A Study of Bayesian Estimation and Comparison of Response Time Models in Item Response Theory

By

Hongwook Suh

Submitted to the graduate degree program in the
Department of Psychology and Research in Education
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

_____
Chairperson

Committee members* _____*

_____*

_____*

_____*

_____*

Date defended: _____

The Dissertation Committee for Hongwook Suh certifies
that this is the approved version of the following dissertation:

A Study of Bayesian Estimation and Comparison of Response Time Models in Item Response Theory

_____
Chairperson

Date approved:_____

Abstract

Response time has been regarded as an important source for investigating the relationship between human performance and response speed. It is important to examine the relationship between response time and item characteristics, especially in the perspective of the relationship between response time and various factors that affect examinee's responses. The purpose of this study was to examine different scoring models using response time data in conjunction with item response models. In this study distinctive response time models incorporated in IRT were compared, and the relationship between item characteristics and examinee ability as well as response time were examined using real and simulated data.

Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods for Thissen's (1983) lognormal response time model, Wang and Hanson's (2005) 4PL RT model, and van der Linden's (2007) hierarchical framework were applied to the investigation of response time on real data. Overall, van der Linden's (2007) hierarchical framework showed the most reasonable outcomes from the real data analysis when it was compared with the 4PL RT and Thissen's models. Compared with Wang and Hanson's (2005) 4PL RT model in the simulated data analysis, the hierarchical framework also showed better results as follows: (1) better recoveries in item and examinee parameter, (2) reasonable explanations in delineating relationships between response time and other related parameters in the model. There were no clear relationships among speed-related parameters across the models when the relationships between the response time-related parameters were investigated across the response time models. This was due to the different definitions and different parameterization procedures of the speed-related parameters based on the response time model.

Acknowledgement

I am grateful to all the people that I met while I have prepared this dissertation. My greatest debt is to Dr. William Skorupski, my academic advisor. He is the one who brought me to the field and provided me unfailing effort and knowledge to pursue this project as my dissertation. I also want to thank each member of my dissertation committee: Dr. Bruce Frey, Dr. Neal Kingston, Dr. Vicky Peyton, and Dr. Kris Preacher. Especially, I would like to thank Kris, who read my writing and corrected every single word. I also want to thank Dr. Kingston, who provided me a financial aid for the last semester at KU. I am also indebted to Bruce and Vicky for their heartful encouragement throughout my study for the last years.

Without the support and cooperation of my family, it would hardly be possible to finish this dissertation. I would like to thank my wife, Sunhyoung Lee, for her loving encouragement and enduring support every day. I also thank my son and daughter, Elliott and Elaine, and my to-be-born daughter for giving me plentiful good reasons to finish dissertation. I also would like to express deep gratitude to my parents and parents-in-law, who supported and encouraged me in their prayers every day since I began to study in the United States.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Appendices

Chapter 1. Introduction

Statement of Problem

Response time on items in a computer–based test enables researchers to study examinees' responses further in test settings and provides valuable information. Response time data allow understanding of examinee behavior from data–based perspectives not previously feasible, and illustrate the important role that these investigations can play in test development, administration, and validation (Schnipke & Scrams, 2002; Zenisky & Baldwin, 2006). Response time has been one of many popular topics in traditional psychological measurement, investigating the relationship between human performance and response speed. Although using response time data is not fully developed in the educational measurement field, it is valuable in understanding human behavior in test settings. It is important to examine the relationship between response time and item characteristics, especially in the perspective of the relationship between response time and various factors that affect examinees' responses.

Response time and test performance have been studied in various ways. Schnipke and Scrams (2002) enumerated the related areas in the measurement field such as scoring models using response time data in conjunction with response data, speed–accuracy relationships, strategy usage, speededness, pacing, predicting finishing times and setting time limits, and subgroup differences. Because several areas are interrelated and quite different perspectives exist depending on the situation, it is not easy to consider only one area without considering the rest. For example, Gulliksen (1950) pointed out two factors of the tests and contrasted power and speed tests. In traditional psychological measurement, response speed and accuracy have been regarded as

interchangeable concepts as accepted in the speed test. However, in the power test situation, speed theoretically is not a related concept; accuracy is independent from response speed. Likewise, speed–accuracy (speed–ability) trade–off and scoring models using response time data also have quite different perspectives when they are applied to speed tests from when they are applied to the power test situation.

Most major standardized achievement tests are power tests, which indicate the goal of testing is to measure how accurately examinees respond to the item rather than how quickly they finish the item. In reality, most tests contain both speed and power components, requiring an assessment of speededness (Rindler, 1979). However, the amount of speededness in operational testing has been underestimated prior to the research on speededness using response time in computer–based testing (Oshima, 1994; Schnipke, 1995; Schnipke & Scrams, 1996). Most tests have multiple choice items, no penalty for incorrectly responded items, and restricted time limits. Therefore, rapid guessing behavior, especially at the end of testing, may be easily attempted by the examinees. The effects of speededness and rapid guessing behavior are highly evident in terms of measurement accuracy. Undetected speededness affects erroneously estimation procedures of item characteristic parameters and examinee true ability parameter (Oshima, 1994). Various research studies have been conducted on speededness by investigating aberrant behaviors (e.g., Schinipke, 1995), strategy usage (e.g., Bontempo & Julian, 1997; Gitomer, Curtis, Glaser, & Lensky, 1987), estimating optimal time to solve the items (e.g., Bridgeman & Cline, 2004), and moderated effort (e.g., Wise & DeMars, 2006). Although each study has indicated a different approach in terms of its focus and design, these studies all contribute to the construction of a nomological validity network for the effect of response time in computer–based testing (Cronbach & Meehl, 1955;

Messick, 1981).

The most commonly observed examinee behavior in testing is accuracy on test items. Although it is not always directly reflected, the score, an examinee receives on the test, is based on their accuracy. Likewise, most psychometric research has focused on scores in some form (Schnipke & Scrams, 2002). Given the primary interest in test scores and the possible effect of speededness, researchers have tried to develop models that use response time in the scoring process (e.g., Roskam, 1987, 1997; Thissen, 1983; van der Linden, 2007; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). Several models have been proposed differing in terms of the assumed response time distributions, the assumed relationship between ability and response speed, and the nature of items for which the model was designed. van der Linden (2006) categorized these models under two distinct approaches: modeling response time in the framework of an item response theory (IRT) and separate models for response time and response for the item. He also stated that, for the educational assessment field, it is pertinent to adopt a response time model integrated in the framework of IRT.

Thissen's model (1983) is one of the oldest models using response time, and has a lognormal distribution of the response time on an item with a two–parameter logistic (2PL) IRT structure. This model has person speed and item speed parameters with a time interpretation. The 2PL IRT response component is regressed on the response time and indicates two sources of relationships: (a) response time and examinee ability, (b) response time and item difficulty. Similarly, Wang and Hanson (2005) proposed the 4PL IRT response time model, which has an examinee and an item slowness parameter in the typical 3PL IRT model. Those two models tried to reflect two different components of examinee data from testing settings. In addition, van der

Linden (2007) proposed a hierarchical modeling framework consistent with the previous two traditions. This model has two separate response and response time models as first level models and the integrated model of their parameters as a second level. Therefore it can be possible to estimate response time and response models independently at the first level as well as identify the relationships between two separate models. More specifically, this hierarchical framework can distinguish the following levels: (1) the within–person level, at which the value of the person parameters are allowed to change over time (e.g., due to a change of strategy or external conditions); (2) the fixed–person level, at which the parameters remain constant; and (3) the level of a population of fixed persons, for which there is a distribution of parameter values across persons (van der Linden, 2007). van der Linden's (2007) hierarchical framework enables one to locate the sources of variability between examinee ability and response time as well as item characteristics and response speed.

Purpose

The purpose of this study is to compare two different scoring models using response time data in conjunction with item response models. Various scoring models incorporating response time have been proposed. However, there are not many studies comparing different orientations on the response time and item response model. Most of the studies using response time models have been focused on model fit to the given data. Although it is not easy to compare models which are founded on different theoretical bases, it is worthwhile considering the potential benefits of using response time information in educational assessment. The results from the analysis of response time allow us to devise appropriate scoring models, secure test validity under the threats of various

factors affecting the assumptions of unidimensional IRT, and further examine human behavior in various test settings.

In this study distinctive response time models incorporated in IRT were compared, and the relationship between item characteristics and examinee ability as well as response time were examined using real and simulated data. Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods for Thissen's (Thissen, 1983) lognormal response time model, Wang and Hanson's (Wang & Hanson, 2005) 4PL RT model, and van der Linden's (van der Linden, 2007) hierarchical framework were applied to the investigation of response time on real data. After the application of those response time models on real data, examinee ability and item characteristic parameters from the item response models, as well as speed–related parameters from response time models, were estimated and used for generating simulated data. Those models were, then, applied to simulated data and compared under various conditions of testing situations by utilizing Bayesian posterior estimates.

Research Questions

The research questions addressed in this study are as follows:

1. Among the 4PL response time (RT), hierarchical framework, and Thissen's model, which is the best method for scoring examinees' item responses when response time data are available on real data?

2. What are the relationships between the response time–related parameters (examinee and item slowness, time intensity and time discrimination parameters) from different models that explain the speed–accuracy trade–off among item characteristics and examinee ability in item responses?

3. Between the 4PL RT model and hierarchical framework, which model is better to use for scoring examinees' responses with response time data under different conditions such as various numbers of examinees, different number of items, and different relationship among item characteristics and examinee ability?

Hypotheses

The 4PL RT model and the hierarchical framework showed successful results in applications to real data as well as simulated ones (e.g., Wang & Hanson, 2005; van der Linden, 2007; Fox, Klein Entink, & van der Linden, 2007). However, Wang & Hanson's (2005) 4PL RT model has several limitations when applied in real situations. Because the 4PL RT model has an assumption of independence between response time and the examinee ability parameters, it is unrealistic in most timed testing environment. Later, Wang (2006) modeled the joint distribution of response time using a 1PL Weibull distribution to extend the 4PL RT model. The joint distribution of a response and response time model enables to remove the independence assumption which the 4PL RT model has; however, it did not show much improvement from the typical IRT models that do not consider response time.

It was hypothesized that van der Linden's (2007) hierarchical framework would fit the data better when there is a positive or negative relationship between item characteristics and examinee's ability parameters. As item difficulty increases, it is assumed that it will take longer for examinees to finish such items than easier ones. Likewise, it is also assumed that high level examinees will complete problem solving processes faster than their low level counterparts. The hierarchical framework allows researchers to estimate item and examinee parameters separately by

distinguishing different models of examinee response time and responses. Thus, identifying various sources of response time latency is available by using the hierarchical framework on response time data. However, it is also assumed that the complex models do not always produce better results than simpler ones do. The principle of parsimony is one of the factors that should be considered when making decisions about model fit and model comparison.

Chapter 2. Literature review

More tests are now being administered on computers, providing easy collection of response times in standard, operational testing settings. As response times are becoming more available, it is more prevalent to make use of this information. Many studies have been done in the area of scoring and parameter estimation procedures utilizing response time data in conjunction with response data (e.g., Roskam, 1987, 1997; Thissen, 1983; van der Linden, 2007; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). This chapter presents a summary of the relevant studies on speed, accuracy, and performance in computer–based tests. It begins with some prerequisite definitions of related concepts, including item response theory, preceding discussions on the relationship of speed and accuracy. Various studies investigating the relationships between ability and speed will be summarized and scoring models with response time data will follow. Finally, for the model parameter estimation procedures, Markov Chain Monte Carlo (MCMC) methods using Gibbs sampling will be introduced.

Power and speed tests

Gulliksen (1950) pointed out two essential factors for the tests: speed and power. A pure power test has items with a range of difficulties and an infinite time limit. The goal of a pure power test is to measure how accurately examinees respond to the items. Because power tests have a time limit long enough to permit everyone to attempt all items, item difficulty is steeply graded and includes items too difficult for anyone to solve, so it is hard to get a perfect score. On the other hand, the goal of a pure speed test is to measure how quickly examinees respond to the items. A test is constructed with easier items and a time limit is so short that no one can finish all the items. On pure speed testing, each person's score directly reflects the speed with which each examinee worked. Anastasi (1976) also defines that a speed test is when the speed of performance determines individual differences. However, both power and speed tests are designed to prevent the achievement of perfect scores.

Historical perspectives on response time analysis

As indicated by Gulliksen's (1950) definitions of power and speed tests, it is generally accepted that there exists interchangeability between speed and ability. Because measuring the time it takes an examinee to process information is deemed indicative of how examinee processed it, researchers had believed speed and accuracy measured the same construct. Spearman (1927) became one of the earliest proponents of the theory that the speed at which an examinee completed a test and the accuracy from the results gave equivalent information. Thus he argued that an examinee's mental ability could be measured on a scale of accuracy, a scale of speed, or some combination of the two constructs (Spearman, 1927). However, the study of these two

constructs on complex tasks did not show that they were same constructs by subsequent researchers (Baxter, 1941; Bridges, 1985; Foos, 1989). Myers (1952) demonstrated that speed and accuracy comprised orthogonal factors in test scores, indicating that an examinee's speed in testing is not related to the examinee's ability. Various other studies also confirmed that Spearman's (1927) interchangeability concept on speed and ability is unrealistic in educational assessment settings (Schnipke & Scrams, 2002).

The speed–accuracy trade–off is one of best known findings in response time research (Luce, 1986). The speed–accuracy trade–off implies that if a person chooses to perform a task at a higher speed rather than a relatively lower speed, their level of accuracy will become lower. It is obvious that the trade–off can be applied either to pure speed tests or pure power tests. However, studies on response time for correct and incorrect responses showed different directions. Bergstrom, Gershon, and Lunz (1994) found that examinees spent more time on items they answered incorrectly than on items they answered correctly. Hornke (2000) also found that relatively longer response times are required to respond to questions that are answered incorrectly. A variety of systematic studies on item response times in computerized adaptive testing found that incorrect answers require much longer processing time than correct answers (Rammsayer, 2004).

It is argued that most wrong responses are from the lower ability group, examinee's lower ability used to relate to relatively longer response time in the marginal analysis (Bergstrom et al., 1994; Hornke, 2000). As explained by Simpson's paradox (Agresti, 2002; Simpson, 1951), taking the ability of examinees into account would result in explaining a somewhat different relationship between response time and response accuracy. Therefore, the relationship among response speed and related examinee characteristics needs to be verified by further examining the relationship

among examinees' ability, item difficulties, and response time simultaneously.

It is reasonable to assume that the relationships between accuracy and speed are not to be correlated without considering other effects derived from item and examinee characteristics. Schnipke and Scrams (2002) pointed out that a great deal of previous research has used confounding measures to investigate the relationship of speed and accuracy. Specifically, examinee speed is easily confounded with item difficulty when it is administered on computer adaptive tests (CAT). More discussion of the relationships between accuracy and speed, ability and response time will be presented in the following sections.

Item response theory

Item response theory (IRT) is a statistical theory about the probability of an examinee responding to an item correctly at a given level of latent proficiency. IRT models specify how test items and examinee responses relate to the abilities of the examinees that are measured by the items in the test (Hambleton & Jones, 1993). Two basic assumptions are required to use these IRT models (Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991). First, a unidimensionality assumption is required, meaning that there is one construct of a given test. The items in a test are considered to be unidimensional when a single factor or trait accounts for a substantial portion of the total test score variance. It is a broad concept which also encompasses local independence and parameter invariance assumptions; item responses are deemed locally independent when examinees' ability is the sole source that affects responses on the items. Second, the item characteristic function or curve (ICC) is needed to form a mathematical representation. It delineates the relationship between examinees' unobserved latent ability and observed test scores

from responses to the items (Hambleton & Jones, 1993; Swygert, 1998).

Many models have been formulated within the general IRT framework; however, usually one, two, or three parameter logistic functions will be considered when the model is applied to dichotomously scored items. In terms of dichotomously scored test items, on which responses are designated either correct or incorrect, all IRT models express the probability of a correct response to a test item as a function of θ, given one or more parameters of the item. The 3PL model is expressed as follows:

$$P(u_{ij}=1|\theta_i) = c_j + \frac{1-c_j}{1+e^{-1.7a_j(\theta_i-b_j)}}, \qquad (1)$$

where $P_{ij}(\theta)$ is the probability that an examinee *i* with ability θ answers test item *j* correctly, which has generally scaled with a mean of 0 and standard deviation of 1. b_j is the item difficulty or location parameter, a_j is the discrimination or slope parameter, which is bounded by 0, and generally ≤ 2.0. c_j is the pseudo guessing or lower asymptote parameter. This is bounded by 0 and 1, and generally ≤ 0.25 depending on the number of alternative answers in the items. Under the typical IRT framework, both the test items and the examinees responding to the items are arrayed on θ from lowest to highest abilities. The position of examinee *i* on θ(denoted θ_i), is usually referred to as the person's ability or proficiency. The position of item *j* on θ, (usually denoted b_j), is termed the item's difficulty. It is expected that the probability of a correct response to item *j* will increase monotonically as (θ_i – b_j) increases. The 1PL and 2PL IRT models are regarded as the constrained forms of the typical 3PL model; the item discrimination parameter is set to 1.0 in the 1PL IRT model; the pseudo guessing parameter is set to zero in the 1PL and 2PL models.

Response time analysis and computerized tests

The availability of item response times, made possible by computerized testing, provides an entirely new type of information about items. Previously, only total testing time and item responses were available. However, in addition to knowing the accuracy with which test takers answer an item, it is now possible to investigate the amount of time examinees spend on each item. This allows one to examine the relationships among examinee ability, item characteristics, and response speed (Schnipke & Scrams, 1999). Various other kinds of information in test settings can be obtained from response time data, such as speededness, pacing, strategies used, and time limit.

