11-26-2013

Extending MPT Models to Rasch MPT: A General Framework, Demonstration, and Applications

Quan Tang

EXTENDING MPT MODELS TO RASCH MPT: A GENERAL FRAMEWORK, DEMONSTRATION, AND APPLICATIONS

by

Quan Tang

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Psychology

The University of Memphis

December 2013

ABSTRACT

Tang, Quan. Ph.D. The University of Memphis. December, 2013. Extending MPT Models to Rasch MPT: A General Framework, Demonstration, and Applications. Major Professor: Xiangen Hu, Ph.D.

Multinomial processing tree (MPT) models, as a family of hierarchical multinomial models tailored by cognitive theories, have been proven to be successful and applied to cognitive psychometrics (Batchelder 1998). Traditional MPT models measure the probability of success for each cognitive stage given their hierarchical relationships. However, this measure neither addresses individual and item difference, nor characterizes the subject's ability and the difficulty of the cognitive stage. In this study, I extend the cognitive stage parameter in MPT models to a Rasch model, and recruit MPT models for a source monitoring paradigm as an example to demonstrate the extension. To evaluate the properties of Rasch MPT models, I conduct systematic simulation studies to test parameter recovery under different conditions including various sample sizes, boundary values of parameters, and missing data. In addition, I use a simple lexical decision experiment and a set of force concept inventory (FCI) multiple-choice questions which are a popular measurement tool in physics teaching research to demonstrate and validate the practical uses of Rasch MPT modeling.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MPT	Multinomial Processing Tree
CTT	Classic Test Theory
EM	Expectation Maximization
ICC	Item Characteristic Curve
IRT	Item Response Theory
I.I.D.	Independent and Identically Distributed
1HTH	One-High-Threshold
1LTH	One-Low-Threshold
2HTH	Two-High-Threshold
MLE	Maximum Likelihood Estimation
MCMC	Markov Chain Monte Carlo
MC	Markov Chain
PMF	Probability Mass Functions
SD	Standard Deviation
Val2.5pc	Value at 2.5 Percentile
Val97.5pc	Value at 97.5 Percentile
TOEFL	Test of English as a Foreign Language
GRE	Graduate Record Examinations
AMT	Amazon Mechanical Turk
FCI	Force Concept Inventory
FF	Free Fall

LIST OF SYMBOLS

$\sum$	Summation
$\prod$	Production
Θ	Parameter vector in classic MPT model setting
D	Data vector
$L(\Theta; D)$	Likelihood function given Θ and D
C	Observed categories
N	Frequencies of observed categories
B	Branch in MPT Models
L	Link on a branch
p	Probability
G	Parameter space
Ψ	Parameter vector in Rasch MPT model setting

CHAPTER 1

INTRODUCTION

1.1 Overview

Multinomial processing tree (MPT) models are a family of statistical models that use serial cognitive processes to depict and analyze categorical data. Formally, MPT models may be regarded as a special family of models in the more general class of parameterized multinomial or product-multinomial models (Stahl and Meiser 2009). MPT models are versatile and have been applied in fields such as cognitive science, medical science, and social science (Hu and Batchelder 1994; Batchelder and Riefer 1999; Schmittmann et al. 2010). MPT models not only share common features of multinomial models, they (1) are hierarchical and in a tree structure; (2) describe a set of serial processes; and (3) may be tailored to different forms according to plausible theories or hypotheses. As a consequence, the development of MPT models has been closely intertwined with the development of paradigms and theories in cognitive psychology (Erdfelder et al. 2009).

A lot of attention has been given to the development of MPT models. For example, researchers have devised various MPT models for paradigms of interests, such as source monitoring experiments (Johnson and Raye 1981; Johnson, Hashtroudi, and Lindsay 1993). These MPT models for source monitoring include a one-high-threshold (1HTH) model (Batchelder and Riefer 1990), a one-low-threshold (1LTH) model (Bayen, Murnane, and Erdfelder 1996; Hu and Batchelder 1994), two-high-threshold (2HTH) MPT models (Bayen, Murnane, and Erdfelder 1996; Yonelinas et al. 1996), and a high-threshold MPT model for more than two old sources (Meiser and Broder 2002). In addition, parameter estimation methods have been extended to Bayesian estimation (Lin and Karabatsos 2006; Klauer 2009; Matzke et al. 2012), from the original EM algorithms for maximum-likelihood estimation (MLE) (Hu and Batchelder 1994).

Furthermore, various model selection methods such as Bayesian information criterion and minimum description length criterion (Myung and Pitt 2004; Wu, Myung, and Batchelder 2010). Also, empirical approaches, such as Hu (2001) and Bayen, Murnane, and Erdfelder (1996), have been applied to MPT models.

Although MPT models are flexible and have been well established with methodologies for data analyses, some insufficiencies exist for the classic setting of MPT modeling. For example, classic MPT modeling assumes that parameters are independent from one another. However, this assumption may be violated for nodes with hierarchical relationship on a branch. An extreme situation is, if the estimate of a parent node (i.e., an earlier cognitive state) is 0, its offspring nodes will not be able to vary freely. Another example is when the classic setting of MPT modeling assumes that subjects in a group have same cognitive abilities for the tasks, and the stimulus in a same type also has the same psychological effect. This assumption is literally described as the independent and identically distributed (i.i.d.) assumption. Only with this assumption, can one aggregate the data collected across the subjects in a group and stimulus items of a same type.

Fortunately, researchers have noticed these issues and proposed some solutions. Klauer (2006), Stahl and Klauer (2007) proposed a latent-class approach to use a higher level discrete distribution to model subject performance on the cognitive states by different “latent classes”. In addition, Klauer (2009) extended this approach to a latent-trait approach that uses a continuous higher level distribution to model subject performance.

Rather than the approaches described above, psychometric models that are developed to measure individual and stimuli differences provide another possible solution. For example, item response theory (Lord and Novick 1968; Rasch 1960) imposes a logistic link function to model the probability of success on a task as the difference of subject ability and task difficulty. However, this

approach has not been applied to the MPT model. Moreover, there exist practical concerns such as whether two models can integrate to perform good estimation for the potential parameters, and under what conditions (or combinations) this approach may or may not be suitable. In the next section, I will summarize the issues of MPT models and existing solutions, which will lead to the theoretical framework to be proposed in this dissertation.

1.2 Problem Statement

There exist two crucial issues for MPT models. One is that MPT models usually assume homogeneity in terms of subject cognitive ability and item difficulty, even though homogeneity between these two often does not exist. The other is that the construct validity of the measure for the subject's performance on a cognitive state is quite arguable. The first issue, is that Classic MPT modeling only takes group differences into account, while ignoring the differences within a group. The second issue is that MPT modeling merely measures a joint outcome of the subject's ability and task difficulty; hence it is hard to interpret its construct validity.

Although some approaches (e.g., latent-class, latent-trait) try to address the individual and item difference issues stated above, they (1) only use a distribution to approximate these differences, without measuring every single individual and/or item; and (2) have the same construct validity issue that the performance on each cognitive state is not explained, and subject ability and item difficulty are intertwined.

Therefore, this dissertation builds up a general framework to address the i.i.d. issue and construct validity issue of classic MPT modeling by using the item response theory (IRT) model approach (Hambleton, Swaminathan, and Rogers 1991; Embretson and Reise 2000; Sijtsma and Junker 2006). This approach models subject ability and task difficulty as independent variables, and estimates

ability and difficulty parameters for each single subject and task, respectively. Furthermore, the performance of this approach under different conditions is evaluated systematically.

1.3 Looking Ahead

This chapter has presented an overview of MPT models. The history, development and some insufficiencies of MPT models were briefly summarized. Also some attempts at solving the efficiencies of MPT models were briefly described, and the issues of these approaches were pointed out. The purpose of this dissertation is to apply the advantages of IRT models to MPT models and examine the performance of the extended model.

Chapter 2 discusses MPT models and IRT models in greater detail. Chapter 3 presents our general framework to extend MPT models to Rasch MPT models and demonstrates Rasch MPT models by using signal detection and source monitoring paradigms. Chapter 4 conducts a systematic evaluation of Rasch MPT models in various conditions. Chapter 5 applies Rasch models to a simple lexical decision experiment and force concept inventory (FCI) test questions. Finally, Chapter 6 discusses the theoretical implications, additional possible applications of Rasch MPT models, and future research.

CHAPTER 2

MPT MODELS AND IRT MODELS

In this chapter, I provide some background for MPT models and IRT models. Section 1 presents the origin of MPT modeling, its advantages compared with former approaches to categorical data, and some typical applications such as source monitoring paradigms; and Section 2 discusses two crucial issues of classic MPT modeling and the necessity of extending to Rasch MPT models.

2.1 Classic Theories of MPT Models

Multinomial processing tree (MPT) models are a family of statistical models for categorical data. MPT models were originally developed to measure latent cognitive processes, such as the capacity to store and retrieve items in memory, or to make inferences and logical deductions, or to discriminate and categorize similar stimuli (Riefer and Batchelder 1988; Batchelder and Riefer 1990). The MPT modeling framework is based on the hypotheses of hierarchical and serial cognitive processes. While such processes are not directly observable, theoretically they can be assumed to interact in certain ways to determine observable behaviors. The goal of multinomial modeling is to identify which underlying factors are important in a cognitive task, explain how those processes combine to create observable behavior, and then use experimental data to estimate the relative contributions of the different cognitive factors. In this way, MPT models can be used as tools to measure unobservable cognitive processes.

MPT modeling, since formally proposed, is always specified with specific paradigms. A typical application of MPT models in cognitive psychology is the MPT models for source monitoring. Source monitoring research is derived from the interest in human source memories. People remember information from two basic sources: that perceived from external sources (stimuli) and that generated by internal processes such as reasoning, imagination, and thought. And people may remember, forget, or mix these memories (Johnson and Raye 1981). There

is a common phenomenon that most people may have experienced: we heard a story from a friend and forgot who told this story, then we share the story back to this friend with interest. Even worse, we may add something to this story by ourselves unconsciously.

To study different kinds of memories, Johnson and Raye (1981) proposed the concept of “reality monitoring”. Reality monitoring refers to the process of distinguishing the memory of a past perception from the memory of past imagination. As an extension of the reality monitoring, the concept of “source monitoring” was proposed by Johnson and her colleagues (Johnson, Foley, and Leach 1988; Johnson, Hashtroudi, and Lindsay 1993; Johnson and Raye 1981). Compared with reality monitoring that focuses on discriminating memories of internally generated information from memories of externally perceived information, source monitoring refers to discriminating different types of internal or external sources, namely, internal source monitoring or external source monitoring (Johnson, Foley, and Leach 1988). For instance, external source monitoring may be interested in discriminating between two externally perceived sources such as statements made by person A or by person B while internal source monitoring concentrates on discriminating the memories of what one thought from what one said. Hence source monitoring is derived and generalized from reality monitoring.

After the concepts of reality monitoring and source monitoring were introduced, quite a number of source monitoring experiments were conducted to test different cognitive models or to measure cognitive capacities of different populations. For example, Harvey (1985) studied how different normal and mentally disordered subjects are able to discriminate their own thoughts and information from external sources, Saegert, Hamayan, and Ahmar (1975) tested if source memory for language is dependent on the nature of the memory task

itself, and Rose et al. (1975) examined whether the phenomenon of accurate source memory for language could be found at complex cognitive levels.

In a typical source monitoring experiment, subjects study items from two or more different sources (Johnson, Hashtroudi, and Lindsay 1993), for example, pictures of the items as source A, and the names of the items as source B. After these items have been studied, a memory test is given, in which the subjects are asked to indicate which source (source A, B or new source) the test items belong to. Data from a group of subjects can be described by a frequency table as in Table 1, where f_{ij} is the counts of j-type responses to i-type source. The row marginal frequency $f_{i\bullet} = \Sigma f_{ij}$ is the total number of i-type source items on the memory test, and $i, j = A, B, C$.

Table 1
Data matrix of a typical source monitoring experiment. Rows represent presentation during learning, columns denote the response of the participants, the cells contain raw frequencies

	Participants' response		
Actual source during learning	"Source A"	"Source B"	"New"
Source A	f_{AA}	f_{AB}	f_{AN}
Source B	f_{BA}	f_{BB}	f_{BN}
New	f_{NA}	f_{NB}	f_{NN}

In early studies on source monitoring, some ad hoc statistical approaches were adapted for separating discriminability of source from overall detectability of old items, such as the Kruskal-Wallis gamma score, identification-of-origin scores, and hit and false-alarm rates for source identification (Batchelder and Riefer 1990). "Discriminability" here means the ability to discriminate the specific old source from other old sources after an item has been detected as an old item in

the source memory test. And “detectability” means the ability to detect an old source item in the test.

The most frequently used method for this type of data analysis is to compute three measures for each subject as shown in equations (1), (2) and (3): hits (H), indicating the rate at which the subject can detect old items correctly; false alarms (F), indicating the rate at which the subject incorrectly reports a distractor item as an old item; and identification-of-origin scores (I), referring to the rate at which the subject discriminates the exact source from all the responded old sources. The equations of these three rates are shown as follow in terms of the frequencies presented in Table 1.

$$H = \frac{(f_{AA} + f_{AB}) + (f_{BA} + f_{BB})}{f_{A\bullet} + f_{B\bullet}} \tag{1}$$

$$F = \frac{f_{NA} + f_{NB}}{f_{N\bullet}} \tag{2}$$

$$I = \frac{f_{AA} + f_{BB}}{(f_{AA} + f_{AB}) + (f_{BA} + f_{BB})} \tag{3}$$

However, about ten years after the concept of source monitoring had been proposed and a multitude of studies had been done, Batchelder and Riefer (1990) noted that there was not a generally accepted measure of the quantities reported in the source-monitoring experiments. In other words, there was not any substantive model to analyze the data of the contingency table obtained from the source-monitoring experiments (see Table 1). For example, the generally used model depicted in equation (1), (2), and (3) failed to look into the internal cognitive processes such that it is impossible to distinguish whether the subject really discriminates the exact old source or answers correctly by guessing, when the subject reports an exact old source (e.g., report source A as source A).

Therefore, Batchelder and Riefer proposed Multinomial Processing Tree (MPT) models for source monitoring experiments as a substantively quantitative

measurement tool for the memory retrieving processes during source monitoring experiment tasks.

Because the response frequencies in source monitoring experiments can be considered as multinomially distributed, it is assumed there are finite numbers of observable categories, $C_1, C_2, ..., C_J$, and there are N total observations. Then N_j is defined as the number of observations in C_j, and $D = (N_1, ..., N_j, ..., N_J)$ is defined to be the data vector of observations for the model. The joint distribution of the data D can be represented by the general multinomial model

$$P(D;\ p_1, ..., p_J) = N! \prod_{j=1}^{J} \frac{p_j^{N_j}}{N_j!}, \tag{4}$$

where p_j is the probability that an observation falls into C_j if the data observations are mutually independent and identically distributed (i.i.d.), and

$$N = \sum_{j=1}^{J} N_j. \tag{5}$$

The general model has the parameter space

$$G_j = \left\{ p = (p_1, ..., p_J) | 0 < p_j < 1, \sum_{j=1}^{J} p_j = 1 \right\}. \tag{6}$$

In addition, a substantive MPT model assigns a parameter to each cognitive event that represents the probability of that event occurring. These events are organized hierarchically according to psychological assumptions or theories, from the very first node to the last, in a tree structure.

Every information source has an MPT model that represents the processing steps (by the parameters), and the categories of the subject's responses. For example, for source A, the first parameter (D_A) in the model is assumed to represent the probability of detecting this source as an old source. Because the detection probabilities for different sources may vary, D_B may be different from D_A. The subsequent step after the detection step is the

discrimination with the parameter d_i as its probability if the subject successfully detects old items, or bias (represented by the parameter b) for responding a non-detected old item as an old item.

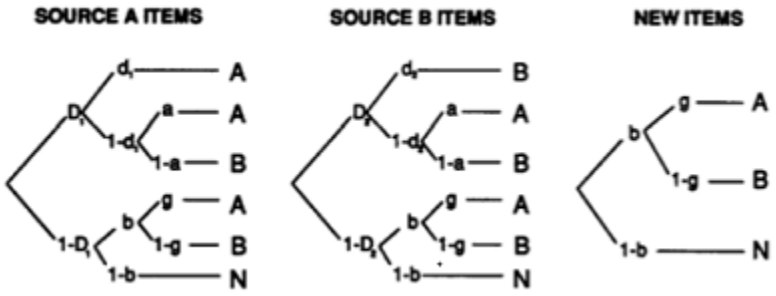

Figure 1
The seven-parameter, joint multinomial model for source monitoring. ($D_1 =$ detectability of the Source A items; $D_2 =$ detectability of the Source B items; $d_1 =$ source discriminability for the Source A items; $d_2 =$ source discriminability for the Source B items; a = guessing that a detected but nondiscriminated item belongs to Source A; b = bias for responding "old" to a nondetected item; g = guessing that a nondetected item belongs to Source A.)

If the subject can detect and discriminate an old item successfully, the response is absolutely correct and this response falls in the cell f_{AA} for source A and in the cell f_{BB} for source B in Table 1. If subjects fail in the detecting or the discriminating step, they guess. And if subjects are "lucky" enough, it is still possible for them to report correctly (e.g., first correctly guess that the item is an old item, and secondly correctly guess its type).

This set of MPT models is called the one high threshold (1HTH) model because in this set of MPT models, only the trees for "old" source items have detection and discrimination steps and the tree for "new" source items (distractors) does not. In contrast, the new items (distractors) are assumed either to be responded to as old items by bias or as new items without bias.

Figure 1 presents the structure of MPT models for source monitoring and the meaning of their parameters. There are 7 parameters in this set of models, with 6 degrees of freedom (3×3 data table with 3 fixed marginal frequencies).

