to be multivariate normal and Lee(1995)’s approach is adopted for implementing the model. Let \( \theta_i \) have a prior \( \theta_i \sim N_m(0, R) \), where \( R \) is a correlation matrix, with 1’s on the diagonal and correlation \( \rho_{ij} \) between \( \theta_i \) and \( \theta_j \), \( i \neq j \) on the off diagonals. Then a Wishart prior is assumed for a covariance matrix \( \Sigma \) corresponding to \( R, \Sigma \sim W_+^{m}(0, 2) \), where \( \Sigma = [\sigma_{ij}]_{m \times m} \) so that \( \rho_{st} = \frac{\sigma_{st}}{\sqrt{\sigma_{ss} \sigma_{tt}}} (s \neq t) \). Let \( \xi_{vj} = (\alpha_j, \gamma_{vj}) \) denote the vector of item parameters for the \( j \)-th item of the \( v \)-th subtest.

With flat priors for \( \alpha_j \) and \( \gamma_{vj} \), i.e., \( \alpha_j > 0 \) and \( p(\gamma_{vj}) \propto 1 \), and introducing random variables \( Z_{vij} \) so that \( Z_{vij} \sim N(\alpha_j \theta_{ui} - \gamma_{vj}, 1) \) and \( y_{vij} = \begin{cases} 1, & \text{if } Z_{vij} > 0 \\ 0, & \text{if } Z_{vij} \leq 0 \end{cases} \), the joint posterior distribution of \((\theta, \xi, Z, R)\) is then:

\[ p(\theta, \xi, Z, R | y) \propto f(y | Z) p(Z | \theta, \xi) p(\xi) p(\theta | R) p(R). \quad (3.6) \]

Likewise, with Gibbs sampling procedures, random samples of parameters \( \theta \) and \( \xi \) can be drawn from the full conditionals derived from the posterior distribution specified in equation 3.6 for this hierarchical model. It has to be noted that in the model, estimates of
“true” correlations $\rho_{ij}$ between abilities $\theta_i$ and $\theta_j$ are made possible through imposing the structural constraint by using the correlation matrix $R$ as the usual variance-covariance matrix.

3.3 Proposed Bayesian IRT models

The current study proposes two classes of IRT models incorporating a general ability as well as several specific ability dimensions, namely, additive and hierarchical models. The key difference lies in their different specifications of the relationship between the general ability and specific abilities.

3.3.1 Additive model

For a $K$-item test containing $m$ subtests, each with $k_v$ multiple choice items, where $v = 1,2,...,m$, $y_{vij}$ is the binary response for the $i$-th examinee on the $j$-th item of the $v$-th subtest. With a 2 parameter probit model, define the probability model $p_{vij} = P(y_{vij} = 1)$ as

$$P(y_{vij} = 1) = \Phi(\alpha_{vij}\theta_{vi} + \alpha_{vj}\theta_{0j} - \gamma_{vj}),$$

(3.7)

where $\theta_{vi}$ is the $i$-th examinee’s ability parameter corresponding to the $v$-th subtest, as is defined in the previous section, $\theta_{0i}$ is the $i$-th examinee ability parameter corresponding to the overall test, $\gamma_{vj}$ is the $j$-th item difficulty in the $v$-th subtest, $\alpha_{vij}$ is that item’s discrimination parameter associated with the general ability and $\alpha_{vj}$ is the item discrimination associated with the specific ability. The probability of endorsing an item correctly is assumed to be determined directly by two latent traits—a general and a specific ability. Due to the nature of the relation between the latent traits, this type of models is referred to as additive models. Denote each examinee’s abilities for all items as $\theta_i = (\theta_{0i}, \theta_{1i}, \theta_{2j},...,\theta_{m})^\top$, vectors of $m+1$ ability parameters, and $\theta = (\theta_1,...,\theta_n)^\top$. Also,
denote \( \xi_{ij} = (\alpha_{0ij}, \alpha_{ij}, \gamma_{ij}) \) the vector of item parameters for the \( j \)-th item of the \( v \)-th subtest and \( \xi = (\xi_1, \ldots, \xi_m) \), where \( \xi_v = (\xi_{v1}, \ldots, \xi_{vk_v}) \). With the assumption of LI, i.e., conditional on \( \theta \) and \( \xi \) the responses are independent, the joint probability function of \( y \), where \( y = [y_{ij}]_{n \times k} \) is

\[
P(y | \theta, \xi) = \prod_{v=1}^{m} \prod_{i=1}^{n} \prod_{j=1}^{k_v} p_{ij}^{y_{ij}} (1-p_{ij})^{1-y_{ij}},
\]

(3.8)

where \( p_{ij} \) is as specified in equation 3.7. For the additive model, several different priors for the ability parameters are assumed and the model specifications are presented in the following sections.

### 3.3.1.1 Model specification 1

Assume that the prior distributions of \( \theta_i, i=1, \ldots, n \) are independent multivariate normal (MVN) with mean \( \mu_i \), where \( \mu_i = (\mu_{0i}, \mu_{1i}, \ldots, \mu_{mi})' \), and covariance matrix \( \mathbf{I} \) (the identity matrix) so the prior probability density function for the abilities is

\[
\varphi_{m+1}(\theta_i; \mu_i, \mathbf{I}) = (2\pi)^{-1/2} \exp \{((\theta_i - \mu_i)'(\theta_i - \mu_i))/2\},
\]

(3.9)

where \( m+1 \) is the dimension of \( \theta_i \). Any unconstrained covariance matrix can be adopted for the prior distribution. However, the identity matrix adopted here for the covariance between the latent abilities is to set a strong prior for the latent abilities, assuming small correlations among them, to get around an indeterminacy problem (see Lee, 1995 for a statement of the problem). Also, for the hyperparameter \( \mu_i, i=1, \ldots, n \), they are assumed to be independent MVN with mean \( 0 \), where \( 0 = (0, \ldots, 0) \), and covariance matrix \( \Sigma \), where \( \Sigma \) is assumed to have an inverse-Wishart distribution \( \Sigma \sim W^{-1}(0, 3) \). So the density function is
The prior distribution of $\xi_{vj}, v=1,\ldots,m, j=1,\ldots,k_v$ are assumed to be independent uniform distributed so that $p(\gamma_{vj}) \propto 1$ and

$$I(\alpha_{0v}) = \begin{cases} 1, & \alpha_{0v} > 0 \\ 0, & \alpha_{0v} \leq 0 \end{cases}, \quad I(\alpha_{0j}) = \begin{cases} 1, & \alpha_{0j} > 0 \\ 0, & \alpha_{0j} \leq 0 \end{cases},$$

(3.11)

when there is no prior information. Also, assume that the prior distributions of $\theta$ and $\xi$ are independent.

Suppose that there is a latent variable $Z_{vij}$ that determines the $i$-th examinee’s performance on item $j, j=1,\ldots,k_v$ so that $y_{vij}=1$ if $Z_{vij} > 0$ and $y_{vij}=0$ if $Z_{vij} < 0$. Further, assume that the conditional distribution of $Z_{vij}$ given $\theta$ and $\xi$ is normal with mean $\alpha_{0v} \theta_{0i} + \alpha_{0j} \theta_{0i} - \gamma_{vij}$ and standard deviation 1. This follows that $y_{vij}$ are independent Bernoulli variables with probability of success given by $p_{vij} = \Phi(\alpha_{0v} \theta_{0i} + \alpha_{0j} \theta_{0i} - \gamma_{vij})$.

The joint posterior distribution of $(\theta, \xi, Z, \Sigma, \mu)$ is then:

$$p(\theta, \xi, Z, \Sigma, \mu | y) \propto f(y | Z)p(Z | 0, \xi)p(\xi)p(\theta | \mu)p(\Sigma)$$

$$\propto | \Sigma |^{-\frac{n}{2}} \prod_{i=1}^{n} \left\{ \exp \left\{ -\frac{1}{2} (\theta - \mu) \Sigma^{-1} (\theta - \mu) \right\} \right\} \prod_{v=1}^{m} \prod_{j=1}^{k_v} \left\{ \exp \left\{ -\frac{1}{2} (Z_{vij} - \eta_{vij})^2 \right\} (I(Z_{vij} > 0)I(y_{vij} = 1) + I(Z_{vij} \leq 0)I(y_{vij} = 0)) \right\}$$

$$\prod_{v=1}^{m} \prod_{j=1}^{k_v} I(\alpha_{0v} > 0)I(\alpha_{0j} > 0)$$

(3.12)

where $\eta_{vij} = \alpha_{0v} \theta_{0i} + \alpha_{0j} \theta_{0i} - \gamma_{vij}$ is the prior mean of $Z_{vij}$. The full conditionals are derived as follows:

1. For variable $Z_{vij}$,
\[ [Z_{ij} | \bullet] \propto f(y_{ij} | Z_{ij}) p(Z_{ij} | \eta_{ij}) \propto \exp \left\{ -\frac{1}{2} (Z_{ij} - \eta_{ij})^2 \right\} I(Z_{ij} > 0) I(y_{ij} = 1) + I(Z_{ij} \leq 0) I(y_{ij} = 0) \]  

(3.13)

So the full conditional of \( Z_{vij} \) denoted as \( Z_{vij} | \bullet \) has as a truncated normal distribution

\[ Z_{vij} | \bullet \sim \begin{cases} 
N(0, \infty)(\eta_{vij}, 1), & \text{if } y_{vij} = 1 \\
N(-\infty, 0)(\eta_{vij}, 1), & \text{if } y_{vij} = 0.
\end{cases} \]  

(3.14)

2. For the person parameters \( \theta_i \),

\[ [\theta_i | \bullet] \propto p(Z | \theta, \xi) p(\theta | \mu) \]

\[ \propto \exp \left\{ -\frac{1}{2} (\theta_i - \mu) (\theta_i - \mu) \right\} \prod_{v=1}^{m} \prod_{j=1}^{k} \exp \left\{ -\frac{1}{2} (Z_{vij} - (\alpha_{0y} \theta_{0y} + \alpha_y \theta_y - \gamma_y))^2 \right\} \]

\[ = \exp \left\{ -\frac{1}{2} (\theta_i - \mu) (\theta_i - \mu) \right\} \exp \left\{ -\frac{1}{2} (A \theta_i - B) (A \theta_i - B) \right\} \]

\[ \propto \exp \left\{ -\frac{1}{2} [\theta_i' (A'A + I) \theta_i - 2(\mu_i + A'B)' \theta_i] \right\}. \]  

(3.15)

Thus the full conditional for \( \theta_i \) has a multivariate normal distribution,

\[ \theta_i | \bullet \sim N_{m+1}((A'A + I)^{-1}(\mu_i + A'B), (A'A + I)^{-1}), \]  

(3.16)

where \( A = \begin{pmatrix} \alpha_{01} & \alpha_1 & 0 & \cdots & 0 \\
\alpha_{02} & 0 & \alpha_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\alpha_{0m} & 0 & 0 & \cdots & \alpha_m \end{pmatrix} \) and \( B = \begin{pmatrix} Z_{yv} + \gamma_1 \\
Z_{yj} + \gamma_2 \\
\vdots \\
Z_{ym} + \gamma_m \end{pmatrix}. \)

3. Then for the item parameters \( \xi_{ij} \),

\[ [\xi_{ij} | \bullet] \propto p(Z | \theta, \xi) p(\xi) \propto \prod_{v=1}^{m} \exp \left\{ -\frac{1}{2} (Z_{vij} - (\alpha_{0y} \theta_{0y} + \alpha_y \theta_y - \gamma_y))^2 \right\} I(\alpha_{0y} > 0) I(\alpha_y > 0) \]

\[ = \exp \left\{ -\frac{1}{2} (Z_{vij} - x_i' \xi_{ij}) (Z_{vij} - x_i' \xi_{ij}) \right\} I(\alpha_{0y} > 0) I(\alpha_y > 0) \]

\[ \propto \exp \left\{ -\frac{1}{2} [\xi_{ij}' (x_i' x_i) \xi_{ij} - 2(Z_{vij} x_i) \xi_{ij}] \right\} I(\alpha_{0y} > 0) I(\alpha_y > 0). \]  

(3.17)

So the full conditional for \( \xi_{ij} \) is

\[ \xi_{ij} | \bullet \sim N((x_i' x_i)^{-1} x_i' Z_{vij} (x_i' x_i)^{-1}) I(\alpha_{0y} > 0) I(\alpha_y > 0), \]  

(3.18)

where \( Z_v = [Z_{vij}]_{nv}, \xi_v = (\xi_{1j}, \ldots, \xi_{vk})', \ x_v = [\theta_0, \theta_v, \gamma], \) and \( \theta_v = (\theta_{01}, \ldots, \theta_{0m})'. \)
\( \theta_v = (\theta_{v1}, \ldots, \theta_{vm})', v = 1, \ldots, m. \)

4. Next, for the hyperparameter \( \mu_i \),

\[
\begin{align*}
[\mu_i | \bullet] & \propto p(\theta | \mu) \propto \exp\left\{-\frac{1}{2} (\theta_i - \mu_i)'(\theta_i - \mu_i)\right\} \exp\left\{-\frac{1}{2} \mu_i' \Sigma^{-1} \mu_i\right\} \\
& \propto \exp\left\{-\frac{1}{2} [\mu_i'(I + \Sigma^{-1}) \mu_i - 2\theta_i' \mu_i]\right\}. \\
\end{align*}
\]

So the full conditional for \( \mu_i \) is distributed as

\[
\mu_i | \bullet \sim N_{m+1}(I + \Sigma^{-1})^{-1} \theta_i, (I + \Sigma^{-1})^{-1}). \tag{3.20}
\]

5. Lastly, for the hyperparameter \( \Sigma \),

\[
\begin{align*}
[\Sigma | \bullet] & \propto p(\mu | \Sigma) p(\Sigma) \propto |\Sigma|^{-\frac{\mu+3+1}{2}} \prod_{i=1}^\mu |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \mu_i' \Sigma^{-1} \mu_i\right\} \\
& = |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \text{tr}(S\Sigma^{-1})\right\}, \\
\end{align*}
\]

Thus the full conditional for \( \Sigma \) is an inverse Wishart distribution,

\[
\Sigma | \bullet \sim W^{-1}(S^{-1}, n + 3), \tag{3.22}
\]

where \( S = \sum_{i=1}^\mu \mu_i \mu_i' \).