Speededness

Speededness is the effect of time limits on the candidate's scores. It is the extent to which a test is affected by time limits, which is measured when the examinee's total incorrect score is equal to the number of items that were not attempted by the examinee (Evans & Reilly, 1972). Bejar (1985) stated that "a test is speeded when some portion of the test–taking population does not have sufficient time to attempt every item in the test within the allocated time." Bontempo and Julian (1997) also defined speededness as "the degree to which the amount of time allowed for test administration affects the rate at which examinees answer items."

Speededness is a closely related concept with other response time related constructs such as pacing, strategy use, and predicting finishing times or setting up time limits. Test speededness is gauged from the perspective of testing environment, while pacing and guessing behaviors are construed from the examinee perspectives. Likewise, strategy use in testing is also related to pacing and test speededness. Schnipke (1995) defined two distinct types of behavior when test

speededness exists: problem solving and rapid guessing behaviors. Just as rapid guessing behavior at the end of a test substantially affects the examinee's ability estimate, test taking strategies, test wiseness, and pacing also need to be examined in the perspective of these two types of test taking behavior.

In reality, most tests contain both speed and power components, requiring assessments of certain amount of speededness (Rindler, 1979). It is argued, therefore, that a test is investigated the degree of speededness instead of existence of speededness or lack thereof (Lu & Sireci, 2007). It is obvious that that the amount of speededness in testing has been underestimated until recent research on speededness conducted using response time in computerized tests (Schnipke & Scrams, 2002). Schaeffer, Reese, Steffen, McKinley, and Mills (1993) examined the average item response time from the Graduate Record Examination (GRE) and concluded that the time limits were sufficient. On the other hand, Bridgeman (2004) found that an examinee who worked at the mean rate for the first 20 items would require 11 more minutes than what was allowed on the GRE.

Speededness in a computerized testing environment has raised significant validity issues in some studies. Oshima (1994) demonstrated that undetected speededness can cause a significant problem on many large–scale standardized tests such as TOEFL and SAT (e.g., Angoff, 1989; Bejar, 1985; Schmidt & Dorans, 1990). Bridgeman (2000) states that time limits may raise equity issues if the limit is imposed for administrative convenience rather than an essential part of what the test is measuring. Bridgeman, Cline, and Hessinger (2003) also concluded that the variation among examinees in the rate of response to test items constitutes an irrelevant source of difficulty in test performance. Irrespective of the definition and directions of the research studies, all research in speededness ended up with one agreement of detrimental results of the validity of

interpretations of test scores (Lu & Sireci, 2007). As indicated by Messick (1981), it is obvious that construct irrelevant variance resulting from test speededness contributes to unreliability and invalidity of test.

Computer adaptive testing (CAT)

Computer adaptive testing (CAT) introduces new dimensions to the speededness issue. Usually, computer–based tests (CBT) implement the same administration and scoring algorithms as typical paper and pencil versions. On the other hand, a CAT modifies the difficulty of a test based on an examinee's responses as a function of the current estimate of ability. However, these procedures may add to the cognitive load of the higher ability examinees, because more difficult items usually demand more time to solve.

Speededness in CAT is connected to the fairness issue because omitting items is no longer an option in CAT. Many studies have demonstrated that item difficulty and response time are positively correlated in CAT (Bergstrom et al., 1994; Bridgeman & Cline, 2004; Chang, 2006; Plake, 1999; Smith, 2000). This is because response time and item difficulty are closely related to critical reasoning and problem solving procedures that increase the number of steps required to answer a problem correctly. The assumption is that successive items become more difficult, it also adds more cognitive load and finally results in spending extra time to solve the item.

Various studies have consistently found that pacing and test taking strategies are also affected by speededness in CAT. Bergstrom et al. (1994) and Bridgeman and Cline (2004) concluded that it took longer for higher ability students to finish the test than lower ability students, because higher ability students are administered more difficult items. Chang (2006) also suggested

the same result, indicating the test becomes more speeded for higher ability students regardless of item types. Specifically, it is noted that higher ability examinees spend much more time on pretest items. This introduces an important piece of information to explain the relationship between ability and speededness in CAT. Because pretest items are not tailored to the examinees based on their relative ability levels, it may be generalized that more able examinees spend more time on all items regardless of whether their responses are right or wrong. Test taking strategies are also confounded by the fact that most CAT implementations prevent the test taker from reviewing previous answers, as well as from omitting answers. Bridgeman and Cline (2004) noted that more rapid guessing behavior is required for the higher ability examinees because they have more time consuming items. Bergstrom et al. (1994) also concluded that the ability and item positions are significant factors in predicting the finishing time of examinees in a within subject model. They suggested that controllable factors such as using figures, item length, and position of keyed correct answers contribute to explaining the variance of response time (Bergstrom et al., 1994).

Relationships between response time and ability

As discussed in the previous section, understanding the relationship between response time and accuracy is important in building appropriate and reasonable models. Results from previous studies indicated that there are distinct patterns among item characteristics, examinee ability, and response time. When items become more difficult, it takes more time for examinees to process (e.g., Bergstrom et al., 1994; Bridgeman & Cline, 2004; Chang, 2006; Plake, 1999; Smith, 2000). Incorrect responses take more time than correct responses (e.g., Bergstrom et al., 1994; Hornke, 2000; Rammsayer, 2004). More able examinees generally take more time to finish items than less

able examinees (e.g., Bergstrom et al., 1994; Bridgeman & Cline, 2004; Chang, 2006; Swygert, 1998). However, there are not many studies regarding systematic explanations of why such relationships exist among these factors.

The relationship between response time and examinee ability is manifested by how those components are modeled in the scoring framework. Various studies have implemented models of response time and item responses based on a range of different scoring methods and response time distributions (Schnipke & Scrams, 2002). Researchers have tried to find models that can be fit with statistical distribution functions with known properties. Normal and lognormal distribution were tested by Thissen (1983), gamma and Weibull distribution have been tested by Tatsuoka and Tatsuoka (1980) and Roskam (1997). These distributions were fit to empirical distribution functions from a computer–based test. Schnipke and Scrams (1997, 2002) found that response time data were best fit by the lognormal distribution for both exploratory and confirmatory samples and provided meaningful interpretations of the data.

van der Linden (2006, 2009) categorized existing response time models into two distinct groups based on the approaches those models have. The first one models response times in the framework of an item response theory (IRT) model. Because response times are modeled in the framework of an IRT model, it is assumed that an interaction exists between the parameters that govern the distributions of the person's response times and response variables for the items. As discussed previously, it is often suggested that more difficult items require more time to be solved. It is also noted that this modeling is based on the speed–accuracy trade–off that has been the focus of much of the psychological literature on response times (Luce, 1983; van der Linden, 2006).

The other group of models discussed by van der Linden (2006, 2009) consists of scoring

models without parametric relationships between response time and the examinee's responses. In this approach, response time distributions are modeled without any parametric consideration of the response variables on the items, in other words, they are assumed to be independent. It is also assumed that speed is not related to the accuracy of an examinee's responses based on an examinee's ability. Results from some of the studies introduced in the previous section suggest this approach is feasible (e.g., Bergstrom et al., 1994; Bridgeman & Cline, 2004; Chang, 2006; Swygert, 1998). Positive as well as no relationships between response time and accuracy have been found in many studies (e.g., Bergstrom et al., 1994; Scrams & Schnipke, 1997; Swygert, 1998; Thissen, 1983).

Schnipke and Scrams (2002) pointed out that the relationship between speed and accuracy depends on the test context and content, and much of the research addressing this issue uses measures of accuracy that are affected by response speed. Thus, response speed is examined with an examinee ability estimate that is already confounded with item difficulty in a given testing situation. Therefore it is important to have response time scoring models in model checking procedures which resolve such problems.

Scoring models using response time

Most psychometric research has focused more on accuracy than speed, although there are many experimental studies that have investigated reaction time in psychology (Schnipke & Scrams, 2002). Research and studies on response times in the educational testing field are limited by practical reasons (e.g., record keeping in operational settings, randomization of ability group). Therefore it was not used much until computerized testing was introduced. However, more tests

are now administered on computer, so it is much easier to collect response time data than before. Accuracy on test items and the score examinees receive on the test based on their ability is the most commonly observed examinee behavior in testing. Early research on scoring models using response time data is closely related to the concept of response time in traditional cognitive psychology. Various models have response speed as a dependent variable and measure the ability of processing skill. These are regarded as distinct models for response time (van der Linden, 2006, 2009). However, these models are appropriate only when items are relatively simple to process and momentary ability is measured by speed of processing, such as a typical speed test in intelligence testing (e.g., processing speed tests in WAIS–IV).

Later models have focused more on empirical response time distribution functions in the response model. Scrams and Schnipke (1997) proposed using response times in standardized tests to compare speed and accuracy as different components of proficiency. These models suggested the way to use both response accuracy and response speed to provide separate measures of performance. More specifically, IRT modeling has been proposed to deal with response time. van der Linden clearly categorized these models as response time models incorporating IRT and IRT models incorporating response time (van der Linden, 2009).

Thissen's (1983) model

Thissen (1983) proposed the response time model which incorporates IRT in it for the first time as follows:

$$\begin{aligned} &\ln T_{ij} = \mu + \beta_j + \tau_i - \rho a_j(\theta_i - b_j) + \varepsilon_{ij}, \\ &\varepsilon_{ij} \sim LN(0, \sigma_j^2), \end{aligned} \tag{2}$$

where $\ln T_{ij}$ is the log response time of examinee *i* to item *j*, μ is the grand mean, β_j is a slowness parameter for item *j*, τ_i is a slowness parameter for examinee *i*, ρ is the regression coefficient for the 2 PL IRT structure on log response time, and ε_{ij} is error term. Specifically, it has person slowness and item slowness parameters as well as the probability of correct response of the examinee to the given item. Therefore this model reflects two different trade–offs; one between the item parameters (item difficulty and slowness) and the other between the person parameters (examinee ability and slowness). The regression term can be interpreted as an index of the direction of the relationships between these two trade–offs (Schnipke & Scrams, 2002). The results from Thissen's study showed that different kind of relationships exist based on the test; explained relationships between examinees' response speed and accuracy were different depending on the characteristics of the test.

Several applications of this model can be found in previous studies. Scrams and Schnipke (1997) applied a 3PL IRT model instead of the 2PL structure as follows:

$$\begin{aligned} &\ln T_{ij} = \mu + \beta_j + \tau_i - \rho Z_{ij} + \varepsilon_{ij}, \\ &Z_{ij} = \ln(c_j + \exp^{a_j(\theta_i - b_j)}) - \ln(1 - c_j), \\ &\varepsilon_{ij} \sim LN(0, \sigma_j^2). \end{aligned} \qquad (3)$$

They applied this model to computer–administered tests of verbal, quantitative, and reasoning skills and found that moderate relationships exist between examinees' response speed and ability as well as item difficulty throughout the different sections of the test. Swygert (1998) used a modified version of Thissen's (1983) model in examining item response time on the GRE CAT. She also found a moderate positive relationship between response speed and examinee proficiency estimates in the two sections of the test. Ingrisone (2008) also used Thissen's (1983) model and

compared a marginal maximum likelihood estimation (MMLE) with a maximum a posteriori (MAP) procedure. Three different simulation studies were conducted and the results of item and person parameter estimates based on MMLE and MAP procedures were found to be consistent and accurate.

Wang and Hanson's (2005) 4PL Response Time model

Wang and Hanson (2005) proposed the 4 PL RT model for item parameter estimation. In this model, response time is incorporated in the parameter estimation procedure as follows:

$$P(x_{ij}=1\left|\theta_i,\tau_i,a_j,b_j,c_j,\beta_j,rt_{ij}\right.)=c_j+\frac{1-c_j}{1+e^{-1.7a_j[\theta_i-(\beta_j\tau_i/rt_{ij})-b_j]}}, \quad (4)$$

where rt_{ij} is the response time by examinee *i* on the item *j*, β_j is the item slowness parameter, and τ_i is the examinee slowness parameter. The item and person slowness parameters determine the rate of increase in the probability of a correct answer as a function of response time. The product of these two slowness parameters determines the rate of probability change with increasing response time for a particular examinee to a particular item.

Later, Wang (2006) modeled the joint distribution of response accuracy and response time using a 1PL Weibull distribution to extend the model. Because Wang and Hanson's (2005) model has an assumption of independence between response time and the examinee ability parameters, it is unrealistic in most timed testing situations (Ingrisone, 2008). The joint distribution of response and response time enables removing this independence assumption; however, it did not show much improvement from the typical IRT models without considering response time. Ingrisone

(2008) extended Wang's (2006) model by applying a 2PL Weibull distribution to the marginal distribution of response time model. Among several estimation methods applied to the item characteristic and examinee true ability parameter, marginal maximum likelihood estimation (MMLE) and maximum a posteriori (MAP) procedures showed that item and examinee parameters were recovered quite well in this model (Ingrisone, 2008).

Hierarchical Framework

van der Linden (2007) introduced the third approach in modeling the response and response time distributions. The hierarchical framework has both response time and typical IRT model as two level–one models and a second level model as a realization of the population model of the two level–one models. Figure 1 shows a graphical representation of the model.

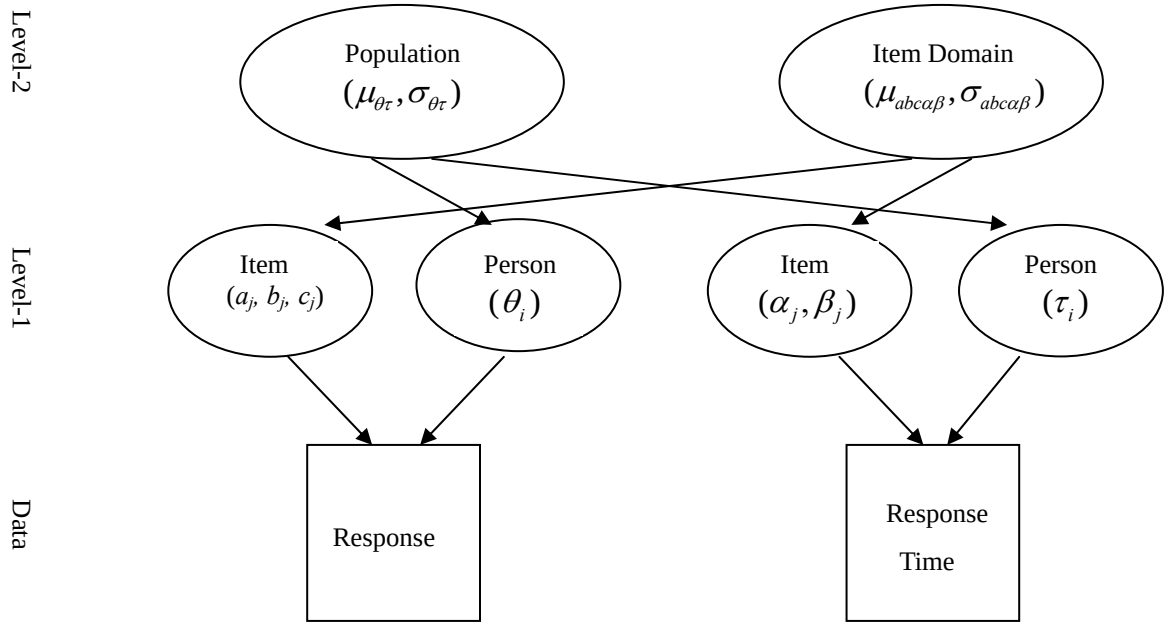


Figure 1. *The hierarchical Framework for modeling speed and accuracy on items (van der Linden, 2007)*

Level-1 response model is typical 3PL IRT model as follows:

$$P(x_{ij}=1\left|\theta_i, a_j, b_j, c_j\right.) = c_j + \frac{1-c_j}{1+e^{-1.7a_j(\theta_i - b_j)}}. \qquad (5)$$

A response time model is a lognormal model as follows:

$$f(t_{ij};\tau_i,\alpha_j,\beta_j) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\{-\frac{1}{2}[\alpha_j(\ln t_{ij} - (\beta_j - \tau_i))]^2\}, \qquad (6)$$

where and t_{ij} is the response time by examinee i on the item j, τ_j is the speed parameter of examinee j, α_i is the time discrimination parameter of item i, and β_j is the time intensity parameter of item j.

The level-2 model has a bivariate normal distribution for examinee's ability and speed parameters and a multivariate normal distribution for the item parameters of response and response time models as follows:

$$\left.\begin{array}{l} (\theta,\tau) \sim N(\mu_P, \Sigma_P), \\ \text{where} \\ \quad \mu_p = (\mu_\theta, \mu_1) \\ \quad \Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\theta^2 \end{pmatrix} \end{array}\right\} \text{Distribution of person parameters}$$

and for item parameters,

$$
\left.
\begin{aligned}
&(a_j, b_j, c_j, \alpha_j, \beta_j) \sim N(\mu_I, \Sigma_I), \\
&\text{where} \\
&\mu_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta) \\
&\Sigma_I = \begin{pmatrix}
\sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\
\sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\
\sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\
\sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\
\sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2
\end{pmatrix}
\end{aligned}
\right\} \text{Distribution of item parameters}
$$

Therefore, the level-1 has each independent response time and response models, but the level-2 has the covariance structure of the parameters of the lower level models. The has a basic assumption that the person operates at constant ability and speed, which indicate that the examinee's true ability and speed levels are constrained by a speed–accuracy trade-off. If the constant level of the examinee's speed is taken, the response-time distribution depends on the speed, and the response times become conditionally independent given speed. However, for a population of examinees, ability and response speed are expected to be dependent; a second-level population model needs to represent the dependency in it (van der Linden, 2006).