Hence, this 7-parameter model is over saturated and the parameters cannot be uniquely estimated, due to the insufficient degree of freedom in the data, unless we eliminate at least one parameter (e.g., we may equate a parameter with another). Figure 2 shows 6 sub-models. In 6a, 6b and 6c sub-models, two parameters are merged into one, based on the hypothesis that the detection rates, the discrimination rates, or the guessing rates of the two sources are equal, respectively. Likewise, 5-parameter sub-models combine another pair of parameters. This paradigm provides 7 sub-models with different corresponding psychological hypotheses that allow us to test the fit of each sub-model.

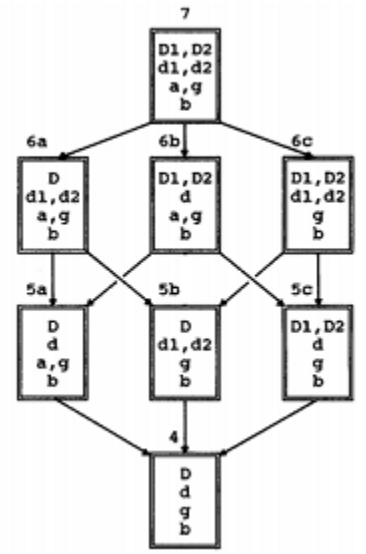

Figure 2
Nested hierarchy for the eight versions of the multinomial model depicted in Figure 1

The MPT models for source monitoring (Batchelder and Riefer 1990) use graphical representation to illustrate the plausible cognitive procedure in the source monitoring test, and explicitly separate the frequencies (including those in the same cell in the data table) to hierarchically organized origins. For example, as introduced previously, equation (3) mixes real discrimination with guessing of an exact old source. When considering the difference between real discrimination

and guessing, f_{AA} in equation (3) can be rewritten as:
$f_{AA}((D_1 d_1) + D_1(1 - d_1)a + (1 - D_1)bg)$. Similarly, f_{BB}, f_{AB} and f_{BA} in equation (3) mixed frequencies from plausibly different origins while MPT models separate these origins into different branches. The MPT models provide an approach to measuring the cognitive processes in source monitoring tasks and test hypotheses of different sub-models under various situations, and they have been applied to source monitoring analyses more and more.

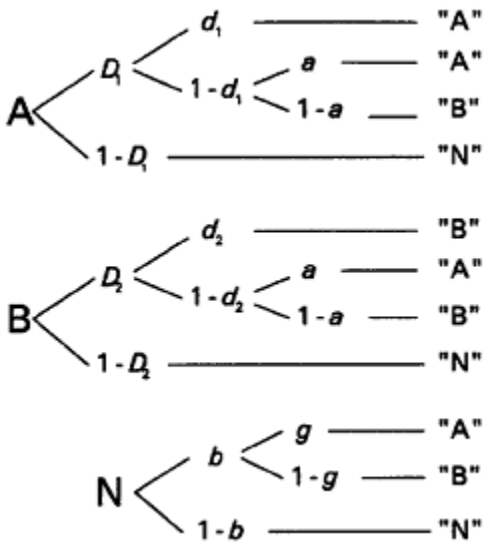

Figure 3
One-low-threshold MPT Model

After 1HTH MPT models were proposed, researchers came up with other MPT models for source monitoring, based on different theories or hypotheses. For example, the 1LTH (one-low-threshold) model assumes that there is only one memory threshold, but an old item will be recognized if it exceeds the threshold, or will be considered as a new item otherwise (see Figure 3) (Bayen, Murnane, and Erdfelder 1996; Hu and Batchelder 1994). Similarly, there are other MPT models such as the 2HTH (two-high-threshold) MPT models which assume both an old item and a new item may have some probability to be recognized, as shown in Figure 4 (Bayen, Murnane, and Erdfelder 1996; Yonelinas et al. 1996).

As well, high-threshold MPT models for more than two old sources (Meiser and Broder 2002) have been applied to source monitoring experiments.

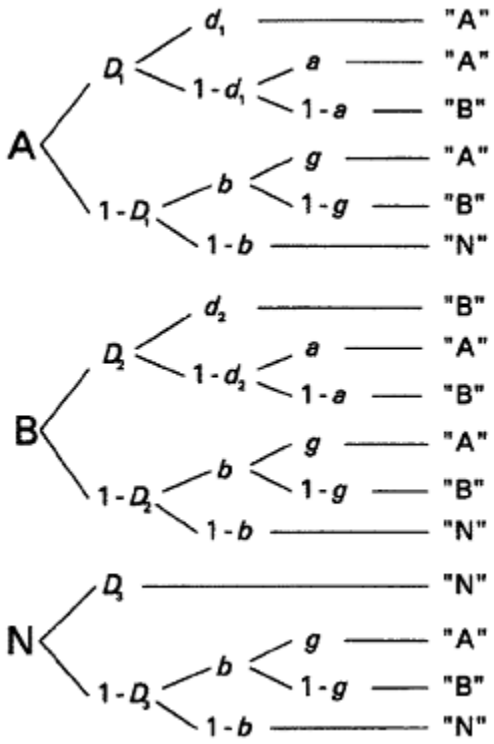

Figure 4
Two-high-threshold MPT Model

Before MPT framework was formally established, some researchers tried to use tree structures to separate subskills involved in multiple-choice questions (Garcia-Perez 1990; Garcia-Perez and Frary 1991; Garcia-Perez 1993). However, these approaches only investigate reasons that lead to “correct answer”, “wrong answer”, and “unanswered” responses, while MPT modeling details the cognitive states/subskills that result in different types of “wrong answers” (e.g., answer “B” or “N” for “A” in the source monitoring experiments).

To utilize MPT models in categorical data analysis, one needs to obtain the probability functions and likelihood from the graphical form of a specified set of MPT models. In the next section, I will discuss basic statistical theories of MPT

models, which is also the essence of the extended MPT model framework I develop in this dissertation.

2.2 Statistical Theories of MPT Models

Since MPT modeling is used to explore the cognitive processes in a cognitive task, the fundamental statistical analysis for MPT models is to estimate the parameters that represent the latent cognitive states.

2.2.1 A Simple Example

Let us consider the following case in which two coins are flipped for one trial each and the final result is recorded. There are 4 observed categories: 2 heads (HH), 2 tails (TT) and 1 head followed by 1 tail (HT), or 1 tail followed by 1 head (TH). The category frequencies are represented by $D = (n_1, n_2, n_3, n_4)$, and the probabilities of these outcomes are represented by b_1, b_2, b_3, and b_4 respectively. The parameter vector is denoted by $\Theta = (\Theta_1, \ldots \Theta_s \ldots \Theta_S) \in \Omega$, where Ω is the parameter space, and $\Theta_s = (\theta_{s1}, \ldots, \theta_{sk} \ldots, \theta_{sK_s})$ refers to the K_s parameters in a group (under a same parent node). This group of parameters indicates the probability of all possible outcomes under a certain condition (i.e., a parent node). Hence the summation of a group of parameters always equals to 1 (i.e., $\sum_{s1}^{sK_s} \theta_{sk} = 1$). In the coin-flipping example, due to the binomial outcomes of each event, there are two parameters (e.g., p and $1 - p$) in a group, and only one is independent. Note that from now on the notations above are for all the coin-flipping examples, unless explicitly indicated otherwise. Figure 5 illustrates this procedure. The frequencies of the final results follow a multinomial distribution with 4 categories and the probabilities of these outcomes are:

$$b_1 = pq, \tag{7}$$

$$b_2 = p(1 - q), \tag{8}$$

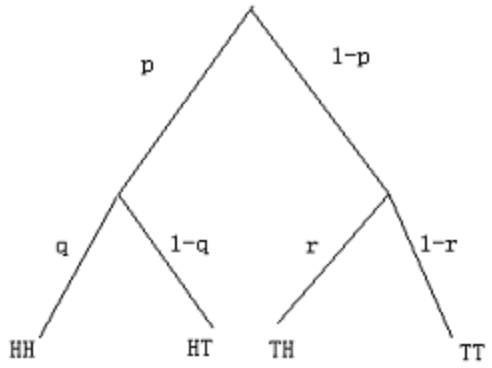

Figure 5
The coin-flipping trials. p = the probability that coin 1 gets a head, q = the probability that coin 2 gets a head if coin 1 gets a head, r = the probability that coin 2 gets a head if coin 1 gets a tail.

$$b_3 = (1-p)r, \tag{9}$$

$$b_4 = (1-p)(1-r). \tag{10}$$

To estimate the parameters in the model in Figure 5, we can use the model's likelihood function:

$$L(\Theta; D) = \frac{n!}{n_1!n_2!n_3!n_4!} b_1^{n_1} b_2^{n_2} b_3^{n_3} b_4^{n_4}. \tag{11}$$

After plugging in Equation 7 - Equation 10, we have

$$\begin{aligned} L(\Theta; D) &= \frac{n!}{n_1!n_2!n_3!n_4!} (pq)^{n_1} (p(1-q))^{n_2} ((1-p)r)^{n_3} ((1-p)(1-r))^{n_4} \\ &= \frac{n!}{n_1!n_2!n_3!n_4!} p^{(n_1+n_2)} (1-p)^{(n_3+n_4)} q^{n_1} (1-q)^{n_2} r^{n_3} (1-r)^{n_4}. \end{aligned} \tag{12}$$

As the simplest scenario of MPT modeling, the coin-flipping example illustrates how to obtain the probability function and likelihood function for the parameters of interest. In the next subsection, I will discuss the statistical theories of classic MPT models with formal notations.

2.2.2 Mathematical Representation of MPT Models.

In Hu and Batchelder (1994), the following mathematical expressions have been developed to represent the MPT models. Let $C_1, ..., C_j, ..., C_J$ denote the observable categories, and $B_{1j}, ..., B_{ij}, ..., B_{I_jj}$ denote the collection of branches whose ending nodes belong to category C_j. In the MPT models for source monitoring (see Figure 1), C_j represents the probability of a categorical response such as A, B or N; B_{ij} represents the probability of a branch in the model such as the first branch of answering A. If the outcomes under a same parent node are binary, $\Theta_s = (\theta_s, 1-\theta_s)$. To be more general, the parameters in a group may be denoted by $(\Theta_s = \theta_{s1}, \ldots, \theta_{sk} \ldots, \theta_{sK_s}) \in \Omega_s = \left\{ [0,1]^S | \sum_{k=1}^{K_s} \theta_{sk} = 1 \right\}$, if there are more than two outcomes in this group (binary). There are S groups, namely $\Theta = (\Theta_1, \ldots \Theta_s, \ldots \Theta_S) \in \Omega = \{\prod_{s=1}^{S} \Omega_s\}$, where Ω is the parameter space, K_s is the number of the parameters nested in the s_{th} group, and $0 < \theta_{sk} < 1$. To estimate the parameters, the first step is to write the mathematical form for the MPT models. In the MPT models, the most basic unit is the link probability $L_{ijl} = (L_{ij1}, ..., L_{ijl}, ..., L_{ijL_{ij}})$, where $l = (1, ..., l_{ij} ..., L_{ij})$ is the lth link on the branch B_{ij}. A link in the MPT models represents the transition probability from one cognitive step to the next. The links then form the branch probability B_{ij} that is the probability from the root node to an ending node of the tree. For example, in the MPT models for source monitoring, the first link in the tree A can be represented as $L_{111} = D_1$, and B_{ij} can be written as the product of the links on this branch, such as $B_{11} = D_1 d_1$. To use a generalized form and facilitate computing, we can present any link probability as the product of all the parameters with their powers:

$$L_{ijl} = \prod_{s=1}^{S} \left(\prod_{k=1}^{K_s} \theta_{sk}^{\alpha_{ijlsk}} \right), \tag{13}$$

$$\sum_{k=1}^{K_s} \theta_{sk} = 1, \tag{14}$$

where the α_{ijlsk} is the summation over links of non-negative integer exponents on θ_{sk}. For instance, in the MPT models for source monitoring, the choices of each step are binary, and the first link under the root node of the tree A can be written as

$L_{111} = D_1^1(1-D_1)^0 d_1^0(1-d_1)^0 a^0(1-a)^0 b^0(1-b)^0 g^0(1-g)^0 D_2^0(1-D_2)^0 d_2^0(1-d_2)^0 = D_1$.

Likewise, we can write a generalized form of the branch probabilities:

$$p_{ij}(\Theta) = Pr(B_{ij};\ \Theta) = c_{ij} \prod_{s=1}^{S} \left(\prod_{k=1}^{K_s} \theta_{sk}^{\alpha_{ijsk}} \right), \tag{15}$$

$$\alpha_{ijsk} = \sum_{l=1}^{L_{ij}} \alpha_{ijlsk}, \tag{16}$$

where $p_{ij}(\Theta)$ is the i_{th} branch probability in the j_{th} category within a tree, and c_{ij} is the product of positive constants on the links in the event that some parameters are set as constants. The use of α_{ijsk} here is to represent the parameters that repeatedly appear on a branch. For example, in the previous coin-flipping example, if the parameters $p = q$, then the power α for p is 2 on B_{11} because $B_{11} = p^2$. Researchers have discussed that the possibility of the constant c_{ij} can arise from the restrictions on some parameters set by the model's hypothesis (Hu and Batchelder 1994; Batchelder and Riefer 1986). In the MPT models, for example, the first branch answering A in the tree A has the probability

$B_{11} = D_1^1(1-D_1)^0 d_1^1(1-d_1)^0 a^0(1-a)^0 b^0(1-b)^0 g^0(1-g)^0 D_2^0(1-D_2)^0 d_2^0(1-d_2)^0 = D_1 d_1$.

At last, the category probability is the summation of the probabilities of the branches going to the same observable response category. For instance, the probability of answering source A as A is $D_1 d_1 + D_1(1-d_1)a + (1-D_1)bg$. Also, this summation can be written in a generalized form as in equation (15)

$$p_j(\Theta) = Pr(C_j;\ \Theta) = \sum_{i=1}^{I_j} \left[c_{ij} \prod_{s=1}^{S} \left(\prod_{k=1}^{K_s} \theta_{sk}^{\alpha_{ijsk}} \right) \right], \tag{17}$$

where

$$\sum_{j=1}^{J} p_j(\Theta) = 1$$

for all $\Theta \in \Omega$. The equations above depict the probability mass functions (PMF) of the MPT models, and the likelihood functions can be obtained from the PMF.

2.2.3 Likelihood Functions of MPT Models.

The likelihood function is the key component in the parameter estimation for MPT models (Hu and Batchelder 1994; Lin and Karabatsos 2006). The joint likelihood of a set of MPT models can be derived from corresponding probability functions and the data provided. Suppose we have a 3×3 data table in which the frequencies are $n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8, n_9$, and their summation is N. The likelihood function for this data given the model is:

$$L = N! \frac{p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4} p_5^{n_5} p_6^{n_6} p_7^{n_7} p_8^{n_8} p_9^{n_9}}{n_1! n_2! n_3! n_4! n_5! n_6! n_7! n_8! n_9!}.$$

Therefore, given the frequency of observations in a category is n_j, the likelihood function for the MPT models is:

$$L(\Theta; < n_j >_{j=1}^{J}) = N! \prod_{j=1}^{J} \frac{[p_j(\Theta)]^{n_j}}{n_j!}, \tag{18}$$

where $p_j(\Theta)$ are given by equation 17, and N is the total number of the observations.

MPT models explore latent cognitive processes in problem-solving by using a hierarchical tree structure, which facilitates understanding of human cognition in a relatively straightforward way. However, there are some insufficiencies in the classic setting of MPT models. Therefore, I will discuss a very important issue in MPT assumption in the next section.

2.3 Crucial Issues of Classic MPT Models and Necessity of the Extension

In spite of the advantages, there exist salient drawbacks due to the hallmark assumptions of classic MPT models. One of the most arguable

assumptions that MPT models stand on is the identical and independently distributed (i.i.d.) assumption. The i.i.d. assumption states that all the subjects within a group are assumed to be equivalent in terms of their cognitive abilities, and all the stimuli in one source are also assumed to be equivalent with respect to their psychological effects to the subjects. This assumption entitles a big advantage to make use of (1) aggregated data (i.e., richer information), therefore yielding more stable estimates for the parameters; and (2) likelihood functions, because parameter estimation methods based on likelihood functions require the observations to be independent (such that the likelihood function may be written as the product of the probability of every single observation), and exchangeable (such that the order of the observations does not matter). Nevertheless, this assumption is quite questionable, because there is no guarantee that subjects within a group have equal cognitive abilities, or stimuli in a set have the same psychological effects. Instead, it is more reasonable to assume there exist individual and item differences.

Another important issue of classic MPT modeling is that the performance on cognitive stages (e.g., D, or g) has not been interpreted in a clear and proper manner. Although researchers tend to interpret some of these parameters as "cognitive ability" (Erdfelder et al. 2009; Kupper-Tetzel and Erdfelder 2012), this is not accurate, because the performance depends on not only the subject's ability, but also the task difficulty.

On the other hand, some researchers have come up with hierarchical MPT models such as latent class MPT or latent trait MPT (Klauer 2006; Stahl and Klauer 2007; Klauer 2009) to model the distribution of individual ability or item effect. While these approaches realize the concern of the i.i.d. issue, the progress is still far from enough, because (1) these approaches do not provide measure of individual abilities and item difficulties, and (2) these approaches still intertwine

item difficulty with subject ability, as classic MPT models. Therefore, psychometric models that are applied to measure individual abilities and item difficulties are necessary to be introduced into MPT modeling to address these issues. In the next section, I will briefly introduce some background to the psychometric models I incorporate into classic MPT models.

2.4 Introduction to IRT Models

In this section, I will introduce some background to classical test theory and IRT models. This includes classical test theory and its drawbacks, IRT and its advantages, and recent variants and development of IRT models.