The Gibbs sampling procedure iteratively updates samples \( Z, \theta, \xi, \mu \) and \( \Sigma \) from the full conditionals shown in equations 3.14, 3.16, 3.18, 3.20 and 3.22 respectively, with starting values \( \theta^{(0)}, \xi^{(0)}, \mu^{(0)} \) and \( \Sigma^{(0)} \). The collection of all these simulated draws from \( p(\xi, \theta | y) \) are then used to summarize the posterior density of item parameters \( \xi \) and ability parameters \( \theta \) (the latent parameters \( Z \) and the hyperparameters \( \mu \) and \( \Sigma \) are not of interest here) and can be used to compute quantiles, moments and other summary statistics. As with standard Monte Carlo, with large enough samples, the posterior means of \( \xi \) and \( \theta \) are considered as estimates of the true parameters.

3.3.1.2 Model specification 2
In the previous section, $\theta_i$ is assumed to have a multivariate prior $\theta_i \sim N_{m+1}(\mu, I)$. It is also reasonable to consider $\theta_i \sim N_{m+1}(\mu, I)$ that all $n$ abilities are multivariate normal with the same mean structure. Then, the full conditionals for $Z$ and $\xi$ are the same as those defined in equations 3.14 and 3.17 and those for $\theta_i$, $\mu$ and $\Sigma$ are slightly different. They are $\theta_i | \bullet \sim N_{m+1}(A'A + I)^{-1}(\mu + A'B), (A'A + I)^{-1}$, $\mu_i | \bullet \sim N_{m+1}((nI + \Sigma^{-1})^{-1}\sum_i \theta_i, (nI + \Sigma^{-1})^{-1})$ and $\Sigma | \bullet \sim W^{-1}((\mu\mu)^{-1}, n+3)$ respectively. In the same fashion, samples of parameters of interest $\theta, \xi$ are iteratively simulated to obtain the posterior estimates.

**3.3.1.3 Model specification 3**

Another approach to parameterizing the additive 2 parameter probit model is to construct the prior for $\theta_i$ so that $\theta_i \sim N_{m+1}(0, R)$, where $R$ is constrained to be a correlation matrix,

$$
\begin{pmatrix}
1 & \rho_{10} & \cdots & \rho_{1m} \\
\rho_{01} & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{m0} & \cdots & \cdots & 1
\end{pmatrix},
$$

with 1’s on the diagonal and correlations between $\theta_i$ and $\theta_s, s \neq t, s, t=0,\ldots,m$, on the off diagonals. The constraints are adopted to get around the model indeterminacy problem as well. With the constraint imposed on $R$, an inverse-Wishart prior is assumed for a transformed covariance matrix $\Sigma$, where

$$
\Sigma = \begin{pmatrix}
\sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0m} \\
\sigma_{10} & \sigma_1^2 & \cdots & \sigma_{1m} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{m0} & \sigma_{m1} & \cdots & \sigma_m^2
\end{pmatrix}
$$

so that \[ \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \] \(i \neq j\), \(p(\Sigma) \propto |\Sigma|^{-\frac{n+3}{2}}\). The full conditionals for $\theta$ and $\Sigma$ are derived as follows:

$$
\theta_i | \bullet \sim N_{m+1}(A'A)^{-1}A'B, (A'A)^{-1}), \quad (3.23)
$$

where $A, B$ are as defined in equation 3.16, and
\[ \Sigma \mid \bullet \sim W^{-1}(S^{*-1}, n + 3), \quad (3.24) \]

where \( S^{*} = \sum_{i=1}^{n} (L \theta_i)(L \theta_i)' \) and \( L = \begin{pmatrix} \prod_{j} \alpha_{0j}^{1/k} & 0 & \cdots & 0 \\ 0 & \prod_{j} \alpha_{1j}^{1/k} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \prod_{j} \alpha_{mj}^{1/k} \end{pmatrix}^{2} \)

The full conditionals for \( Z \) and \( \xi \) are the same as those defined in equations 3.14 and 3.17.

Then, Gibbs sampling is used to obtain samples of \( Z, \theta, \xi \), and \( \Sigma \).

### 3.3.2 Hierarchical model

The proposed hierarchical model incorporating a general ability and several specific ability dimensions differs from the additive model in that the general ability does not directly affect the examinee’s response to a test item. Instead, it affects the response through each of the specific ability.

Again, for a \( K \)-item test consists of \( m \) subtests, each containing \( k_v \) multiple choice items, where \( v = 1, 2, \ldots, m \), let \( y_{vij} \) denote the binary response for the \( i \)-th examinee on the \( j \)-th item of the \( v \)-th subtest, where \( i=1,2,\ldots,n \) and \( j=1,2,\ldots,k_v \). Then the proposed 2 parameter hierarchical model is defined as follows:

\[ P(y_{vij} = 1) = \Phi(\alpha_{vij} \theta_{vi} - \gamma_{vij}), \quad (3.25) \]

where \( \alpha_{vij} \) and \( \gamma_{vij} \) are item parameters and \( \theta_{vi} \) is the specific ability parameter corresponding to the \( v \)-th subtest. The general form of this model is the same as the multi-unidimensional model presented in equation 3.4 and hence has exactly the same

\footnote{For a detailed illustration of the full conditionals and derivation of the \( L \) matrix, please refer to Appendix A.}
likelihood function as defined in equation 3.5. Based on different beliefs of the relation between the general and specific abilities, two hierarchical models are proposed in the study.

3.3.2.1 Hierarchical model 1

First, let’s assume that the specific abilities are linear combinations of the general ability so that \( \theta_{ij} = \beta_i \theta_{0i} + \varepsilon_{ij} \), where \( \varepsilon_{ij} \sim N(0,1) \). With priors \( \alpha_{ij} > 0 \), \( p(\gamma_{ij}) \propto 1 \), \( \theta_{0i} \sim N(0,1) \), \( \beta_i \sim N(1,1) \), random variables \( Z_{vij} \) are again introduced so that

\[
y_{ij} = \begin{cases} 1, & \text{if } Z_{vij} > 0 \\ 0, & \text{if } Z_{vij} \leq 0 \end{cases}
\]

Let \( \theta_v = (\theta_{01}, \theta_{02}, \ldots, \theta_{0m})' \) denote the vector of specific abilities for the \( i \)-th person and \( \xi_{vij} = (\alpha_{ij}, \gamma_{ij}) \) the vector of item parameters for the \( j \)-th item of the \( v \)-th subtest. The full conditionals for \( Z, \theta, \theta_0, \xi, \) and \( \beta \), where \( \beta = (\beta_1, \ldots, \beta_m) \), are derived as follows:

\[
Z_{vij} \mid \bullet \sim \begin{cases} N(0,\infty)(\alpha_{ij} \theta_{v} - \gamma_{ij}, 1), & \text{if } y_{ij} = 1 \\ N(-\infty,0)(\alpha_{ij} \theta_{v} - \gamma_{ij}, 1), & \text{if } y_{ij} = 0 \end{cases}
\] (3.26)

\[
\theta_v \mid \bullet \sim N_m((A' A + I)^{-1}(\beta' \theta_0 + A' B), (A' A + I)^{-1}),
\] (3.27)

where \( A = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \alpha_m \end{pmatrix} \) and \( B = \begin{pmatrix} Z_{1i} + \gamma_1 \\ Z_{2i} + \gamma_2 \\ \vdots \\ Z_{mi} + \gamma_m \end{pmatrix} \)

\[
\theta_{0i} \mid \bullet \sim N((1 + \beta \beta')^{-1}(\beta' \theta_0), (1 + \beta \beta')^{-1})
\] (3.28)

\[
\xi_{vij} \mid \bullet \sim N((x_v' x_v)^{-1} x_v' Z_v, (x_v' x_v)^{-1}) I(\alpha_{ij} > 0) I(\alpha_{ij} > 0),
\] (3.29)

where \( Z_v = [Z_{vij}]_{n \times k_v} \), \( \xi_v = (\xi_{v1}, \ldots, \xi_{vk_v})' \), \( x_v = [\theta_v, \cdots, 1]' \), and \( \theta_v = (\theta_{v1}, \ldots, \theta_{vm})' \), \( v = 1, \ldots, m \).

\[
\beta \mid \bullet \sim N((\theta_0' \theta_0 + I)^{-1}(\theta_0' \theta_0 + I), (\theta_0' \theta_0 + I)^{-1}),
\] (3.30)
where $\theta_0 = (\theta_{01}, \ldots, \theta_{0n})'$, $\theta = (\theta_1, \ldots, \theta_n)'$, $I$ is a $m \times m$ identity matrix and $1$ is a row vector with $m$ 1's, i.e., $1 = (1, \ldots, 1)$. Then, with the full conditionals in closed forms, Gibbs sampling is used to update the samples to obtain the posterior estimates for the parameters of interest.

### 3.3.2.2 Hierarchical model 2

Alternatively, it can be assumed that the general ability is a linear combination of all the specific abilities so that $\theta_{0i} = \sum \beta_i \theta_{si} + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. Assume $\alpha_{ij} > 0$, $p(\gamma_{ij}) \propto 1$, $\beta_i \sim N(1, 1)$, and $\theta_i \sim N(0, R)$, where $R$ is a correlation matrix as described in Section 3.2.2. Then an inversed Wishart prior is assumed for the unconstrained covariance matrix $\Sigma$ transformed from $R$, $\Sigma \sim W^{-1}(0, 2)$. The full conditional distributions for $\theta_i$, $\theta_{0i}$, $\beta$, and $\Sigma$, where $\beta = (\beta_1, \ldots, \beta_m)$ and $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{in})'$, are derived as follows:

$$
\begin{align*}
\theta_i | \bullet & \sim N_m((R^{-1} + A'A + \beta'\beta)^{-1}(\beta'\theta_{0i} + A'\beta), (R^{-1} + A'A + \beta'\beta)^{-1}), \\
\theta_{0i} | \bullet & \sim N(\sum \beta_i \theta_{si}, 1) \\
\beta' | \bullet & \sim N((\theta' \theta + 1)^{-1}(\theta'\theta_{0} + 1'), (\theta' \theta + 1)^{-1}), \\
\Sigma | \bullet & \sim W^{-1}(\sum_{i=1}^{n} (L\theta_i)(L\theta_i)'),
\end{align*}
$$

where $A$ and $B$ are as defined in equation 3.27.

$$
\begin{align*}
(\prod_j \alpha_{ji}) \theta_{0i} & = 0, \\
(\prod_j \alpha_{kj}) \theta_{0k} & = 0, \\
(\prod_j \alpha_{kj}) \theta_{0j} & = 0, \\
(\prod_j \alpha_{kj}) \theta_{0i} & = 0.
\end{align*}
$$

where $S^* = \sum_{i=1}^{n} (L\theta_i)(L\theta_i)'$ and $L = \left( \begin{array}{cccc} (\prod_j \alpha_{ji})^{1/e_i} & 0 & \ldots & 0 \\
0 & (\prod_j \alpha_{kj})^{1/e_j} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & (\prod_j \alpha_{kn})^{1/e_n} \end{array} \right)$ (see Lee).
(1995) for a detailed derivation of the full conditional for $\Sigma$ and iterative procedure updating the correlation matrix $R$). The full conditional distributions for $Z$ and $\xi$ are the same as those derived in 3.26 and 3.29. Similarly, Gibbs procedure is used to iteratively update posterior samples for the parameters of interest.