Bayesian estimation in IRT

Bayesian inference enables us to fit a probability model to data and to summarize the result by a probability distribution on the parameters of the model, as well as on unobserved quantities such as predictions for new observations (Gelman, Carlin, Stern, & Rubin, 2003). For further application of Bayesian procedures, the core principles of Bayesian inference need to be discussed. The centerpiece of this framework is Bayes' theorem, as follows:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}, \quad (7)$$

where $p(B|A)$ is the posterior probability of *B* given *A*, $p(A|B)$ is the conditional probability of *A* given *B*, and $p(B)$ is the prior probability of *B*. Equation (7) can be extended when we accept $p(A)$ as the marginal probability of event *A* as follows:

$$p(A) = \sum_{B_i \in S_B} p(A|B_i)p(B_i). \quad (8)$$

Therefore the marginal probability of event *A* is computed as the sum of conditional probability of *A* under all event of B_i in the sample space. The summation represents an accumulation across all possible outcomes of event *B* and thus can also be taken as the probability of *A*, *P(A)*. This is the process of using the known value of the data and the basic property of conditional probability, resulting in the posterior distribution of the given data. From Bayes' theorem it is known that a representation of the conditional probability of one event given another provides an explanation in terms of the opposite conditional probability (Kim & Bolt, 2007). Lynch (2007) also stated that "the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty."

Bayes' theorem expressed in terms of a probability density function appears as:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} = \frac{f(X|\theta)f(\theta)}{\int f(X|\theta)f(\theta)d\theta}, \quad (9)$$

where $f(\theta|X)$ is the posterior distribution for the parameter θ, $f(X|\theta)$ is the sampling density for the data *X*, and $f(X)$ is the marginal probability of the data *X*. The sampling density is

proportional to the likelihood function, and the denominator of (9) has a role of scaling the posterior density to make it a proper density, otherwise Bayes' theorem for probability distributions is simply stated as:

$$Posterior \propto Likelihood \times Prior . \quad (10)$$

When fitting an item response model to data, it is necessary to obtain information about parameters of the item response model from the response data of the examinees. From the perspective of Bayes' theorem, this information is expressed as the relative likelihood of particular parameter values for the model given the observed item response data. The three–parameter logistic model (3PL) introduced in equation (1) presents the probability of an examinee responding to an item correctly as a function of the examinee ability, item difficulty, item discrimination, and guessing parameters. The joint distribution of all variables when there are N examinees and J items in the test is presented as:

$$\begin{aligned} &P(\theta_1,\cdots,\theta_N,a_1,\cdots,a_J,b_1,\cdots,b_J,c_1,\cdots,c_J \,|\, X_{11},\cdots,X_{NJ}) \\ &\propto \prod_{i=1}^{N}\prod_{j=1}^{J} P(X_{ij} \,|\, \theta_i,a_j,b_j,c_j)P(\theta_i)P(a_j)P(b_j)P(c_j) . \end{aligned} \quad (11)$$

The joint posterior density in the left hand side of (11) is used to determine estimates of the model parameters. To evaluate it requires knowledge about the quantities on the right hand side. The quantities of $P(a_j)$, $P(b_j)$, $P(c_j)$ are the prior densities of the model parameters and can be thought of as indicating the relative likelihoods of particular parameter values prior to data collection. The likelihood of the item response data given all of the model parameters is expressed as $P(X_{ij} \,|\, \theta_i,a_j,b_j,c_j)$ and it is defined by the item response model along with its associated assumptions of local independence and exchangeability. The quantity in the denominator is not

written in (11) and is regarded as a constant for a fixed data set. It is often referred to as a normalizing constant since its value generally makes a proper density. This proportionality relationship is often the basis for sampling procedures that underlie MCMC, when it is possible to evaluate the relative likelihoods of different sets of parameter values even if the exact form of the posterior density cannot be determined (Kim & Bolt, 2007; Lynch, 2007).

Markov Chain Monte Carlo (MCMC) method

Markov Chain Monte Carlo (MCMC) methods have offered many advantages such as convenience of implementation and software availability. MCMC methods provide an opportunity to sample from multivariate densities that are not easily sampled from by implementing maximum likelihood methods using a laborious EM (Expectation Maximization) algorithm. A fundamental difference between MCMC and other popular estimation techniques, such as maximum likelihood (ML) estimation, lies in the emphasis on Bayesian inference on estimating distributions. Kim and Bolt (2007) contrasted that Bayesian estimation has "a potentially richer description of the parameter estimate distribution than is usually provided in ML estimation." MCMC methods have expanded the opportunity to experiment with new models needed for specialized measurement applications (Kim & Bolt, 2007; Lynch, 2007).

Kim and Bolt (2007) described the basic MCMC approach applied to IRT estimation. It provides a way for sampling from one or more dimensions of a posterior distribution and moving throughout the entire support of a posterior distribution. According to Lynch (2007) MCMC methods "utilized the process of sampling by breaking these densities down into more manageable univariate or multivariate densities." Because the MCMC estimation results in the reproduction of

the posterior distribution of interested parameters, iterative procedures of samplings from observations based on this distribution are important. These procedures imply that by sampling enough observations, it becomes possible to determine characteristics of the distribution. Those characteristics, captured in the form of mean and variance, can be the basis for model parameter estimates for given data. The precise mechanism by which sampling is conducted may vary based on the known features of the posterior distribution. However, once an appropriate sampling procedure is determined, computing corresponding characteristics of the generated sample make it possible to have relevant posterior distributions.

The use of MCMC estimation for IRT models was introduced by Patz and Junker (1999a) and has since been used to estimate a variety of models. When item parameter estimates are treated as known, interest centers on estimating examinee ability parameters. Likewise, when examinee parameters are treated as known, interest centers on estimating item parameters. More generally, both examinee and item parameters can be estimated concurrently. After an IRT model is chosen and priors have been specified for all model parameters, sampling procedure for updating posterior distribution begins. The objective of MCMC is to define a mechanism by which observations can be sampled from the joint posterior density of model parameters shown in (8), making the iterative process conducted under MCMC methods considerably different from that conducted in the ML procedure. MCMC procedures enable us to have representative posterior distribution of the model parameters rather than a converged point estimate of the model parameters. Gilks, Richardson, and Spiegelhalter (1996) provide a more general explanation about the method on various models and Patz and Junker (1999b) describe an application on IRT in detail.

Gibbs sampler

Kim and Bolt (2007) described the Gibbs sampler as follows;

> "a mechanism by which sampling can be performed with respect to smaller numbers of parameters, often one at a time. The Gibbs sampler samples with respect to univariate conditional distributions of the model parameters. Unlike the full joint posterior distribution, the conditional distributions, denoted as $f(\xi_k \mid X, \xi_{-k})$, represent the posterior distribution of a single model parameter (ξ_k) conditional upon the data (X) and all other model parameters (ξ_{-k})".

Therefore, after all the other parameters are known, Gibbs sampling enables each parameter to be sampled individually based on its conditional distribution. In other words, the full conditional density for a parameter needs to be known only up to a normalizing constant, and it allows one to use the joint density with the other parameters set at their current values. Gibbs sampling involves ordering the parameters and sampling from the conditional distribution for each parameter given the current updating process. This makes Gibbs sampling relatively simple for most problems in which the joint density is reduced to known forms for each parameter once all other parameters are treated as fixed (Lynch, 2007).

A generic Gibbs sampler follows the following iterative process (e.g., Kim, 2001; Lynch , 2007; Rowe, 2003):

0. Assign a vector of starting values as an initial value for the parameter vector:

$\xi^{j=0} = S$.

1. Set $j = j + 1$, where j indicates the iteration count.

At the j^{th} iteration define $\xi^{(j+1)} = (\xi_1^{(j+1)}, \xi_2^{(j+1)}, \xi_3^{(j+1)}, \cdots, \xi_{k-1}^{(j+1)}, \xi_k^{(j+1)})$ by the values from following procedures:

2. Sample ξ_1^{j+1} from $p(\xi_1 \left| \xi_2^j, \xi_3^j, \cdots, \xi_{k-1}^j, \xi_k^j \right.)$.

3. Sample ξ_2^{j+1} from $p(\xi_2 \left| \xi_1^{j+1}, \xi_3^j, \cdots, \xi_{k-1}^j, \xi_k^j \right.)$.

4. Sample ξ_3^{j+1} from $p(\xi_3 \left| \xi_1^{j+1}, \xi_2^{j+1}, \cdots, \xi_{k-1}^j, \xi_k^j \right.)$.

$\vdots$

k. Sample ξ_k^{j+1} from $p(\xi_k \left| \xi_1^{j+1}, \xi_2^{j+1}, \cdots, \xi_{k-1}^{j+1}, \xi_k^j \right.)$.

k+1. Return to step 1.

In Gibbs sampling procedure each step draws random sample from the associated conditional posterior distribution. After drawing j^{th} iteration of the sample, there will be $\xi^1, \xi^2, \xi^3, \cdots, \xi^{j-1}, \xi^j$ samples of the parameter estimates. A pre–specified number of first samples is called “burn–in”, it will be discarded and remaining samples will be kept and used for calculating the mean and the standard deviation values for posterior distribution of the samples (Kim, 2001; Lynch, 2007).

Checking model convergence

Monitoring the simulated states of the Markov chain is an important procedure for checking model convergence. Theoretically, the Markov chain should converge to a stationary distribution so that the sampled observations can be regarded as a sample from the posterior distribution of the

model parameters. The rate at which this convergence occurs can vary depending on several factors as follows: (a) high correlations between adjacent states, (b) sampling algorithms, and (c) identification problems with the model. When there are relative high correlations between states, a slow rate of convergence occurs; therefore, a very large number of iterations is necessary. The selections of the sampling algorithm and problems in identification with the models also will affect model convergence in the MCMC procedures (Kim & Bolt, 2007; Lynch, 2007).

It is possible to determine whether an MCMC run has been successful by detecting convergence. Observations of the history plots of the chain, autocorrelation between the states, and the posterior density plots of the estimated parameters are usually made. Various diagnostic indices can also be applied to observations from the chain to evaluate the likelihood of convergence. Kim and Bolt (2007) described how these indices are calculated in detail. One of the diagnostics is Geweke's (1992) criterion; a z–score is computed from the sampled states for each parameter in this approach. The z–score for a given parameter is defined by taking the difference between the mean of the first 10% of states, and the mean of the last 50% of states, and dividing by their pooled standard deviation. Z–values within a range of non–significance can be taken as evidence of convergence. Another criterion explained in Kim and Bolt (2007) is the Raftery and Lewis criterion which considered the number of samples needed to estimate quantiles of the posterior with sufficient precision. When the index, *I*, indicates greater than 5.0, the increase in the number of sampled states needed to reach convergence due to autocorrelations in the chain (Raftery & Lewis, 1992). When multiple chains are applied, the Gelman and Rubin criterion can be used. There is a strong likelihood of convergence if the chains demonstrated the same stationary distribution, which is reflected by a large overlap in their sampling histories. The

Gelman and Rubin test is based on a comparison of (a) the pooled between chain variances and (b) within chain variances for each parameter. If the *R* value is approaching to 1.0, it is indicated that stability for the chains are assumed (Gelman & Rubin, 1992).

Checking model goodness of fit and comparison

IRT models require that several assumptions be met by the data including local independence and specific forms of the item response function. When these assumptions are not appropriately satisfied, inferences regarding the nature of the items and examinees can be erroneous, and the potential advantages of IRT are not attained. It is therefore crucial to check the adequacy of the fit of the chosen IRT model to item responses. Several fit statistics have been proposed within the frequentist framework (e.g., Orlando & Thissen, 2003; Yen, 1981), but it is difficult to find a universally accepted model fit checking method, and this still remains an underdeveloped area in IRT (Sinharay, Johnson, & Stern, 2006).

Several Bayesian model goodness of fit indices are available. Among them posterior predictive model checking (PPMC) is one of the general strategies in the IRT context and it is also popular Bayesian model diagnostic tool (Gelman et al., 2003; Kim & Bolt, 2007). Various studies showed the applications of this index in several conditions by checking the plausibility of posterior predictive replicated data against observed data (Albert & Ghosh, 2000; Glas & Meijer, 2003; Hoijtink, 2001; Hoijtink & Molenaar, 1997; Janssen, Tuerlinckx, Meulders, & DeBoeck, 2000; Rubin & Stern, 1994; Scheines, Boomsma, & Hoijtink, 1999; Sinharay, 2005; Sinharay & Johnson, 2003; Sinharay et al., 2006; van Onna, 2003).

Beyond studies of absolute model fit, other approaches can be used for model comparison.

Model comparison and selection procedures are implemented without evaluating the degree of fit in an absolute sense. There are several criteria that identify which of the models provide a better fit to the data. Among several indices, the Deviance Information Criterion (DIC) is easily used and calculated as follows:

$$DIC_{(Model)} = \overline{D(\theta)} + p_D = D(\bar{\theta}) + 2 \times p_D , \qquad (12)$$

where $\overline{D(\theta)}$ is a Bayesian measure of fit (posterior mean deviance), $D(\bar{\theta})$ is the deviance of the posterior model, and p_D is the number of free parameter which accounts for the expected decrease in deviance attributable to the added parameters of the more complex model. DIC is an index for model comparison similar to the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwartz, 1978) (Spiegelhalter, Thomas, Best, & Lunn, 2003). As with AIC and BIC, the model with the smallest value of DIC would indicate the better model to the observed data set. Estimation of the DIC index can be requested within the WinBUGS program (Spiegelhalter et al., 2003).

Chapter 3. Methods

In this chapter response time models incorporated in IRT were compared and the relationship between item characteristics and examinee ability as well as response time were examined using real and simulated data. In study 1, Thissen's (Thissen, 1983) response time model, Wang and Hanson's (Wang & Hanson, 2005) 4PL RT model, and van der Linden's (van der Linden, 2007) hierarchical framework were applied to the investigation of response time on real data through Bayesian estimation using MCMC. In study 2, Wang and Hanson's (Wang & Hanson, 2005) 4PL RT model, and van der Linden's (van der Linden, 2007) hierarchical framework will be applied on simulated data. These models are explored further, using simulated data with known generating parameters to help understand how they might behave under some different conditions typically encountered in applied testing situations: varied test lengths, sample sizes, and the extent of relationships between item and examinee speed related parameters.

Study 1

Data

Real data from a pool of nationally standardized English verbal tests administered in 2007 was used. Response data as well as response time for 33 discrete items were collected during computer based test administration. Examinees' responses were coded dichotomously and response time data were recorded to 1-second precision on a computer based test. The recorded response time for an item was the total time spent on the item during all attempts to that item. The test consists of operational and field test items in multiple choice formats.

There were relatively few examinees showed peculiar responses. In order to identify potential outliers from the examinees, responding in an extremely short time span (e.g., finishing the test within 10 minutes out of 60 minutes), no responses after completing a few items (e.g., 5 items out of 33) were eliminated from the analysis. Out of the 978, only 3 examinees (0.3%) showed peculiar responses, further analyses were conducted on the response and response time data collected from 975 examinees.

Estimation methods

Three response time models were used for estimating item and examinee parameters; (1) Thissen's (1983) log normal response time model, (2) Wang and Hanson's (2005) 4PL response time model, and (3) van der Linden's (2007) hierarchical framework. All three models based on IRT response model for response data, it can be applied either 2PL or 3PL IRT model. In study 1, total 6 models (3 response time models with 2PL as well as 3PL IRT sub-model) were investigated by using Bayesian posterior parameter estimates.

In order to implement Bayesian posterior parameter estimation, it needs to specify their prior distributions. It is important to choose the strength of the priors in MCMC; if the prior has narrower variance, it is more informative in guiding the algorithm. Therefore the priors on item and examinee parameters were set to be relatively large and intended as less informative so that the given data can drive the posterior distributions. Starting values are also needed for each parameter to define the first state of the Markov chain and those values for each model parameters were randomly generated using the WinBUGS computer program (Spiegelhalter et al., 2003). The following priors were described based on 3PL IRT model for response data in Thissen's (1983)

response time model:

$\theta_i \sim N(0,1), \quad i = 1, \cdots, N$

$a_j \sim LN(0,1), \quad j = 1, \cdots, J$

$b_j \sim N(0,1), \quad j = 1, \cdots, J$

$c_j \sim Beta(5,17), \quad j = 1, \cdots, J$

$\tau_i \sim N(0,1), \quad i = 1, \cdots, N$

$\beta_j \sim N(0,1), \quad j = 1, \cdots, J$

$\mu \sim N(0,2)$

$\rho \sim N(0,1)$

where N is the total number of examinees, J is the total number of items, *a, b,* and *c* are the item discrimination, difficulty, and pseudo guessing parameters, respectively; θ is the person ability parameter; τ_i and β_j are an item and examinee slowness parameters, respectively; μ is general mean of response time; and ρ is a regression coefficient of IRT structure on log of response time.