2.4.1 Classical Test Theory

In the 1960s, item response theory (Rasch 1960; Lord and Novick 1968) was proposed to address the insufficiencies in the classical test theory (CTT) which assumes

$$X_i = T_i + \varepsilon_i, \tag{19}$$

where X_i denotes the observed score of testee i, T_i represents corresponding true score, and ε_i stands for the random error (Lord 1952; Traub 1997; Sijtsma and Junker 2006). In the CTT framework, for a fixed testee i, the expected value of ε_i is assumed to be 0, that is, $E(\varepsilon_i) = 0$. Therefore, the expected value across the testees in a population also is 0. Since $E(\varepsilon_i) = 0$, $T_i = E(X_i)$.

In spite of simplicity of CTT model and its assumptions, a crucial drawback of CTT is that examinee characteristics and test characteristics cannot be separated. According to the equation of CTT, both the observed score and the true score depend on the joint effect of the testee's ability and the test item difficulty. Hence ability and difficulty can only be interpreted in the context of each other. Details of the shortcomings of the CTT are discussed in Hambleton, Swaminathan, and Rogers (1991).

2.4.2 Item Response Theory

To solve these issues, IRT tries to separate item difficulty from ability (also known as latent trait) of the examinee. IRT framework entails three basic assumptions (Sijtsma and Junker 2006): (1) a unidimensional trait (ability) denoted by θ, (2) local independence of items; (3) the probability of the response of a testee to an item, which can be modeled by the item characteristic function/curve (ICF, or ICC).

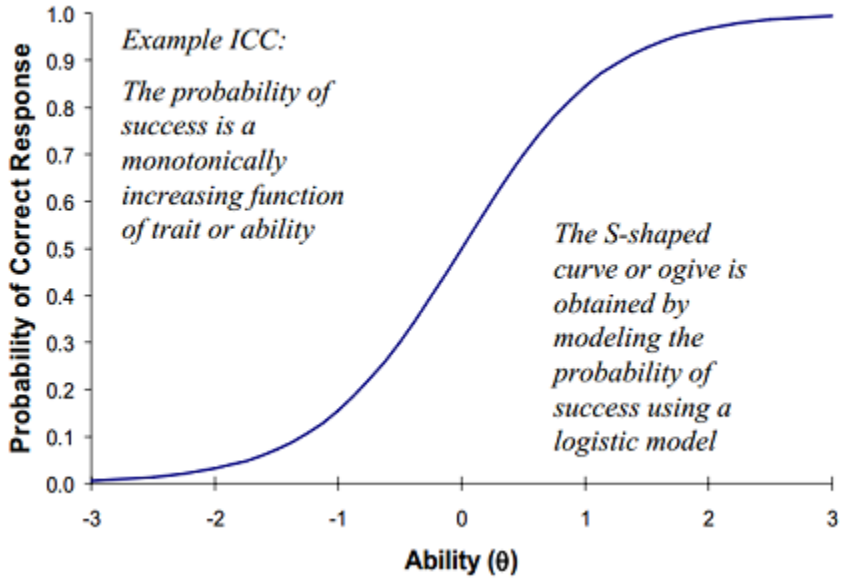

Figure 6
An Example of The ICC for A Given Item.

The unidimensional trait assumption indicates that only one trait (e.g., mathematics) is measured in a set of test items. “Local independence” is related to the unidimensionality assumption, which means item responses are independent of one another, given ability. This assumption is a hallmark of IRT, because only if the response to an item neither relies on the response of any previous items, nor influences subsequent items, can we simply write the probability of seeing the overall response (i.e., likelihood, which will be discussed

in following paragraphs) as the product of the probability of each response. At last, ICC imposes a logistic link function to model the relationship between examinee ability and item difficulty, as in equation 20

$$Pr\{X_{ni}\} = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}}, \tag{20}$$

where $Pr\{X_{ni}\}$ is the probability for examinee n to give correct answer to item i, θ_n is the ability of examinee n, and δ_i is the difficulty of item i. Equation 20 is the simplest form of IRT model, and it is also known as 1-parameter logistic (1PL) IRT, or Rasch model. There exist arguments that Rasch modeling uses a different approach to conceptualize the relationship between data and theory, although this is out of the scope of this paper. See Andrich (1989) for details. To enable the linear combination of θ and δ, IRT references these two parameters to the same scale. Accordingly, ICC is shown in Figure 6.

2.4.3 Variants and Recent Development of IRT Models

In addition to the basic 1-PL IRT model, more parameters have been introduced to investigate the effect of other factors such as discriminability and guessing rate. These models are known as 2-PL IRT and 3-PL IRT models, shown in equation 21:

$$\begin{aligned} Pr\{X_{ni}\} &= \frac{e^{\alpha_i(\theta_n - \delta_i)}}{1 + e^{\alpha_i(\theta_n - \delta_i)}}, \\ Pr\{X_{ni}\} &= \gamma_i + (1 - \gamma_i)\frac{e^{\alpha_i(\theta_n - \delta_i)}}{1 + e^{\alpha_i(\theta_n - \delta_i)}}, \end{aligned} \tag{21}$$

where α represents the slope of the ICC curve (i.e., discrimination rate) and γ is the guessing rate for multiple-choice questions. These two additional parameters are both item parameters, and describe the discriminability, and the probability to get the right answer by guessing for an item (Embretson and Reise 2000). For example, if α is small, the ICC curve is flat, and the probability to get the correct answer increase smoothly. Also, if an item has 4 choices, then $\gamma = 0.25$. The 2-PL

IRT can also be considered as a special form of the 3-PL IRT, when the item has no chance to guess (i.e., $\gamma_i = 0$).

There are some other variants of IRT models, such as normal-ogive IRT which uses a cumulative normal rather than logistic function as the link function (Sijtsma and Junker 2006), IRT models for polynomial responses (Samejima 1969), and partial credit IRT (Masters 1982). These variants try to address more generalized test conditions, or give more reasonable explanation for scoring.

2.4.4 Some Concerns with IRT Modeling

There exist some concerns with IRT modeling. One is the model's performance given different sample sizes. Researchers have different suggestions or arguments on sample size issues. For example, Linacre (1990) suggests 50 subjects for the Rasch model for accurate parameter estimates of ability. Other researchers, such as Tsutakawa and Johnson (1990), Orlando and Marshall (2002), Thissen and Steinberg (2002), have different suggestions for enough sample sizes to obtain adequate estimates for Rasch model and other versions of IRT models. This indicates that the Rasch model may have a different performance with different sample sizes. Another concern is the model's performance when the parameter values are close to the model's boundaries. This is because the ICC curve of the Rasch model is S-shaped, which means its discriminability around the high and low bounds is weaker than in other areas. Moreover, the estimates for the individual parameters may be more sensitive to missing data, because estimating of individual person ability or item difficulty cannot use mean/median like we do in aggregated data. These concerns, of course, will be inherited by models derived from IRT models, and I will discuss and test them in chapter 4.

In the next section, I will discuss the connections of MPT models and test theories, and the natural reasons to extend classic MPT models to a new framework.

2.5 Connections – MPT models and Test Theories

MPT models and test theories (including CTT and IRT) emphasize different facets of the cognitive performance. MPT models stress on the hierarchical relationships of the cognitive processes involved in a cognitive task, while test theories articulate the measuring of the overall ability solving a problem. These two methodologies, actually, have connections and complementarity to each other.

Foremost, if we consider every single cognitive state in an MPT model as a subtask, the concept of "probability to succeed" for classic MPT models is equivalent to the "true score" classic test theory, in that both of these two models are based on the same assumption, as shown in Equation 19. Therefore, classic MPT models only measure the joint effect of subject's ability and item difficulty as the classic test theory does, without looking into the essence of this joint effect. This is a marked issue and has never been articulated and solved by current methodologies for MPT models. Therefore, the theory of CTT is applied to classic MPT models to measure subjects' ability at a group (of subjects) level.

Other than the similarity and connections, classic MPT modeling and test theories can help enhance each other. On one hand, classic MPT modeling is not capable of detecting individual differences. This is because the basic assumptions of classic MPT models assume both stimulus and subjects are homogeneous within their groups. Classic MPT modeling is much simplified by this model setting and takes advantage of aggregated data (i.e., stability of the estimates and cumulative data from similar experiments). However, these assumptions are quite arguable and some researchers including Klauer (2006, 2009) have attempted to

address this issue by using methods such as adding hyperparameters to describe the distributions of the individual performances. On the other hand, test theories focus on the subject's overall performance of the whole task. This prevents it from possible applications to more general and realistic situations that involve multiple and structural subtasks and/or abilities. For example, in diagnostic tests, questions often involve various learning contents that are in a hierarchically structured form. Nevertheless, test theories fail to look into the latent cognitive processes of the examinees during problem-solving (Chipman, Nichols, and Brennan 1995), even though this is crucial for assessment purposes. Although all students may give wrong answers, they very likely experience different cognitive processes (e.g., different strategies or subskills). Therefore, it is also necessary to explore more detailed components that lead to final responses for better understanding of students' ability and issues. In the next section, I will outline the integration of MPT models and IRT models.

CHAPTER 3

EXTENDING CLASSIC MPT MODELS TO RASCH MPT MODELS

In this section, I first use a simple signal detection example to give basic senses of Rasch MPT modeling, then formalize notations and the theoretical framework of Rasch MPT models. I further use a source monitoring example to demonstrate this extension, and finally, I introduce the Bayesian inference I use for Rasch MPT model parameter estimation and the implementation in WinBUGS (http:// www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

3.1 A Simple Signal Detection Example

In the signal detection paradigm, subjects are asked to answer if they perceive a specific signal, which may or may not exist. There are four possible results when answering the signal. If the signal is positive (i.e., the signal occurs) and the subject answers correctly, then the answer is considered as a “hit”, otherwise the answer is a “miss”. If the signal is negative (i.e., the signal does not occur) and the subject answers correctly, the answer is considered as a “correct rejection”, otherwise the answer is a “false alarm” (Peterson, Birdsall, and Fox 1954). If we use MPT framework, the signal detection paradigm can be illustrated in Figure 7.

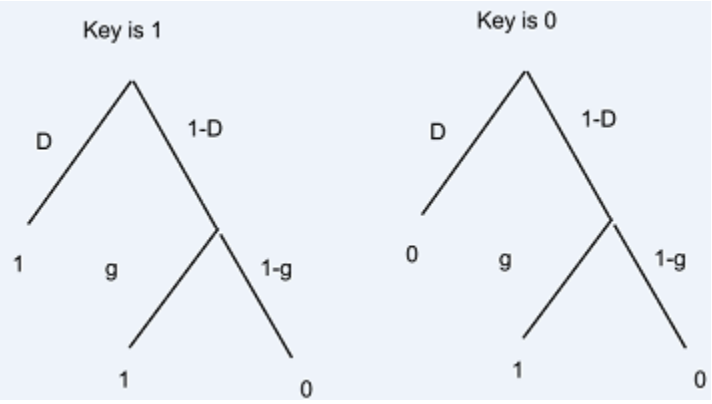

Figure 7
MPT models for signal detection paradigm.

Suppose there are $(1, ..., m, ..., M)$ subjects and $(1, ..., k, ..., K)$ signals (or non-signals). As an example, for the real signals, the correct answer is 1, then for subject m and signal k, the link probabilities leading to 2 observable outcomes are:

$$L_{111km} = D_{km} = \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \tag{22}$$

$$L_{221km} = g_{km} = \frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})}, \tag{23}$$

where L_{111km} means this parameter is on the first link of the first branch, in the first response category of the whole cognitive process, for the k_{th} signal and m_{th} subject. Similarly, L_{221km} means this parameter is on the second link of the second branch, in the first category, for the k_{th} signal and m_{th} subject.

Further, the branch probabilities are:

$$B_{11km} = D_{km} = L_{111km} = \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \tag{24}$$

$$B_{21km} = (1 - D_{km})g_{km} = \left[\frac{1}{1 + exp(\theta_{1m} - \delta_{1k})}\right] \times \left[\frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})}\right] \tag{25}$$

$$B_{31km} = (1 - D_{km})(1 - g_{km}) = \left[\frac{1}{1 + exp(\theta_{1m} - \delta_{1k})}\right] \left[\frac{1}{1 + exp(\theta_{2m} - \delta_{2k})}\right], \tag{26}$$

where B_{11km} means branch probability for the first branch in the first response category for signal k and subject m.

Finally, the category probabilities are:

$$\begin{aligned} C_{1km} &= B_{11km} + B_{21km} \\ &= D_{km} + (1 - D_{km})g_{km} \\ &= \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} + \frac{1}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})} \\ &= \frac{exp(\theta_{1m} - \delta_{1k})[1 + exp(\theta_{2m} - \delta_{2k})] + exp(\theta_{2m} - \delta_{2k})}{[1 + exp(\theta_{1m} - \delta_{1k})][1 + exp(\theta_{2m} - \delta_{2k})]} \end{aligned} \tag{27}$$

$$
\begin{aligned}
C_{2km} &= B_{31km} \\
&= (1 - D_{km})(1 - g_{km}) \\
&= \frac{1}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{1}{1 + exp(\theta_{2m} - \delta_{2k})},
\end{aligned} \tag{28}
$$

where C_{1km} means the category probability of the first response category for signal k and subject m.

From this simple example, we can see the advantages and disadvantages of classic MPT models and Rasch models discussed in the previous chapter. The classic MPT modeling depicts an interpretable and intuitive explanation for the cognitive processing structure in the signal detection task. However, it stops at this level, without exploring the underlying reasons for the really interesting cognitive factors that may cause the performance differences. On the other hand, the Rasch modeling cannot even tell what exact ability or difficulty is being measured, even though it does tell whether one person is overall better/worse than another in the signal detection task, or one signal is overall harder/easier than another signal. Once we plug in the Rasch model into MPT models, we are able to benefit from the advantages of both MPT and Rasch modeling, and clearly state how people may process the signals and give final responses in the signal detection task.

In the next section, I will formally introduce the notation and framework of Rasch MPT modeling, followed by the demonstration of Rasch MPT modeling for the 1HTH MPT for source monitoring paradigm.

3.2 Notations and Theoretical Framework of Rasch MPT Models

Suppose there are $(1, ..., s, ..., S)$ parameter groups for the parameter vector $\Psi = (\Psi_1, ..., \Psi_s..., \Psi_S)$. Ψ corresponds to the parameter vector Θ in the classic MPT modeling, which represents all the cognitive states in a specific task. Since we will plug the Rasch model into each of the parameters in Ψ, which

enriches the meaning of these cognitive states, a similar but different parameter vector Ψ is recruited to take the place of Θ. For Ψ_s, there are s possible mutually exclusive outcomes. For simplicity purposes, I start from binomial results (ψ_s or $1-\psi_s$) for each state, and use the Rasch model to represent the simplest form of IRT models. Suppose there are $\{1, ..., m, ..., M\}$ subjects, and $\{1, ..., k, ..., K\}$ test items; Suppose there are $\{1, ...j_k, ..., J_k\}$ possible answers (i.e., observable categories) for item k, $\{1, ...i_{jk}, ..., I_{jk}\}$ branches nested in the jk^{th} category, and $\{1, ...l_{ijk}, ..., L_{ijk}\}$ links (i.e., states) nested in the ijk^{th} branch. Suppose state s in item k has δ_{sk} as its state difficulty parameter. Suppose subject m has θ_{sm} as the ability parameter for state s (θ_{sm} is independent from item k). The idea of integrating classic MPT models with Rasch model is to extend each cognitive state (e.g., detection step D in Figure 1) to a Rasch model. This is feasible because the state parameters represent the probability of success on this state, while the Rasch model uses a logistic link function to model the relationship between the success probability of a task and the difference of subject ability and task difficulty.

The first step for Rasch MPT modeling is to plug in Rasch models to a cognitive state (i.e., a link):

$$\begin{aligned} Pr[L_{lijkm}] &= \prod_{s=1}^{S} \psi_{skm}^{\alpha_{lijs}} (1-\psi_{skm})^{\beta_{lijs}} \\ &= \prod_{s=1}^{S} \left[\frac{exp(\theta_{sm}-\delta_{sk})}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\alpha_{lijs}} \left[\frac{1}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\beta_{lijs}}, \end{aligned} \quad (29)$$

where the exponent α_{lijs} is the power of ψ_{skm} and β_{lijs} is the corresponding power of $1-\psi_{skm}$. These two exponents can either be 1 or 0, and $\sum_{s=1}^{S}[\alpha_{lijs}+\beta_{lijs}]=1$. In Equation 29, a general form that multiplies all the parameters in the parameter vector is used because this form contains all the possible parameters in such that we do not need to use specific forms for each of them, but only use this form to

represent a specific parameter by setting its exponent as 1, while all others as 0. For instance, in Equation 22, because all other parameters except L_{111km} do not appear on the first link, their exponents are all 0, which lead to a constant 1 in the multiplication, making it needless to write out.

The next step is to multiply the links that appear on the same branch to get branch probability:

$$\begin{aligned} Pr[B_{ijkm}] &= \prod_{s=1}^{S} \psi_{skm}^{\alpha_{ijs}} (1-\psi_{skm})^{\beta_{ijs}} \\ &= \prod_{s=1}^{S} \left[\frac{exp(\theta_{sm}-\delta_{sk})}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\alpha_{ijs}} \left[\frac{1}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\beta_{ijs}} \end{aligned} \tag{30}$$

Here, α_{ijs} and β_{ijs} have similar meaning with α_{lijs} and β_{lijs}, while they sum up over the branch, instead of the link.