3.4 Method

The two classes of IRT models with one general ability and several specific ability dimensions are assessed under different simulated situations. The posterior estimates are compared with the values obtained from bifactor analysis by using TESTFACT to compare the two estimation procedures. Furthermore, the proposed models are implemented to a subset of the CBASE English data via the Gibbs sampling procedure illustrated earlier. R statistics is adopted to assess convergence. In addition, to evaluate the Monte Carlo uncertainty, batch mean and standard errors obtained based on the posterior estimates from a single long chain are examined. Finally, Bayesian model comparisons are performed to compare the proposed model with the UIRT or the multi-unidimensional IRT model.

3.4.1 Sample

To illustrate the procedure, a subset of CBASE English subject data is used. The overall exam contains 41 English items, which consist of 16 multiple choice questions on writing and 25 on reading/literature. The data used in this study were from college students who took the LP form of CBASE in years 2001 and 2002. After removing all those who attempted the exam multiple times and removing missing responses, a sample of 1,231 examinees was randomly selected.

3.4.2 Assessing convergence and Monte Carlo variances
Two difficulties arise from the simulation procedures: how long to burn-in the simulation and how to account for the less precise inference resulting from within-sequence correlations. To address the first problem, one can simulate multiple chains with starting values widely dispersed throughout the parameter space (e.g., random draws from prior distribution). In other words, one can compare the variance within and between simulated chains until they are roughly equal by using the R statistic (Gelman, Carlin et al., 2004). Suppose $\psi$ (e.g., either $\xi$ or $\theta$ for the IRT model) is the parameter of interest. With $K$ chains of length $2N$ and the first $N$ being discarded, the $N$ by $K$ samples of $\psi$ can be denoted as $\psi_{ij}$ where $i=1,\ldots,N$, and $j=1,\ldots,K$. Define the between sequence variance as $B = \frac{N}{K-1} \sum_{j=1}^{K} (\bar{\psi}_j - \bar{\psi})^2$ and within sequence variance as $W = \frac{1}{K} \sum_{j=1}^{K} s_j^2$, where $s_j^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\psi_{ij} - \bar{\psi}_j)^2$. The R statistic for assessing convergence is then estimated as $\hat{R} = \frac{1}{\sqrt{W}} \left( \frac{\text{vár}(\psi | y)}{W} - 1 \right)$, where $\text{vár}(\psi | y) = \frac{N-1}{N} W + \frac{1}{N} B$.

If $\hat{R} > 1$ then further simulations should improve convergence.

From the Gibbs sampling procedure, one obtains a single chain with length $N$ and can estimate the mean $E(\psi | y)$ with $\hat{E}(\psi | y) = \frac{1}{N} \sum_{i=1}^{N} \psi^{(i)}$ and the variance with $\text{vár}_{\text{id}}(\hat{\psi}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\psi^{(i)} - \hat{\psi}_N)^2$. However, with the Gibbs procedure, the latter is obviously an underestimate of the Monte Carlo variance of $\hat{\psi}$ since there is inevitably autocorrelation among the samples in one MCMC chain. The less precise inference generated from autocorrelation can be removed by keeping every $k$-th sample from each chain. However, such sub-sampling of Markov chains always increases the variance of the sample mean estimates. Patz & Junker (1999) suggested a method for estimating
Monte Carlo standard errors by simulating several chains from independent starting values and then calculating the standard deviations of the chains’ sample means (MC SE’s) (1999). A better approach to get a more precise estimate of MC variance is batching (Carlin & Louis, 2000). One can divide a single long chain of length \( N \) into \( m \) successive batches of length \( a \ (N=m*a) \) with batch means \( B_1, B_2, \ldots B_m \). Then, the expectation is estimated by \( \hat{\psi}_N = \overline{B} = \frac{1}{m} \sum_{i=1}^{m} B_i \) and the variance by

\[
\text{var}_{\text{batch}} (\hat{\psi}_N) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (B_i - \overline{B})^2.
\]

As long as \( a \) is large enough that the correlation between batches is small and \( m \) is large enough, this is a reliable estimate of MC variance.

### 3.4.3 Bayesian model choice techniques

With several candidate models, it is natural to compare them using likelihood ratio tests or other information criteria from the frequentist’s perspective. Likewise, in the Bayesian framework, model comparison/selection is made possible with several criteria, among which, Bayes factors, Bayesian deviance and posterior predictive model checks are to be considered in the study.

#### 3.4.3.1 Bayes factor

When a set of \( s \) different Bayesian hierarchical models \( M_1, \ldots, M_s \) are considered, the Bayes factor for comparing model \( M_i \) and \( M_j \) is defined as

\[
BF = \frac{p(y \mid M_i)}{p(y \mid M_j)},
\]

where

\[
p(y \mid M_i) = \int p(y \mid \theta) p(\theta \mid M_i) d\theta
\]

is the marginal probability of the data \( y \) (also referred to as the prior mean of the likelihood), and \( p(\theta \mid M) \) is the prior density for the unknown parameters under the specific model \( M \). This is the Bayesian analogue of the likelihood
ratio between two models, and describes the evidence provided by the data in favor of \( M_i \) over \( M_j \). The Bayes factor allows comparison of non-nested models and ensures consistent results for model comparisons, but is usually difficult to calculate due to the difficulty in exact analytic evaluation of the marginal density of the data. Some approximation methods, such as Laplace integration, the Schwarz criterion, and reversible jump, etc. have been proposed and developed (see Kass & Raftery (1995) for a detailed description). In more complex modeling situations, the method using MCMC provides another approximation for the marginal density. Although it is unstable, researches show that it often produces results that are accurate enough for interpreting the Bayes factors (e.g., Carlin & Chib, 1993) and therefore it is used in this study.

To estimate the marginal density, one can draw MCMC samples of the parameters, \( \xi^{(g)} \), \( \ldots \), \( \xi^{(G)} \) and \( \theta^{(1)} \), \( \ldots \), \( \theta^{(G)} \) so that \( p(y \mid M) \) is approximated as \( \frac{1}{G} \sum_{g=1}^{G} L(y \mid \theta^{(g)}, \xi^{(g)})^{-1} \). It is defined as the harmonic mean of the likelihood values (Newton & Raftery, 1994).

In addition, the Bayes factor is not well defined for improper priors, which leads to the Lindley Paradox (Kass & Raffery, 1995). To overcome this problem, Aitkin (1991) proposed a posterior Bayes factor \( PBF = \frac{p^*(y \mid M_i)}{p^*(y \mid M_j)} \), where

\[
p^*(y \mid M) = \int_{\theta} L(y \mid \theta) p(\theta \mid y, M) d\theta
\]

is the posterior mean of the likelihood. To approximate this marginal density, one can also use the posterior samples so that

\[
p^*(y \mid M) = \frac{1}{G} \sum_{g=1}^{G} L(y \mid \theta^{(g)}, \xi^{(g)})
\]

3.4.3.2 Bayesian deviance
Bayesian deviance information criterion (DIC) was introduced by Spiegelhalter et al. (1998) who generalized the classical information criteria to one that is based on the posterior distribution of the deviance. This criterion is defined as $\text{DIC} = \overline{D} + p_D$, where $\overline{D} \equiv E_{\theta_0}(D) = E(-2 \log L_{\theta_0}(y \mid \theta))$ is the posterior expectation of the deviance and $p_D = E_{\theta_{\overline{D}}}(D) - D(E_{\theta_{\overline{D}}}(\theta)) = \overline{D} - D(\overline{\theta})$ is the effective number of parameters (Carlin & Louis, 2000). Further, $D(\overline{\theta}) = -2 \log(L(y \mid \overline{\theta}))$, where $\overline{\theta}$ is the posterior mean. To compute Bayesian DIC, MCMC samples of the parameters, $\xi^{(1)}, \ldots, \xi^{(G)}$ and $\theta^{(1)}, \ldots, \theta^{(G)}$, can be drawn with the Gibbs procedure, then $\overline{D}$ could be approximated as

$$\overline{D} = \frac{1}{G} \left(-2 \log \prod_{g=1}^{G} L(y \mid \theta^{(g)}, \xi^{(g)})\right).$$

Generally the more complicated models tend to provide better fit. Hence, penalizing for number of parameters makes DIC a more reasonable measure to use. However, unlike the Bayes factor, DIC is not invariant to parameterization and sometimes can produce unrealistic results.

### 3.4.3.3 Posterior predictive model checks

Among the methods proposed for model checking, posterior predictive checking is easy to carry out and interpret in spite of its limitation in being conservative (Sinha ray & Stern, 2003). The basic idea is to draw simulated values from the posterior predictive distribution of replicated data, $y_{\text{rep}}$, $p(y_{\text{rep}} \mid y) = \int \int p(y_{\text{rep}} \mid \xi, \theta) p(\xi, \theta \mid y) d\xi d\theta$, and compare them to the observed data $y$. If the model fits, then replicated data generated under the model should look similar to the observed data. A test statistic $T(y_{\text{rep}}, (\xi, \theta))$ is chosen to define the discrepancy between the model and the data. If there are $L$ simulations from the posterior distribution of $(\xi, \theta)$, one $y_{\text{rep}}$ can be drawn from the predictive distribution for each simulated $(\xi, \theta)$ so there are $L$ draws from the joint
posterior distribution \( p(y^{rep}, \xi, \theta \mid y) \). It is then easy to compare the realized test statistics \( T(y, (\xi^i, 0^i)) \) with the predictive test statistics \( T(y^{rep}, (\xi^i, 0^i)) \) by plotting the pairs on a scatter plot. Alternatively, one can calculate the probability or posterior predictive p-value (PPP-value) (Sinharay & Johnson, 2003) that the replicated data could be more extreme than the observed data: 

\[
p_B = \Pr(T(y^{rep}, (\xi, 0)) \geq T(y, (\xi, \theta) \mid y)).
\]
CHAPTER 4.

RESULTS

This chapter is organized as follows. Section one summarizes the results from the simulation studies conducted using several additive models as well as the TESTFACT program. Section two shows the simulation results for the two hierarchical models. In the third section, the proposed additive and hierarchical models are implemented to the CBASE data and model comparisons are carried out to compare different IRT models, i.e., the unidimensional, multi-unidimensional and the proposed additive and hierarchical models.

4.1 Simulation Studies for the Proposed Additive Model

For the proposed additive model, each test item is assumed to measure two ability dimensions, namely, a general and a specific ability dimension, directly. This is reflected in the probability function of the model defined in equation 3.7:

\[ P(y_{ij} = 1) = \Phi(\alpha_{0i} \theta_{0i} + \alpha_{1i} \theta_{si} - \gamma_y). \]

The additive nature of the latent traits in the model leads to a potential problem of indeterminancy when item parameter and person abilities are estimated simultaneously. In the hierarchical Bayesian framework, although informative priors are specified for the ability parameters to help the convergence of Markov chains, it is still uncertain how the Bayesian additive model performs in various scenarios as well as compared with the TESTFACT program, which adopts marginal maximum likelihood method to estimate parameters in bifactor models. To evaluate the additive model, five simulations were conducted. First, 41 item parameters, \( \alpha_0, \alpha_1, \gamma \) were randomly simulated from a uniform
distribution so that $P(\alpha_0) > 0$, $P(\alpha_i) > 0$ and $P(\gamma) \neq 1$, and they were used in all the five simulation studies. In the simulations, tests with one general ability and two specific abilities were considered, i.e., $m=2$. Then, 300 ability parameters, $\theta_i$, were simulated from $N_{m+1}(0, R_0)$, where $R_0$ is a correlation matrix, which was specified differently based on various actual scenarios, so that:

**Simulation 1** assumed zero correlations between all latent abilities in the model so that the correlation matrix was $R_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

**Simulation 2** specified that the general ability correlated highly with the two specific abilities whereas no relation was assumed between the specific abilities, $R_0 = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0 \\ 0.6 & 0 & 1 \end{pmatrix}$.

**Simulation 3** was the opposite to the second simulation. No correlation was assumed between the general ability and each specific ability whereas there was a fairly high correlation between the two specific ability dimensions, i.e., $R_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.6 \\ 0 & 0 & 1 \end{pmatrix}$.

**Simulation 4** specified the scenario where the general ability was highly correlated with only one specific ability and all other correlations were set as zeros, $R_0 = \begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

---

Note here, $m$ was set to two, so $\theta_i = (\theta_{0i}, \theta_{1i}, \theta_{2i})$. 

Simulation 5 specified that all three latent traits were supposed to be moderately correlated and the correlation coefficients were set to be equal, \[ R_0 = \begin{pmatrix} 1 & 0.5 & 1 \\ 0.5 & 0.5 & 1 \\ 1 & 0.5 & 1 \end{pmatrix} \]

Next, the response matrix \( Y \) was simulated based on person ability and item parameters in each simulation. With the simulated \( Y \), four Bayesian additive models together with TESTFACT were used to recover parameters as well as latent structures. The four additive models differ in their priors for \( \theta_i \) and are briefly described as follows:

**Known prior:** In this model, the ability parameters were assumed to have a multivariate normal distribution, \( N_{m+1}(0, R_0) \), where \( R_0 \) was exactly the same as the actual correlation matrix as specified in each simulation. This model can only be adopted in the simulation studies where the relationship between the latent abilities is known. In real test situations where the actual latent structure is not certain, it is not possible to adopt models with known prior. Additionally, with the known correlation matrix set as the prior, this model also helps with detection of any computational problem in the Bayesian additive models.