The following priors were used for the Wang and Hanson's response time model, based on the suggestions by Wang and Hanson (2005):

$\theta_i \sim N(0,1), \quad i = 1, \cdots, N$

$a_j \sim LN(0,1), \quad j = 1, \cdots, J$

$b_j \sim N(0,1), \quad j = 1, \cdots, J$

$c_j \sim Beta(5,17), \quad j = 1, \cdots, J$

$\tau_i \sim U(0,2), \quad i = 1, \cdots, N$

$\beta_j \sim U(0,10), \quad j = 1, \cdots, J$

where N is the total number of examinees, J is the total number of items, *a, b*, and *c* are the item discrimination, difficulty, and pseudo guessing parameters, respectively; θ is the person ability parameter, as in regular 3PL model; τ_i is an item slowness parameter; and β_j is an examinee slowness parameter.

van der Linden's (2007) hierarchical framework incorporate the 3PL IRT model as a level-1 response model as follows:

$$P(x_{ij} = 1 \left| \theta_i, a_j, b_j, c_j \right.) = c_j + \frac{1 - c_j}{1 + e^{-1.7 a_j (\theta_i - b_j)}}.$$

For priors for the population and item models in the previous chapter, it is recommended to use normal inverse-Wishart prior distributions denoted as follows:

$$\sum\nolimits_P \sim Inverse - Wishart(\sum\nolimits_{P_0}^{-1}, v_{P_0}),$$

$$\mu_P \left| \sum\nolimits_P \right. \sim MVN(\mu_{P_0}, \sum\nolimits_P \Big/ k_{P_0}),$$

$$\sum\nolimits_I \sim Inverse - Wishart(\sum\nolimits_{I_0}^{-1}, v_{I_0}),$$

$$\mu_I \left| \sum\nolimits_I \right. \sim MVN(\mu_{I_0}, \sum\nolimits_I \Big/ k_{I_0}),$$

where $v_{P_0}, k_{P_0}, v_{I_0},$ and k_{I_0} are corresponding degrees of freedom parameters for respective Wishart distributions (Gelman et al., 2003). To reflect low model confidence, corresponding degrees of freedom parameters should be set low. In the case of the prior based on other than previous documented works, there is one way to treat this as the size of a pseudo sample as the number of

parameter estimates such as setting $v_{P_0} = 2, k_{P_0} = 1, v_{I_0} = 4,$ and $k_{I_0} = 1$. A prior with such a small pseudo sample size is quite vague, allowing the data to drive the solution. In addition, the following priors were also used based on van der Linden (2007) and Fox et al. (2007):

$$\mu_P = (\mu_\theta, \mu_\tau) = (0, 0),$$

$$\Sigma_P^{-1} = \begin{pmatrix} 1 & 10 \\ 10 & 1 \end{pmatrix},$$

$$\mu_I = (\mu_a, \mu_b, \mu_\alpha, \mu_\beta) = (1, 0, 1, 0),$$

$$\Sigma_I^{-1} = \begin{pmatrix} 1 & 10 & 10 & 10 \\ 10 & 1 & 10 & 10 \\ 10 & 10 & 1 & 10 \\ 10 & 10 & 10 & 1 \end{pmatrix}.$$

Checking model convergence and DIC

Model convergence diagnostics were used to determine the number of iterations for burn-in and the number of post-burn-in. Burn-in iterations were discarded and only post-burn-in iterations were used to estimate the posterior distributions for parameter estimates. In this study, graphical diagnostics such as monitoring history plots, autocorrelation, and posterior distribution of parameter estimates were conducted. The Gelman Rubin convergence diagnostic index was also used for checking the model convergence. After model convergences were confirmed, DIC values from the models were used for model comparison in this study.

Followed by checking model convergences and DIC comparisons among different models, parameter estimates from the three models were compared and examined through the Pearson product-moment correlation. Examining the relationships between related parameter estimates

across the models is important in response time data perspective. Because it is hard to examine true relationships between parameters in population through real data analysis, consistent results from the response time models can be regarded as a proxy of population relationships. Each model has speed related parameters as well as typical item and examinee ability parameters, therefore, investigation of the relationship between these estimates were also available to give further understanding of the given response time data in conjunction with item response analysis.

Study 2

Data generation

Simulated data were generated for study 2. The examinee ability parameters were randomly generated from $N(0,1)$, the item discrimination, item difficulty, and lower asymptote parameters were generated from distributions as follows: $LN(0,.5)$, $N(0,1)$, $Beta(5,17)$, respectively for each item parameter. The generated item parameters are displayed in Table A1 in Appendix.

Response time data also were generated with Thissen's model (1985) conjunction with item and examinee parameters from the 2PL item response model as follows:

$$\ln T_{ij} = \mu + \beta_j + \tau_i - \rho a_j(\theta_i - b_j) + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim LN(0, \sigma_j^2),$$

In order to minimize compounding sources of the speed related variables, overall mean response time (μ), examinee (τ_i) and item slowness parameter (β_j) were set to 0.0s. Natural log of response time data were randomly generated from N(0, 0.5).

Factors of investigation

The design of simulation study included two test lengths, 30 and 60 items; four sample sizes, 100, 500, 1000, and 2000; and three distributions of regression coefficient of Thissen's model in equation (1). Test lengths and sample sizes for examinees reflect that test has moderate to long items in the tests; relatively small to large examinee samples for those tests. Seven levels of regression coefficients, -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9 indicate a range of relationships between examinee ability and response time as well as item difficulty and response time. Positive values of ρ indicate that as examinee's ability increases the response latency decreases; as the item difficulty increases the response latency increases. Likewise, negative values of ρ imply that there are reverse relationships between response time and ability, response time and item difficulty.

A total of 56 different conditions were simulated and two response time models from Bayesian estimation methods were implemented. The simulated data were generated and calibrated 30 times for each of the fifty six conditions. A typical 3PL IRT model was also implemented on the generated data to determine whether there was improvement in estimation procedure when the response time data was considered.

Measured criteria

Assessment of the response time models was based on retrieval of item and examinee ability parameters. The degree to which the response time models recovered the known item and examinee ability parameters were evaluated through descriptive statistics, bias, root mean square error (RMSE) and the Pearson product moment correlation. The means and standard deviation of these error indices are computed across 30 replications. Bias and RMSE were calculated for the

sets of 30 replications as follows:

$$Bias(\hat{\delta}_j) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_{rj} - \delta_j) \quad (13)$$

$$RMSE(\hat{\delta}_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_{rj} - \delta_j)^2} \quad (14)$$

The correlations between the true and estimated item parameters were also computed for each item parameter for each replication. The means and standard deviation of these correlations across replications were computed.

To compare between response time models the deviance information criteria (DIC) and the relative efficiency were calculated. To quantify the amount of improvement attributable to simultaneous estimation, relative efficiency was computed. Relative efficiency is available from mean squared error (MSE). de la Torre and Patz (2005) used calculated relative efficiency from the ratio of MSE from each estimation methods as follows:

$$Relative\ Efficiency = \frac{MSE_{model1}}{MSE_{model2}} = \frac{\frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_{rj}-\delta_j)^2}{\frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_{rk}-\delta_k)^2}. \quad (15)$$

A ratio greater than 1 indicates that given interested estimation method, which is the denominator, has higher efficiency compared to the other method.

Chapter 4. Results

In this chapter the results from the real data and the simulation study are discussed. In the results of Study 1, the three response time models discussed in the previous chapter were applied to the real data. To begin with, the overall descriptions of data are presented and item responses were analyzed via classical testing theory as well as IRT. The response time models were compared through investigations of item and response time parameter estimation. In the results of Study 2, Wang and Hanson's 4PL RT model, and van der Linden's hierarchical framework were applied to the simulated data and item and examinee parameter estimates were examined in various conditions.

Study 1 results

Preliminary Data Analysis

Table 1 indicated descriptive statistics for responses and response times, and Figure 2 showed the distribution of total score and total response times from the real data. Approximate Normal distributions for those data were assumed. Slightly negatively skewed response times were shown, however, this is not an uncommon case when it is a timed testing (Schnipke & Scrams, 1997; Schnipke, Scramps, & van der Linden, 2001).

Among 975 examinees, 917 (94.1%) showed complete responses in the test and every item was reached by more than 95% of examinees. The proportion of missing responses increased as examinees approached the end of the test. The last item of the test showed the biggest proportion of missing responses (3.2%). Descriptive statistics for missing responses are displayed in Table 2.

Table 1

Descriptive statistics for responses and response times (n=975, item=33)

	Mean	Median	SD	Min	Max	Skewness	Kurtosis
Total Score	20.613	21	4.902	4	32	-0.319	-0.234
Total Time	1610.39	1604	284.219	694	2888	0.251	0.841

Table 2

Frequencies for missing responses and response times (n=975)

Omitted response	Frequency	Percent	Cumulative Percent
0	917	94.1	94.1
1	35	3.6	97.7
2	8	0.8	98.5
3	3	0.3	98.8
4	2	0.2	99.0
5	7	0.7	99.7
7	3	0.3	100.0

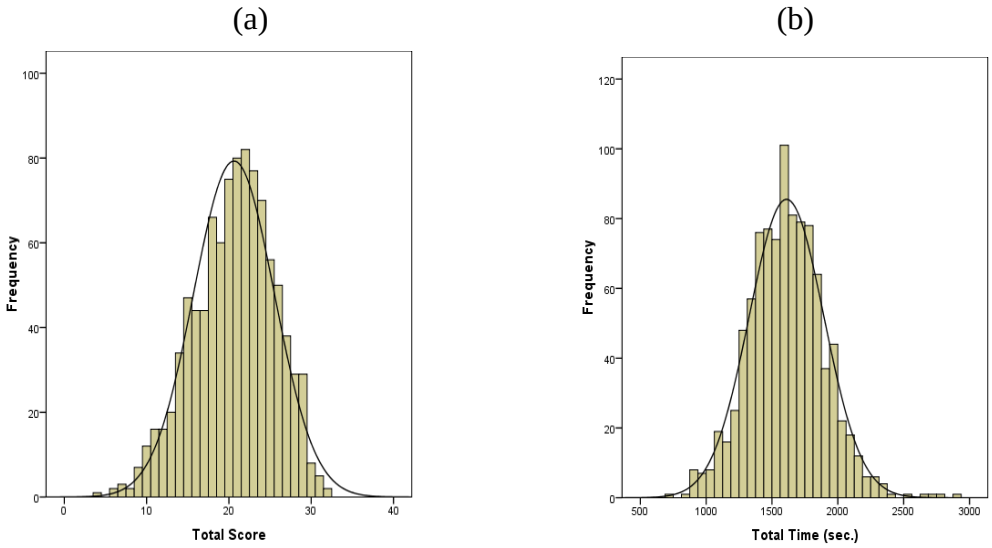


Figure 2. *Histograms of total score (a) and total response time (b).*

Table 3 shows the means and the standard deviations for item parameter estimates from the CTT and IRT estimation methods. Discrimination index and biserial, point biserial correlation were calculated by using the top 27% of the examinees as a high ability group and the bottom 27% of the examinees as a low ability group. Overall, more than 62% of the examinees showed correct responses to the given items. IRT parameter estimates were obtained from BILOG-MG computer program (Zimowsky, Muraki, Mislevy, & Bock, 1996). The mean of the examinee's true ability parameters centered around 0.0 in both 2PL and 3PL models. The item difficulty and discrimination indices showed quite different when guessing parameters were estimated. The mean of the item discrimination parameter estimates is 0.413; the mean difficulty parameter estimate is -0.853 from the 2PL IRT model. The mean item discrimination parameter estimate is 0.63; the mean difficulty parameter is -0.007 from the 3PL IRT model.

Table 3
Means and standard deviations for item parameter estimates from the CTT and IRT methods

	Parameter	Mean	SD
CTT	Proportion of correct response	0.625	0.156
	Discrimination	0.341	0.092
	Biserial Correlation	0.433	0.985
	Point Biserial correlation	0.326	0.069
2PL IRT	Examinee true ability (θ)	0.000	0.897
	Item discrimination (a)	0.413	0.143
	Item difficulty (b)	-0.856	1.160
3PL IRT	Examinee true ability (θ)	0.000	0.903
	Item discrimination (a)	0.630	0.339
	Item difficulty (b)	-0.007	1.221
	Item guessing (c)	0.254	0.064

Figure 3 contains the plot of eigenvalues from the inter item correlation matrix produced by factor analysis for the given data. 7.5% of total variance was explained by the first factor. Test reliability value from the Chronbach's alpha was 0.737, and it is also indicated that each item contributed to the test evenly; differences in the Cronbach's alpha values were small when each item was deleted as shown in Table 4.

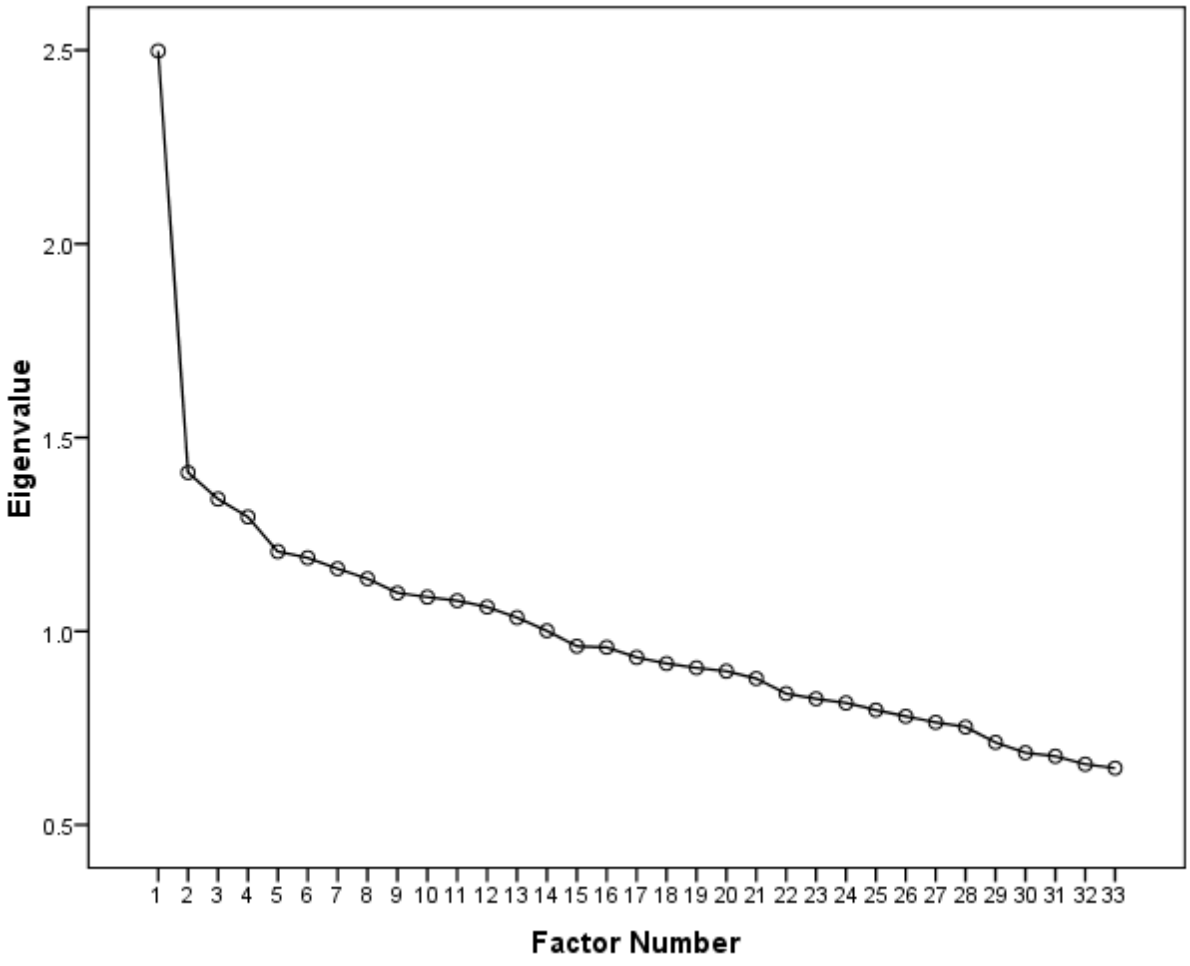


Figure 3. *Scree plot of eigenvalues from factor analysis.*

Table 4
Item-total score correlation coefficients and reliability indices

Item Number	Corrected Item-Total correlation	Chronbach's alpha if item deleted
1	.204	.732
2	.183	.734
3	.176	.734
4	.160	.735
5	.306	.727
6	.105	.738
7	.168	.735
8	.154	.736
9	.297	.727
10	.312	.727
11	.233	.731
12	.108	.739
13	.315	.728
14	.154	.736
15	.254	.730
16	.210	.732
17	.304	.727
18	.151	.736
19	.300	.727
20	.344	.725
21	.332	.725
22	.292	.727
23	.266	.729
24	.167	.735
25	.184	.734
26	.214	.732
27	.324	.725
28	.271	.729
29	.333	.726
30	.291	.728
31	.363	.724
32	.236	.731
33	.287	.728

Each individual item parameter estimates from the CTT and IRT models are presented in Table 5, and descriptive statistics for each item response time are also presented in Table 6.