Then, we sum up the branches in the same category to get the categorical probability:

$$\begin{aligned} Pr[C_{jkm}] &= \sum_{i=1}^{I_{jkm}} \prod_{s=1}^{S} \left[\psi_{skm}^{\alpha_{ijs}} (1-\psi_{skm})^{\beta_{ijs}}\right] \\ &= \sum_{i=1}^{I_{jkm}} \prod_{s=1}^{S} \left[\frac{exp(\theta_{sm}-\delta_{sk})}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\alpha_{ijs}} \left[\frac{1}{1+exp(\theta_{sm}-\delta_{sk})}\right]^{\beta_{ijs}} \end{aligned} \tag{31}$$

The final step is to establish the likelihood function. Given the available data in practice, only category responses can be observed (i.e., the subjects only report a test item as “A”, “B”, or “N”). Namely, although we have link probability and branch probability functions, some links are not observable (because they result in a same observed category as other links do), and hence not applicable for the likelihood for parameter estimation. This means, only the category probabilities may be used for the likelihood function for parameter estimation. The joint likelihood function based on observed categorical frequencies is:

$$L(\Theta; < n_{jkm} >_{j=1}^{J_k}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \left[N! \prod_{j=1}^{J_k} \frac{[Pr[C_{jkm}(\Theta)]]^{n_{jkm}}}{n_{jkm}!} \right], \tag{32}$$

where n_{jkm} is the observed frequency of a category, N is the total frequency distributing multinomially, and $Pr[C_{jkm}(\Theta)]$ is the category probability defined in Equation 31.

Seemingly, this parameter setting doubles the number of parameters. However, if we consider individual difference and/or item difference, Rasch MPT models may have fewer parameters. This is because Rasch models assume the difficulty δ for a given task is invariant across all the subjects, and ability θ for a given subject is also the same across all the tasks that measure the same latent trait (e.g., ability for detecting old items). Provided the same condition (that assumes individual and item differences), any cognitive state success probability is different from one another, hence the parameter number increases sharply as the number of subjects and items/cognitive states increase. Usually, researchers only count S parameters in the parameter vector. However, if the homogeneity assumption for subject or task does not hold, the total parameters in classic MPT models should be timed by M and/or K. Table 2 is a comparison of parameter numbers given the setting of classic MPT and Rasch MPT models.

Table 2
Number of parameters in classic MPT and Rasch MPT Models

Subject Group/Task Pool	Classic MPT	Rasch MPT
Homo/Homo $(M = K = 1)$	S	$2S$
Hetero/Homo $(M > 1, K = 1)$	$M \times S$	$(M + 1) \times S$
Homo/Hetero $(M = 1, K > 1)$	$K \times S$	$(1 + K) \times S$
Hetero/Hetero $(M > 1, K > 1)$	$M \times K \times S$	$(M + K) \times S$

Table 2 shows the number of parameters in classic MPT and Rasch MPT models in the conditions of homogeneous subjects and homogeneous tasks, heterogeneous subjects and homogeneous tasks, homogeneous subjects and heterogeneous tasks, and heterogeneous subjects and tasks.

In the next section, I will demonstrate Rasch MPT modeling in the source monitoring paradigm.

3.3 Demonstration of Rasch MPT Models

In a typical source monitoring experiment, the answers can be classified into three types. The first type is correct answers, second is incorrect but related answers, and last is incorrect and unrelated answers (refer to Figure 1 for details). For source A, these three types of answers are A, B, and N, respectively.

Suppose in a 5-parameter class MPT (see Figure 2 sub-model 5a), $\psi_{skm} = (\psi_{1km}, ..., \psi_{5km}) = (D_{km}, d_{km}, b_{km}, a_{km}, g_{km})$. If we assume a unified tree structure, in which each source tree (including old and new sources) consists of the same structure with all the possible cognitive states. By using this general representation, different types of the sources can be written in the same form, and specified by their own parameter setting (e.g., the tree for a new source can be considered as $D_{km} = a_{km} = 0$).

The link probabilities of the first and second links on the first branch and first category for item k and subject m can be written as:

$$Pr[L_{111km}] = D_{km} = \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})}, \tag{33}$$

$$Pr[L_{211km}] = d_{km} = \frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})}, \tag{34}$$

where θ_{1m} and δ_{1k} represent the ability of the m_{th} person, and the difficulty of the k_{th} source item in the first cognitive task (i.e., D), respectively.

After the link probability is established, we can obtain branch probabilities. For example, the branch probability for the first and second branches in the first category for item k and subject m can be written as:

$$Pr[B_{11km}] = D_{km}d_{km} = \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})} \tag{35}$$

$$\begin{aligned} Pr[B_{21km}] &= D_{km}(1 - d_{km})a_{km} \\ &= \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{1}{1 + exp(\theta_{2m} - \delta_{2k})} \times \frac{exp(\theta_{4m} - \delta_{4k})}{1 + exp(\theta_{4m} - \delta_{4k})} \end{aligned} \tag{36}$$

And similarly, the category probability of the first category for item k and subject m can be written as:

$$\begin{aligned} Pr[C_{1km}] &= Pr[B_{11km}] + Pr[B_{21km}] + Pr[B_{41km}] \\ &= D_{km}d_{km} + D_{km}(1 - d_{km})a_{km} + (1 - D_{km})b_{km}g_{km} \\ &= \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{exp(\theta_{2m} - \delta_{2k})}{1 + exp(\theta_{2m} - \delta_{2k})} + \\ &\quad \frac{exp(\theta_{1m} - \delta_{1k})}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{1}{1 + exp(\theta_{2m} - \delta_{2k})} \times \frac{exp(\theta_{4m} - \delta_{4k})}{1 + exp(\theta_{4m} - \delta_{4k})} + \\ &\quad \frac{1}{1 + exp(\theta_{1m} - \delta_{1k})} \times \frac{exp(\theta_{3m} - \delta_{3k})}{1 + exp(\theta_{3m} - \delta_{3k})} \times \frac{exp(\theta_{5m} - \delta_{5k})}{1 + exp(\theta_{5m} - \delta_{5k})} \end{aligned} \tag{37}$$

For the likelihood function, we have:

$$L = \prod_{m=1}^{M} \prod_{k=1}^{K} N! \prod_{j=1}^{J_k} \frac{\left[\sum_{i=1}^{I_{jkm}} \prod_{s=1}^{S} \left(\frac{exp(\theta_{sm} - \delta_{sk})}{1 + exp(\theta_{sm} - \delta_{sk})} \right)^{\alpha_{ijs}} \left(\frac{1}{1 + exp(\theta_{sm} - \delta_{sk})} \right)^{\beta_{ijs}} \right]^{n_{jkm}}}{n_{jkm}!}. \tag{38}$$

Rasch MPT models introduce two advantages over the classic MPT models. One is that Rasch MPT models measure ability and difficulty of each subject on each cognitive state (subtask) in a precise manner. The other is that Rasch MPT models model every single performance by independent subject ability and subtask difficulty, while classic MPT models are not able to. The

relationship between Rasch MPT models and classic MPT models can be viewed as a counterpart of the classic test theory and IRT model pair, because as introduced previously, the classic test theory model mixes the subject ability and item difficulty, therefore can only estimate their interactive performances.

In the next section, I will discuss the likelihood function and parameter estimation of Rasch MPT models. Specifically, I will use the Bayesian inference to estimate the parameters in the Rasch MPT models. In addition, I will systematically evaluate the performance of Rasch MPT models in various conditions.

3.4 Estimation Using Bayesian Inference

In this section, I discuss the background of Bayesian inference, including its origin, rationale, some advantages and potential issues. Then, I recruit Bayesian inference to recover the parameter values generated by various simulation conditions. This is to systematically evaluate the performance of Rasch MPT models and conclude the situations in which Rasch MPT models may or may not be an appropriate measurement tool.

3.4.1 Theories of Bayesian Inference

The Bayesian inference is derived from the concept of Bayesian probability, the basic idea of which is that any given probability should be a conditional probability (posterior probability), impacted by the prior probability. Therefore information obtained is connected with prior information, and will influence the prediction. The Bayesian parameter estimation method can be considered as an alternative of the maximum likelihood estimation (MLE) . The two most important differences between Bayesian and traditional Frequentists' perspective are 1) whether prior knowledge about the studied objects is involved and, 2) whether the estimate of a parameter is a fixed value or a distribution (Carlin and Louis 2009). In Bayesian probability theory, given observed data and a hypothesis, the

posterior probability is proportional to the product of the likelihood function and the prior probability. The likelihood function represents the information from the data and the model, while the prior specifies the hypothesis before the data was observed:

$$Pr(\Theta|D) = \frac{Pr(D|\Theta)\ Pr(\Theta)}{Pr(D)} \quad , \tag{39}$$

where Θ is a parameter vector and D is the data. $Pr(\Theta)$ is the prior probability of Θ, $Pr(D|\Theta)$ is the conditional probability of observing the data given Θ, namely, $Pr(D|\Theta)$ is the likelihood. $Pr(D)$ is the marginal probability of D, and finally $Pr(\Theta|D)$ is the posterior probability of Θ. The meaning of $Pr(\Theta|D)$ is the probability that the hypothesis is true, given the data and the previous belief about Θ (the prior). So equation (39) can be rewritten as:

$$Pr(\Theta|D) = \frac{Pr(D|\Theta)\ Pr(\Theta)}{\sum Pr(D|\theta_i)Pr(\theta_i)} \quad , \tag{40}$$

where θ_i is every single possible value of Θ if the distribution of Θ is discrete, or

$$Pr(\Theta|D) = \frac{Pr(D|\Theta)\ Pr(\Theta)}{\int_\Omega Pr(D|\tilde{\Theta})Pr(\tilde{\Theta})d\tilde{\Theta}} \quad , \tag{41}$$

where Ω is the parameter space, if the distribution of Θ is continuous (Hoff 2009). Therefore, the most important components of Bayesian formula are the prior distribution and the likelihood function.

Let us again consider the coin-flipping example introduced in section 2.2.1. The Bayesian inference for the posterior of the parameter vector is: $Pr(\Theta|D)$, and $Pr(D|\Theta)$ here is the likelihood function $L(\Theta; D)$ as given in equation (11), and $Pr(\Theta)$ is a prior distribution of the independent parameter vector $\Theta = (p, q, r)$ assigned by the researcher (say, a beta distribution $B_\Theta(\alpha_\Theta, \beta_\Theta)$), and $\int_\Omega Pr(D|\tilde{\Theta})Pr(\tilde{\Theta})d\tilde{\Theta}$ is the integration of the probabilities of the observed data given the range of the parameter vector (here is from 0 to 1). Therefore, the Bayesian inference equation for the coin-flipping example is:

$$Pr(\Theta|D) = \frac{b_1^{n_1} b_2^{n_2} b_3^{n_3} b_4^{n_4} B_\Theta(\alpha_\Theta, \beta_\Theta)}{\int_\Omega b_1^{n_1} b_2^{n_2} b_3^{n_3} b_4^{n_4} B_\Theta(\alpha_\Theta, \beta_\Theta)\ d\Theta}. \tag{42}$$

If we plug in equations (7)–(10), we have:

$Pr(\Theta|D) = \frac{p^{(n_1+n_2)}(1-p)^{(n_3+n_4)} q^{n_1}(1-q)^{n_2} r^{n_3}(1-r)^{n_4} B_p(\alpha_p,\beta_p) B_q(\alpha_q,\beta_q) B_r(\alpha_r,\beta_r)}{\int_0^1 \int_0^1 \int_0^1 p^{(n_1+n_2)}(1-p)^{(n_3+n_4)} q^{n_1}(1-q)^{n_2} r^{n_3}(1-r)^{n_4} B_p(\alpha_p,\beta_p) B_q(\alpha_q,\beta_q) B_r(\alpha_r,\beta_r) dpdqdr}$.

After simplifying the equation above, we have:

$$\begin{aligned} Pr(\Theta|D) &= \frac{1}{B(\alpha_p,\beta_p)} \frac{p^{n_1+n_2+\alpha_p-1}(1-p)^{n_3+n_4+\beta_p-1}}{B_p(n_1+n_2+\alpha_p, n_3+n_4+\beta_p)} \\ &\quad \frac{1}{B(\alpha_q,\beta_q)} \frac{q^{n_1+\alpha_q-1}(1-q)^{n_2+\beta_q-1}}{B_q(n_1+\alpha_q, n_2+\beta_q)} \\ &\quad \frac{1}{B(\alpha_r,\beta_r)} \frac{r^{n_3+\alpha_r-1}(1-r)^{n_4+\beta_r-1}}{B_r(n_3+\alpha_r, n_4+\beta_r)} \\ &= \frac{B(\alpha_1-1,\beta_1-1) B_p(\alpha_1-1,\beta_1-1)}{B(\alpha_p,\beta_p) B_p(\alpha_1,\beta_1)} \\ &\quad \frac{B(\alpha_2-1,\beta_2-1) B_q(\alpha_2-1,\beta_2-1)}{B(\alpha_q,\beta_q) B_q(\alpha_2,\beta_2)} \\ &\quad \frac{B(\alpha_3-1,\beta_3-1) B_r(\alpha_3-1,\beta_3-1)}{B(\alpha_r,\beta_r) B_r(\alpha_3,\beta_3)}, \end{aligned} \tag{43}$$

where $\alpha_1 = n_1 + n_2 + \alpha_p$, $\beta_1 = n_3 + n_4 + \beta_p$, $\alpha_2 = n_1 + \alpha_q$, $\beta_2 = n_2 + \beta_q$, $\alpha_3 = n_3 + \alpha_r$, $\beta_3 = n_4 + \beta_r$. These equations indicate that the posterior distribution of the parameters is still in the beta distribution family when the prior distribution is conjugate with the likelihood function, and how prior information impacts the posterior distribution.

Although the equation of Bayesian inference is simple, the real computation may be quite difficult because of the integration in the equation, especially when there are many parameters or there are latent variables and incomplete data. Therefore an approximation method named Markov chain Monte Carlo (MCMC) may be used to obtain the approximation of the posterior distribution through iterative algorithms such as the Gibbs sampler and the Metropolis algorithm (Hoff 2009).

3.4.2 Implementation in WinBUGS

In this study, I use WinBUGS to implement the Gibbs sampler algorithm to achieve MCMC estimation for Rasch MPT models. The Gibbs sampler is a technique for generating random variables from the marginal distribution directly, in situations where the conditional distributions of each parameter can be acquired when all the others are fixed. This algorithm does not have to calculate the density, which is difficult to compute in complex cases. Rather than compute or approximate a (marginal) distribution directly, the Gibbs sampler allows us to effectively generate a sample sequence from this distribution without requiring its density.

In the next chapter, I systematically test the performance of Rasch MPT models in different conditions (e.g., different combinations of ability and difficulty values and ranges) for better understanding of the properties of Rasch MPT models.

CHAPTER 4

SIMULATION EVALUATION FOR RASCH MPT MODELS

To understand more properties of Rasch MPT models, I conduct a systematic simulation study that evaluates the performance of Rasch MPT model in different conditions.

Although Rasch MPT modeling addresses the key issues of classic MPT modeling (i.e., violation of subjects/stimulus homogeneity and mixing of subject ability with item difficulty), some issues of Rasch modeling (as discussed in section 2.4.4) may be inherited. Hence there are two main goals of this simulation study. One goal is to examine the performance of Rasch MPT models in different conditions that may take place in reality, including different subject sample sizes and number of items, missing data, and parameter boundary conditions. The other goal is to detect if basic theoretical assumptions may be violated by the setting of the model. Because the Rasch MPT model does not assume homogeneity of items or subjects as classic MPT models do, only parameter independence will be examined. In addition, given that 1HTH classic MPT models have been successful in various source monitoring scenarios (Harvey 1985; Saegert, Hamayan, and Ahmar 1975; Rose et al. 1975), I will use the structure proposed in the 1HTH MPT model to simulate the data.

4.1 Model Performance in Different Conditions

I first tested the parameter recovery given different subject sample sizes, including small, medium, and large sample sizes. The reason for this test is, as discussed in the previous section, the Rasch model may perform differently with different sample sizes. Therefore, there is a need to investigate the parameter estimation accuracy of Rasch MPT models given different sample sizes. For this test, I chose 10 questions with 10 subjects, 20 questions with 20 subjects, and 40 questions with 40 subjects as small, medium, and large sample sizes. Note that the number of subjects could be any arbitrary positive integer and do not

necessarily have to be the same as the number of questions. I used this setting because the probability for getting a correct answer is the function of the difference of θ and δ. Hence both the number of subjects and questions had the same effect on the sample size. Choosing small, medium and large sample sizes for both task items and subjects made their combinations (e.g., 10 items and 10 subjects) to be typical small, medium and large samples. In this study, no mixed combinations (e.g., 10 items with 40 tasks) were tested, and these situations will be discussed in the final discussion.

To illustrate the parameter recovery results, I will use the estimates for the medium sample size as an example. In the medium sample size setting, I chose 20 subjects ($M = 20$), with ability θ mean as 0.5, and standard deviation as 0.5. The MPT model I imposed to simulate data is 5a 1HTH Model in which 5 cognitive states (parameters) are (D, d, a, g, b) (see Figure 2, for details). In addition I recruited a uniform distribution with a range from -0.5 through 1.5 as an approximate non-informative prior, because this prior gives approximately the whole range (95.5%) of possible true values a flat distribution. Table 3 shows the estimates for ability of the first 5 subjects on 5 cognitive states.