**Model 1:** The prior for the person abilities was \( \theta_i \sim N_{m+1}(\mu_i, I) \) and \( \mu \sim N_{m+1}(0, \Sigma) \). The detailed full conditionals for all parameters in the model are defined in chapter 3.

**Model 2:** This model was similar to the previous one with the prior as \( \theta_i \sim N_{m+1}(\mu_i, I) \) and \( \mu \sim N_{m+1}(0, \Sigma) \).

**Model 3:** In this model, \( \theta_i \sim N_{m+1}(0, R) \), where \( R \) was a correlation matrix. The model is also introduced in chapter 3 and the detailed full conditionals are derived in Appendix A.
For the above four additive models, Gibbs sampling was implemented where 7,000 iterations were obtained with the first 2,000 as burn-in. The posterior estimates were obtained as the average of the Gibbs samples and the results for the five simulation studies are summarized in the following two sections.

4.1.1 Results with the item parameters

With 41 items, four model specifications and five simulations, the amount of information was too large to present in one single table. Hence the correlations between the estimated parameters with their true values as well as the sum of their squared differences were adopted to illustrate parameter recovery. Denote $\alpha_0$ and $\alpha_1$ vectors of the item discriminations associated with the general ability and specific abilities respectively. Figure 1 plots the correlations between the recovered and true item parameters using the four Bayesian additive models as well as TESTFACT in each simulation.

Consider each simulation separately and it is observed that:

In Simulation 1, all correlations were around or above 0.9, with those for $\gamma$ being the highest (about 0.98). The recovered item parameters using the four Bayesian models and TESTFACT were almost equally highly correlated with their actual values. A close examination of the correlation coefficients suggest that among the four Bayesian models, the one with known prior performed a little better than Models 1, 2 or 3.

In Simulation 2, the correlations, especially those for $\alpha_0$ and $\alpha_1$ dropped noticeably. The correlations between posterior and actual $\gamma$ were still above 0.95. The Bayesian model with known prior performed the best, which was not surprising, for the prior for ability parameters was set to be the same as the actual situation. The posterior estimates of abilities were pushing toward that direction, and hence the item parameters could be
recovered well. In addition, Model 1 and Model 3 performed better than Model 2 or TESTFACT.

**Figure 1.** Plots of the correlations between the actual and recovered item parameters ($a_0$, $a_1$ and $\gamma$)
using 4 different additive models or TESTFACT for the five simulation studies.

In Simulation 3, again, $\gamma$ was recovered the best, with correlations all above 0.95. The recovery for $\alpha_0$ and $\alpha_1$ was found to be similar except for Model 3, where $\alpha_0$ did not seem to be recovered well at all (with correlation about 0.54). The model with known prior performed the best, slightly better than Model 1, which was slightly better than Model 2. TESTFACT did not recover the item parameters as well as Models 1 or 2.

The plots for Simulations 4 and 5 show similar pattern, i.e., $\gamma$ was recovered the best and the model with known prior performed better than Models 1, 2 or 3, which were better than TESTFACT.

Correlations describe the overall relationship between the recovered parameters and their actual values. To illustrate how large the estimated item parameters deviate from their true values, sums of their squared differences were obtained for $\alpha_0$, $\alpha_1$ and $\gamma$ from each model and summarized in Table 1. Generally, $\alpha_1$ and $\gamma$ were fairly stable across models and from simulation to simulation, except that TESTFACT always had the largest squared deviation values in all five simulations. $\alpha_0$ had quite obvious fluctuations. In Simulations 1 and 3, Models 1, 2 and 3 had small deviation values for $\alpha_0$, which was close to those from the model with known priors, whereas in Simulations 2, 4 and 5, those values were considerably larger for Models 1, 2, 3 than that with known priors. Further, among the three Bayesian models, i.e., Models 1 to 3, Model 1 always had the smallest deviation values. Plus, the deviation values for $\gamma$ were uniformly smaller than those for $\alpha_0$ or $\alpha_1$. 

Table 1.
Sum of squared difference between the actual and recovered item parameters ($\alpha_0$, $\alpha_1$ and $\gamma$) using different additive models or TESTFACT under five simulated scenarios.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Known prior</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>TESTFACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1</td>
<td>$\alpha_0$</td>
<td>1.2908</td>
<td>1.1758</td>
<td>1.4425</td>
<td>1.3229</td>
</tr>
<tr>
<td>Simulation 1</td>
<td>$\alpha_1$</td>
<td>1.6755</td>
<td>1.2842</td>
<td>1.7182</td>
<td>1.5919</td>
</tr>
<tr>
<td>Simulation 1</td>
<td>$\gamma$</td>
<td>0.8507</td>
<td>0.9084</td>
<td>0.8092</td>
<td>0.8705</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>$\alpha_0$</td>
<td>2.6298</td>
<td>12.7078</td>
<td>14.1511</td>
<td>12.2820</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>$\alpha_1$</td>
<td>1.2461</td>
<td>1.1167</td>
<td>3.1337</td>
<td>1.4364</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>$\gamma$</td>
<td>0.6548</td>
<td>0.9221</td>
<td>1.0280</td>
<td>0.9918</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>$\alpha_0$</td>
<td>0.6055</td>
<td>0.9571</td>
<td>1.4610</td>
<td>4.2944</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>$\alpha_1$</td>
<td>1.4015</td>
<td>1.1907</td>
<td>1.1950</td>
<td>1.4982</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>$\gamma$</td>
<td>0.5147</td>
<td>0.5422</td>
<td>0.5585</td>
<td>0.5577</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>$\alpha_0$</td>
<td>1.8322</td>
<td>6.3476</td>
<td>6.9956</td>
<td>7.1878</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>$\alpha_1$</td>
<td>2.9846</td>
<td>1.4949</td>
<td>1.5486</td>
<td>1.5707</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>$\gamma$</td>
<td>0.5712</td>
<td>0.6188</td>
<td>0.6127</td>
<td>0.6258</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>$\alpha_0$</td>
<td>2.0847</td>
<td>7.3830</td>
<td>9.1526</td>
<td>9.1310</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>$\alpha_1$</td>
<td>1.2290</td>
<td>1.2343</td>
<td>1.2214</td>
<td>1.2061</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>$\gamma$</td>
<td>0.4089</td>
<td>0.5062</td>
<td>0.4490</td>
<td>0.4107</td>
</tr>
</tbody>
</table>

From the above observations, the following conclusions can be drawn:

- The $\gamma$ parameters are always well recovered and hence they are not affected by various actual structures existing in the latent traits.

- On the contrary, $\alpha_0$ and $\alpha_1$ are affected by the actual structure of the abilities because they are slopes for the corresponding abilities in the model. It is further noticed that when there is no correlations between the general ability and each specific ability, $\alpha_0$ and $\alpha_1$ are recovered well, as shown in Simulation 1 and Simulation 3. However, when the general ability is correlated with the specific abilities, the slopes are less well recovered. Further, a comparison between Simulations 2, 4, and 5 indicates that the higher the correlations between the general ability $\theta_0$ and specific abilities $\theta_1$ or $\theta_2$, the less well the item parameters are recovered.
• TESTFACT works well only when there is no correlations between all latent ability dimensions although it performed no better than the fully Bayesian models, as is shown in Simulation 1.

• When comparing the four Bayesian additive models, it is easily seen that the model with known prior performs the best in all five simulations. This further confirms that no computational problem occurred during the implementation of the Gibbs sampling procedure. Moreover, Model 1 performs consistently better than Models 2 or 3. Finally, the models with fully Bayesian methodology are found to recover item parameters better than TESTFACT, where MML is implemented for parameter estimation.

These results are based on the item parameters. It would also be interesting to check the performance of each model on the recovery of the ability parameters and the recovery of their relations.

4.1.2 Results with the person ability parameters

Likewise, the posterior estimates for the abilities were obtained and correlated with the corresponding true values. The accuracy of the recovered person abilities was not assessed, however, for in educational testing situations, one is usually interested only in the relative values of $\theta$, instead of the true values for different examinees. That is, how each person performs compared with other examinees. In addition, the correlations between the posterior ability estimates were obtained as well to describe how well each model recovered the latent structure. Table 2 summarizes all the correlation results, where the underlined values (correlations between the actual abilities and posterior estimates) are to check if the abilities were recovered well whereas the bolded ones (correlations between the posterior ability estimates) are to check if the latent trait
structure could be recovered well in each simulation. It has to be noted that TESTFACT only reports the estimates for the general ability. Hence, the correlations between the actual and recovered abilities were computed only for the general ability (last column in Table 2).

A few remarks can be made after a close examination of and comparison between the correlations of interest for the models considered in the five simulations.

♦ The model with known prior always outperforms Models 1, 2 or 3 in either recovering the person abilities or the relations among those abilities, as was expected and noted earlier. In addition, as far as the general ability is concerned, the Bayesian additive models are recovering the ability parameters equally well or even better compared with TESTFACT.

♦ When there are no correlations between the general ability and each specific ability, both ability parameters and their relations get recovered better (Simulations 1 and 3) than when there are correlations between them (Simulations 2 and 5). Among the three models, Model 1 performs clearly better than Model 3 or Model 2, as shown in Simulation 3. Further, the higher the true correlations, the worse the Bayesian additive models recover the ability parameters as well as their correlations.

♦ In general, $\theta_2$ is recovered better than $\theta_1$ and the correlation between $\theta_0$ and $\theta_2$ is recovered better than that between $\theta_0$ and $\theta_1$ as well. This could be due to the reason that there are more items and hence more information in the second subtest than the first.

♦ In all five simulation studies, the general ability $\theta_0$ is recovered better than both specific abilities no matter what model or method is used.
More interestingly, when there are correlations between the general ability and specific abilities, the recovered correlations between their posterior estimates are pushing toward zero.

Table 2.
Correlations between the actual person ability parameters \((\theta_0, \theta_1, \theta_2)\) and posterior estimates for the ability parameters \((\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)\) using 4 different additive models or TESTFACT under five simulated scenarios.
The three additive models considered in this section were similar with slightly different model specifications. Having relatively better parameter as well as dimensional structure recovery in the simulation studies, Model 1 was chosen as a representative additive model in the subsequent analyses.

4.2 Simulation Studies for the Proposed Hierarchical Model

Different from the additive model, the proposed hierarchical model assumes that an examinee’s response to a particular item is directly affected by a specific ability, which is either a linear function of the general ability or linearly combines to form the general ability. This probability function of the model is defined in equation 3.25:

\[ P(y_{ij} = 1) = \Phi(\alpha_i \theta_i - \gamma_{ij}), \]

which is much simpler than the additive model whose probability function is defined in 3.7 and hence avoids the potential problem of model indeterminancy. The two hierarchical models are based on different beliefs on the relationship between the general ability and specific abilities. That is, Model 1 assumes that each specific ability is a linear function of the general ability, \( \theta_i = \beta_i \theta_{0i} + \epsilon_i \), whereas Model 2 assumes that the general ability is a linear combination of the specific abilities, \( \theta_{0i} = \sum_i \beta_i \theta_i + \epsilon_i \). In Model 2, the correlations between the specific ability dimensions are realized through the specification of the prior, \( \theta_j \sim N(0, R) \). On the other hand, the correlations between the specific abilities are realized through their relationships with the general ability in Model 1. Due to the different assumptions on latent ability structures, each model was found to perform better in recovery than the other in its true situations. However, in order to compare the two hierarchical models, none of their true situations were considered. Instead, seven simulations were conducted where the correlations between
the abilities were specified differently. In the seven simulations, tests with one general ability and two specific abilities were considered, i.e., \( m=2 \). Similar to the procedure in the previous section, 41 item parameters, \( \alpha \) and \( \gamma \), were randomly generated from a uniform distribution so that \( P(\alpha) > 0 \) and \( P(\gamma) \propto 1 \). Then, 300 ability parameters, \( \theta_i \), were simulated from \( N_{m+1}(0, R_0) \) to construct the response matrix \( Y \). In the seven simulations, \( R_0 \) was specified differently so that:

Simulation 1 assumed zero correlations between all latent abilities in the model.

Simulation 2 assumed zero correlation between the general ability and each specific ability whereas a moderate correlation (0.6) between the two specific ability dimensions,

\[
R_0 = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0.6 \\
0 & 0 & 1
\end{pmatrix}.
\]

Simulation 3 specified the scenario where the general ability was highly correlated (0.8) with only one specific ability and all other correlations were set as zeros,

\[
R_0 = \begin{pmatrix}
1 & 0.8 & 0 \\
0.8 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

Simulation 4 assumed high correlations between the general ability and the two specific abilities whereas no correlation between the two specific abilities,
Simulation 5 was similar to the previous simulation except that there was a 0.5 correlation between the two specific abilities, \( R_0 = \begin{pmatrix} 1 \\ 0.8 \\ 0.6 \end{pmatrix} \). 