Table 5
Item parameter estimates from the CTT and IRT models

Item	CTT		2PL IRT		3PL IRT		
number	*P*	*disc*	*a*	*b*	*a*	*b*	*c*
1	.81	.20	0.38	-2.47	0.38	-1.91	0.22
2	.42	.32	0.29	0.73	0.55	1.60	0.25
3	.75	.23	0.31	-2.21	0.39	-0.98	0.30
4	.64	.27	0.25	-1.43	0.35	0.23	0.32
5	.59	.46	0.47	-0.50	0.59	0.08	0.20
6	74	.19	0.21	-3.13	0.23	-1.44	0.29
7	.57	.31	0.27	-0.70	0.45	0.93	0.34
8	.43	.31	0.25	0.67	0.54	1.83	0.30
9	.53	.48	0.47	-0.16	0.61	0.41	0.19
10	.80	.31	0.57	-1.71	0.60	-1.28	0.22
11	.78	.28	0.42	-2.01	0.46	-1.23	0.28
12	.46	.24	0.20	0.49	0.51	2.24	0.36
13	.85	.28	0.67	-1.89	0.72	-1.48	0.24
14	.35	.29	0.25	1.53	0.58	2.20	0.25
15	.67	.34	0.40	-1.17	0.52	-0.23	0.28
16	.60	.34	0.31	-0.84	0.40	0.30	0.27
17	.50	.49	0.47	-0.02	0.60	0.47	0.17
18	.63	.28	0.24	-1.38	0.31	0.12	0.29
19	.42	.46	0.46	0.47	0.61	0.87	0.15
20	.80	.35	0.66	-1.52	0.72	-1.17	0.19
21	.57	.50	0.51	-0.38	1.79	0.59	0.37
22	.46	.46	0.44	0.27	1.76	0.92	0.30
23	.80	.27	0.47	-1.97	0.51	-1.39	0.23
24	.53	.33	0.26	-0.25	1.11	1.15	0.40
25	.46	.32	0.29	0.34	0.41	1.23	0.21
26	.88	.18	0.45	-2.86	0.50	-2.17	0.27
27	.39	.47	0.52	0.62	0.69	0.92	0.12
28	.63	.39	0.42	-0.83	0.59	0.11	0.28
29	.82	.34	0.71	-1.60	0.80	-1.20	0.21
30	.78	.34	0.55	-1.61	0.60	-1.17	0.19
31	.73	.46	0.66	-1.12	0.80	-0.62	0.23
32	.75	.32	0.41	-1.75	0.47	-0.98	0.25
33	.48	.45	0.42	0.15	0.64	0.82	0.22

Table 6

Descriptive statistics for item response times

Item number	Response time			Log of response time		
	Mean (SD)	Skewness	Kurtosis	Mean (SD)	Skewness	Kurtosis
1	61.88 (34.42)	1.34	3.01	3.96 (0.64)	-1.29	4.19
2	63.22 (41.20)	1.62	4.34	3.93 (0.73)	-1.17	3.79
3	57.86 (35.69)	1.50	4.16	3.87 (0.66)	-0.83	2.55
4	32.74 (34.41)	2.07	7.24	2.88 (1.23)	-0.43	-0.62
5	47.48 (35.97)	2.59	11.03	3.62 (0.74)	-0.72	2.33
6	57.30 (36.70)	1.48	3.35	3.84 (0.69)	-0.89	3.27
7	64.09 (42.90)	1.71	4.34	3.94 (0.73)	-1.15	4.05
8	35.44 (34.78)	2.03	6.27	3.03 (1.17)	-0.61	-0.23
9	69.11 (51.40)	2.30	8.65	3.96 (0.84)	-1.39	3.93
10	53.37 (35.67)	2.71	16.33	3.75 (0.79)	-1.71	5.30
11	57.98 (37.45)	1.94	7.42	3.81 (0.87)	-1.89	5.07
12	46.70 (35.37)	2.00	7.03	3.54 (0.89)	-1.35	3.41
13	46.02 (31.41)	2.70	19.34	3.56 (0.87)	-1.63	3.71
14	39.51 (47.32)	2.82	11.41	2.96 (1.37)	-0.52	-0.51
15	51.89 (37.57)	2.31	9.29	3.70 (0.76)	-0.98	2.63
16	50.27 (35.06)	2.25	9.60	3.67 (0.80)	-1.55	5.20
17	55.42 (36.26)	1.60	4.49	3.79 (0.76)	-1.21	3.16
18	42.73 (33.80)	1.82	5.22	3.43 (0.91)	-1.09	2.56
19	26.49 (29.50)	2.86	13.53	2.70 (1.18)	-0.37	-0.53
20	52.66 (39.96)	2.86	15.44	3.71 (0.77)	-0.97	2.99
21	54.79 (32.17)	2.38	12.18	3.84 (0.65)	-1.73	7.94
22	63.17 (37.85)	2.52	17.83	3.96 (0.69)	-1.76	6.86
23	49.48 (29.46)	1.57	6.60	3.69 (0.77)	-1.79	5.52
24	32.23 (33.39)	1.98	6.49	2.82 (1.32)	-0.57	-0.57
25	63.28 (38.94)	1.62	4.62	3.94 (0.74)	-1.50	4.19
26	45.06 (27.58)	2.39	14.10	3.60 (0.75)	-1.87	6.38
27	43.19 (29.75)	2.50	12.57	3.52 (0.80)	-1.59	4.84
28	25.83 (28.15)	2.56	10.53	2.64 (1.24)	-0.46	-0.56
29	45.06 (34.63)	2.34	8.86	3.54 (0.81)	-1.04	2.83
30	42.41 (34.88)	2.99	18.29	3.45 (0.85)	-0.98	2.56
31	51.42 (35.75)	1.54	4.68	3.64 (0.92)	-1.48	3.13
32	45.97 (34.27)	1.59	3.98	3.49 (0.97)	-1.31	2.25
33	41.22 (47.24)	2.73	11.55	3.08 (1.27)	-0.45	-0.44
Total	48.98 (38.02)	2.16	9.21	3.54 (0.99)	-1.16	3.07

Response time models implementation

Thissen's (Thissen, 1983) lognormal response time model, Wang and Hanson's (Wang & Hanson, 2005) 4PL RT model, and van der Linden's (van der Linden, 2007) hierarchical framework were applied to the investigation of response time on the real data through Bayesian estimation using the MCMC method. Those six models, three response time models applied in both 2PL IRT and 3PL IRT as response data models, were implemented using WinBUGS program. First, the model convergence was checked using various graphical methods as well as a diagnostic index. Second, the response time models were compared on response data as well as response time information.

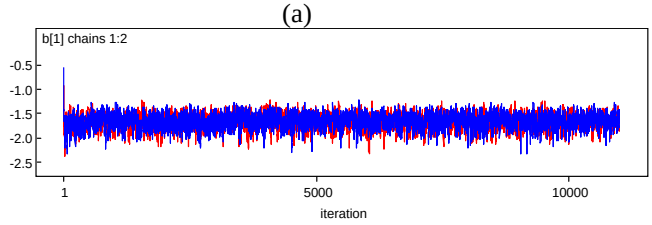
Convergence Check

The model convergence check was conducted by using the Gelman-Rubin diagnostic as well as graphical diagnostic methods. The Gelman-Rubin ratios are available when multiple chains are applied in the model specification and estimation. This study used two chains to calculate the Gelman-Rubin ratios for the item and examinee parameter estimates. Table A2 through Table A4 in Appendix show the Gelman-Rubin diagnostics for the item parameter estimates calculated as the average of the values from 2,000 post burn-in after 8,000 burn-in iterations were discarded. Figure 4 and Figure 5 demonstrate representative item parameter estimates and examinee ability parameter estimates from all the six response time models after implementing the MCMC estimation.

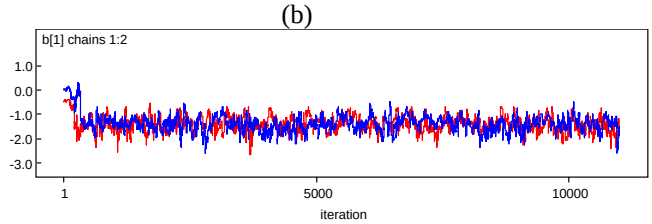
The Gelman-Rubin ratio indicated that all the values for item parameter estimates were around 1.0 which is showing the evidence of the model convergence. The history graphs and the

posterior density plots also showed 2 chains of these six models quickly reached an acceptable convergence on the stationary distribution for all the items. Thus, a conservative burn-in of 8,000 iterations and 2,000 post burn-in iterations were used in implementing all the models in this study. Figure B1 through B6 in Appendix show some exemplary posterior density plots for item parameter and examinee ability parameter estimates in the six response time models.

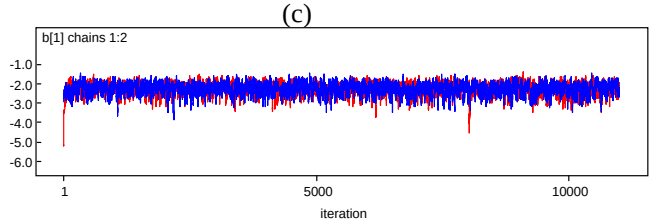
(a)



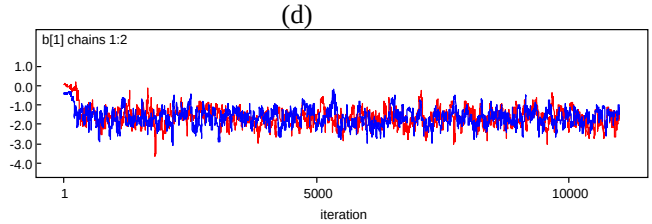
(b)



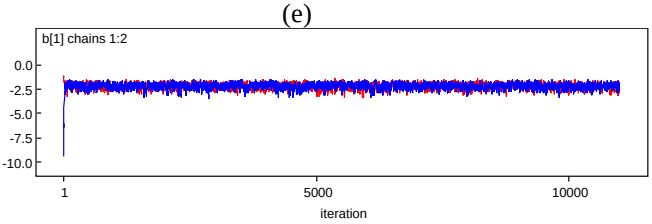
(c)



(d)



(e)



(f)

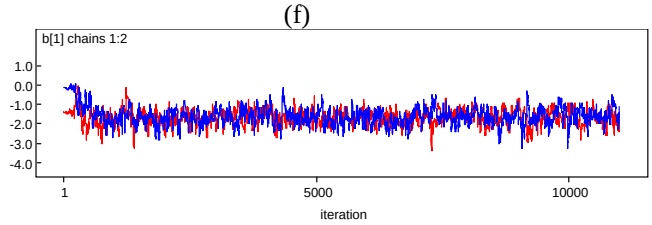


Figure 4. *Some representative history plots of the item difficulty parameter estimates.*
Note. (a) Thissen's model (2PL); (b) Thissen's model (3PL); (c) Wang & Hanson's 4PL RT model (2PL); (d) Wang & Hanson's 4PL RT model(3PL); (e) van der Linden's hierarchical framework (2PL); (f) van der Linden's hierarchical framework (3PL)

(a)

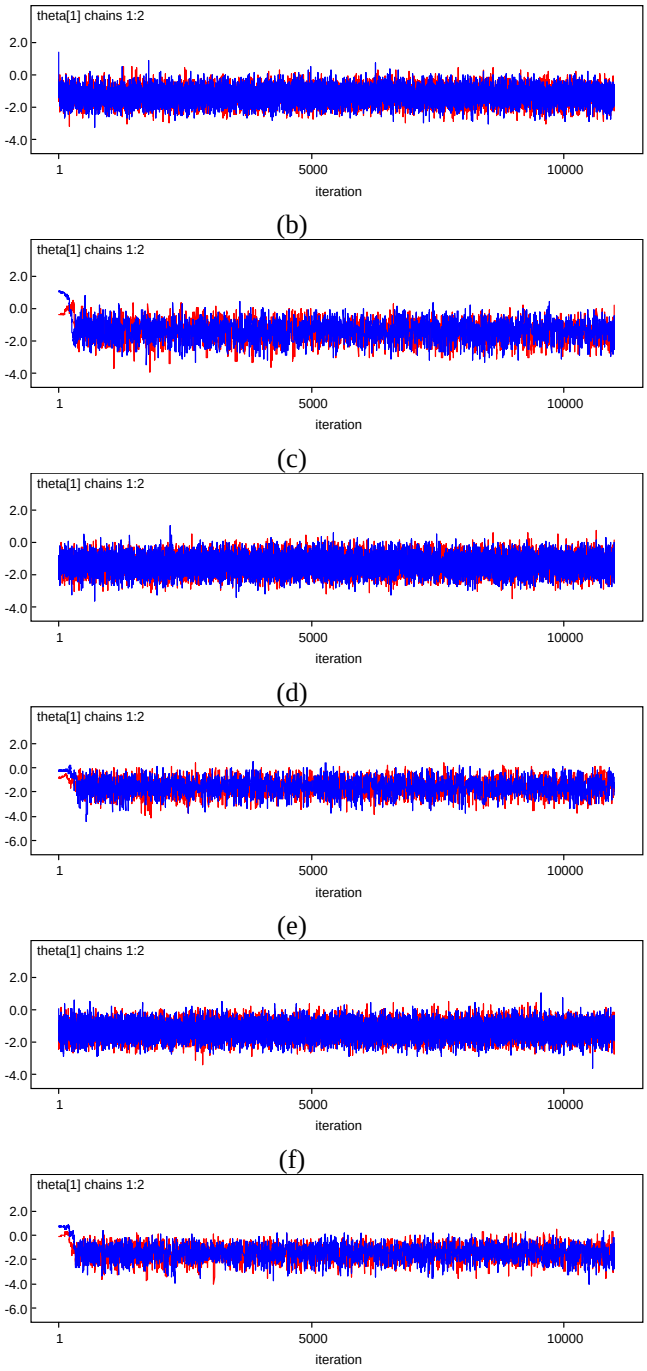


Figure 5. *Some representative history plots of the examinee true ability parameter estimates. Note.* (a) Thissen's model (2PL); (b) Thissen's model (3PL); (c) Wang & Hanson's 4PL RT model (2PL); (d) Wang & Hanson's 4PL RT model(3PL); (e) van der Linden's hierarchical framework (2PL); (f) van der Linden's hierarchical framework (3PL)

Model goodness of fit and comparison

The DIC values from the models were obtained from the posterior means after an additional 1,000 iterations were applied. Table 7 showed the DIC values from the six response time models according to the response and response time distribution. Marginally, 3PL application models showed better fit by having lower DIC values than 2PL applications; hierarchical framework showed the lowest values among three models.

Table 7.
DIC values from the response time models

		Response	Response time	Total
4PL RT model	2PL	36983.3	82504.1	119487.4
	3PL	36915.4	82498.7	119414.1
Hierarchical framework	2PL	36997.9	81504.4	118502.3
	3PL	36896.0	81499.3	118395.3
Thissen's model	2PL	37085.6	85058.4	122144.0
	3PL	37025.0	85189.0	122214.0

Comparison of parameter estimates

Item parameter estimates from the six response time models are displayed in Tables A5 through A7 in Appendix. Descriptive statistics for item and examinee parameter estimates from the six models are presented in Table 8. The means and standard deviations of these models showed similar results across the models.

Table 8.
Means and standard deviations for the item and examinee parameter estimates from the response time models

		Item				Examinee		
		a	b	c	alpha	beta	theta	tau
4PL RT model	2PL	0.719 (0.247)	-0.951 (1.119)			3.378 (2.392)	0.033 (0.823)	0.988 (0.181)
	3PL	0.672 (0.463)	-0.249 (1.127)	0.245 (0.069)		4.695 (2.481)	0.010 (0.821)	0.997 (0.154)
Hierarchical framework	2PL	0.745 (0.261)	-0.776 (1.060)		1.492 (0.527)	3.422 (0.043)	0.024 (0.831)	-0.117 (0.198)
	3PL	0.678 (0.416)	-0.051 (1.062)	0.245 (0.060)	1.492 (0.527)	3.423 (0.043)	0.004 (0.838)	-0.115 (0.199)
Thissen's model	2PL	0.721 (0.251)	-0.729 (1.073)			0.062 (0.519)	0.023 (0.835)	0.003 (0.355)
	3PL	0.622 (0.248)	-0.009 (1.079)	0.238 (0.053)		0.056 (0.438)	0.002 (0.833)	0.001 (0.285)

Table 9 shows the correlation coefficients between item difficulty parameter estimates. Table 10 displays the correlation coefficients between examinee ability parameter estimates among the models. The item and examinee parameter estimates from the 2PL and 3PL IRT model also inserted for comparison with the response time models. Correlations for item parameter estimates in the same application among the models ranged as follows: a) .772~.995 for the item discrimination; b) .915~.995 for the item difficulty; c) .724~.971 for the lower asymptote. Both correlations for item difficulty and correlations for examinee ability were ranged from .915 through .995 and they showed comparable results among the six models and also indicated that these models were pointing the same direction. The results also showed there were higher correlations in the same application of the response model (e.g., 3PL applications) than the counterpart (e.g., 2PL applications).

Table 9.
Correlations between the item difficulty parameter estimates among the models

		4PL RT		Hierarchical framework		Thissen's model		IRT	
		2PL	3PL	2PL	3PL	2PL	3PL	2PL	3PL
4PL RT model	2PL								
	3PL	.953							
Hierarchical framework	2PL	.993	.934						
	3PL	.949	.987	.944					
Thissen's model	2PL	.977	.928	.969	.935				
	3PL	.946	.983	.938	.994	.942			
IRT model	2PL	.987	.925	.997	.935	.955	.927		
	3PL	.937	.979	.938	.994	.915	.986	.935	

Table 10.
Correlations between the examinee true ability parameter estimates among the models

		4PL RT		Hierarchical framework		Thissen's model		IRT	
		2PL	3PL	2PL	3PL	2PL	3PL	2PL	3PL
4PL RT model	2PL								
	3PL	.986							
Hierarchical framework	2PL	.993	.979						
	3PL	.983	.989	.990					
Thissen's model	2PL	.980	.967	.983	.974				
	3PL	.987	.986	.992	.995	.989			
IRT model	2PL	.990	.976	.997	.988	.980	.989		
	3PL	.977	.983	.985	.995	.969	.990	.989	

Examining further the relationships between item response time and other parameter estimates, the Person product moment correlations were investigated. Correlations for the speed related parameter estimates among the models are shown in A10 and A11 in Appendix. These correlations were showing somewhat different directions across the response time models. The correlations between the hierarchical framework and Thissen's model showed high ranged .882~.962 in the item speed parameter estimates. The 4PL RT showed relatively lower correlations in these speed parameter estimates ($r_{\hat{\beta}\hat{\beta}} = .322 \sim .431$). The correlations for the examinee ability parameter estimates also showed a similar pattern. The hierarchical framework and Thissen's models showed closer to each other in the magnitude of the relationship; the speed parameters from 4PL RT models showed weak correlations with the parameters from the other response time models ($r_{\hat{\tau}\hat{\tau}} = -.359 \sim .182$).

Comparison of response time related parameter estimates

All the Response time models in this study have item and examinee response speed parameters that explain relationship between item difficulty and response time, and relationship between the examinee true ability and response speed. Thissen's model has a rho parameter, a regression coefficient of the 2PL IRT structure on response time, which is also indicating overall response latency and the item and examinee parameter estimates. Thissen's models showed that 0.236 of $\hat{\rho}$ value from the 3PL application and 0.309 from the 2PL application model. Thus, those positive values imply that overall response latency is increased as item difficulty increases; response latency is increased as examinee ability decreases.