Table 3
Rasch MPT Model Recovery for Ability Parameters

Parameter	True value	Mean	SD	MC error	Val2.5pc	Val97.5pc
θ[1,1]	0.59	0.5071	1.067	0.01892	-1.588	2.574
θ[1,2]	-0.60	-0.6769	0.8115	0.01357	-2.226	0.9646
θ[1,3]	1.40	1.362	0.8854	0.01446	-0.4418	3.053
θ[1,4]	0.54	0.5087	0.9949	0.007028	-1.417	2.477
θ[1,5]	-0.93	-0.9619	0.6419	0.0108	-2.311	0.2025
θ[2,1]	0.79	0.7738	0.9795	0.01921	-1.151	2.735
θ[2,2]	-0.08	-0.1146	0.8284	0.01526	-1.755	1.553
θ[2,3]	1.32	1.286	0.886	0.01629	-0.4823	2.994
θ[2,4]	0.58	0.4966	0.9971	0.007212	-1.474	2.448
θ[2,5]	-0.32	-0.3723	0.6809	0.01285	-1.85	0.8113
θ[3,1]	0.80	0.7211	0.9981	0.01994	-1.218	2.667
θ[3,2]	-0.02	-0.05546	0.855	0.01822	-1.715	1.629
θ[3,3]	1.39	1.323	0.8675	0.01439	-0.4675	2.991
θ[3,4]	0.58	0.5044	0.9944	0.00714	-1.448	2.441
θ[3,5]	-0.40	-0.4159	0.7042	0.01457	-1.949	0.8279
θ[4,1]	0.91	0.8932	0.9717	0.01908	-0.9828	2.807
θ[4,2]	0.19	0.1353	0.8494	0.01731	-1.569	1.816
θ[4,3]	1.31	1.225	0.8818	0.01329	-0.5344	2.964
θ[4,4]	0.56	0.4964	0.9926	0.006471	-1.457	2.442
θ[4,5]	-0.07	-0.07727	0.6941	0.01454	-1.583	1.16
θ[5,1]	0.88	0.7883	0.9744	0.0193	-1.079	2.717
θ[5,2]	0.04	0.0352	0.8619	0.01743	-1.708	1.76
θ[5,3]	1.27	1.254	0.8661	0.01355	-0.4637	2.904
θ[5,4]	0.58	0.4971	1.003	0.007482	-1.489	2.473
θ[5,5]	-0.33	-0.2395	0.7063	0.01398	-1.769	1.004

On the other hand, I chose 20 source items ($K = 20$) and set the cognitive state difficulty parameter δ with a mean of 0, and standard deviation as 0.5. I recruited a uniform distribution ranging from -1 to 1, which is also an approximately non-informative prior because it covers nearly the whole range (95.5%) of the possible true values with an equal probability. Table 4 shows the estimates for ability of the first 5 subjects on 5 cognitive states.

Table 4
Rasch MPT Model Recovery for Difficulty Parameters

Param	True value	Mean	SD	MC error	Val2.5pc	Val97.5pc
δ[1,1]	0.02	-0.05086	1.046	0.02045	-0.07252	2.02
δ[1,2]	1.08	0.9993	0.8421	0.01807	1.028	2.599
δ[1,3]	-0.7	-0.8723	0.8612	0.01464	-0.8743	0.8367
δ[1,4]	0.14	-0.004513	0.9964	0.006168	-0.008115	1.966
δ[1,5]	1.42	1.358	0.6527	0.01159	1.32	2.747
δ[2,1]	-0.13	-0.1724	1.014	0.01892	-0.1823	1.778
δ[2,2]	1.04	0.9449	0.8046	0.01473	0.9558	2.548
δ[2,3]	-0.75	-0.7874	0.8851	0.01396	-0.8016	0.9939
δ[2,4]	0.10	-0.003527	0.9968	0.00693	-7.34E-05	1.969
δ[2,5]	1.34	1.174	0.6536	0.01162	1.141	2.575
δ[3,1]	-0.06	-0.2373	1.035	0.021	-0.251	1.798
δ[3,2]	1.14	0.9427	0.7944	0.01469	0.9329	2.547
δ[3,3]	-0.62	-0.7738	0.9072	0.0153	-0.8051	1.102
δ[3,4]	0.11	0.005661	0.9925	0.007226	0.003701	1.938
δ[3,5]	1.33	1.166	0.6524	0.01242	1.122	2.527
δ[4,1]	-0.18	-0.2426	0.9767	0.0196	-0.2283	1.691
δ[4,2]	0.47	0.4137	0.8695	0.01792	0.414	2.168
δ[4,3]	-0.67	-0.8113	0.8813	0.01422	-0.8118	0.9799
δ[4,4]	0.19	-2.67E-04	1	0.007394	0.002247	1.953
δ[4,5]	0.81	0.7748	0.7068	0.01437	0.7046	2.286
δ[5,1]	0.04	-0.05502	1.037	0.02054	-0.05506	2.003
δ[5,2]	0.94	0.8182	0.869	0.01917	0.837	2.416
δ[5,3]	-0.73	-0.8572	0.8762	0.0152	-0.8833	0.9068
δ[5,4]	0.03	-0.002215	1.005	0.007596	-0.01152	1.964
δ[5,5]	1.24	1.187	0.6475	0.01282	0.01908	2.548

The “True value” column in Table 3 and 4 represents the real parameter values generated by the normal distributions I chose for θ and δ. The following columns from “Mean” through “Val97.5pc” are the MCMC sample descriptions of these statistics for the posterior, such as the mean of the posterior, or the value at 97.5 percentile. These descriptions help us understand the distribution of the posteriors. When evaluating the performance of the parameter recovery, we should first check whether the ”Mean” (i.e., the point estimate of the parameter true value) is close to the true value. This is because the basic goal of checking parameter recovery is to test if the estimate of a true parameter value may be close to it. Also, we check if the true value is in the range between “Val2.5pc” and “Val97.5pc”. This can be considered as a minimum requirement of an acceptable recovery. This is because, if the true value falls out of this range, it is either in the upper 2.5% or the lower 2.5% area, namely out of the middle 95% area of the estimated distribution. This means we cannot accept that the true value as a data point belongs to the estimated distribution, at .05 significance level. From Table 3 and 4, we may see that most of the true parameter values were recovered well by posterior means, and all of them are recovered in the range between “Val2.5pc” and “Val97.5pc”, which is consistent with our expectation.

Similarly, I tested the small and large sample size, and a summarization table is given in Table 5. In this table, I used three measures to evaluate the performance of Rasch MPT modeling under different sample sizes. The basic one is whether the true parameter value is in the estimated range. The second measurement is the mean of the absolute difference between the true values and their point estimates. The last measurement is the mean of the standard deviation of each estimate.

Table 5
Summarization of Rasch MPT Model Parameter Recovery

Sample Size	True value in range	Mean of difference	Mean of SD
Small	Yes	0.1413	1.315
Medium	Yes	0.0984	0.8779
Large	Yes	-0.0517	0.6325

As expected, all the true parameter values are recovered in the estimated range, and as the sample size increases, the precision of the estimates is improved significantly.

The second step is to test Rasch MPT models to see if they have poor validity in extreme conditions for θ and δ, usually ranging $[-4, 4]$ if standard normal distributions are assumed (because this range covers over 99.99% area of the possible values of a standard normal distribution). In theory, the discrepancy between δ and θ may be 8 (i.e., $4 - (-4)$). However, in practice, it is very unlikely to test extremely high ability testees with extremly low difficulty tasks. Therefore, the extreme condition I tested here is the absolute difference between θ and δ (i.e., $|\theta - \delta|$) ranges from 3 through 4 (as shown in Figure 6). I generated 250 combinations ($S = K = 5$, $M = 10$) that satisfied this boundary condition, and Table 6 shows the corresponding performance of the Rasch MPT model. From the table we can see that the true value is still in the estimated range, however with much worse precision (i.e., larger difference from the true value and SD), compared with the parameter recovery results in Table 5. This may partly be caused by the low discriminability of the IRT models under boundary conditions, as well as small sample sizes (which is realistic for boundary values).

The third step is to test the Rasch MPT model parameter for recovery performance given partially missing values. This tests the reliability of Rasch MPT estimation in case of partly missing data. Some simple ways used for missing

Table 6
Summarization of Rasch MPT Model Parameter Recovery

$\lvert\theta - \delta\rvert$	True value in range	Mean of difference	Mean of SD
$[3, 4]$	Yes	0.2432	1.821

values in Rasch/IRT model estimation include ignoring them or marking them as incorrect answers (Holman and Glas 2005). However, this is quite problematic because these methods may introduce severe bias to the estimates, especially marking them as incorrect (Rose, von Davier, and Xu 2011). On the other hand, some complex methods such as treating missing values based on additional assumptions (DeMars 2002) are out of the scope of this study. Hence I recruited a relatively straightforward way that used the observed response probability of each category to generate random responses to impute missing values. That is, for example, if a testee responded to 90 out of 100 stimuli, I used the observed probability of the responses to 90 stimuli (e.g., 0.5 for correct answer, 0.4 for incorrect but related answer, and 0.1 for incorrect and unrelated answer) to generate random responses to the last 10 stimuli. So first I use the large sample generated for the previous sample size test and remove 10% responses from 10% of testees, then 20% responses from 20%, and finally 30% responses from 30% testees.

Table 7 shows a summarization of the parameter recovery in these three conditions. This table shows that some of the parameters involving missing values were not recovered in the estimated range, so the percentage of true values in the estimated range is presented. In addition, the percentage for the true values recovered in the estimated range is in terms of the whole parameter vector. We can see that the performance (true values in range and mean difference from true value) worsened significantly as the proportion of missing values increased.

Although the mean of SD did not change much, this is because the sample size was the same after the missing values were imputed. One should note that, the way used to impute the missing values in this study may need to be improved. For example, in reality, the unanswered questions are usually too hard for the students, hence these questions should have lower (even much lower) probability to be “correct”, or “incorrect but relevant”. The main reason for the missing value setting here is to test the robustness of the Rasch MPT’s parameter recovery, hence we stay away from the arguments of the reasons for the missing values.

Table 7
Rasch MPT Parameter Recovery for Missing Data

Missing Data	True value in range	Mean of difference	Mean of SD
10%	99.25%	0.0652	0.5513
20%	95.5%	0.0884	0.5486
30%	88.5%	0.1325	0.5602

4.2 Parameter Correlation Check

The way I tested the parameter independence was to generate 100 random non-aggregate data samples, to estimate the parameters using the 5a sub-model structure in (Batchelder and Riefer 1990), to use correlation tests to check if there exist parameter correlations, and to exhibit in a correlation table. These samples were not used to test the parameter recovery, hence the data in each data table cell will be random numbers ranging from 1 to 200 (with a restriction that the diagonal frequency in the aggregate frequency table is dominant), which is usual for empirical studies. Also, I used small samples with $M = K = 10$ to control the computing load.

Table 8 shows the mean of correlation coefficients between ability parameters, and Table 9 shows the mean correlation between difficulty

parameters. The critical value at $\alpha = .05$ level given sample size ≈ 100 is .195. Hence in these simulated samples, no significant correlation was detected. However, this test only examined linear correlation, and potential complex correlation was not examined.

Table 8
Summarization of Rasch MPT Model Ability Parameter Correlation

Ability Parameters	$\theta[,1]$	$\theta[,2]$	$\theta[,3]$	$\theta[,4]$	$\theta[,5]$
$\theta[,1]$	1	0.14	0.09	0.09	0.06
$\theta[,2]$		1	0.09	0.06	0.05
$\theta[,3]$			1	0.1	0.09
$\theta[,4]$				1	0.05
$\theta[,5]$					1

Table 9
Summarization of Rasch MPT Model Difficulty Parameter Correlation

Difficulty Parameters	$\theta[,1]$	$\theta[,2]$	$\theta[,3]$	$\theta[,4]$	$\theta[,5]$
$\theta[,1]$	1	0.16	0.11	0.1	0.08
$\theta[,2]$		1	0.1	0.08	0.07
$\theta[,3]$			1	0.11	0.09
$\theta[,4]$				1	0.07
$\theta[,5]$					1

Overall, these simulations tested the performance of MPT models under different conditions, including different sample sizes, different portions of missing values, and parameter correlations. These tests showed us which conditions are best for using Rasch MPT models. For example, in our case, we should avoid using small sample sizes such as 10 items and 10 subjects, and we should be cautious if the missing values reach the portion of 30% in total. Also, the

parameter independence assumption should be held. The simulation tests of the Rasch MPT models may help us better understand the properties of the Rasch MPT models, and assist researchers in applying Rasch MPT models to the empirical research. For an instance, in Harvey (1985), 20 manic patients, 20 schizophrenic patients, and 10 normal subjects were recruited for the source monitoring experiments. Our simulation study indicates that 10 normal subject may be insufficient to get reliable and accurate estimates of their cognitive abilities if the research plans to apply Rasch MPT model to measure these subjects. Can we pull all these subjects together to get a larger sample size? Obviously we cannot, because the other two groups are not reasonable references for normal subjects. In other words, the measure of abilities and difficulties are based on the comparison of each subject/task to other subjects/tasks, and hence the measure is a relative measure. However, we should be aware that these tests were based on the specific MPT structure (i.e., sub-model 5a in 1HTH), and more combinations of the conditions exist (e.g., 10 items with 20 subjects). Therefore, fully understanding the performance of a model needs more exploration.

More detailed information about the simulation will be attached in the appendix, including the code used for data simulation and parameter estimation, the data simulated, and the estimates for different conditions tested in this section.

Next, I will use a simple lexical decision experiment and a set of physics test data as applications to discuss the uses of Rasch MPT modeling.

CHAPTER 5

APPLICATIONS OF RASCH MPT MODELS

In this chapter, I will use Rasch MPT models to analyze the empirical data sets obtained from two experiments. The purposes of these analyses are not to explore or discuss the nature of the psychological phenomena involved in these experiments, but to validate the Rasch MPT modeling, and demonstrate its applications in real cognitive studies.

5.1 A Lexical Decision Making Experiment

Traditional memory experiments usually only report aggregated or averaged response data for subjects in one group and stimulus of one source. Therefore, a source monitoring experiment is needed to obtain non-aggregated empirical data to apply Rasch-MPT models. Here I conducted a simple lexical decision making experiment conceived in Link (1982) to test the examinee's response to detect words and non-words (pseudo words). This experiment has been waived by the IRB at the University of Memphis. Please see Appendix for details.

5.1.1 Method

Participants. Twenty workers (anonymous participants who work on experiments/surveys to earn money) on Amazon mechanical turk (AMT) were recruited to finish the lexical decision task. This was an online experiment and only anonymous responses were needed. Hence no personal information was collected.

Design and Materials. In this experiment, subjects were given a list of words. Some of these words were real words, while some were pseudo words. The task was to report whether a stimulus word was a real word or not. Because different subjects may have had different lexicons, they were hypothesized to possess different ability, hence different performance in the test. Also, different words may have had different difficulties (e.g., a commonly used word vs. a rarely used word,

or a random-combination of letters vs. “sanny”). I sampled pseudo words from online sources http://ibbly.com/Pseudo-words.html, as well as added some randomly-combined pseudo words. The sampled words are listed in Table 10. There were 40 words in total, with 20 real words and 20 non-words. The real words were selected from the vocabularies of the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examinations (GRE).

Table 10
Word List for Lexical Decision Experiment

unshott	wave*	chine*	celants	obvious*
sanny	thriste	gement*	hambo*	alies
binated	borato*	unded	estival*	latent*
indigo*	melopepon*	nary*	pennag	ambiguous*
zigant	abandon*	inworm	priole	implicit*
abasement*	enship	heters	refute*	multive
simos	anoes	paradox*	fane*	thwards
lavish*	nauses	wittes	helm*	selfies

Note.Words with an asterisk are real words.

The subjects were assumed to experience the following cognitive processes to output the observed responses: firstly they were to attempt to detect (θ_1) if a word was a real word. If they failed to detect, an additional guessing step (θ_2) was attempted. I modeled θ_1 and θ_2 by Rasch model, and measured the detection and guessing ability of the subjects, and corresponding subtask difficulties.

To make sure the cognitive processes depicted in Figure 8 were valid, I first asked the subject “Do you think this a real word?”; then ”If you think this is a real word, do you know its meaning?”. These two questions guaranteed that the subject would undergo the cognitive processes as illustrated in Figure 8 by first trying to detect (ψ_1) if the word existed in his/her lexicon, and only to guess (ψ_2)

after detection had failed. For simplicity, only the responses to real words were counted (but blinded to the subjects).

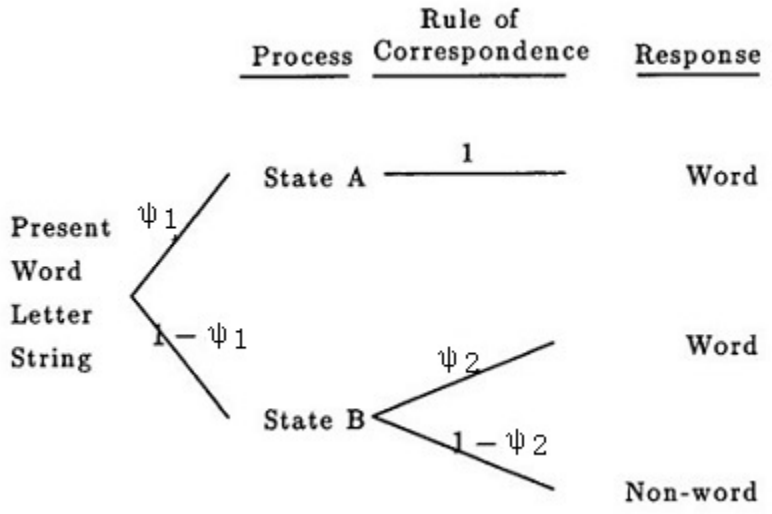

Figure 8
A word recognition experiment

5.1.2 Analysis

There were 3 observed categories for each item (because the subjects were asked if they knew the meaning, correct answers by guessing can be split from those from real recognition) and they are represented by 3 branches in Figure 8. Every participant responded to 40 items, while only responses to 20 real words were used to estimate the ability and difficulty parameters. The ability parameter estimation results for the first 10 participants are shown in Table 11. Similarly, the difficulty parameter estimation results for the first 10 real words are shown in Table 12.