Simulation 6 set a slightly higher correlation (0.7) between the two specific abilities, 

\[
R_0 = \begin{pmatrix} 1 \\ 0.8 \\ 0.6 \end{pmatrix}, 
\]

with everything else the same as Simulations 4 and 5.

Simulation 7 had an even higher correlation (0.9) between the specified abilities, 

\[
R_0 = \begin{pmatrix} 1 \\ 0.8 \\ 0.6 \end{pmatrix} 
\]

With the simulated \( Y \), the Gibbs sampling procedure was implemented with the two hierarchical models to obtain posterior estimates of the item as well as person parameters. 7,000 iterations were obtained with the first 2,000 set as burn-in. The simulation results are summarized in the following two sections.

4.2.1 Results with the item parameters

For the 41 simulated item parameters, the correlations between the estimated parameters with their true values as well as the sums of their squared differences were again obtained to describe how well each model recovered the item parameters. Figure 2 plots the correlations between the recovered and true item parameters using the two hierarchical models in each simulation.

It is obvious from the figure that \( \gamma \) was more stable than \( \alpha \). Both models had uniformly high correlations for \( \gamma \) (all about 0.99) in the seven simulations. The correlations for \( \alpha \)
were relatively more fluctuating and the values were within the range of (0.92, 0.95).

Specifically, for \( \alpha \), the two models had similar correlations in Simulations 1, 2, 3, 4 and 6 whereas in Simulations 5 and 7, Model 2 had slightly higher correlations than Model 1.

Across the 7 simulations, Simulations 2 and 4 had the highest correlations for \( \alpha \).

![Figure 2](image.png)

**Figure 2.** Plot of the correlations between the actual and recovered item parameters (\( \alpha \) and \( \gamma \)) using the 2 hierarchical models for the seven simulation studies.

Furthermore, the sums of the squared differences between the estimated and actual item parameters were obtained for \( \alpha \) and \( \gamma \) from each model and summarized in Table 3.

Again, \( \gamma \) was stable across the seven simulations for the two models, with sum of squared differences around 0.3. On the other hand, those for \( \alpha \) varied. In Simulations 1 and 4, the two models had similar total deviation values for \( \alpha \). In Simulations 2, 3 and 5, Model 1 had lower values than Model 2. Then in Simulations 6 and 7, Model 1 had much larger sums of squared deviations than Model 2 (4.3235 compared with 0.7823 in the last simulation). When comparing the deviations for \( \alpha \) across the seven simulations, the total deviation values for Model 1 were consistently increasing from Simulation 4 to
Simulation 7. However, with Model 2, those deviation values were the lowest in Simulations 1 and 4 (about 0.5), and the highest in Simulation 3 (1.3087).

It can hence be summarized from the above observations that:

♦ Item difficulty parameters, $\gamma$, are very stable and can be recovered extremely well in different scenarios.

♦ Item discriminations, $\alpha$, are less stable. For both models, they are better recovered in Simulations 1 and 4, where zero correlation between the specific abilities is assumed and the correlations between the general ability and each specific ability are of similar magnitude.

♦ As far as item discriminations are concerned, Model 2 works fine under different situations. However, Model 1 seems not to work well in situations where there are nonzero correlations between the general ability and specific abilities and they are smaller than the correlation between the specific abilities (e.g., Simulations 6 and 7). Actually, the smaller the correlations between the general ability and specific abilities than the correlation between the specific abilities, the less likely Model 1 recovers $\alpha$ well.

Table 3.
Sum of squared difference between the actual and recovered item parameters ($\alpha$ and $\gamma$) for Models 1 and 2 under seven simulated scenarios.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
<th>Simulation 5</th>
<th>Simulation 6</th>
<th>Simulation 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.5267</td>
<td>0.6961</td>
<td>0.8013</td>
<td>0.6802</td>
<td>0.8772</td>
<td>1.8844</td>
<td>4.3235</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.2712</td>
<td>0.2614</td>
<td>0.3731</td>
<td>0.3278</td>
<td>0.3725</td>
<td>0.3164</td>
<td>0.3265</td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.5299</td>
<td>1.0340</td>
<td>0.8298</td>
<td>0.5951</td>
<td>1.3087</td>
<td>0.9803</td>
<td>0.7823</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.2684</td>
<td>0.2737</td>
<td>0.3909</td>
<td>0.3099</td>
<td>0.3618</td>
<td>0.3163</td>
<td>0.3261</td>
</tr>
</tbody>
</table>

4.2.2 Results with the person ability parameters
Next, the posterior person abilities were correlated with their actual values (they are shown as the underlined values in Table 4) to describe how well the abilities were recovered. The correlations between the recovered abilities were calculated as well (the bolded values in Table 4) to examine how well each model recovers the latent structure.

Table 4.
Correlations between the actual person ability parameters ($\theta_0$, $\theta_1$, $\theta_2$) and posterior estimates for the ability parameters ($\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$) for Models 1 and 2 under seven simulated scenarios.

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
<th>Simulation 5</th>
<th>Simulation 6</th>
<th>Simulation 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_0$</td>
<td>-0.1054</td>
<td>0.004</td>
<td>0.4283</td>
<td>0.0861</td>
<td>-0.9088</td>
<td>0.5015</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.7884</td>
<td>-0.0282</td>
<td>0.6772</td>
<td>0.0044</td>
<td>0.5696</td>
<td>1.0004</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0628</td>
<td>-0.0141</td>
<td>0.0609</td>
<td>0.6082</td>
<td>0.9185</td>
<td>0.0861</td>
</tr>
</tbody>
</table>

Checking how well the person abilities were recovered, one can see from the table that both models recovered the specific abilities equally and consistently well, with
correlations all above 0.8, in all the seven simulations. However, they did not recover $\theta_0$ well in Simulations 1, 2 and 3, especially in the first two simulations where the correlations were about 0. Further, in Simulations 4 through 7, the correlation between the recovered and actual general ability reduced from over 0.8 to 0.64. Then with the bolded correlations, both models performed equally well in recovering the correlation between the specific abilities. It is the correlations between $\theta_0$ and $\theta_1$ or $\theta_2$ that had more variations. In effect, they were not well recovered in Simulations 1, 2 and 3. But in the last four simulations, the recovery of the dimensional structure was quite satisfactory. The results of the two models did not differ much except in Simulation 4, where Model 2 recovered the correlations between the abilities better than Model 1.

With the above observations, several conclusions can be drawn as to the recovery of the person abilities as well as the underlying structure:

♦ No matter what the true situation is, the specific abilities and the correlation between the specific abilities are always recovered well using either model.

♦ The hierarchical model works well when each specific ability is moderately or highly related to the general ability. Since the general ability depends on the data as well as on the specific abilities, it incorporates more variations, i.e., those from the data and those from the specific abilities. When it is not or less highly related to the specific abilities (even if one of the correlations is high), little information could be used to estimate it. And hence neither the general ability nor the structure underlying the general ability and specific ones gets well recovered.

♦ When the general ability has moderate to high correlations with the specific abilities, the magnitude of the correlation between the specific abilities determines how well
the general ability is recovered. Specifically, the larger that correlation, the less likely the hierarchical model recovers the general ability well.

The two hierarchical models made different assumptions on the underlying ability structures. Although they performed similarly in the simulation studies, both were used in the later sections to fit the real data and carry out model comparisons. Models 1 and 2 were denoted as hierarchical model 1 and hierarchical model 2 respectively in the later sections.

4.3 Bayesian Model Choice with CBASE Data

In this section, the three proposed models, namely, the additive model, hierarchical model 1 and hierarchical model 2, were fitted to the CBASE English data via the Gibbs sampling procedure. Again, 7,000 iterations were obtained with the first 2,000 set as burn-in. To assess the convergence, the R statistic was calculated for the item parameters in each model with multiple chains and they were all found to be close to 1, suggesting that stationarity had been reached within the simulated Monte Carlo chains for the three fitted models. Then, the posterior estimates of the item parameters were obtained as the average of the Gibbs samples and are displayed in Table 5. The table also reports the Monte Carlo (MC) uncertainty estimated using batching. A close examination of the uncertainty values indicates that those for item difficulties, $\gamma$, estimated from the three models, were generally very similar, which further confirms that difficulty parameters are independent of the way the latent abilities are related. In addition, when compared with the two hierarchical models, the additive model had slightly larger SE's for the discrimination parameters, $\alpha_0$ and $\alpha_1$. But the differences were rather small.
Generally, all the standard errors for estimating the posterior means of the item parameters were small. It can be interpreted that, for example, with the additive model, an approximate 99% MC interval for the true posterior expectation for the second item’s discrimination parameter associated with the general ability was $0.3656 \pm 3 \times (0.0055)$, suggesting that MC estimate of this posterior mean was good to about two digits of
accuracy. Hence, the item parameters using the three Bayesian models were estimated with little error.

4.3.1 Model comparisons

As stated earlier, the three formerly fitted models represent different beliefs on the relationship between general ability and specific abilities. One has to determine which model describes CBASE English data better. Hence, model comparisons were carried out by using Bayesian model comparison techniques. In addition to comparisons among the three proposed models, their goodness-of-fit was evaluated relative to the much simpler unidimensional (UIRT) model and multi-unidimensional model as well.

Because Bayes factors do not produce meaningful results with non-informative priors, all the flat priors specified in the models were changed to be informative. Then, with the five candidate models, 7000 iterations were obtained with 2000 burn-ins. The results with Bayes factors, Bayesian DICs and predictive model checks are summarized in what follows.

To obtain Bayes factors, the marginal densities \( p(y | M) \) and \( p^*(y | M) \) were approximated using MCMC and are displayed in the first two columns of Table 6. Due to the reason that all the likelihoods for the data went down to zero, the values shown in the two columns of the table were some constant multiply of \( p(y | M) \) or \( p^*(y | M) \), as is noted below the table, so that when computing Bayes factors, the constant cancelled out. Bayes factors (BF) and posterior Bayes factors (PBF) are ratios of the marginal densities for comparing two models \( M_i \) and \( M_j \), i.e., \( BF = \frac{p(y | M_i)}{p(y | M_j)} \),

\[
PBF = \frac{p^*(y | M_i)}{p^*(y | M_j)},
\]

and they provide evidence in favor of \( M_i \) to \( M_j \). As a BF or PBF
beyond 100 indicates decisive evidence in favor of $M$, the additive model was found to be the best among the five models. Taking the ratio of its marginal density with that for any other models resulted in BF or PBF estimates beyond 100. Likewise, the hierarchical model 2 proved to be better than the hierarchical model 1, the unidimensional model or the multi-unidimensional model based on the BF values. Moreover, there was much evidence against the unidimensional model when comparing it to either the multi-unidimensional model or any of the three proposed models.

Table 6 also shows the Bayesian deviance results, where smaller values indicate better model fit. Among the five IRT models, the additive model had the smallest DIC and expected posterior deviance ($\bar{D}$) and the hierarchical model 2 had the second smallest values. Therefore, the additive model provided the best goodness-of-fit to the data compared with other models, even after penalizing for a large effective number of parameters ($pD=1817.3$), which is shown in the last column of the table. On the other hand, the unidimensional model was relatively worse than any of the multiple-ability models. The results are generally consistent to those obtained with Bayes factors.

Table 6. Approximated marginal densities of the data and Bayesian deviances for the five IRT models.

| Model                  | $p(y|M)$ | $p^*(y|M)$ | DIC   | $\bar{D}$    | $D(\bar{θ})$ | $p_0$ |
|------------------------|----------|------------|-------|--------------|--------------|-------|
| Unidimensional         | 1.2254E-224 | 8.55E-308  | 55639 | 54548        | 53457       | 1090.6|
| Multi-unidimensional   | 4.2856E-163 | 1.04E-207  | 55571 | 54160        | 52750       | 1410.5|
| Additive model         | 156      | 107.5633   | 55135 | 53318        | 51501       | 1817.3|
| Hierarchical model 1   | 8.0348E-177 | 4.6805E-220 | 55571 | 54188        | 52805       | 1382.7|
| Hierarchical model 2   | 2.568E-143 | 2.83E-215  | 55386 | 54121        | 52656       | 1464.6|

Note: 1. The reported values are $p(y|M)\times\exp(26840)$
2. The reported values are $p^*(y|M)\times\exp(26460)\times4000$

Next, the posterior predictive model checking procedure was implemented to compare the five candidate models. To do so, a test statistic had to be chosen for describing the
discrepancy between the model and data. For this analysis, the odds ratio was adopted for measuring association among item pairs, $T(y, (\xi, \theta)) = OR_{ij} = \frac{n_{1i}n_{0j}}{n_{0i}n_{1j}}$, which has been reported powerful for detecting unidimensionality in the data (Sinharay & Johnson, 2003). Hence, for each fitted model, based on each pair of $(\xi, \theta)$ samples, a $y_{\text{rep}}$ was simulated and the replicated discrepancy $T(y_{\text{rep}}, (\xi, \theta))$ was computed to be compared with the actual discrepancy. The tail-area PPP-values ($p_{B}$) were estimated as the proportion of the simulated samples for which $T(y_{\text{rep}}, (\xi, \theta)) \geq T(y, (\xi, \theta))$, i.e.,

$$p_{B} = \sum_{i=1}^{I} I(T(y_{\text{rep}}, (\xi, \theta)) \geq T(y, (\xi, \theta))).$$

Figure 3 summarizes the extreme PPP-values for the odds ratios with each model (here $\alpha = .05$ was used as the critical level). It is immediately clear that with far fewer extreme replicated odds ratios, the additive model performed much better than the other four models. In particular, the numbers of extreme PPP-values for the five models, namely, the unidimensional, multi-unidimensional, additive and two hierarchical models, were 39, 38, 15, 38 and 37, respectively. The additive model had far fewer extreme PPP-values and was considered to be the best among the four models. Moreover, the PPP-values for the other three models were very close, with the unidimensional model having a few more of those PPP-values.