Both correlations between the item difficulty and the item speed parameter estimates and the

examinee true ability and the examinee speed parameter estimates are displayed in Table 11. The correlations between the item difficulty and the item speed indicated that there were negative relationships in all of the response time models, however, interpretations of these values showed different directions based on the response time models. As the item difficulty increases the item slowness decreases in the 4PL RT models and Thissen's models. However, the hierarchical framework has a response time intensity parameter; a negative correlation implies that the more difficult items tend to be less time intensive. Scatter plots of item difficulty and item speed parameter estimates are displayed in Figure B7 in Appendix.

Correlations between the examinee true ability and the examinee speed parameter estimates showed another complicated results. The hierarchical framework and 4PL RT models showed negative correlations, while positive relationships were shown in Thissen's models. It is pertinent to have a positive relationship in Thissen's models because of an examinee slowness parameter instead of an examinee speed. This result is also pointing the same direction as described in the overall relationship ($\hat{\rho}$) of the IRT structure and response time. The hierarchical framework also showed the same direction as Thissen's models did. 4PL RT models showed weak correlations but having opposite direction between examinee ability and examinee speed; the more able examinees tended to take the exam more faster. Scatter plots of examinee ability and examinee speededness (slowness) parameter estimates are displayed in Figure B8 in Appendix.

Table 11.
Correlations between the item difficulty (b) and item speed (β) parameter estimates (N=33); correlations between the examinee true ability ($\hat{\theta}$) and speed parameter ($\hat{\tau}$) estimates (N=975)

	4PL RT		Hierarchical framework		Thissen's model	
	2PL	3PL	2PL	3PL	2PL	3PL
$r_{\hat{b}\hat{\beta}}$ (1)	-.344	-.424	-.273	-.301	-.629	-.575
$r_{\hat{\theta}\hat{\tau}}$ (2)	-.169	-.318	-.291	-.290	.792	.638

Note. (1) All correlations are significant at $\alpha = .05$; (2) all correlations are significant at $\alpha = .01$.

In order to investigate further the relationships among response speed, item characteristics and examinee ability, average response times on items and examinees are examined. Both correlations between item parameter estimates and the mean of item response time and examinee parameter estimates and response time are displayed in Table 12.

Correlations between response time and item difficulty indicated comparable results across the response time models. However, correlations between response time and item speed parameter estimates indicated contrasting results in the magnitude of relationship across the models. Hierarchical framework and Thissen's models showed strong relationship between the two estimates (Hierarchical framework: $r_{\hat{\beta}RT(2PL)} = .950$, $r_{\hat{\beta}RT(3PL)} = .950$; Thissen's models: $r_{\hat{\beta}RT(2PL)} = .767$, $r_{\hat{\beta}RT(3PL)} = .873$), while 4PL RT models indicated a somewhat weak relationship ($r_{\hat{\beta}RT(2PL)} = .234, r_{\hat{\beta}RT(3PL)} = .233$).

Both correlations between response time and examinee ability and response time and examinee speed parameter estimates showed a similar pattern. Comparable results in the relationship between response time and examinee ability were shown across the response time

models. 4PL RT models showed almost no relationships between examinee speed parameter estimates and mean examinee response time $\left(r_{\hat{\theta}RT(2PL)} = -.032;\ r_{\hat{\theta}RT(3PL)} = -.079\right)$. The hierarchical framework showed highest correlations among the models ($r_{\hat{\theta}RT(2PL)} = -.751;$ $r_{\hat{\theta}RT(3PL)} = -.751$). Although Thissen's models showed positive correlations between examinee ability and examinee speed but it indicated the same direction of the relationship when the concept of the parameter in the model was considered. Thissen's models have an examinee slowness parameter and it showed the same direction with those results from the hierarchical framework.

Table 12.
Correlations among the item parameters and mean response time (N=33); correlations among the examinee parameters and response time (N=975)

	4PL RT		Hierarchical framework		Thissen's model	
	2PL	3PL	2PL	3PL	2PL	3PL
$r_{\hat{b}RT}$	-.147	-.105	-.127	-.143	-.085	-.179
$r_{\hat{\beta}RT}$	.234	.233	.950	.950	.767	.873
$r_{\hat{\theta}RT}$	.143	.133	.156	.150	.198	.165
$r_{\hat{\tau}RT}$	-.032	-.079	-.751	-.751	.618	.695

Note. All correlations are significant at $\alpha = .05$

Study 2 results

The design of Study 2 included four sample sizes (100, 500, 1,000, and 2,000 examinees), two test lengths (30 and 60 items), and seven different conditions of the relationship (ρ=-0.9, -0.6, -0.3, 0.0, 0.3, 0.6, and 0.9) between the 2PL IRT structure and response time from Thissen's model. All the results are obtained from 2,000 posterior burn-in iterations after 8,000 iterations of burn-in. First, obtained DIC values were compared to examine the overall model goodness of fit between two response time models. The item and examinee parameter estimates were compared through examining bias and RMSE values. The estimates from a typical 3PL IRT model were also compared with those of the two response time models to measure improvement from the estimation without considering response time data. Finally, the Pearson product-moment correlation coefficients were examined to compare parameter estimates in the two response time models.

DIC comparison

Table 13 showed mean DIC values for the 4PL RT model and hierarchical framework in various conditions of size of the examinees, number of the items, and strength of the relationships between the IRT structure and response time. Overall, the hierarchical framework showed lower DIC values than the 4PL RT model throughout the conditions. However, it is obvious that the main difference between two models is due to the DIC values from the response time data distributions. The DIC values from the response time distributions in 4PL RT models showed about 2~4 times more than those of hierarchical framework. When the DIC values from the response data distributions were focused, they showed comparable results; the 4PL RT model showed slightly

better fit. For response data distributions the 4PL RT model showed better fit in all of the marginal conditions of the size of the examinees, the 30 items condition, and negative relationships between the IRT structure and response time. The hierarchical framework showed better fit in the 60 items condition and the following conditions: $\rho = 0.0$, 0.3, 0.6, and 0.9. Thus, the hierarchical framework showed better fit when there are either no or positive relationships between the IRT structure and response time.

Table 13.

DIC values for the 4PL RT model and hierarchical framework

		4PL RT			Hierarchical Framework		
		Response	Response time	Total	Response	Response time	Total
Examinees	100	5315.239	42580.988	47896.224	5323.890	10232.769	15556.660
	500	26043.763	210284.195	236327.995	26050.971	50498.912	76549.878
	1000	52311.092	422333.764	474644.857	52316.302	101194.657	153510.933
	2000	104475.890	843590.847	948066.828	104484.028	202254.116	306738.152
Items	30	32422.845	243892.942	276315.799	32441.459	59882.469	92323.918
	60	61650.146	515501.955	577152.153	61646.137	122207.758	183853.894
Rho	-0.9	46993.025	429504.952	476498.140	47038.308	95413.096	142451.399
	-0.6	47021.172	377520.707	424541.733	47043.109	91408.557	138451.643
	-0.3	47039.833	346780.932	393820.746	47043.929	88280.312	135324.234
	0.0	47047.089	337267.701	384314.896	47047.544	86995.960	134043.488
	0.3	47050.904	348963.497	396014.659	47044.352	88311.975	135356.335
	0.6	47051.715	381887.137	428938.852	47043.878	91443.175	138487.063
	0.9	47051.734	435957.212	483008.805	47045.464	95462.720	142508.179
Total		47036.496	379697.448	426733.976	47043.798	91045.114	138088.906

Parameter recovery analysis

A series of recovery analyses were conducted to determine the extent to which the generating parameters could be recovered from the simulated data sets. The recovery analyses considered two issues, recovery of the simulated item parameters and latent ability of the examinees. The recoveries of the item parameters and examinee true ability were assessed using bias and RMSE values between the generating parameters and parameter estimates. Relative efficiency values were also used to measure the efficiency of the given model over the counterpart model by applying MSE values.

Item parameter recovery

Table 14 and Figure 6 indicate the mean bias for item parameters among the two response time models and the 3PL IRT model. Overall, hierarchical framework showed the lowest mean bias in absolute term in all of the three marginal conditions. A similar pattern is shown in RMSE values for item parameters. Figure 7 and Table 15 indicate that the hierarchical framework shows the lowest mean RMSE values in all of the three marginal conditions; 4PL RT models showed better item parameter recoveries than the 3PL IRT model.

Table 14.

Mean bias for the item parameters in the 3 models

		4PL RT			Hierarchical Framework			3PL IRT		
		a	b	c	a	b	c	a	b	c
Examinees	100	.307	.016	-.050	.286	-.124	-.037	.408	-.283	-.149
	500	.384	.008	-.031	.364	-.087	-.021	.420	-.178	-.159
	1000	.411	.015	-.054	.391	-.109	-.039	.457	-.213	-.156
	2000	.474	.018	-.035	.461	-.057	-.026	.486	-.191	-.156
Items	30	.317	.036	-.042	.295	-.093	-.031	.390	-.251	-.263
	60	.470	-.007	-.043	.456	-.095	-.031	.495	-.182	-.047
Rho	-0.9	.403	.036	.008	.385	-.073	.007			
	-0.6	.400	.033	.008	.382	-.076	.007			
	-0.3	.396	.024	.007	.377	-.085	.006			
	0.0	.391	.015	.007	.373	-.094	.006			
	0.3	.389	.002	.007	.371	-.107	.006			
	0.6	.389	-.001	.007	.371	-.109	.006			
	0.9	.388	-.008	.007	.370	-.117	.006			
Total		.394	.014	-.042	.376	-.094	-.031	.443	-.216	-.155

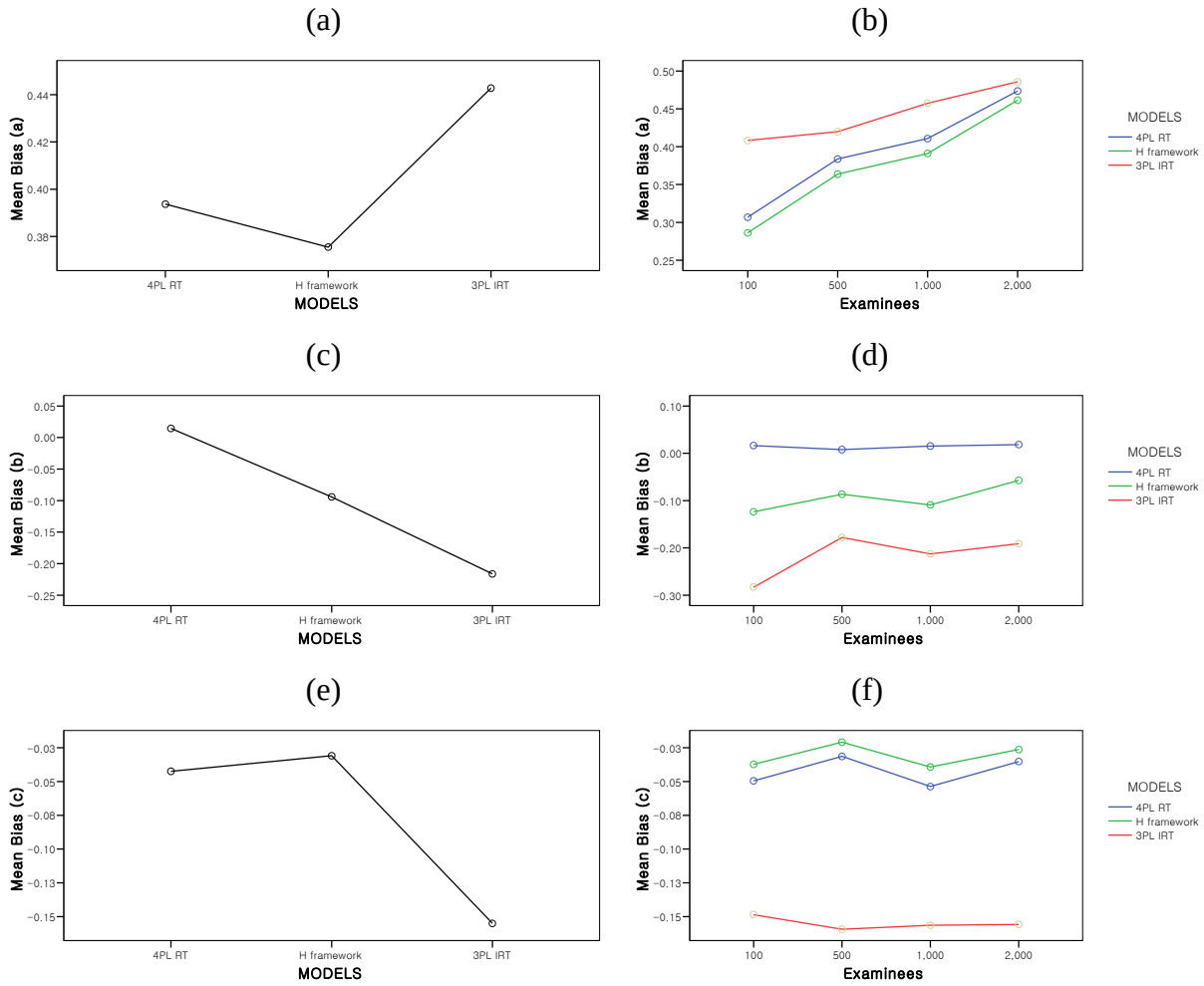


Figure 6. *Mean bias for the item parameters in the 3 models*

Note. (a)-(b) Mean bias for the item discrimination parameter; (c)-(d) mean bias for the item difficulty parameter; (e)-(f) mean bias for the item guessing parameter

Table15.

Mean RMSE for the item parameters in the 3 models

		4PL RT			Hierarchical Framework			3PL IRT		
		a	b	c	a	b	c	a	b	c
Examinees	100	.409	.412	.086	.391	.404	.082	.543	.584	.177
	500	.490	.317	.088	.466	.297	.082	.510	.398	.186
	1000	.529	.300	.084	.508	.307	.078	.529	.401	.182
	2000	.554	.303	.082	.538	.303	.075	.547	.384	.179
Items	30	.451	.333	.086	.425	.317	.079	.468	.443	.270
	60	.540	.334	.084	.527	.338	.080	.597	.441	.092
Rho	-0.9	.507	.350	.088	.488	.344	.083			
	-0.6	.503	.350	.087	.484	.344	.081			
	-0.3	.497	.330	.085	.478	.325	.079			
	0.0	.492	.333	.084	.473	.327	.078			
	0.3	.490	.326	.084	.471	.320	.078			
	0.6	.490	.329	.084	.470	.324	.078			
	0.9	.489	.316	.084	.469	.310	.078			
Total		.496	.333	.085	.476	.328	.079	.532	.442	.181

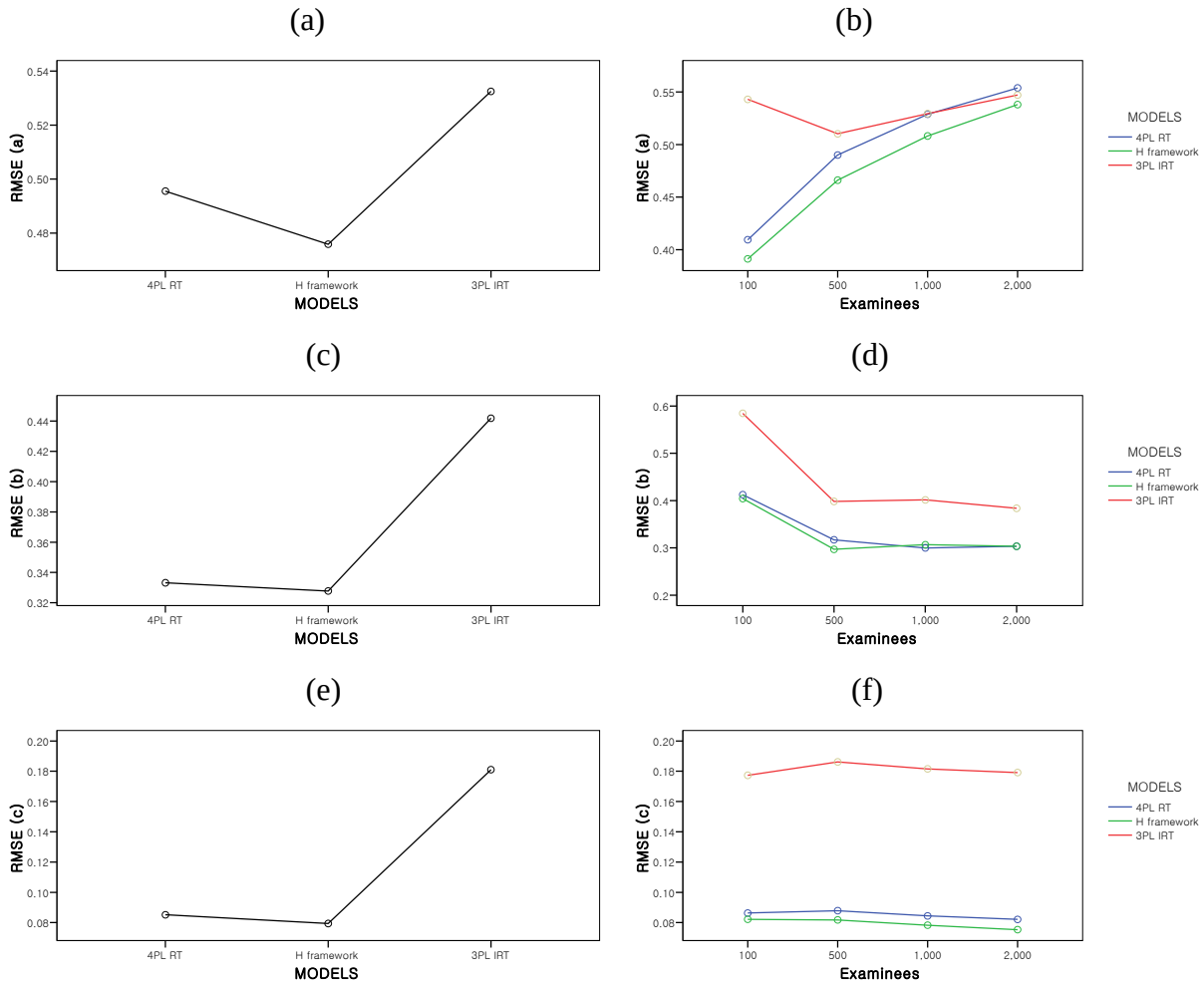


Figure 7. Mean *RMSE for the item parameters in the 3 models*

Note. (a)-(b) Mean RMSE for the item discrimination parameter; (c)-(d) mean RMSE for the item difficulty parameter; (e)-(f) mean RMSE for the item guessing parameter

Relative efficiency from MSE values were displayed in Table 16. It also suggested that hierarchical framework and 4PL RT models showed better results than the 3PL IRT model in all of the conditions; when it is compared with the 4PL RT model, the hierarchical framework is showing greater than 1.0 in all of the conditions except for item difficulty (*b*) parameter in the 60 items condition.