In Table 11, parameter θ is the ability parameter. For example, $\theta[1,1]$ means the first subject's first cognitive stage ability (i.e., real knowledge about a word), and $\theta[1,2]$ refers to this person's second cognitive stage ability (i.e., guessing of a word). The most important value is the (posterior) mean estimate of

θ, which is the point estimate of the ability of a person. The following estimates include the standard deviation of the posterior distribution (SD), the computational accuracy of the mean (MC error), the 2.5th percentile of the simulations as an approximation of the lower endpoint of the 95$\%$ credible interval (Val2.5pc), and he 97.5th percentile of the simulations, an approximation of the upper endpoint of the 95$\%$ credible interval (Val97.5pc). Accordingly, Table 12 shows these statistical descriptions for the difficulty parameter δ. In addition, Table 12 associates the parameter names with they corresponding words to help understand the difficulty of the words measured in the experiment.

Although we have no information about the participant's lexical ability, we can see from Table 12 that common words (e.g., "wave", "obvious", and "latent")

Table 11
Rasch MPT Model Recovery for Ability Parameters

Parameter	Mean	SD	MC error	Val2.5pc	Val97.5pc
$\theta[1,1]$	0.49201	0.95813	0.02390	-0.17069	0.86641
$\theta[1,2]$	0.15670	1.18689	0.01420	0.08443	0.86640
$\theta[2,1]$	1.16490	0.63082	0.01625	0.63286	1.81296
$\theta[2,2]$	1.18808	1.23789	0.01759	1.00928	1.50685
$\theta[3,1]$	1.09908	1.32295	0.01816	0.37649	1.93414
$\theta[3,2]$	0.99776	1.40054	0.01008	0.54149	1.43419
$\theta[4,1]$	1.02336	1.11547	0.01487	0.99540	1.28203
$\theta[4,2]$	0.37684	0.69983	0.01374	-0.05147	0.64254
$\theta[5,1]$	0.84587	0.72042	0.02732	0.80122	1.77050
$\theta[5,2]$	0.67778	1.16518	0.01554	0.02191	0.94548
$\theta[6,1]$	0.74671	0.87175	0.01314	0.31356	1.74387
$\theta[6,2]$	1.01237	0.98078	0.01254	0.67080	1.60876
$\theta[7,1]$	0.52394	1.09473	0.01383	-0.20272	1.41116
$\theta[7,2]$	1.36517	1.33957	0.01604	0.94077	2.08771
$\theta[8,1]$	0.31276	1.14293	0.01732	-0.25270	0.73638
$\theta[8,2]$	0.80073	0.76611	0.01877	0.28228	1.09531
$\theta[9,1]$	0.89355	1.01755	0.01720	0.04166	1.72363
$\theta[9,2]$	1.69200	0.92437	0.01477	1.03077	1.78606
$\theta[10,1]$	0.95778	1.13888	0.01939	0.95030	1.38445
$\theta[10,2]$	0.34948	0.80708	0.02208	-0.60395	0.95713

had low difficulty estimates, while some rare words had much higher difficulty estimates. This observation may be more clearly indicated by the correlation between the average difficulty score of a word and its corresponding “Ngram” at the Google Book (https://books.google.com/ngrams/info). The ”Ngram” value simply means the percentage of a word or phrase used in all the books collected in Google Books. I use the “Ngram” value for the latest available year (i.e., year 2000) for each of the 20 real words (except “melopepon”, which means “any of various kinds of squash” but cannot be found in Google Books). The “Ngram” values of each word, the average difficulty score, and their correlation are presented in Table 13. The table shows high (negative) correlation between the difficulty to know a word and its “Ngram” value in Google Book, which is

Table 12
Rasch MPT Model Recovery for Difficulty Parameters

Word	Param	Mean	SD	MC err	Val2.5pc	Val97.5pc
wave	$\delta[1,1]$	-1.55719	0.78968	0.01354	-2.02530	-1.23242
wave	$\delta[1,2]$	-0.92573	0.93447	0.02102	-1.87454	-0.22884
chine	$\delta[2,1]$	0.81509	1.16977	0.01327	0.44451	1.05017
chine	$\delta[2,2]$	1.17862	0.91599	0.02101	0.85042	1.36517
obvious	$\delta[3,1]$	-0.86424	0.99383	0.01806	-1.32587	0.00747
obvious	$\delta[3,2]$	-2.54810	1.07362	0.00787	-3.49773	-2.03007
gement	$\delta[4,1]$	1.99292	1.27647	0.00765	1.76236	2.69579
gement	$\delta[4,2]$	0.56398	0.87124	0.00549	-0.19942	0.82711
hambo	$\delta[5,1]$	1.45682	1.07175	0.01533	0.67641	1.78512
hambo	$\delta[5,2]$	1.35032	1.16816	0.01513	1.26869	1.85221
borato	$\delta[6,1]$	2.54566	0.95470	0.02135	2.13929	3.48895
borato	$\delta[6,2]$	1.87423	1.23579	0.02301	1.50933	2.37889
estival	$\delta[7,1]$	1.43356	0.94554	0.01850	1.31266	1.71697
estival	$\delta[7,2]$	1.31051	1.10770	0.01230	0.96333	1.87400
latent	$\delta[8,1]$	-0.77044	1.31110	0.02066	-1.51363	-0.15253
latent	$\delta[8,2]$	-1.06988	1.14773	0.01744	-1.87383	-0.89729
indigo	$\delta[9,1]$	0.58483	1.05817	0.01988	0.51106	1.15928
indigo	$\delta[9,2]$	1.13528	1.09728	0.00576	0.44713	1.23418
melopepon	$\delta[10,1]$	1.00707	0.91630	0.02245	0.35439	1.22783
melopepon	$\delta[10,2]$	1.09501	1.11506	0.02486	1.00711	1.82342

significant at .999 level. The correlation between the difficulty to guess a real word as a word shows a lower, yet still significant (negative) correlation with the "Ngram" value. This implies that Rasch MPT models may be used to measure potential sub ability and difficulty. Moreover, because δ_1 reflects the familiarity of a word in a person's mind, it is reasonable to have higher correlation with the Ngram value. Rather, δ_2 may be impacted by other factors (e.g., a person's understanding of morphology), hence has lower correlation with the Ngram value.

Table 13
Correlation Between Ability Scores and "Ngram" Values

Word	δ_1	δ_2	Ngram
wave	-3.26661	-1.02192	0.0045000%
chine	3.331255	1.729232	0.0000143%
obvious	-3.80933	-0.29908	0.0054600%
gement	3.299576	0.088019	0.0000009%
hambo	3.020747	0.020135	0.0000001%
borato	2.396213	0.73968	0.0000001%
estival	2.834477	0.998978	0.0000008%
latent	-2.02712	-0.35634	0.0006555%
indigo	1.48704	0.132873	0.0001304%
melopepon	3.977155	2.435091	0.0000000%
nary	0.593494	0.073102	0.0000320%
ambiguous	-1.0904	-0.419	0.0010270%
abandon	-1.33509	-0.67572	0.0010890%
implicit	-1.55917	-0.20332	0.0015010%
abasement	1.965743	0.067682	0.0000208%
refute	1.060002	0.516078	0.0002224%
paradox	-1.92311	-1.29572	0.0008282%
fane	2.9825	0.289294	0.0000073%
lavish	-0.19934	-0.01597	0.0002657%
helm	1.70584	0.656816	0.0001391%

Note. Correlation between δ_1 and Ngram is $r(19) = -0.77$, $p < .001$ (i.e., significant at .999 level). Correlation between δ_2 and Ngram is $r(19) = -0.46$, $p < .05$ (i.e., significant at .95 level).

In this experiment, the experiment questions were straightforward and the cognitive processes were explicitly regulated. However, in more real and complex situations, cognitive processes are usually unobservable. Therefore, I used a more generalized application to a physics concept test to demonstrate the uses of Rasch MPT models.

5.2 A Generalized Application to Multiple-Choice Questions

In this section, I used the multiple choice questions for Force Concept Inventory (FCI) (Hestenes, Wells, and Swackhamer 1992) from the DeepTutor project (provided by Dr. Vasile Rus). This resource included (1) A 30-question FCI test paper; (2) Documentation that maps the answers of each question to the force concepts; (3) 217 college students' answers to each question.

Because different questions may involve different cognitive processes (i.e., different MPT structures). I first sampled one question and came up with a hypothetical MPT structure to depict the cognitive process for solving this question. Then I used the classic MPT modeling approach to validate the structure(s) by aggregate subject and question data (i.e., test the model's goodness-of-fit to the aggregate data). After the structure was validated, I plugged in Rasch models to measure abilities and subtask (i.e., conceptions) difficulties. Below is a sample question (Table 14), the answer-to-conception mapping (Figure 15), and a hypothetical MPT tree corresponding to this question (Figure 9).

Figure 15 shows the mapping of each answer to the underlying Newtonian force concept(s). Each concept or misconception is represented by a code defined in the FCI. For example, G3 means the belief that "heavier objects fall faster", which is a misconception. In contrast, 5G means the correct concept that "objects fall with the same acceleration regardless of mass". The LP level means the learning progress level on each concept, while FF_L1 refers to the lowest level on the "free fall" (FF) concept, and FF_L6 means the highest level on this concept.

Table 14
Sample Question 1 of FCI

1. Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time. The time it takes the balls to reach the ground below will be: (a) about half as long for the heavier ball as for the lighter one. (b) about half as long for the lighter ball as for the heavier one. (c) about the same for both balls. (d) considerably less for the heavier ball, but not necessarily half as long. (e) considerably less for the lighter ball, but not necessarily half as long.

Table 15
Mapping from Answers to Concepts for Sample Question 1

Answer	FCI Coding	LP Level
a	G3: Heavier objects fall faster	FF_L1: When air resistance is not important, objects of different masses fall at different rates.
b		FF_L1: When air resistance is not important, objects of different masses fall at different rates.
c	5G	FF_L1: Objects fall with the same acceleration regardless of mass.
d	G3: Heavier objects fall faster	FF_L6: When air resistance is not important, objects of different masses fall at different rates.
e		FF_L1: When air resistance is not important, objects of different masses fall at different rates.

Figure 9 shows a hypothetical tree structure for the cognitive process that the students could use to get their final observed responses. Other than the two parameters (“5G” and “G3”) introduced in the concept-question mapping table, there are three additional parameters specified in this tree to help depict the whole cognitive processes. In Figure 9, 5G means the student had the correct

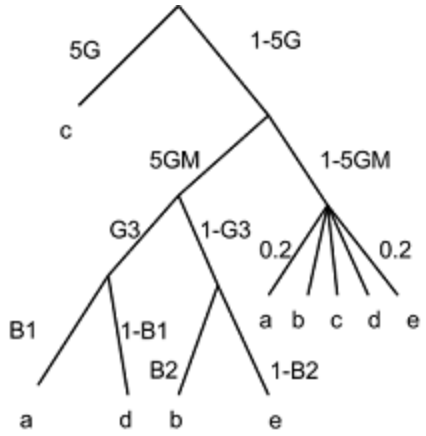

Figure 9
A hypothetical tree structure for sample question 1

concept “5G”, 5GM means the student had misconception(s) about “5G”, and G3 represents the misconception “G3”. We hypothesized that from the root of the tree, if the student had “5G”, the student will obviously get the correct answer (answer c). If the student did not possess “5G”, 5GM may or may not be in mind. If the student had “G3”, this misconception will lead to the answers a or d, depending on the response bias parameter “B1”. However, if “G3” was not the student’s misconception, answer b or e would be observed, depending on the bias parameter “B2”. Finally, if the student had no idea about the concepts involved in the question, a random choice would be made, which means each answer had a probability of 0.2 to be observed.

5.2.1 Model Structure Validation

Given a hypothetical MPT structure, the first step was to use aggregate data to validate the tree structure. This step was crucial because only a tree structure that can be used to represent most subjects’ cognitive process has the potential to further measure ability and difficulty. Therefore, this first step was exactly the same as the procedure in the classic MPT model parameter estimation and model goodness-of-fit test.

The model structure may be validated based on different amounts of information. For example, in the sample question, the

concept(s)/misconception(s) we were interested in were 5G, 5GM, and G3. To test the model fit, we needed at least 5 (= $3 + 1 + 1$) observed categories (i.e., if we only had 4 observed categories, the parameters may have been estimated, but the goodness-of-fit could not be tested because the model would be saturated). Although we had 5 alternative options in each question, we still had to fix the probabilities for parameter B to make the model testable. It was quite subjective to set a constant to B and this setting may possibly have led to a bad goodness-of-fit value. In my trial, that set B as 0.5, $\chi^2(1) = 9.77$, which was unacceptable. The observed frequencies (a = 27, b = 32, c = 104, d = 40, e = 14) also implied that B1 was probably less than 0.5 (a = 27 vs. d = 40), while B2 was greater than 0.5 (b = 32 vs. d = 14). This implied that the bias to a versus d (and b versus e) should not be 0.5. So an ideal way was to set B1 and B2 as free, which demanded more degrees of freedom. To acquire additional information, one could use other questions that involve the same concepts/misconceptions (or say parameters) as in Figure 9. Therefore, I used another question as shown in Table 16, and the mapping from answers to FCI concepts as shown in Table 17.

Table 16
Sample Question 2 of FCI

2.	The two metal balls of the previous problem roll off a horizontal table with the same speed. In this situation: (a) both balls hit the floor at approximately the same horizontal distance from the base of the table. (b) the heavier ball hits the floor at about half the horizontal distance from the base of the table than does the lighter ball. (c) the lighter ball hits the floor at about half the horizontal distance from the base of the table than does the heavier ball. (d) the heavier ball hits the floor considerably closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance. (e) the lighter ball hits the floor considerably closer to the base of the table than the heavier ball, but not necessarily at half the horizontal distance.

Table 17
Mapping from Answers to Concepts for Sample Question 2

Answer	FCI Coding	LP Level
a	5G	FF_L6: Objects fall with the same acceleration regardless of mass.
b	G3: Heavier objects fall faster	FF_L1: When air resistance is not important, objects of different masses fall at different rates.
c		FF_L1: When air resistance is not important, objects of different masses fall at different rates.
d	G3: Heavier objects fall faster	FF_L1: When air resistance is not important, objects of different masses fall at different rates.
e		FF_L1: When air resistance is not important, objects of different masses fall at different rates.

Apparently, sample question 2 has the same tree structure because it involves the same concepts/misconceptions and uses a different scenario to describe the question and adjusts the order of the answers. Therefore, I use these two trees jointly to acquire more degrees of freedom to estimate the parameters $(5G, 5GM, G3, B1, B2)$. The observed frequencies for question 2 were a = 74 (key), b = 54, c = 32, d = 49, e = 8. The estimated parameters were 5G = 0.41, 5GM = 0.65, G3 = 0.65, B1 = 0.58, B2 = 0.82, and $\chi^2(3) = 4.95$, which was smaller than the critical value 7.81 at $\alpha = .05$ level.

5.2.2 Measure of Ability and Item difficulty

After the tree structure was validated, the next step was to plug Rasch models into the classic MPT model to measure θ for the subjects, and δ for the concepts. This step has been illustrated a few times in the previous simulation evaluation chapter, so I will present the estimates of parameters for interesting

concepts (5G and G3, as provided in the concept-question mapping in Tables 15 and 17) in Table 18 and Table 19.

Table 18
Rasch MPT Model Estimates for Ability Parameters in FCI

Parameter	Mean	SD	MC error	Val2.5pc	Val97.5pc
$\theta[1,1]$	1.8825	1.2368	0.0138	1.2991	2.8308
$\theta[1,2]$	0.1157	1.2370	0.0211	-0.8006	0.8739
$\theta[2,1]$	0.8361	1.2371	0.0311	-0.0629	1.7897
$\theta[2,2]$	0.6798	1.2452	0.0329	-0.0086	0.8843
$\theta[3,1]$	-0.8924	1.2417	0.0237	-0.9702	0.0264
$\theta[3,2]$	1.3265	1.2408	0.0270	0.3269	2.0314
$\theta[4,1]$	-0.9013	1.2465	0.0288	-1.0771	-0.1644
$\theta[4,2]$	0.5135	1.2420	0.0189	-0.2045	1.3844
$\theta[5,1]$	0.8362	1.2381	0.0332	0.7930	1.6188
$\theta[5,2]$	0.6796	1.2369	0.0221	-0.2842	1.4139

Table 19
Rasch MPT Model Estimates for Difficulty Parameters in FCI

Concept	Parameter	Mean	SD	MC error	Val2.5pc	Val97.5pc
5G in Q1	$\delta[1,1]$	0.4369	1.2413	0.0159	0.1457	1.2312
G3 in Q1	$\delta[1,2]$	-0.2165	1.2433	0.0330	-1.1889	0.0707
5G in Q2	$\delta[2,1]$	1.4361	1.2418	0.0305	0.5829	1.8539
G3 in Q2	$\delta[2,2]$	-1.1231	1.2441	0.0248	-1.6347	-0.6199

Table 18 shows the first 5 students' standardized ability scores in the population (217 observations). The parameter $\theta[,1]$ refers to the ability score on 5G, which is the correct concept, and $\theta[,2]$ represents the "ability" on G3, which is the misconception that heavier objects fall faster. These scores can be roughly considered as the likelihood of a student to possess this concept/misconception.

In other words, if a student got a higher score on a concept or misconception, she or he is more likely to possess this concept/misconception. So we can find some interesting information from the estimates. For example, the third student and the forth student had similarly low scores on 5G, but differ on G3. This implies that the two students may have had different level on the misconception G3, and the third student may have deeper belief on G3. Therefore, we can discover the information of the students' ability from several aspects: (1) How well a student masters a concept (or how deep a student believes a misconception); (2) How well a student compares to another student or the average of the class, with respect to the mastery of a concept; (3) How well a class master a concept (upper 5%, lower 5%, SD, skewness, etc). Certainly, we may also do the same descriptive and inferential statistical analyses for the aggregated data (e.g., evaluating and comparing subgroups, classes, or schools etc).