Although the four non-additive models were similar in their numbers of extreme PPP-values, they showed somewhat different prediction errors. In the plots, the extreme PPP-values above 0.975 were differentiated from those below 0.025. The former indicates that the fitted model overpredicts the odds ratios between two items and the latter suggests underprediction. Hence, the unidimensional model mostly underpredicted the odds ratios within the two clusters and overpredicted the odds ratios between the items.
from each set. The underpredictions were outcomes of associations among the items beyond the model’s prediction capacity and the overpredictions were because some of the examinees showed higher ability for one dimension and lower ability for the other, which was not in accordance with the assumption of the unidimensional model that all examinees at a specific ability level should get all items correct within their ability level. On the other hand, the multi-unidimensional model did not show as many underpredictions within the two clusters, especially in “writing”. It performed similarly for items in the “reading\literature” cluster as the unidimensional model. However, the multi-unidimensional model underpredicted more item pairs between the two clusters. Note that these underpredictions were simply due to the reason that the examinee’s two abilities were more similar than what the multi-unidimensional model specifies, i.e., high ability in one dimension does not have to be associated with similarly high ability in the other. In addition, the two hierarchical models displayed similar patterns of prediction errors as the multi-unidimensional model.

Therefore, in conclusion, with Bayesian model checking techniques, the five candidate IRT models were evaluated as to which model provided a better description of, and hence a better goodness-of-fit to the CBASE data. The results from Bayes factors, Bayesian deviances and posterior predictive checks all provided strong evidence in favor of the proposed additive model, which was believed to fit the data conceivably better than the other four candidate models. On the contrary, the unidimensional model provided relatively the worst description of the data. Consequently, for the CBASE English data, the model comparison results did not support the more stringent unidimensionality assumption either.
Figure 3. Plots of extreme tail-area PPP-values for odds ratios with the five IRT models.
CHAPTER 5.
DISCUSSION AND CONCLUSION

This chapter contains four major sections. Section one summarizes the findings by discussing the results on the simulation studies as well as model comparisons for the proposed additive and hierarchical models. The differences between the IRT models, including the unidimensional, multi-unidimensional and the proposed additive and hierarchical models, are also illustrated graphically to describe their structural differences in the latent abilities. A conclusion is drawn based on the findings in Section two. Section three discusses the findings and the meaning of the results, together with the implications for the study. And finally, directions for the future studies are given in Section four.

5.1 Summary of Findings

The findings for the simulation studies as well as model comparisons for the \textit{CBASE English} data are summarized in the following sections.

5.1.1 Simulation studies

Simulation studies were conducted to assess the proposed IRT models incorporating one general ability and several specific ability dimensions under various actual test situations. From the simulation results summarized in Sections 4.1 and 4.2, it is suggested that the additive and hierarchical models proposed in the study apply when different beliefs are held on the underlying structure of the latent ability dimensions.

5.1.1.1 The additive model
Intuitively, the additive model induces additivity between general ability and specific abilities, proposing that the probability of endorsing an item correctly is a generalized linear function of the general ability and a specific ability. As described in Section 4.1, four additive models were used to obtain posterior estimates of the parameters under five simulated situations. The models were compared in their recovery of the model parameters as well as in the structure of the underlying latent abilities. The results were further compared with those from TESTFACT to compare two different estimating procedures, Gibbs sampling vs. EM algorithm. Based on the results with the simulation studies, the following conclusions are summarized:

First, the item difficulty, $\gamma$, is independent of the ability parameters, and hence it is not affected by different specifications of the actual structure underlying the latent traits. On the contrary, the item discrimination parameters, $\alpha_0$ and $\alpha_1$, are associated with ability parameters and hence less stable. In effect, the discrimination corresponding to the general ability, $\alpha_0$, is the least stable and is sensible to the influence of the correlations between $\theta_0$ and $\theta_1$ or $\theta_2$. Moreover, both item and person ability parameters are recovered better when the general ability is orthogonal to each of the specific abilities. This is due to the reason that the additive model specifies a generalized linear function of the general ability and a specific ability. The multicollinearity problem, i.e., high correlations between the general ability and each specific ability, problem associated with the linear models, would affect the accuracy of parameter estimation.

Second, the general ability $\theta_0$ is recovered better than all the specific abilities no matter which model specification or method was used in the five simulations. This can be explained by the reason that the general ability $\theta_0$ is measured by all test items whereas
each of the specific ability is measured by a subset of the total test items. Consequently, there is more information for $\theta_0$.

Third, the additive models implemented with Gibbs sampling are found to work well when there are no or low correlations between the general ability and each specific ability. This is also the case with the TESTFACT program. Also, Model 1 seems to be the best model among the three additive models, i.e., Models 1, 2 and 3. One can recall that the three models use different methods to get around the model indeterminacy problem, i.e., Models 1 and 2 use a relatively strong prior, identity matrix, whereas Model 3 adopts constrained covariance matrix. The better recovery of Model 1 in the simulation studies suggests using informative priors is preferred to setting constraints for the additive model. Moreover, when comparing the proposed models with TESTFACT, it is clear that TESTFACT performs less well than some proposed additive models, especially Model 1.

Finally, it is observed that the recovered correlations between the posterior estimates are pushing toward zero when the general ability and specific abilities are not orthogonal. The stronger they are related, the more the recovered values are pushing to zero. A close examination of the model specifications indicates that this is because of the prior for $\theta_i$. The identity matrix chosen as the variance-covariance matrix in the multivariate normal prior and small values in $\Sigma$ sampled from the inverse-Wishart distribution result in the posterior variance-covariance matrix having diagonal elements relatively larger than those on the off-diagonals. Hence, under the influence of small off-diagonal values, the covariance or correlation between the posterior estimates of the person abilities is pushed toward zero.

5.1.1.2 The hierarchical model
Rather different from the additive model, the proposed hierarchical model assumes that the ability dimensions form a hierarchy so that the general ability is underlying the specific abilities, which in turn are underlying the test items. In the study, two hierarchical models are proposed with different beliefs on the relationship between the general ability and specific ability dimensions. Model 1, as is graphically illustrated in Figure 4a, assumes that every specific ability is a linear function of the general ability. On the other hand, Model 2, as is represented in Figure 4b, assumes that the general ability is a linear combination of all the specific abilities.

From the simulation studies, it is suggested that the two models are similar, with Model 1 being slightly less satisfactory in certain situations. The following general points are summarized regarding the proposed hierarchical models.

Figure 4. Graphical illustrations of the two proposed hierarchical IRT models (a) Model 1 (b) Model 2. Circles represent latent traits, and squares represent observed items.

First, for both hierarchical models, the item difficulty parameters (the intercepts) are stable and always well recovered. However, item discriminations, which are the slopes associated with the specific abilities, are not as stable.

Next, as far as the person abilities are concerned, the specific ones, together with their relations, are always recovered well with both models.
Third, taking the recovery of both item and person ability parameters into consideration, one can see that the two models do not differ much. Both models work better in the circumstances when the specific abilities are moderately or highly related to the general ability. This is intuitive since the model specifies linear relationships between the general and specific abilities. With higher correlation between the general and specific abilities, more information is gained from the specific ability estimates, and hence there is more accuracy in estimating the general ability as well as its relationship with those specific ones.

Finally, the correlation between the specific abilities makes a difference as well. Actually, the higher the specific abilities are correlated, the less likely the general ability parameter is estimated accurately. Therefore, it can be concluded that the hierarchical model is ideal if the specific abilities, being highly correlated with the general ability, have no or low correlations among themselves.

5.1.1.3 Comparing the additive and hierarchical models

It can be noted that in Sections 4.1 and 4.2, four simulations had exactly the same correlation matrices $R_0$. Therefore the results from the four simulations are compiled in this section to further compare the performance of the three additive models (Models 1, 2 and 3 in Section 4.1) and two hierarchical models (Models 1 and 2 in Section 4.2). The four simulations are in the same order as those described in Section 4.1, in which no correlations were specified between the latent abilities in Simulation 1,

$$\begin{pmatrix}
1 \\
0.8 & 1 \\
0.6 & 0 & 1
\end{pmatrix}$$

and $R_0 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0.6 \\ 0 & 0 & 1 \end{pmatrix}$, $R_0 = \begin{pmatrix} 1 & 0.8 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ for Simulations 2, 3 & 4 respectively.
First, we consider item parameters. For the purpose of simplicity, only sums of squared deviations are considered. The values are reorganized and displayed in Table 7. Generally, one can easily conclude from the table that the item parameters are consistently well recovered with the two hierarchical models. On the other hand, the additive model only recovers them well in Simulations 1 and 3. Those high deviation values for the discrimination parameters are due to the additive nature of the model, which makes it more complex and less likely to recover the discrimination parameters $\alpha_0$ well when there are correlations between the general ability and specific abilities. Apart from that, the difficulty parameters $\gamma$ are more stable and thus get better recovered relative to item discriminations in both types of proposed models.

Table 7.
Sum of squared difference between the actual and recovered item parameters using for 3 additive models and 2 hierarchical models under four simulated scenarios.