Table 16.

Relative efficiency for the item parameters in the 3 models

		3PL IRT/4PL RT			3PL IRT/Hierarchical Framework			4PL RT/ Hierarchical Framework		
		a	b	c	a	b	c	a	b	c
Examinees	100	1.778	1.823	5.571	1.936	1.966	5.571	1.089	1.000	1.000
	500	1.086	1.298	5.375	1.200	1.610	6.143	1.105	1.143	1.143
	1000	1.004	1.630	5.857	1.088	1.482	6.833	1.084	1.167	1.167
	2000	0.977	1.358	5.857	1.034	1.437	6.833	1.058	1.167	1.167
Items	30	1.058	1.549	9.125	1.196	1.856	12.167	1.130	1.198	1.333
	60	1.214	1.597	1.143	1.274	1.526	1.333	1.050	0.956	1.167
Rho	-0.9							1.078	1.059	1.143
	-0.6							1.079	1.065	1.143
	-0.3							1.081	1.067	1.167
	0.0							1.083	1.073	1.167
	0.3							1.083	1.076	1.167
	0.6							1.084	1.066	1.167
	0.9							1.084	1.089	1.167
Total		1.073	1.327	2.129	1.118	1.348	2.291	1.042	1.015	1.076

Table 17 through Table 20 show the results from the two three-way multivariate analysis of variances to investigate further the recoveries of each of the item parameters. All of the main effects of the 3 factors (estimation models, the numbers of examinees, and item numbers) as well as the interaction effects indicated statistically significant differences on the two measured criteria across the three item parameters. The MANOVA results indicate that the omnibus F-test was significant and the model accounted for a very large amount of variance ($bias: F_{(6,7148)} = 6913.79, p < .001,\ partial\,\eta^2 = .853;\ RMSE: F_{(6,7148)} = 3853.415, p < .001,\ partial\,\eta^2 = .764$). For effect size measures, all of the main effects and the interaction effects showed large

effects that have values greater than .135, except a 3-way interaction effect ($bias$: $F_{(18,10109)} = 21.052, p < .001,\ partial\ \eta^2 = .034$; $RMSE$: $F_{(6,7148)} = 5.066, p < .001,\ partial\ \eta^2 = .008$).

Table 17.

The MANOVA results for the bias of the item parameters

Source	*Wilks' Lambda*	*F*	*Hypothesis df*	*Error df*	*p-value*	*Partial* η^2
Model	.022	6913.787	6	7148	<.001	.853
Examinee	.195	924.020	9	8698	<.001	.420
Item	.047	24396.728	3	3574	<.001	.953
Model*Examinee	.626	101.176	18	10109	<.001	.145
Model*Item	.066	3457.934	6	7148	<.001	.744
Model*Examinee *Item	.901	21.052	18	10109	<.001	.034

Table 18.

The post hoc comparison results for the bias of the item parameters

Dependent variable	*Model*	*Mean difference*	*Standard error*	*p-value*
Bias(a)	H – 4PL	-.018	.0007	<.001
	H – 3PL	-.067	.0015	<.001
	4PL – 3PL	-.049	.0015	<.001
Bias(b)	H – 4PL	-.109	.0017	<.001
	H – 3PL	.122	.0034	<.001
	4PL – 3PL	.231	.0034	<.001
Bias(c)	H – 4PL	.012	.0003	<.001
	H – 3PL	.124	.0005	<.001
	4PL – 3PL	.113	.0005	<.001

Table 19.

The MANOVA results for the RMSE of the item parameters

Source	*Wilks' Lambda*	*F*	*Hypothesis df*	*Error df*	*p-value*	*Partial* η^2
Model	.056	3853.415	6	7148	<.001	.764
Examinee	.414	422.612	9	8698	<.001	.255
Item	.059	19023.672	3	3574	<.001	.941
Model*Examinee	.717	70.049	18	10109	<.001	.105
Model*Item	.068	3470.072	6	7148	<.001	.739
Model*Examinee *Item	.975	5.066	18	10109	<.001	.008

Table 20.

The post hoc comparison results for the RMSE of the item parameters

Dependent variable	*Model*	*Mean difference*	*Standard error*	*p-value*
RMSE(a)	H – 4PL	-.020	.0009	<.001
	H – 3PL	-.057	.0018	<.001
	4PL – 3PL	-.037	.0018	<.001
RMSE(b)	H – 4PL	-.005	.0042	.591
	H – 3PL	-.114	.0084	<.001
	4PL – 3PL	-.109	.0084	<.001
RMSE(c)	H – 4PL	-.006	.0002	<.001
	H – 3PL	-.102	.0004	<.001
	4PL – 3PL	-.096	.0004	<.001

Examinee true ability parameter recovery

Table 21 and Figure 8 show the bias and RMSE values for the examinee ability parameter in the models. Overall, the examinee true ability parameters for hierarchical framework recovered better than the other two models by showing the lowest mean bias in absolute term and mean RMSE values (*bias*=-.005; *RMSE*=.422). Parameters for the 3PL IRT models were recovered slightly better than the 4PL RT model.

Table.21

Bias and RMSE for the examinee true ability parameter in the 3 models

		4PL RT		Hierarchical Framework		3PL IRT	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
Examinees	100	-.087	.448	-.040	.422	-.065	.451
	500	.007	.435	.020	.419	-.064	.414
	1000	-.016	.436	-.006	.421	.002	.416
	2000	.005	.443	.006	.426	.003	.417
Items	30	-.013	.503	-.001	.480	-.051	.477
	60	-.033	.377	-.010	.363	-.011	.372
Rho	-0.9	-.029	.467	-.006	.449		
	-0.6	-.027	.450	-.005	.432		
	-0.3	-.026	.440	-.006	.421		
	0.0	-.023	.433	-.004	.415		
	0.3	-.021	.430	-.007	.411		
	0.6	-.019	.434	-.004	.416		
	0.9	-.017	.428	-.004	.409		
Total		-.023	.440	-.005	.422	-.031	.424

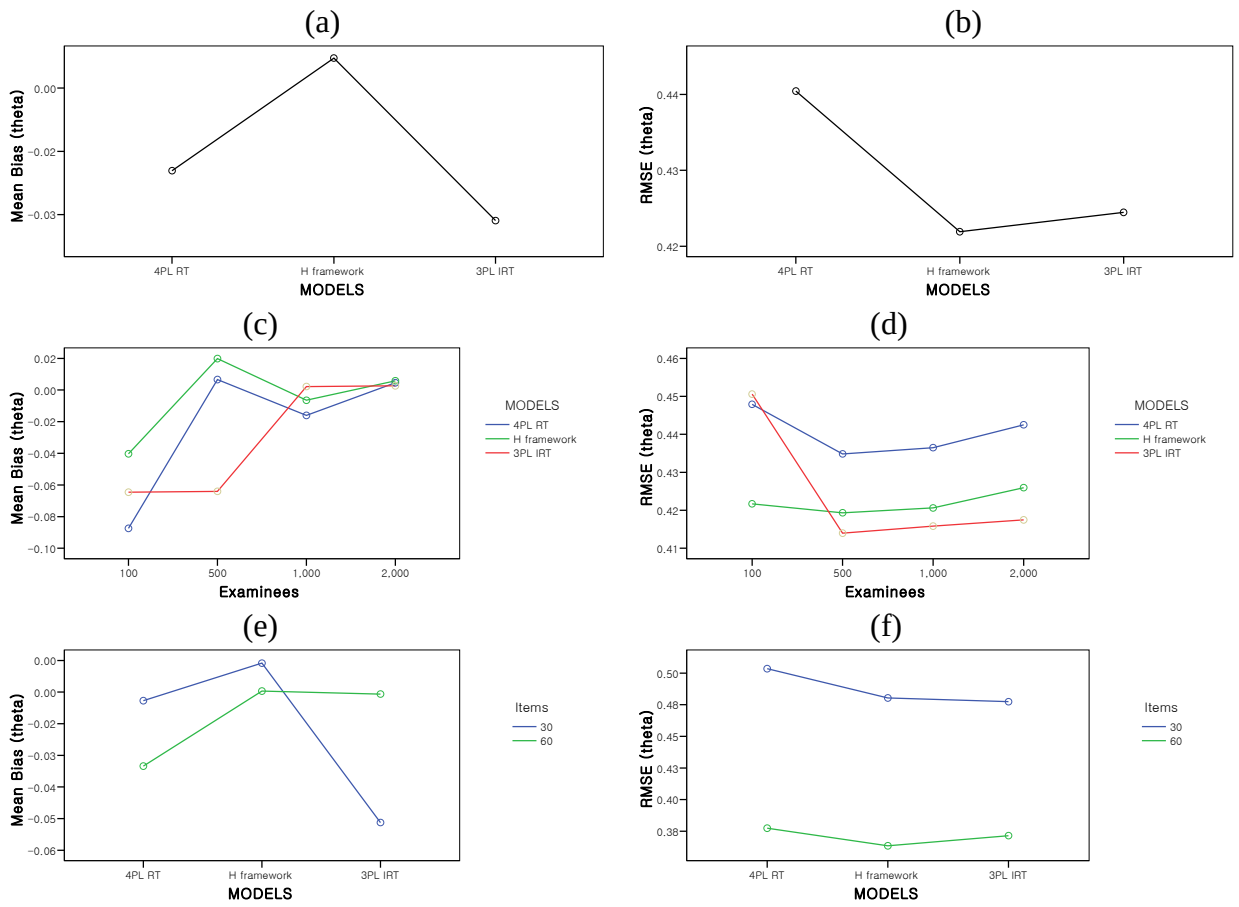


Figure 8. *Bias and RMSE for the examinee true ability parameter in the 3 models.*

Note. (a)-(b) Bias and RMSE for the examinee ability parameter; (c)-(d) Bias and RMSE based for the examinee ability parameter on the number of the examinees; (e)-(f) Bias and RMSE for the examinee ability parameter based on the number of the items

Table 22 Shows relative efficiency values from MSE and suggests that the 3PL IRT model is the most efficient model overall. MSE values for hierarchical framework were lower than the 3PL IRT model in the conditions of the 60 items and the 100 examinees; the hierarchical framework was shown better results than the 4PL RT model in all of the conditions.

Table 22.

Relative efficiency for the examinee ability parameter in the 3 models

		3PL IRT/4PL RT	3PL IRT/Hierarchical Framework	4PL RT/Hierarchical Framework
Examinees	100	0.972	1.084	1.115
	500	0.888	0.956	1.077
	1000	0.893	0.967	1.082
	2000	0.881	0.957	1.086
Items	30	0.888	0.979	1.103
	60	0.945	1.015	1.074
Total		0.909	0.993	1.092

A three-way MANOVA was conducted to determine the effect of 3 factors (the estimation models, the number of the examinees, and the items) on the three measured criteria (bias, MSE, RMSE). The MANOVA results confirmed that hierarchical framework was shown to be the best model in the examinee parameter recovery. All of the main effects were statistically significant, post hoc analyses to the MANOVA for the estimation models were conducted using Bonferroni method. Although the relative efficiency indicated the 3PL IRT model was the best recovered model, differences from the hierarchical framework were not statistically significant. These results are summarized in Table 23 and 24 for the MANOVA and the post hoc procedures respectively.

Table 23.

The MANOVA results for the measured criteria of the examinee true ability

Source	*Wilks' Lambda*	*F*	*Hypothesis df*	*Error df*	*p-value*	*Partial* η^2
Model	.022	6913.787	6	7148	<.001	.853
Examinee	.195	924.020	9	8698	<.001	.420
Item	.047	24396.728	3	3574	<.001	.953
Model*Examinee	.626	101.176	18	10109	<.001	.145
Model*Item	.066	3457.934	6	7148	<.001	.744
Model*Examinee *Item	.901	21.052	18	10109	<.001	.034

Table 24.

The post hoc comparison results for the measured criteria of the examinee true ability

Dependent variable	*Model*	*Mean difference*	*Standard error*	*p-value*
Bias(θ)	H – 4PL	-.018	.0007	<.001
	H – 3PL	-.067	.0015	<.001
	4PL – 3PL	-.049	.0015	<.001
RMSE(θ)	H – 4PL	-.109	.0017	<.001
	H – 3PL	.122	.0034	<.001
	4PL – 3PL	.231	.0034	<.001
MSE(θ)	H – 4PL	.012	.0003	<.001
	H – 3PL	.124	.0005	<.001
	4PL – 3PL	.113	.0005	<.001

The recovery of the examinee true ability parameter was investigated further by using 10 θ categories. The examinees that have smaller or greater than the absolute value of 2.0 in θ were grouped in $\theta 1$ and $\theta 10$ respectively; eight more θ categories were generated in between -2.0 and 2.0 in step of 0.5. Figure 9 and 10 show bias and RMSE values of these 10 θ categories. The 3PL IRT model showed the least bias in absolute term throughout the examinee ability groups. RMSE values also displayed a similar pattern, however, the 4PL RT and hierarchical framework showed comparable or lower RMSE values in the middle ability groups ($\theta 3$ through $\theta 7$). The mean bias and RMSE values of the examinee true ability parameters are displayed in Table A14 and A15 in Appendix. Figure B7 in Appendix also shows differences in RMSE values of the 3 models in each of the θ groups in more detail.

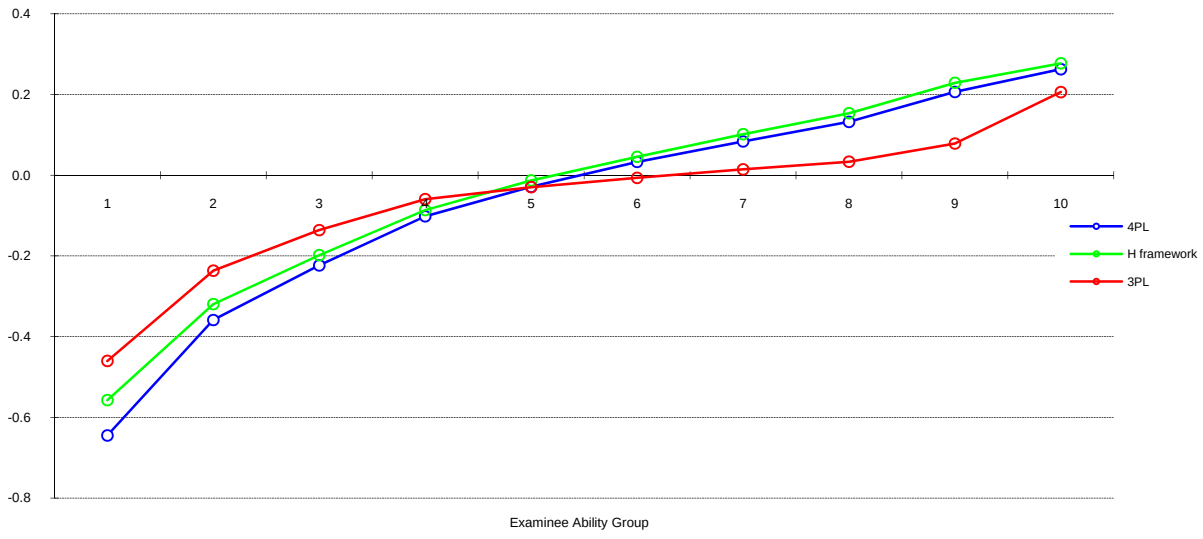


Figure 9. *Bias for the examinee true ability parameter based on the examinee ability groups.*

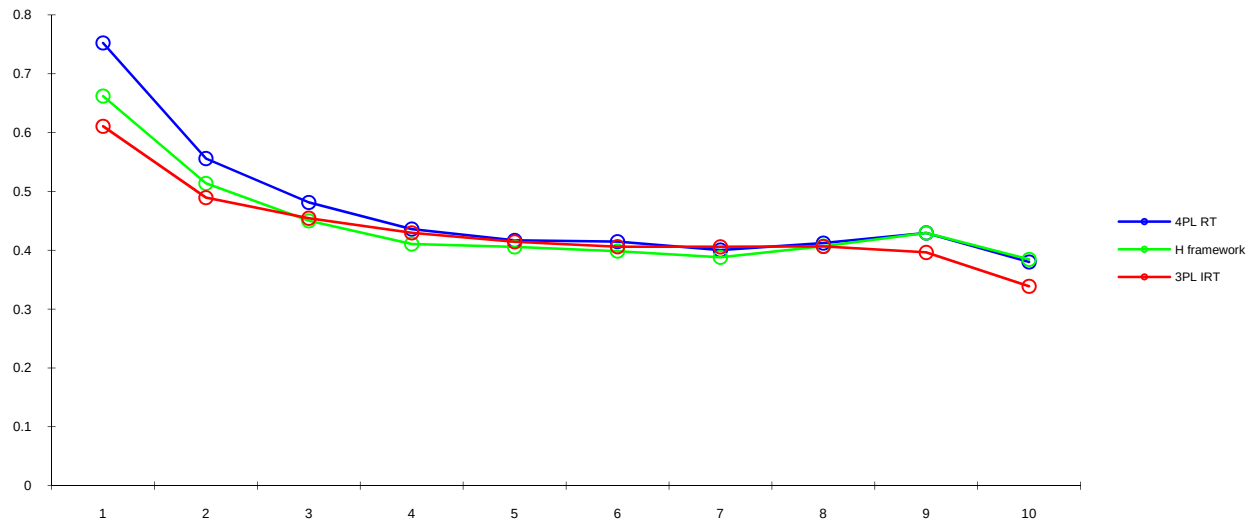


Figure 10. *RMSE for the examinee true ability parameter based on the examinee ability groups.*

A three-way MANOVA was conducted again to determine the effect of 3 factors (the estimation models, the number of examinees, and the items) on the RMSE values in the 10 θ categories. The results confirmed that the hierarchical framework and 4PL RT model showed comparable results in the examinee parameter recovery based on the examinee ability groups. All of the main effects were statistically significant in the MANOVA results, post hoc analyses to the MANOVA for the estimation models were conducted using Bonferroni method. Although the bias and the RMSE values favored the 3PL IRT model, most of the differences from the hierarchical framework and 4PL RT model were not significant in the examinee ability group analysis. The 3PL IRT model showed significant differences from the 2 response time models in $\theta 9$ and $\theta 10$ categories. These results are summarized in Table 25 and 26 for the MANOVA and the post hoc procedures respectively.