In Table 19, although question 1 and question 2 involved the same concept/misconception in the FCI, they showed different difficulties for the students to succeed on 5G ($\delta[1,1] < \delta[2,1]$), however more students' were more likely to possess G3 in question 2. Therefore, this analysis shows that question 1 and question 2 were not equally difficult to the students, not only with respect to the correct answer, but with respect to different misconceptions. Actually, if we look into these two questions, we can find that the first question was more straightforward than the second, because only vertical motion (1 dimension) was involved in question 1, and both vertical and horizontal motions (2 dimensional) were involved in question 2. Likewise, we may conduct evaluation and comparisons for a specific concept, between individual concepts, and groups of concepts (given some concepts can be grouped based on some relationships).

This application in FCI data analysis shows that Rasch MPT modeling has the potential to depict the students' ability, the concepts' difficulty, and various

comparisons. Hence this measure may potentially help understand both the students and the learning tasks in a much deeper and broader way, compared to what a general learning diagnosis does. In the final chapter, I will overall discuss the advantages and disadvantages of Rasch MPT models, as well as some future work.

CHAPTER 6

DISSCUSIONS AND FUTURE WORK

6.1 Advantages and Disadvantages of Rasch MPT Models

As a combination of the cognitive modeling and the psychometric modeling, Rasch MPT modeling possesses obvious advantages. First, it looks deeper into hypothesized cognitive states (e.g., detection, and discrimination), compared with classic MPT models that only depict these states. Namely, Rasch MPT models measure not only different subjects' performance on a cognitive state, but the underlying reasons for these differences (i.e., due to their abilities and task difficulties). Second, Rasch MPT models stand on a more reliable foundation in that neither subjects nor stimuli are assumed identical. This is more reasonable to real situations, especially when we lack accurate information about the subjects and the stimuli. Last, Rasch MPT models integrate two successful models in psychology to their advantages, while offsetting their respective drawbacks.

The integration of classic cognitive models and classic psychometric models can be very helpful in psychometrics. For example, MPT models point out that people may have different processing paths on latent cognitive processes, even though they report the same answer. A subject who gives the same number of correct answers but has a different number of related or unrelated wrong answers actually has different performance in their cognitive processes on the tasks. For example, suppose subject 1 gave the same correct answer of "A"s as subject 2 did. However, subject 1 also gave 10 "B"s with 10 "N"s, while subject 2 gave 1 "B" with 19 "N"s. This may imply that subject 1 actually is more likely to have partial knowledge about the correct answer, compared with subject 2. This argument challenges current evaluation systems that only count the correct answers given by the examinees, and implies that we should consider both correct answers and wrong answers, even in multiple-choice problems. In

addition, Rasch MPT models also help in other related fields. For example, in the student model in intelligent tutoring systems, cognitive diagnostic tests help understand students abilities and problem-solving strategies. For better understanding, more detailed information is needed. Rasch MPT models provide the possibility of microscopic diagnosis for students' cognitive abilities (however, it is also obvious that this kind of diagnosis relies on specific tasks, therefore different cognitive models may be applied to corresponding tasks).

Although Rasch MPT models confer various advantages, we should take note that as we try to improve the precision of measurement, less information is assigned to each single data point and parameter (because we now have many more data points than aggregated data!). Therefore, even though we have enough degrees of freedom to estimate the parameters, less information implies less stable estimates. This can be a reason for large differences between some of the estimates and true values in Table 3 and 4. However, if heterogeneity of the subjects and/or stimuli is the case, it is inappropriate to aggregate data even if more information goes to every data point and parameter.

6.2 Future Study

This study proposes a general framework for Rasch MPT modeling, evaluates its performance, and applies it to empirical studies. This general framework includes a simple example of signal detection tasks, the formal mathematical definition of Rasch MPT modeling, and the demonstration in a real MPT model. The evaluation consists of tests of the performance of a Rasch MPT model under different sample size conditions, different missing data conditions, and whether its parameter independence assumption holds. At last, the empirical application uses two real experiments (a lexical experiment and an FCI experiment) to validate and demonstrate the use of Rasch MPT models in real cognitive studies. The work done in this study provides a formal introduction to

Rasch MPT modeling, how it performs under different conditions, and how it may be applied to psychological studies. However, to understand and apply Rasch MPT models to more practical uses, futher detailed research work needs to be conducted. Future research may be conducted from several aspects: (1) More parameters such as the discriminability parameter in IRT modeling may be added to get more information about the item difficulty. Also some other interesting parameters, such as demographic factors and motivation factors, may possibly be used to model the ability parameter in a linear or generalized linear form. Of course, if we put more parameters into the model, we may have less information for each parameter, hence worse precision for the estimates. (2) Rasch MPT modeling usually uses estimated probability, rather than observed probability, for each cognitive state. A comparison between these two estimations is needed to make researchers aware of the precision of Rasch MPT modeling. (3) More sample conditions, such as a small number of items with a large number of subjects, or vice versa, maybe tested to get more knowledge about the performance of Rasch MPT modeling in these sample size conditions. (4) There are several reasons I did not use random missing values in this study. First, in real situations, most students answer all multiple-choice questions because they can always give an answer. Second, huge computational burden occurs if missing values are fully random, because if missing values exist for every student and each item, a separate imputation needs to be done for each student and item. Besides, discussion about different reasons for missing values is beyond the scope of this study. Missing data used in this study accounts for 10% of the responses from 10% of the subjects. In other words, it appears that about 10% of the subjects with lower ability gave up 10% of the questions. However, another possibility is that missing values are totally random. Some discussion and comparison of these issues regarding Rasch models can be found in Holman and

Glas (2005), DeMars (2002), Rose, von Davier, and Xu (2011). However, similar studies are also needed for further Rasch MPT modeling research.

REFERENCE

Andrich, David. 1989. Distinctions between assumptions and requirements in measurement in the social sciences. In J.A Keats, R. Taft, R.A Heath, and S. Lovibond, editors, *Mathematical and Theoretical Systems*. Elsevier Science Publishers.

Batchelder, W. H., and D. M. Riefer. 1999. Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review*, 6:57–86.

Batchelder, William. 1998. Multinomial processing tree models and psychological assessment. *Psychological Assessment*, 10(4):331–344.

Batchelder, William, and David Riefer. 1990. Multinomial processing models of source monitoring. *Psychological Review*, 97(4):548–564.

Batchelder, William. H., and D. M. Riefer. 1986. The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, 39:129–149.

Bayen, Ute, Kevin Murnane, and Edgar Erdfelder. 1996. Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:197–215.

Carlin, Bradley P., and Thomas A. Louis. 2009. *Bayesian Methods for Data Analysis*. CRC Press, Boca Raton, FL, 3rd edition.

Chipman, Susan, Paul Nichols, and Robert Brennan. 1995. Introduction. In P. D. Nichols, S. F. Chipman, and R. L. Brennan, editors, *Cognitively diagnostic assessment*. Erlbaum.

DeMars, Christine. 2002. Missing data and IRT item parameter estimation. Presented as the Annual meeting of the American Educational Research Association, Chicago, IL.

Embretson, Susan, and Steven. Reise. 2000. *Item Response Theory for Psychologists*. Psychology Press, New York, NY.

Erdfelder, Edgar, Tina-Sarah Auer, Benjamin Hilbig, Andre Abfalg, Morten Moshagen, and Lena Nadarevic. 2009. Multinomial processing tree models: A review of the literature. *Zeitschrift fur Psychology / Journal of Psychology*, 217:108–124.

Garcia-Perez, Miguel. 1990. A comparison of two models of performance in objective tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology*, 43:73–91.

Garcia-Perez, Miguel. 1993. In defence of “none of the above”. *British Journal of Mathematical and Statistical Psychology*, 46:213–229.

Garcia-Perez, Miguel, and R. B. Frary. 1991. Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology*, 44:45–73.

Hambleton, Ronard., H. Swaminathan, and Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Sage Publications, Inc, Newbury Park, CA.

Harvey, Philip. 1985. Reality monitoring in mania and schizophrenia. *The Journal of Nervous and Mental Disease*, 173:67–72.

Hestenes, David, Malcolm Wells, and Gregg Swackhamer. 1992. Force concept inventory. *The Physics Teacher*, 30:141–158.

Hoff, Peter. 2009. *A First Course in Bayesian Statistical Methods*. Springer, New York, NY.

Holman, Rebecca, and Cees A. W. Glas. 2005. Modeling nonignorable missing data mechanism with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58:1–17.

Hu, Xiangen. 2001. Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, 45:603–634.

Hu, Xiangen, and William H. Batchelder. 1994. The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59(1):21–47.

Johnson, Marcia, Mary Ann Foley, and Kevin Leach. 1988. The consequences for memory of imagining in another person's voice. *Memory and Cognition*, 16(4):337–342.

Johnson, Marcia, Shahin Hashtroudi, and Stephen Lindsay. 1993. Source monitoring. *Psychological Bulletin*, 144(1):3–28.

Johnson, Marcia, and Carol Raye. 1981. Reality monitoring. *Psychological Review*, 88:67–85.

Klauer, Carl. 2006. Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, 71(1):7–31.

Klauer, Carl. 2009. Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75(1):70–98.

Kupper-Tetzel, Carolina III, and Edgar Erdfelder. 2012. Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, 20(1):37–47.

Lin, Hua, and George Karabatsos. 2006. A bayesian approach to the multinomial processing tree model for the analysis of a multiple-choice examination of pharmacy knowledge. Midwest Social And Administrative Pharmacy Conference, University of Minnesota, MN.

Linacre, John Michael. 1990. Sample size and item calibrations stability. *Rasch Measurement Transactions*, 7(4):328.

Link, William. 1982. Correcting response measures for guessing and partial information. *Psychological Bulletin*, 92(2):469–486.

Lord, Frederic. 1952. *A Theory of Test Scores*. Psychometrika, Richmond, VA.

Lord, Frederic, and Melvin Novick. 1968. *Statistical theories of mental test scores.* Addison-Wesley Pub. Co., Reading, MA.

Masters, Geoff. 1982. A rasch model for partial credit scoring. *Psychometrika*, 47:149–174.

Matzke, D., C. V. Dolan, W. H. Batchelder, and E.-J. Wagenmakers. 2012. Hierarchical multinomial processing tree models for the pair-clustering paradigm with heterogeneity in participants and items. The 45th annual meeting of the Society for Mathematical Psychology, Columbus, OH.

Meiser, T., and A. Broder. 2002. Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1):116–137.

Myung, I. J., and M. A. Pitt. 2004. Model comparison methods. *Methods in Enzymology*, 383:351–366.

Orlando, M., and G. N. Marshall. 2002. Differential item functioning in a spanish translation of the ptsd checklist: detection and evaluation of impact. *Psychological Assessment*, 14(1):50–59.

Peterson, W. W., T. G. Birdsall, and W. C. Fox. 1954. The theory of signal detectability. pages 171–212. Proceedings of the IRE Professional Group on Information Theory.

Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press, Chicago, IL.

Riefer, David, and William Batchelder. 1988. Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95:318–339.

Rose, Norman, Matthias von Davier, and Xueli Xu. 2011. Modeling nonignorable missing data with item response theory (IRT). Technical report, Educational Testing Service, Princeton, NJ.

Rose, Robert, Patricia Rose, Nelson King, and Alicia Perez. 1975. Bilingual memory for related and unrelated sentences. *Journal of Experimental Psychology: Human learning and Memory*, 1:599–606.

Saegert, oel, Else Hamayan, and Hana Ahmar. 1975. Memory for language of input in polyglots. *Journal of Experimental Psychology: Human Learning and Memory*, 5:607–613.

Samejima, Fumiko. 1969. Estimation of latent ability using a response pattern of graded scores. In *Psychometric Monograph No. 17*, Richmond, VA.

Schmittmann, Verena, Conor Dolan, Maartje Raijmakers, and William Batchelder. 2010. Parameter identification in multinomial processing tree models. *Behavior Research Methods*, 42(3):836–846.

Sijtsma, Klaas, and Brian Junker. 2006. Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, 33:75–102.

Stahl, Christoph, and Carl Klauer. 2007. HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior research methods*, 39(2):267–273.

Stahl, Christoph, and Thorsten Meiser. 2009. New directions in multinomial modeling. *Zeitschrift fur Psychology / Journal of Psychology*, 217.

Thissen, David, and Lynne Steinberg. 2002. A taxonomy of item response models. *Psychometrika*, 51(4):567–577.

Traub, Traub. 1997. Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4):8–14.

Tsutakawa, Robert, and Jane Johnson. 1990. The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, 55:371–390.

Wu, Hao, Jay Myung, and William Batchelder. 2010. On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, 54:291–303.

Yonelinas, Andrew, Ian Dobbins, Michael Szymanski, Harpreet Dhaliwal, and Ling King. 1996. Signal-detection, threshold, and dual-process models of recognition memory: Rocs and conscious recollection. *Consciousness and Cognition*, 5(4):418–441.

APPENDIX A
Mathematical Details

A.1 Bayesian Inference for The Coin-flipping Example

The Bayesian inference equation for the coin-flipping example is:

$$Pr(\Theta|D) = \frac{b_1^{n_1} b_2^{n_2} b_3^{n_3} b_4^{n_4} B_\Theta(\alpha_\Theta, \beta_\Theta)}{\int_\Omega b_1^{n_1} b_2^{n_2} b_3^{n_3} b_4^{n_4} B_\Theta(\alpha_\Theta, \beta_\Theta)\ d\Theta}. \tag{44}$$

If we plug in equations (7)–(10), we have:

$\frac{p^{(n_1+n_2)}(1-p)^{(n_3+n_4)} q^{n_1}(1-q)^{n_2} r^{n_3}(1-r)^{n_4} B_p(\alpha_p,\beta_p) B_q(\alpha_q,\beta_q) B_r(\alpha_r,\beta_r)}{\int_0^1\int_0^1\int_0^1 p^{(n_1+n_2)}(1-p)^{(n_3+n_4)} q^{n_1}(1-q)^{n_2} r^{n_3}(1-r)^{n_4} B_p(\alpha_p,\beta_p) B_q(\alpha_q,\beta_q) B_r(\alpha_r,\beta_r) dpdqdr}$. According to the definition of the Beta distribution:

$$B_p(\alpha_p, \beta_p) = \frac{p^{(\alpha_p-1)}(1-p)^{(\beta_p-1)}}{B(\alpha_p, \beta_p)} \quad , \tag{45}$$

where $B(\alpha_p, \beta_p)$ is the beta function and $B(\alpha_p, \beta_p) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp$. In addition, $B_p(\alpha_p, \beta_p)$, $B_q(\alpha_q, \beta_q)$, or $B_r(\alpha_r, \beta_r)$ are Beta functions for p, q, or r exclusively (e.g., $B_p(\alpha_p, \beta_p)$ is a function for p only, not containing p and r). So these Beta functions may be considered as constants when integrating on other parameters (e.g., when integrating on p, we only consider $\int_0^1 p^{(n_1+n_2)}(1-p)^{(n_3+n_4)}$ as a function of p to be integrated). Therefore, we can further obtain

$$\begin{aligned} Pr(\Theta|D) &= \frac{1}{B(\alpha_p, \beta_p)} \frac{p^{n_1+n_2+\alpha_p-1}(1-p)^{n_3+n_4+\beta_p-1}}{B_p(n_1+n_2+\alpha_p, n_3+n_4+\beta_p)} \\ &\quad \frac{1}{B(\alpha_q, \beta_q)} \frac{q^{n_1+\alpha_q-1}(1-q)^{n_2+\beta_q-1}}{B_q(n_1+\alpha_q, n_2+\beta_q)} \\ &\quad \frac{1}{B(\alpha_r, \beta_r)} \frac{r^{n_3+\alpha_r-1}(1-r)^{n_4+\beta_r-1}}{B_r(n_3+\alpha_r, n_4+\beta_r)} \\ &= \frac{B(\alpha_1-1, \beta_1-1) B_p(\alpha_1-1, \beta_1-1)}{B(\alpha_p, \beta_p) B_p(\alpha_1, \beta_1)} \\ &\quad \frac{B(\alpha_2-1, \beta_2-1) B_q(\alpha_2-1, \beta_2-1)}{B(\alpha_q, \beta_q) B_q(\alpha_2, \beta_2)} \\ &\quad \frac{B(\alpha_3-1, \beta_3-1) B_r(\alpha_3-1, \beta_3-1)}{B(\alpha_r, \beta_r) B_r(\alpha_3, \beta_3)}, \end{aligned} \tag{46}$$

where $\alpha_1 = n_1 + n_2 + \alpha_p$, $\beta_1 = n_3 + n_4 + \beta_p$, $\alpha_2 = n_1 + \alpha_q$, $\beta_2 = n_2 + \beta_q$, $\alpha_3 = n_3 + \alpha_r$, $\beta_3 = n_4 + \beta_r$.