<table>
<thead>
<tr>
<th></th>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>1.1758</td>
<td>12.7078</td>
<td>0.9571</td>
<td>6.3476</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1.2842</td>
<td>1.1167</td>
<td>1.1907</td>
<td>1.4949</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.9084</td>
<td>0.9221</td>
<td>0.5422</td>
<td>0.6188</td>
</tr>
<tr>
<td>Additive Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>1.4425</td>
<td>14.1511</td>
<td>1.4610</td>
<td>6.9956</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1.7182</td>
<td>3.1337</td>
<td>1.1950</td>
<td>1.5486</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.8092</td>
<td>1.0280</td>
<td>0.5585</td>
<td>0.6127</td>
</tr>
<tr>
<td>Additive Model 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>1.3229</td>
<td>12.2820</td>
<td>4.2944</td>
<td>7.1878</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1.5919</td>
<td>1.4364</td>
<td>1.4982</td>
<td>1.5707</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.8705</td>
<td>0.9918</td>
<td>0.5577</td>
<td>0.6258</td>
</tr>
<tr>
<td>Hierarchical Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.5267</td>
<td>0.6802</td>
<td>0.6961</td>
<td>0.8013</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.2712</td>
<td>0.3278</td>
<td>0.2614</td>
<td>0.3731</td>
</tr>
<tr>
<td>Hierarchical Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.5299</td>
<td>0.5951</td>
<td>1.0340</td>
<td>0.8298</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.2684</td>
<td>0.3099</td>
<td>0.2737</td>
<td>0.3909</td>
</tr>
</tbody>
</table>
Table 8.
Correlations between the actual person ability parameters ($\theta_0$, $\theta_1$, $\theta_2$) and posterior ability estimates ($\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$) for 3 additive models and 2 hierarchical models in four simulated scenarios.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\theta_0$</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\hat{\theta}_0$</th>
<th>$\hat{\theta}_1$</th>
<th>$\hat{\theta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Additive Model 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.8899</td>
<td>0.1032</td>
<td>0.1313</td>
<td>0.9458</td>
<td>0.3585</td>
<td>0.1799</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.0144</td>
<td>0.7604</td>
<td>-0.0642</td>
<td>0.7338</td>
<td>0.6625</td>
<td>-0.2421</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1778</td>
<td>-0.0395</td>
<td>0.8288</td>
<td>0.6526</td>
<td>-0.3097</td>
<td>0.6773</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.1244</td>
<td>1.0000</td>
<td>-0.0034</td>
<td>0.0834</td>
<td>0.1032</td>
<td>0.0781</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.1049</td>
<td>-0.0141</td>
<td>1.0000</td>
<td>0.16</td>
<td>-0.4351</td>
<td>0.0821</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1244</td>
<td>1.0000</td>
<td>-0.0034</td>
<td>0.0834</td>
<td>0.1032</td>
<td>0.0781</td>
</tr>
<tr>
<td><strong>Additive Model 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.8912</td>
<td>0.1044</td>
<td>0.1483</td>
<td>0.9458</td>
<td>0.3585</td>
<td>0.1799</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.0218</td>
<td>0.7624</td>
<td>-0.0781</td>
<td>0.7562</td>
<td>0.6268</td>
<td>-0.1753</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1901</td>
<td>-0.0576</td>
<td>0.8315</td>
<td>0.6199</td>
<td>-0.2651</td>
<td>0.7225</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.1386</td>
<td>-0.0057</td>
<td>0.2424</td>
<td>-0.3409</td>
<td>0.1934</td>
<td>0.1538</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.1411</td>
<td>-0.0537</td>
<td>0.2094</td>
<td>-0.3409</td>
<td>0.1873</td>
<td>0.1538</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1411</td>
<td>-0.0537</td>
<td>0.2094</td>
<td>-0.3409</td>
<td>0.1873</td>
<td>0.1538</td>
</tr>
<tr>
<td><strong>Additive Model 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.8911</td>
<td>0.1077</td>
<td>0.1601</td>
<td>0.9443</td>
<td>0.381</td>
<td>0.2489</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.0242</td>
<td>0.7615</td>
<td>-0.0847</td>
<td>0.7351</td>
<td>0.6646</td>
<td>-0.1767</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1962</td>
<td>-0.0636</td>
<td>0.8336</td>
<td>0.648</td>
<td>-0.2719</td>
<td>0.7078</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.144</td>
<td>0.144</td>
<td>0.2477</td>
<td>0.2199</td>
<td>1</td>
<td>0.1934</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.1411</td>
<td>-0.0537</td>
<td>0.2477</td>
<td>-0.3409</td>
<td>0.1873</td>
<td>0.1538</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.1411</td>
<td>-0.0537</td>
<td>0.2477</td>
<td>-0.3409</td>
<td>0.1873</td>
<td>0.1538</td>
</tr>
<tr>
<td><strong>Hierarchical Model 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>-0.1054</td>
<td>-0.1144</td>
<td>-0.0141</td>
<td>0.8091</td>
<td>0.677</td>
<td>0.5696</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.7884</td>
<td>0.8063</td>
<td>-0.0013</td>
<td>0.7725</td>
<td>0.8411</td>
<td>0.0198</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0628</td>
<td>-0.0062</td>
<td>0.913</td>
<td>0.3375</td>
<td>0.0321</td>
<td>0.9103</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.9815</td>
<td>1.0000</td>
<td>-0.0043</td>
<td>0.9251</td>
<td>0.0215</td>
<td>0.5804</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.0861</td>
<td>0.0043</td>
<td>0.3594</td>
<td>0.0215</td>
<td>-0.8594</td>
<td>0.1221</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0861</td>
<td>0.0043</td>
<td>0.3594</td>
<td>0.0215</td>
<td>-0.8594</td>
<td>0.1221</td>
</tr>
<tr>
<td><strong>Hierarchical Model 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>-0.1035</td>
<td>-0.1164</td>
<td>-0.0149</td>
<td>0.8759</td>
<td>0.6803</td>
<td>0.5712</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.6675</td>
<td>0.8056</td>
<td>0.0000</td>
<td>0.6374</td>
<td>0.8421</td>
<td>0.0216</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.5029</td>
<td>-0.0047</td>
<td>0.913</td>
<td>0.6207</td>
<td>0.0364</td>
<td>0.9108</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.8317</td>
<td>1.0000</td>
<td>0.0003</td>
<td>0.7581</td>
<td>0.0277</td>
<td>0.8481</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.5617</td>
<td>0.0083</td>
<td>0.6726</td>
<td>0.0277</td>
<td>0.9407</td>
<td>0.6187</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.8317</td>
<td>1.0000</td>
<td>0.0003</td>
<td>0.7581</td>
<td>0.0277</td>
<td>0.8481</td>
</tr>
</tbody>
</table>

For the person abilities, the correlations between the actual and recovered values as well as the correlation between the estimated abilities are shown in Table 8 for the 3 additive and 2 hierarchical models. Considering the recovery of the ability parameters as well as the underlying structure of the ability dimensions, one can see that the additive models work well in Simulations 1 and 3 when there is no correlation between the general ability $\theta_0$ and specific abilities $\theta_1$ or $\theta_2$ whereas the hierarchical models work well in Simulation 1 and 2.
2 when the specific abilities, moderately or highly correlated with the general ability, are not correlated among themselves. Moreover, the general ability always gets better recovered in the additive model, for all items contain information for the general ability while the specific abilities are based on only a subset of the test items. On the contrary, due to the reason that the general ability underlies the specific ability dimensions, it is less well recovered than the specific abilities in the hierarchical model.

To sum it up, the proposed two classes of models have different features and they work well under different circumstances. With more parameters, the additive model is more complex than the hierarchical model. The additive model applies in the situation where the general ability is not at all related to neither of the specific abilities. Interestingly, the hierarchical model works well in a totally opposite scenario, that is, only when the general ability is moderately or strongly related to all of the specific abilities.

5.1.2 Model comparisons

The differences between the unidimensional, the multi-unidimensional and the proposed additive and hierarchical IRT models can be illustrated graphically.

**Unidimensional IRT model.** This model is the usual IRT model in the literature, in which all items are measuring a common latent trait, as is shown in Figure 5a.

**Multi-unidimensional IRT model.** In this model, there are several latent traits, each of which is associated with two or more test items. An example is shown in Figure 5b.

**Proposed additive IRT model.** This model consists of one general ability and several group ability dimensions as in the previous model. Figure 5c shows an example. All items are measuring two traits simultaneously, a general ability and a specific ability dimensions.
Proposed hierarchical IRT model. Similar to the previous model, this model has one general ability and several specific abilities. However, each specific ability either forms a linear function of the general ability (Figure 4a) or linearly combines to form the general ability (Figure 4b) so that the items measure the general ability dimension indirectly through the group ability dimensions.

Figure 5. Four classes of IRT models: (a) the unidimensional IRT model, (b) the multi-unidimensional IRT model, (c) the proposed additive IRT model, (d) the proposed hierarchical IRT model. Circles represent latent traits, and squares represent observed items.

The four classes (or five, with two in the hierarchical type) IRT models make different assumptions on the latent ability dimensions. A particular model is usually adopted when the researcher holds certain belief or there is theoretical evidence on the actual dimensional structure for a test. However, when the latent structure is not clear, one can consider several candidate models and compare their goodness-of-fit to the test data.
this study, we illustrate model comparisons for a subset of CBASE English data and compare among the proposed models as well as compare them with the conventional unidimensional and/or multi-unidimensional models. Three Bayesian model checking techniques are adopted for model comparisons, namely, Bayes factors, Bayesian deviances and posterior predictive model checks. All of them show similar results. That is, for the CBASE data, the additive model provides the best description (it is conceivably better than other models using any of the three Bayesian methods), and the hierarchical model 2 the second-best, among the five candidate models. It has to be noted that the additive model is the most complicated one among the five candidate models and it has certainly more parameters. We know that DIC penalizes for large number of parameters, for models with more parameters always fit data better than those without. Therefore, a smaller DIC value suggests that the additive model still performs the best after controlling for its large number of parameters.

On the other hand, the unidimensional model, the most specific and hence the simplest IRT model among the five candidate models, is the least desirable for the data. The fact that this one-dimension model does not fit as well as any other multiple-dimension models further suggests that its model assumptions, particularly, the unidimensionality assumption, cannot be assumed.

5.2 Conclusions

Based upon the findings, we can conclude that IRT-based models incorporating both general ability and specific abilities can be developed. In fact, the proposed IRT-based additive model, using an MCMC procedure, performs consistently better in parameter and dimensional structure recoveries than the bifactor analysis implemented in TESTFACT in various test situations. This further supports the advantage of the fully
Bayesian methodology over the conventional marginal maximum likelihood (MML) method in parameter estimations.

Implementing the two types of proposed IRT models to various simulated situations, one can conclude that they work well in quite different situations. Generally, the additive model applies when the general ability is not related to any of the specific abilities, whereas the hierarchical model performs well when the general ability is moderately or highly correlated with the specific ones.

The proposed additive and hierarchical models are implemented to the CBASE English data via Gibbs sampling procedure with little estimating error. This suggests both general ability and specific ability dimensions can be estimated in one implementation with enough accuracy. As far as the CBASE data are concerned, the proposed models, especially the additive model, provide a better description to the data than the conventional unidimensional or multi-unidimensional model. Moreover, the proposed additive model describes the data better than the proposed hierarchical model.

5.3 Discussion and Implications

It is well accepted that due to the complexity of the reality, all the theoretical models are just simplified approximations of the real world. Some models represent the reality better than others. Therefore, it is vitally important to find the model(s) providing the most complete description of the data. In testing situations where IRT models are used for parameter estimation as well as some other applications, one has to decide the dimensional structure for the latent abilities in order to choose an appropriate model and hence obtain reliable estimates of person abilities. Usually, a unidimensional model is adopted by assuming one latent ability. However, this assumption is more likely to be violated in real situations because the test items are not always measuring a single trait.
This point can be easily seen from the findings of the current study. Model comparisons indicate that the unidimensional model describes the CBASE data the worst compared with models with multiple-dimensions. Therefore, using the one-dimensional model for the CBASE English test is not validated. The actual dimensionality for the test proves to be more complex than a single latent trait. The model with one general and two specific ability dimensions is the closest to the reality among the models considered. In particular, the first 16 test items measure the overall English ability and a writing ability, and the last 25 items measure the overall ability together with a reading/literature ability. All items are affected by both general and a specific ability simultaneously and directly.

The proposed models offer a better way to represent the test situations not realized in existing models. By incorporating a general ability and several specific ability dimensions in one model, the examinee’s composite score as well as sub-scores can be obtained with one single implementation. The fully Bayesian methodology employed in the study considers dependencies among variables and sources of uncertainty. It is found that the proposed additive models have better recovery than the TESTFACT program. Hence, the fully Bayesian method is proved to be more accurate and efficient in parameter estimation compared with the common MML method with EM algorithm.

It has to be noted that the four types of IRT models, namely, the unidimensional model, the multi-unidimensional model, the proposed additive model and the proposed hierarchical model, are making different assumptions and should be adopted based on different beliefs on the latent structure of the ability dimensions. In the situations when it is a priori clear that certain abilities are being measured, one can use a particular model. However, in other situations when the test dimension is not clear at all, finding the model adequacy by comparing several candidate models provides evidence for the satisfaction of the model assumption on the dimension structure.
The additive model is more complex than the hierarchical model in that all test items measure a general ability and a specific ability simultaneously. The indeterminacy problem associated with the additive model can be overcome by either setting constraints on the covariance matrix or choosing relatively more informative priors. From the study, the latter is found to work better. In setting up the priors, one has to note that they comply with the researcher’s prior belief on the underlying structure of the latent ability dimensions. This perspective peculiar in Bayesian inferences also allows test developers to pre-determine the latent structure and to design tests accordingly.

5.4 Recommendations for Future Studies

It is concluded that the proposed additive model provides the most complete description to the CBASE data. Based on the simulation results, it may be assumed that the general ability has zero or low correlations with the two specific abilities for the CBASE data. However, the actual relationship between the general ability and each of the two specific ability dimensions is not yet clear. Simulation studies are needed in future studies to further investigate the actual relations between the latent abilities.

In the current study, odds ratios were adopted as a discrepancy measure when using the predictive model checking technique. Other test statistics could also be considered, such as item test biserial correlations and observed score distribution, among others. The choice of discrepancy measures is crucial with the method. Some measure may fail to detect the differences between models, such as item proportion-correct (Sinhary & Johnson, 2003). However, one has to note that this procedure has been criticized for being conservative and the PPP-value is not uniformly distributed under the null hypothesis (Sinhary & Stern, 2003). Future studies can adopt other methods for comparing models, such as looking at the Bayesian residuals as proposed by Albert &
Chib (1993). Additionally, in the study, Bayes factors were approximated because of the difficulty with the exact analytic evaluation for complicated hierarchical Bayesian models. The harmonic mean of the likelihood, which is used to approximate the marginal likelihood of the data using MCMC methods, converges to the correct value as the chain length goes to infinity. However, it does not satisfy a Gaussian central limit theorem because the model parameter may take an occasionally occurred value with small likelihood, which results in a large effect on the final result. Future studies may adopt other methods to approximate Bayes factors, such as Laplace method, which yields accurate approximations and is computational efficient (Kass & Raftery, 1995).

Finally, the current study focused on 2 parameter normal ogive IRT models with no more than two ability dimensions. Future studies can consider other IRT models, such as 3 parameter models, logistic models or models with 3 or more ability dimensions.
APPENDIX A.