APPENDIX B
Computational Environment and Code
B.1 Configuration of The Computer Used for The Simulation Study

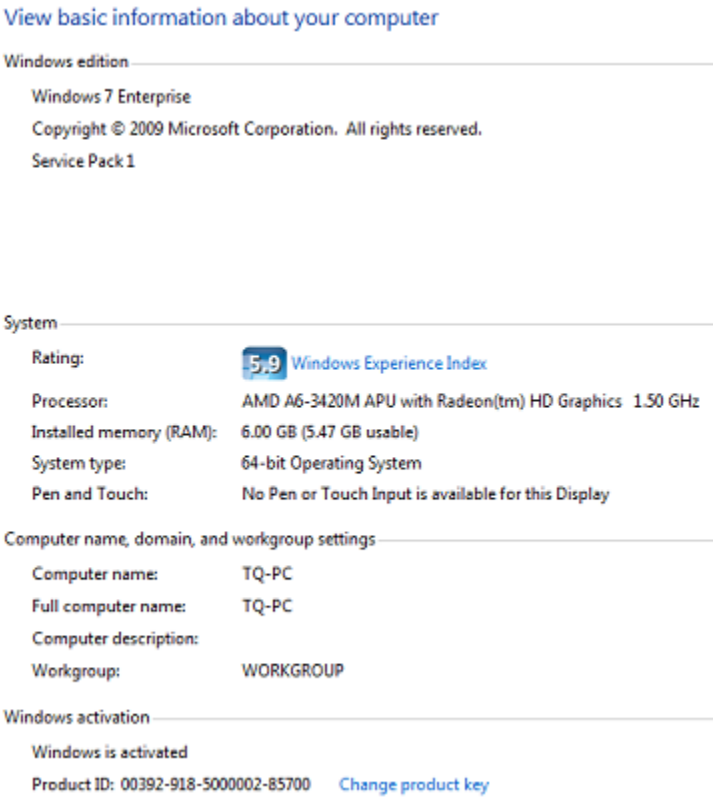

Figure 10
Configuration of The Computer for Simulation and Parameter Estimation Studies

B.2 R Code Used to Simulate Rasch MPT Model Data

```
install.packages("knitr")
library("MASS")
set.seed(100)

N<-50
K<-50
S<-5

theta <- array(0,dim=c(N,S))
delta <- array(0,dim=c(K,S))
psi<-array(0,dim=c(N, K, S))
#response<-array(0,dim=c(N,K, 3))

D<-d<-b<-a<-g<-array(0,dim=c(N,K))
p<-array(0,dim=c(N,K,3))

for (n in 1:N)
  {for (s in 1: S)
    {
    theta[n, s]<- rnorm(1, mean=0.5, sd=0.5)
     }
   }  # prior distribution for student abilities

for (k in 1:K)
  {for (s in 1: S)
    {
    delta[k, s]<- rnorm(1, mean=0, sd=0.5)
     }
   } # prior distribution for item difficulties

#for (n in 1:N) {for (s in 1: S){theta[n, s]<- 0.5}}  # prior distribution for student abilities
#for (k in 1:K) {for (s in 1: S){delta[k, s]<- 0.3}}  # prior distribution for item difficulties
for (n in 1 : N )
  {  # Total number of students: N

    for (k in 1 : K)
      { # Total number of items: K

        for (s in 1 : S)
          {
            x<-theta[n, s] - delta[k, s]
            psi[n, k, s] <- exp(x)/(1+exp(x))  # logit transform
          }
    D[n,k]<-temp<-psi[n, k, 1]
    d[n,k]<-temp<-psi[n, k, 2]
    b[n,k]<-temp<-psi[n, k, 3]
    a[n,k]<-temp<-psi[n, k, 4]
    g[n,k]<-temp<-psi[n, k, 5]

    #tree structure
    p[n, k, 1] <- (D[n,k]*d[n,k]) + (D[n,k]*(1-d[n,k])*g[n,k]) + ((1-D[n,k])*b[n,k]*g[n,k])
    p[n, k, 2] <- (D[n,k]*(1-d[n,k])*(1-g[n,k])) + ((1-D[n,k])*b[n,k]*(1-g[n,k]))
    p[n, k, 3] <- (1-D[n,k])*(1-b[n,k])

    #Simulated Observations (Suppose 1 is correct answer, 2 is related wrong answer,
        #3 is #unrelated wrong answer)

    response<-rmultinom(N*K, size=1, prob=c(p[n,k, 1],p[n,k, 2],p[n,k, 3]))

  }
}
adjresponse<-t(response)
write.table(adjresponse, "C:/response5050.txt", sep=",",col.names = F, row.names = F)
write.table(delta, "C:/delta5050.txt", sep=",",col.names = F, row.names = F)
write.table(theta, "C:/theta5050.txt", sep=",",col.names = F, row.names = F)
```

B.3 R Code Used to Implement Bayesian Analyses for MPT Models

```
#Simulate data
n<-6 #number of parameters
sn<-10 #number of simulated data sets
rslt<-array(1:(sn*n),c(sn,n))
for (q in 1:sn){
  #x<-c(rbeta(n,1,1)) #generate random values for parameters x[1]~x[5] are D1,D2,d,g,b
  x<-rep(0.5,n)
  p<-array(1:9, dim=c(3,3))
  p[1,1] <- (x[1]*x[3]) + (x[1]*(1-x[3])*x[5]) + ((1-x[1])*x[6]*x[5])
  p[1,2] <- (x[1]*(1-x[3])*(1-x[5])) + ((1-x[1])*x[6]*(1-x[5]))
  p[1,3] <- (1-x[1])*(1-x[6])
  p[2,1] <- (x[2]*(1-x[4])*x[5]) + ((1-x[2])*x[6]*x[5])
  p[2,2] <- (x[2]*x[4]) + (x[2]*(1-x[4])*(1-x[5])) + ((1-x[2])*x[6]*(1-x[5]))
  p[2,3] <- (1-x[2])*(1-x[6])
  p[3,1] <-  x[6]*x[5]
  p[3,2] <-  x[6]*(1-x[5])
  p[3,3] <- (1-x[6])

  A<-100
  B<-100
  N<-200
  simdata<-c(A*p[1,1:3],B*p[2,1:3],N*p[3,1:3])
  mean<-sd<-rep(NA,n)

  ############################################################
  TEMP2 <<- array(0)

  initiate<-function() {
    #
    OUTPUT <<- "MPT7.out"    # Name of analysis output file
    Tdraws <<- "MPT7.sam"    # Output files of parameter draws
    #
    #N   <<-  c(23,22,35, 9, 45,26, 7, 10,63)     # Category observations (N11,N12,N13,N14,N21,N22)
    N   <<-  simdata
    Ntot  <<-  c(80,80,80,80,80,80,80,80,80) #Total N per category system
    K.Group <<- c(  1,  1,  1,  2,  2,  2,  3,  3,  3)  # Use this to label the K groups of
    #multinomial distributions
    #
    S   <<- 6       # Number of GPT parameters
    prior.a  <<-  c(1,1,1,1,1,1)  # Beta priors for GPT parameters, shape a
    prior.b  <<-  c(1,1,1,1,1,1)  #     "       "            "            b        , shape b
    #
    T.lbl  <<- c("D1","D2","d","g","b") # sub model 5c
    C.lbl  <<- c("R=S | I=S","R=T | I=S","R=N | I=S",  "R=S | I=T","R=T | I=T","R=N | I=T",
                 "R=S | I=N","R=T | I=N","R=N | I=N") #Input and Response pairs
    Notes1  <<- "Source monitoring analysis"
    Notes2  <<- "Batchelder & Riefer (1990; Psych rev) p. 557 Schizo-TD 3x3 data"
    #
    itstart <<- 500    # Treat iterations 1 to 500 as burn-in
    itend  <<-20000
    #
    ########################################
    Tstart  <<- T0 <<- rep(.5,S) # Starting parameter values
    ## Pbin  <<- rep(0,count.rows(N)) ##
    Pbin  <<- rep(0,length(N))
    iter <<- 0
    s <<- 0
    K  <<-  max(K.Group)
    return()}

  GPT <- function(Ts,S,P,N){
```

```
D1   <-  Ts[1]
D2 <- Ts[2]
d1 <- Ts[3]
d2    <-    Ts[4]
b.  <- Ts[5]
g. <- Ts[6]
#
p11 <- (D1*d1) + (D1*(1-d1)*g.) + ((1-D1)*b.*g.)
p12 <- (D1*(1-d1)*(1-g.)) + ((1-D1)*b.*(1-g.))
p13 <- (1-D1)*(1-b.)
p21 <- (D2*(1-d2)*g.) + ((1-D2)*b.*g.)
p22 <- (D2*d2) + (D2*(1-d2)*(1-g.)) + ((1-D2)*b.*(1-g.))
p23 <- (1-D2)*(1-b.)
p31 <-  b.*g.
p32 <-  b.*(1-g.)
p33 <- (1-b.)
P <- c(p11,p12,p13,p21,p22,p23,p31,p32,p33)   #category probabilities
return(P)}

draw.GPT <- function(T0s,N,Prior.a,Prior.b,S) {
#
s <<- ifelse((s+1)>S,1,s+1) #from the 1st parameter to the next, if finished a round,
#then from the 1st again
draw <- runif(1) #generates random deviates for Uniform distri~, n is number of observations
#
L0 <- GPT(T0,S,P,N)  #give the P vector/category probabilities to L0
L0 <- prod(L0^N)     #likelihood function, p11^N1*p12^N2****p33^N9
GPT0 <- log(prod(L0,T0^(prior.a-1),(1-T0)^(prior.b-1))) #log-posterior function with default base e
#
T1 <- replace(T0,s,draw) #replace value of T0 with No. s value of draw, here use
#a random value generated by runif(1) to replace the former T0 value
L1 <- GPT(T1,S,P,N)
L1 <- prod(L1^N)
GPT1 <- log(prod(L1,T1^(prior.a-1),(1-T1)^(prior.b-1)))
#
accept <- GPT1-GPT0
accept <- ifelse(accept>0,0,accept)  # if GPT1>GPT0, then assign 0 to "accept", else assign GPT1-GPT0
accept <- ifelse(runif(1)<exp(accept),1,0)
T0 <- if(accept==1) T1 else T0
decision <- ifelse(accept,"Accept theta","Reject theta")
#
cat(decision,s,fill=T)
#
return(T0)

}
#
iterate<-function() {
iter <<- iter+1
cat("======== ITERATION     ", iter, " ==========",fill=T)
for (i in 1:S) {
T0 <<- draw.GPT(T0,N,Prior.a,Prior.b,S)

TEMP2<<-T0
}
P0 <- GPT(T0,S,P,N)
TEMP2<<-P0
if(iter>=itstart) Pbin  <<- Pbin + (P0/(itend-itstart+1))
outinfo <- round(c(T0),15)
#write(outinfo,file=Tdraws,ncol=S,append=T)
return(Pbin)
}
```

```
clean <<- function(){
  C.lbl <<- col <<- group <<- itend <<- iter <<- itstart <<- K <<- K.group <<- N <<- Notes1
  <<- Notes2 <<- Ntot <<- c()
  OUTPUT <<- Pbin <<- prior.a <<- prior.b <<- ptile <<- ptile1 <<- Ref.Prior <<- S <<- s
  <<- T.lbl <<- T0 <<- Tdraws <<- c()
  ThetaDraws <<- ThetaDraws.burnin <<- Tstart <<- K.Group <<- c()
  return()}

Bayesian2go<-function() {
  Result<<-array(0)
  initiate()
  for (i in 1:itend)
  {
    Result <<- iterate()

    #print("======== ITERATION   RESULT  FOR N ARRAY ========== ")
    #print(Result)

    print("======== ITERATION   RESULT  FOR PARAMETERS ========== ")

    if (i==1) TE<<-T0 else TE<<-TE+T0
    TF<<-TE/i
    print(TF)

  }
}

Bayesian2go()
rslt[q,]<-TF

}
mean<-apply(rslt,2,mean) #column mean
sd<-apply(rslt,2,sd) #column sd
#
```

B.4 WinBUGS Code Used to Implement Bayesian Analyses for Rasch MPT Models

```
model {  # Simple Rasch MPT in BUGS

  for (n in 1 : N) { # Total number of students: N

    for (k in 1 : K) { # Total number of items: K

      for (s in 1 : S) {
        logit(psi[n, k, s]) <- theta[n, s] - delta[k, s]  # logit transform
      }
      D[n,k]<-psi[n, k, 1]
      d[n,k]<-psi[n, k, 2]
      b[n,k]<-psi[n, k, 3]
      a[n,k]<-psi[n, k, 4]
      g[n,k]<-psi[n, k, 5]

      #tree structure
      p[n, k, 1] <- (D[n,k]*d[n,k]) + (D[n,k]*(1-d[n,k])*g[n,k]) + ((1-D[n,k])*b[n,k]*g[n,k])
      p[n, k, 2] <- (D[n,k]*(1-d[n,k])*(1-g[n,k])) + ((1-D[n,k])*b[n,k]*(1-g[n,k]))
      p[n, k, 3] <- (1-D[n,k])*(1-b[n,k])

      #Simulated Observations (Suppose 1 is correct answer,
      #2 is related wrong answer, 3 is #unrelated wrong answer)

      response[n,k,1:3]~dmulti(p[n,k,1:3], 1)

    }
  }
  # Prior distributions for unknown parameters
for (n in 1:N) { for (s in 1: S){theta[n, s] ~ dunif(0,3)}}#prior distribution for student abilities
for (k in 1:K) {for (s in 1: S){delta[k, s] ~ dnorm(0,1)}}#prior distribution for item difficulties

}
#################Import Simulated Data#################
list(
  N=10,
  K=10,
  S=5,
  response =structure( .Data=c(
    51,34,15,45,38,17,52,34,14,59,32,9,
    59,31,10,55,31,14,49,41,10,56,33,11,
    48,40,12,56,34,10,55,32,13,58,33,9,
    53,38,9,52,32,16,54,33,13,56,32,12,
    54,31,15,45,37,18,56,32,12,47,38,15,
    41,41,18,57,29,14,55,33,12,46,39,15,
    50,37,13,59,29,12,50,26,24,54,37,9,
    50,42,8,50,37,13,44,41,15,51,33,16,
    47,45,8,61,26,13,58,31,11,52,34,14,
    60,29,11,59,29,12,56,30,14,48,38,14,
    58,34,8,62,26,12,47,36,17,51,40,9,
    55,32,13,65,29,6,55,33,12,54,36,10,
    57,33,10,63,30,7,45,37,18,47,41,12,
    57,30,13,48,35,17,48,35,17,53,35,12,
    55,35,10,56,30,14,58,29,13,53,35,12,
    56,31,13,61,31,8,53,35,12,47,34,19,
    63,29,8,53,38,9,49,40,11,60,25,15,
    59,30,11,51,34,15,58,35,7,55,32,13,
    62,24,14,52,34,14,44,40,16,52,32,16,
    55,32,13,57,31,12,56,30,14,53,35,12,
    58,29,13,54,34,12,49,39,12,63,27,10,
    53,35,12,44,45,11,50,37,13,61,33,6,
    48,44,8,48,39,13
  ), .Dim=c(10,10,3),)
)
```

APPENDIX C
Empirical Studies
C.1 Responses Data from The Lexical Experiment

Table 20: Responses Data from The Lexical Experiment for A Sample Subject

WORDS	Worker	Answer	Correct/Incorrect
borato	A2QLSHXNCHBRN4	Non-word	In
celants	A2QLSHXNCHBRN4	Non-word	
anoes	A2QLSHXNCHBRN4	Non-word	
gement	A2QLSHXNCHBRN4	Non-word	In
enship	A2QLSHXNCHBRN4	Non-word	
pennag	A2QLSHXNCHBRN4	Non-word	
estival	A2QLSHXNCHBRN4	Non-word	
fane	A2QLSHXNCHBRN4	Non-word	In
unshott	A2QLSHXNCHBRN4	Non-word	
zigant	A2QLSHXNCHBRN4	Non-word	
wittes	A2QLSHXNCHBRN4	Non-word	
unded	A2QLSHXNCHBRN4	Non-word	
inworm	A2QLSHXNCHBRN4	Non-word	
heters	A2QLSHXNCHBRN4	Non-word	
thwards	A2QLSHXNCHBRN4	Non-word	
simos	A2QLSHXNCHBRN4	Non-word	
abasement	A2QLSHXNCHBRN4	Non-word	In
chine	A2QLSHXNCHBRN4	Non-word	In
hambo	A2QLSHXNCHBRN4	Non-word	In
thriste	A2QLSHXNCHBRN4	Non-word	
nauses	A2QLSHXNCHBRN4	Non-word	
multive	A2QLSHXNCHBRN4	Non-word	
helm	A2QLSHXNCHBRN4	Word	
indigo	A2QLSHXNCHBRN4	Word	
obvious	A2QLSHXNCHBRN4	Word	
refute	A2QLSHXNCHBRN4	Word	
sanny	A2QLSHXNCHBRN4	Word	
alies	A2QLSHXNCHBRN4	Word	
priole	A2QLSHXNCHBRN4	Word	
ambiguous	A2QLSHXNCHBRN4	Word	
abandon	A2QLSHXNCHBRN4	Word	
paradox	A2QLSHXNCHBRN4	Word	
wave	A2QLSHXNCHBRN4	Word	
implicit	A2QLSHXNCHBRN4	Word	
lavish	A2QLSHXNCHBRN4	Word	
binated	A2QLSHXNCHBRN4	Word	
selfies	A2QLSHXNCHBRN4	Word	
latent	A2QLSHXNCHBRN4	Word	
melopepon	A2QLSHXNCHBRN4	Word	

C.2 Empirical Data from the FCI Experiment

Table 21
Sample Data of the FCI Experiment

Teacher	School	StudentID	Classroom	Course	Q1_pre	Q2_pre
1	1	3	11	3	0	0
1	1	5	11	3	0	0
1	1	6	11	3	0	0
1	1	7	11	3	1	0
1	1	8	11	3	0	0
1	1	9	11	3	0	0
1	1	10	11	3	1	0
1	1	11	11	3	0	1
1	1	14	11	3	0	0
1	1	17	11	3	0	1
1	1	18	11	3	0	0
1	1	19	11	3	0	0
1	1	20	11	3	0	0
1	1	22	11	3	1	0
1	1	23	11	3	0	0
1	1	24	11	3	1	1
1	1	26	11	3	1	1
1	1	3	12	2	1	1
1	1	5	12	2	0	0
1	1	6	12	2	1	1
1	1	7	12	2	0	0
1	1	9	12	2	1	1
1	1	10	12	2	1	1
1	1	12	12	2	0	0
1	1	13	12	2	0	0
1	1	19	12	2	0	1
1	1	20	12	2	0	0
1	1	21	12	2	0	0
1	1	22	12	2	1	0
1	1	23	12	2	0	0
1	1	25	12	2	1	1
1	1	27	12	2	0	0
1	1	28	12	2	1	1
1	1	29	12	2	1	1
1	1	31	12	2	0	1
1	1	32	12	2	0	0