Model specification for an additive model (Model 3 in Section 4.1) and the derivation of the L matrix

The proposed additive 2 parameter normal ogive model is defined as

\[ P(y_{ij} = 1) = \Phi(\alpha_{vij} \theta_{vij} + \alpha_{vij} \theta_{vij} - \gamma_{vij}) , \]

and the joint probability function of \( y \) is written as

\[ P(y | \theta, \xi) = \prod_{v=1}^{m} \prod_{i=1}^{n} \prod_{j=1}^{k} p_{vij}^{y_{vij}} (1-p_{vij})^{1-y_{vij}} , \]

as showed in equations 3.7 and 3.8.

Assume that the prior distribution of \( \theta_i, i=1,\ldots,n \) are independent multivariate normal (MVN), \( N_{m+1}(0, R) \) so the prior probability density function for the abilities is

\[ \varphi_{m+1}(\theta; 0, R) = (2\pi)^{-1/2} \exp\{\theta_i ' R^{-1} \theta_i / 2\} , \]  

(D.1)

Where \( m+1 \) is the dimension of \( \theta_i \), \( R \) is a constrained correlation matrix with 1’s on the diagonal, and \( \rho_{st} \) between \( \theta_i \) and \( \theta_t \), \( s \neq t \), \( s,t=0,\ldots,m \), on the off diagonals. With constraint imposed on \( \Sigma \), an inverse-Wishart prior is assumed for an unconstraint covariance matrix \( \Sigma^* \), i.e., \( \Sigma^* \sim W^{-1}(0, 3) \). The covariance matrix \( \Sigma = [\sigma_{st}]_{m \times m} \) corresponds to unconstrained abilities \( \theta^* \) such that \( \theta^* \sim N_{m+1}(\mu, \Sigma) \) and their relationship with the constrained \( \theta \) or \( \Sigma \) is

\[ \theta^* - \mu = L \theta \]

\[ \Sigma = L R L^\top \]

where \( L \) is a triangular matrix.

The prior distribution of \( \xi_{vij}, v=1,\ldots,m, j=1,\ldots,k, \) are assumed to be independent uniform distributed so that \( p(\gamma_{vij}) \propto 1 \) and

\[ I(\alpha_{vij}) = \begin{cases} 1, & \alpha_{vij} > 0 \\ 0, & \alpha_{vij} \leq 0 \end{cases} \]

Again, introduce the latent variable \( Z_{vij} \) that determines the examinee’s performance on each item. The joint posterior distribution of \( (\theta, \xi, Z, \Sigma) \) is then:
\[ p(\theta, \xi, Z, \Sigma | y) \propto f(y | Z) p(Z | \theta, \xi) p(\xi) p(\theta | R) p(\Sigma) \]
\[
\propto \Sigma^{-\frac{1}{2}} \prod_{i=1}^{n} \left| R_{i} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left( (Z_{ij} - \eta_{ij})^2 \right) \right\} \]
\[
\prod_{v=1}^{k_v} \prod_{j=1}^{m_v} \exp\left\{ -\frac{1}{2} (Z_{vij} - \eta_{vij})^2 \right\} (I(Z_{vij} > 0)I(y_{vij} = 1) + I(Z_{vij} \leq 0)I(y_{vij} = 0)]
\]
where \( \eta_{vij} = \alpha_{vij} - \gamma_{vij} \) is the prior mean of \( Z_{vij} \). The full conditionals are derived as follows:

1. For variable \( Z_{vij} \):
\[
[Z_{vij} | \cdot] \propto f(y_{vij} | Z_{vij}) p(Z_{vij} | \eta_{vij}) \propto \exp\left\{ -\frac{1}{2} (Z_{vij} - \eta_{vij})^2 \right\} (I(Z_{vij} > 0)I(y_{vij} = 1) + I(Z_{vij} \leq 0)I(y_{vij} = 0) \] (D.3)
So the full conditional of \( Z_{vij} \), denoted as \( Z_{vij} | \cdot \) has a truncated normal distribution
\[
Z_{vij} | \cdot \sim \begin{cases} 
N(\eta_{vij}, \Gamma), & \text{if } y_{vij} = 1 \\
N(\eta_{vij}, \Gamma), & \text{if } y_{vij} = 0 
\end{cases} \] (D.4)

2. For the person parameters \( \theta \):
\[
[\theta | \cdot] \propto p(Z | \theta, \xi) p(\theta | R)
\]
\[
\propto \exp\left\{ -\frac{1}{2} \left( (\theta - \Theta) R^{-1} \theta \right) \right\} \prod_{i=1}^{m} \prod_{j=1}^{k} \exp\left\{ -\frac{1}{2} (Z_{ij} - (\alpha_{ij} + \alpha_{ij} \beta_{ij} - \gamma_{ij}))^2 \right\}
\]
\[
= \exp\left\{ -\frac{1}{2} \left( (\theta - \Theta) R^{-1} \theta \right) \right\} \exp\left\{ -\frac{1}{2} (\Theta \Theta - (\Theta B)^T (\Theta B)) \right\} \] (D.5)

Thus the full conditional for \( \theta \) has a multivariate normal distribution,
\[
\theta | \cdot \sim N_{m+1}((A^T A + R^{-1})^{-1} A^T B, (A^T A + R^{-1})^{-1}) \] (D.6)

where \( A = \begin{pmatrix} 
\alpha_{01} & \alpha_{1} & 0 & \cdots & 0 \\
\alpha_{02} & 0 & \alpha_{2} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\alpha_{0m} & 0 & 0 & \cdots & \alpha_{m} 
\end{pmatrix} \) and \( B = \begin{pmatrix} 
Z_{11} + \gamma_{1} \\
Z_{12} + \gamma_{2} \\
\vdots \\
Z_{1m} + \gamma_{m} 
\end{pmatrix} \)

3. Then for the item parameters \( \xi_{vij} \):
\[
[\xi_{vij} | \cdot] \propto p(Z | \theta, \xi) p(\xi) \propto \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} (Z_{vij} - (\alpha_{vij} + \alpha_{vij} \beta_{vij} - \gamma_{vij})^2 \right\} I(\alpha_{vij} > 0)I(\alpha_{vij} > 0)
\]
\[
= \exp\left\{ -\frac{1}{2} (Z_{vij} - x_{vij}) (Z_{vij} - x_{vij}) \right\} I(\alpha_{vij} > 0)I(\alpha_{vij} > 0)
\]
\[
\propto \exp\left\{ -\frac{1}{2} (\xi_{vij} - x_{vij} \xi_{vij} - 2(Z_{vij} x_{vij}) \xi_{vij} \right\} I(\alpha_{vij} > 0)I(\alpha_{vij} > 0) \] (D.7)
So the full conditional for \( \xi_{vij} \) is
\[ v_{ij} \mid \xi_i \sim N((x_v'x_v)^{-1}x_v', (x_v'x_v)^{-1})I(\alpha_{0v} > 0)I(\alpha_v > 0) \tag{D.8} \]

where \( Z_v = [Z_{vij}]_{m \times k} \), \( \xi_i = (\xi_{i1}, \ldots, \xi_{ik})' \), \( x_v = [\theta_0, \theta_v, 1] \), and \( \theta_v = (\theta_{0v}, \ldots, \theta_{kv})' \), \( \theta_v = (\theta_{1v}, \ldots, \theta_{mv})' \), \( v = 1, \ldots, m \).

4. Lastly, for the hyperparameter \( R \):
To obtain \( R \), one has to work on the unconstrained \( \Sigma \) first.

The full conditional for \( \Sigma \) is an inverse-Wishart distribution,
\[ \Sigma \mid \bullet \sim W^{-1}(S^{-1}, n + 3), \tag{D.10} \]
where \( S = \sum_{i=1}^{n} (\theta_i^* - \mu)^*(\theta_i^* - \mu)^* = \sum_{i=1}^{n} L \theta_i (L \theta_i)^* \).

To derive the triangular matrix \( L \), we first consider the probability functions with unconstrained \( \theta^* \), i.e., \( P(y_i = 1) = \Phi(A^* \theta_i^* - \gamma^*) \). With certain constraints imposed on \( A^* \) and \( \gamma^* \), this function is the same as \( P(y_i = 1) = \Phi(A \theta_i - \gamma) \), where \( A \) and \( \gamma \) are free whereas \( \theta \) is constrained to have a mean \( \theta \) and variance-covariance \( \Sigma \) (which is as defined earlier).

For the purpose of illustration, assume \( m = 2 \), \( A = \begin{pmatrix} \alpha_{01} & \alpha_1 & 0 \\ \alpha_{02} & 0 & \alpha_2 \end{pmatrix} \) and
\[ A^* = \begin{pmatrix} \alpha_{01}^* & \alpha_1^* & 0 \\ \alpha_{02}^* & 0 & \alpha_2^* \end{pmatrix}. \]
With the unconstrained \( \theta^* \), certain constraints have to be placed for the item parameters, e.g., \( \prod_{j=1}^{k} \alpha_{0j}^* = 1 \), \( \prod_{j=1}^{k} \alpha_{1j}^* = 1 \), \( \prod_{j=1}^{k} \alpha_{2j}^* = 1 \). Since \( \theta^* - \mu = L \theta \), it is easy to show \( A^* (L \theta + \mu) - \gamma^* = A^* L \theta + A^* \mu - \gamma^* \). Then we can have \( A^* L = A \) and \( A^* \mu - \gamma^* = \gamma \).

To solve for the \( L \) matrix, where \( L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{12} & l_{22} & 0 \\ l_{13} & l_{23} & l_{33} \end{pmatrix} \), we have
\[ \begin{pmatrix} \alpha_{01}^* & \alpha_1^* & 0 \\ \alpha_{02}^* & 0 & \alpha_2^* \end{pmatrix} \begin{pmatrix} l_{11} & 0 & 0 \\ l_{12} & l_{22} & 0 \\ l_{13} & l_{23} & l_{33} \end{pmatrix} = \begin{pmatrix} \alpha_{01} & \alpha_1 & 0 \\ \alpha_{02} & 0 & \alpha_2 \end{pmatrix}. \]
Hence a system of equations can be derived from the previous equation:
\[
\begin{align*}
\alpha_{1}^* l_{22} &= \alpha_1 \\
\alpha_{2}^* l_{33} &= \alpha_2 \\
\alpha_{1}^* l_{33} &= \alpha_2 \\
\alpha_{2}^* l_{23} &= 0 \\
\end{align*}
\]
Use \( \prod_{j=1}^{k} \alpha_{0j}^* = 1 \), \( \prod_{j=1}^{k} \alpha_{1j}^* = 1 \),
\[
\begin{align*}
\alpha_{01}^* l_{11} + \alpha_{1}^* l_{12} &= \alpha_{01} \\
\alpha_{02}^* l_{11} + \alpha_{2}^* l_{13} &= \alpha_{02} \\
\end{align*}
\]
\[ \prod_{j=1}^{k_2} \alpha_{2j}^* = 1 \] to solve for the six elements in the \( L \) matrix, i.e., \( l_{11}, l_{12}, l_{13}, l_{22}, l_{23}, l_{33} \). It’s easy to show that

\[
\begin{align*}
  l_{22}^{k_1} &= \prod_{j=1}^{k_1} \alpha_{1j} \\
  l_{13}^{k_2} &= \prod_{j=1}^{k_2} \alpha_{2j}
\end{align*}
\]

to show that

\[
  l_{13}^{k_2} = \prod_{j=1}^{k_2} \alpha_{2j}.
\]

However, \( l_{11}, l_{12}, l_{13} \) are not easy to solve with

\[
  l_{23} = 0
\]

\[
\begin{align*}
  \prod_{j=1}^{k_1} (\alpha_{01j} - \alpha_{01j}^* l_{11}) &= (l_{12})^{k_1} \\
  \prod_{j=1}^{k_2} (\alpha_{02j} - \alpha_{02j}^* l_{11}) &= (l_{13})^{k_2}
\end{align*}
\]

In the study, to solve for \( L \) we set \( l_{12} = 0 \) and \( l_{13} = 0 \) to obtain

\[
  l_{11}^{k_1} = \prod_{j=1}^{k_1} \alpha_{01j} \prod_{j=1}^{k_2} \alpha_{02j}
\]

This specification for \( L \) matrix can be generalized to the situations where \( m > 2 \).

Consequently, once \( L \) is determined, \( \mathbf{R} \) can be easily computed using

\[
\mathbf{R} = L^{-1} \Sigma (L^{-1})^T
\]

after \( \Sigma \) is obtained from the Gibbs sampler.
REFERENCES


VITA

Yanyan Sheng was born March 8, 1975, in Tongling, a small city besides the Yangtze River in China. After attending public schools there, she received the following degrees: B.A. in English from Anhui University (1997); M. A. in Educational Psychology from the University of Missouri-Columbia (2002); M. A. in Statistics from the University of Missouri-Columbia (2004). She is becoming a member of the Department of Educational Psychology and Special Education in Southern University of Illinois, Carbondale.