Table 25.

The MANOVA results for the RMSE of the examinee ability based on the ability groups

Source	*Wilks' Lambda*	*F*	*Hypothesis df*	*Error df*	*p-value*	*Partial* η^2
Model	.883	22.794	20	7134	<.001	.060
Examinee	.259	204.015	30	10470	<.001	.363
Item	.422	488.849	10	3567	<.001	.578
Model*Examinee	.929	4.442	60	18693	<.001	.012
Model*Item	.963	6.873	20	7134	<.001	.019
Model*Examinee *Item	.977	1.416	60	18693	.019	.004

Table 26.

The post hoc comparison results for the RMSE of the examinee true ability based on ability groups

Dependent variable	*Model*	*Mean difference*	*Standard error*	*p-value*
RMSE(θ1)	H – 4PL	-.090	.0093	<.0001
	H – 3PL	.051	.0186	.0172
	4PL – 3PL	.142	.0186	<.0001
RMSE(θ2)	H – 4PL	-.042	.0041	<.0001
	H – 3PL	.024	.0082	.0101
	4PL – 3PL	.066	.0082	<.0001
RMSE(θ3)	H – 4PL	-.031	.0030	<.0001
	H – 3PL	-.005	.0060	1.0000
	4PL – 3PL	.027	.0060	<.0001
RMSE(θ4)	H – 4PL	-.026	.0026	<.0001
	H – 3PL	-.019	.0051	.0006
	4PL – 3PL	.006	.0051	.6220
RMSE(θ5)	H – 4PL	-.011	.0035	.0044
	H – 3PL	-.009	.0070	.6483
	4PL – 3PL	.002	.0070	1.0000
RMSE(θ6)	H – 4PL	-.016	.0036	<.0001
	H – 3PL	-.008	.0072	.8767
	4PL – 3PL	.009	.0072	.7203
RMSE(θ7)	H – 4PL	-.012	.0031	<.0001
	H – 3PL	-.018	.0062	.0115
	4PL – 3PL	-.005	.0062	1.0000
RMSE(θ8)	H – 4PL	-.005	.0023	.0601
	H – 3PL	.000	.0046	1.0000
	4PL – 3PL	.006	.0046	.6449
RMSE(θ9)	H – 4PL	-.000	.0042	1.0000
	H – 3PL	.033	.0084	.0003
	4PL – 3PL	.033	.0084	.0002
RMSE(θ10)	H – 4PL	.004	.0033	.5862
	H – 3PL	.046	.0066	<.0001
	4PL – 3PL	.041	.0066	<.0001

Note. Mean difference is significant at α=.0017.

Correlations between item parameters and estimates

Table 27 and Figure 11 show the Pearson product moment correlation coefficients between the item parameters and estimates in the various conditions across 30 replications. Overall the mean correlations between the item parameter and estimates from the 3 models were high; the highest correlation were shown in hierarchical framework in all of the item parameters ($r_{a\hat{a}} = .761;\ r_{b\hat{b}} = .941;\ r_{c\hat{c}} = .496$). The hierarchical model also showed consistent results throughout various conditions of the ρ parameter when it was compared to the 4PL RT model. These results are displayed in Figure 12. It was also noted that the 3PL IRT model showed lower correlations in the 100 examinees condition comparing to response time models. However, in the other examinees conditions the results were comparable to those of the response time models.

Table 27.

Correlation between the item parameters and estimates in the 3 models

		4PL RT			Hierarchical Framework			3PL IRT		
		a	b	c	a	b	c	a	b	c
Examinees	100	.600	.895	.419	.619	.905	.410	.512	.871	.366
	500	.738	.941	.496	.773	.950	.487	.745	.947	.443
	1000	.796	.953	.535	.831	.956	.532	.825	.955	.513
	2000	.845	.960	.602	.878	.965	.617	.877	.965	.594
Items	30	.713	.931	.515	.732	.936	.479	.731	.936	.467
	60	.776	.944	.511	.782	.945	.509	.748	.933	.491
Rho	-0.9	.678	.919	.546	.766	.942	.499			
	-0.6	.721	.930	.528	.760	.935	.496			
	-0.3	.748	.937	.513	.760	.943	.497			
	0.0	.762	.941	.507	.760	.940	.499			
	0.3	.769	.943	.501	.760	.942	.498			
	0.6	.768	.941	.497	.759	.942	.492			
	0.9	.767	.949	.497	.760	.944	.492			
Total		.745	.937	.513	.761	.941	.496	.740	.934	.479

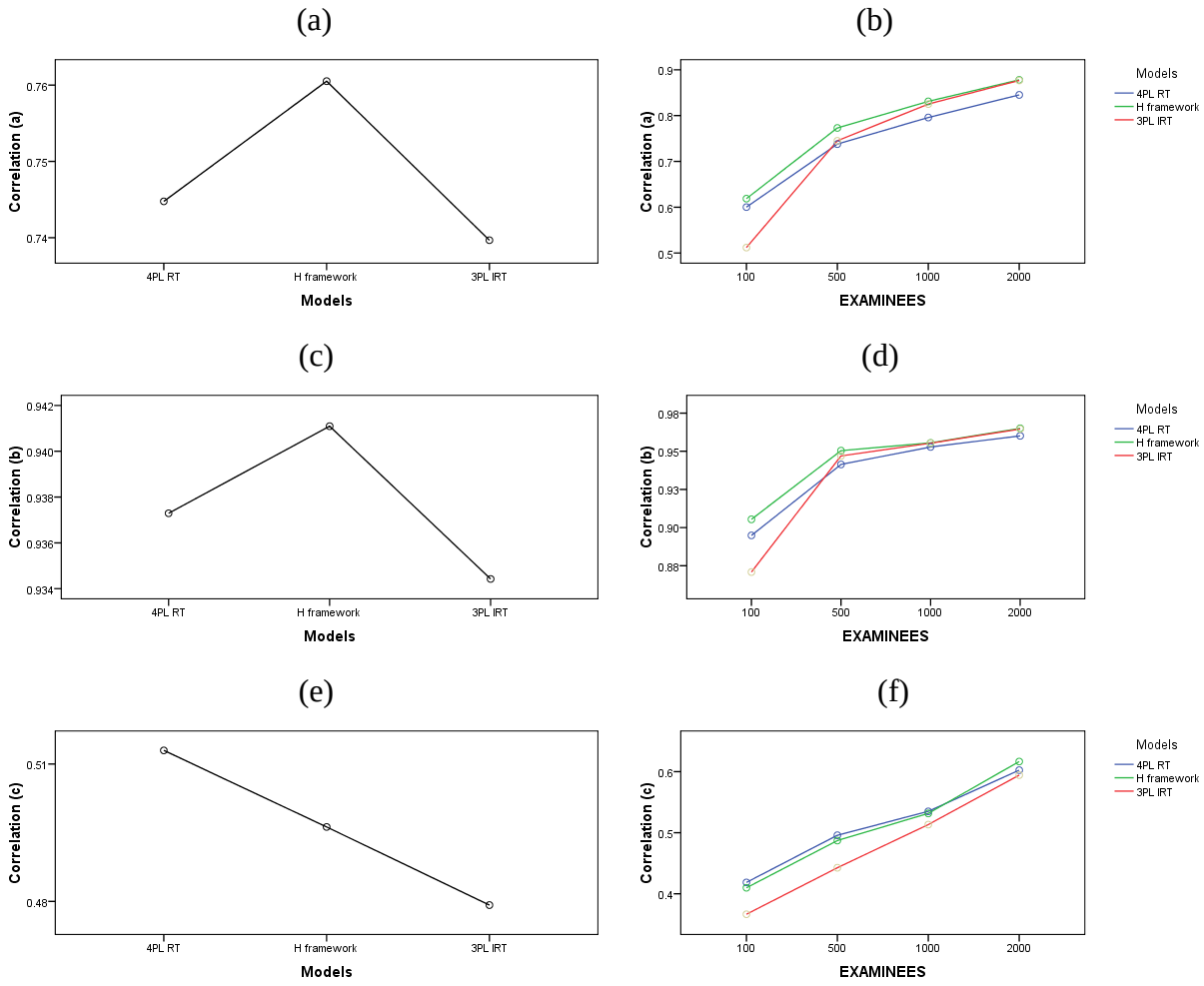


Figure 11. *Correlation between item parameters and estimates in the 3 models*

Note. (a)-(b) Correlation coefficients for the item discrimination parameter; (c)-(d) Correlation coefficients for the item difficulty parameter; (e)-(f) Correlation coefficients for the lower asymptote

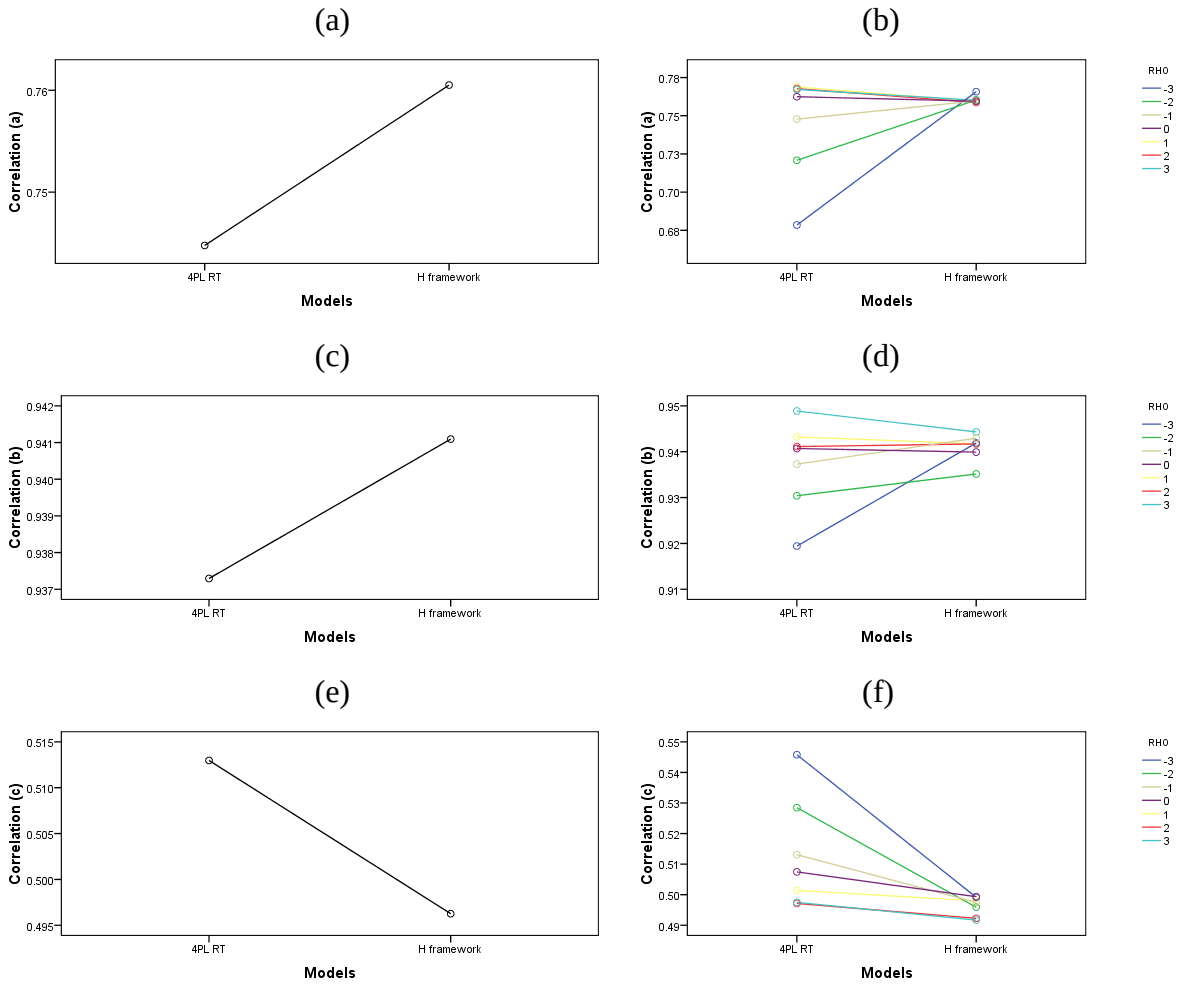


Figure 12. *Correlation between item parameters and estimates in the 2 response time models*

Note. (a)-(b) Correlation coefficients for the item discrimination parameter; (c)-(d) Correlation coefficients for the item difficulty parameter; (e)-(f) Correlation coefficients for the lower asymptote

Correlation between examinee parameters and estimates

Table 28 and Figure 13 show the Pearson product moment correlations between the examinees true ability parameter and the estimates in the various conditions. Overall the 3PL IRT model showed the highest correlation among the models ($r_{\theta\hat{\theta}} = .876$). The hierarchical framework showed higher correlation than the 3PL IRT model in the condition of the 2,000 examinees. When it is compared to the 4PL RT model shown in (c) and (d) of the Figure 13, the hierarchical framework showed consistent correlations throughout of the seven conditions of the ρ parameter.

Table 28.

Correlation between the examinee true ability parameter and estimates in the 3 models

		4PL RT	Hierarchical Framework	3PL IRT
Examinees	100	.818	.833	.853
	500	.857	.868	.885
	1,000	.856	.867	.884
	2,000	.850	.901	.881
Items	30	.798	.815	.837
	60	.893	.898	.914
Rho	-0.9	.800	.871	
	-0.6	.829	.867	
	-0.3	.843	.867	
	0.0	.853	.865	
	0.3	.860	.871	
	0.6	.867	.859	
	0.9	.864	.869	
Total		.845	.862	.876

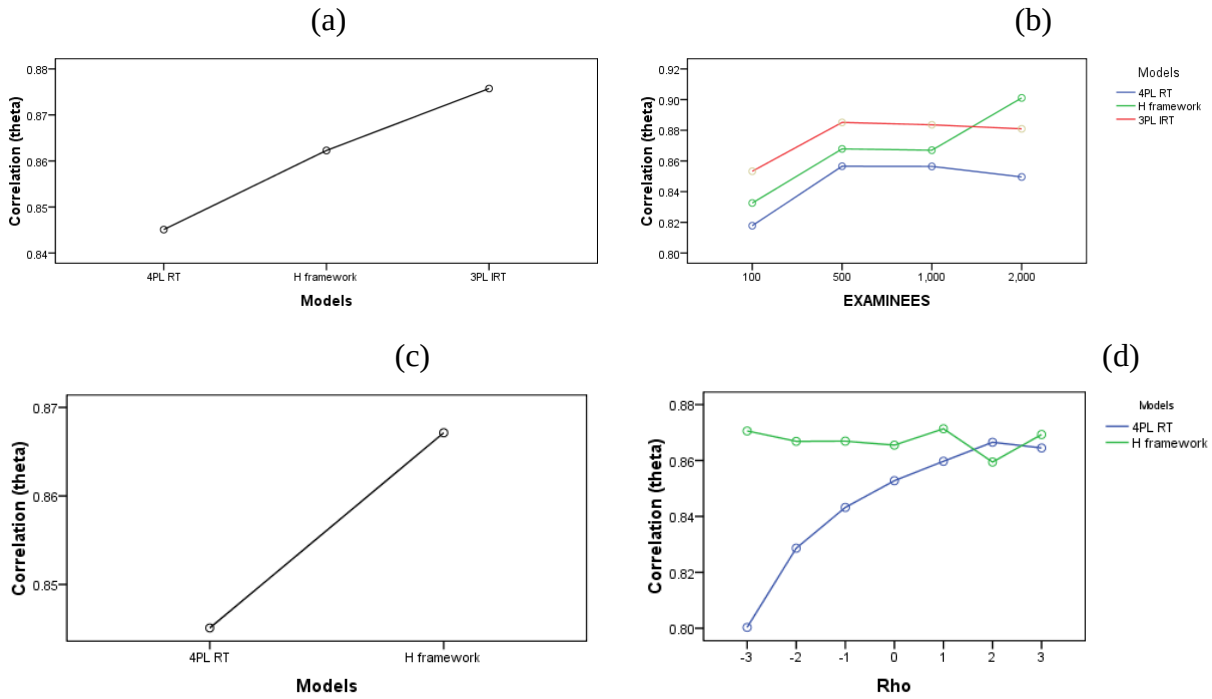


Figure 13. *Correlation between the examinee true ability parameters and estimates in the 3 models (a, b); correlation between examinee the examinee true ability parameters and estimates in the 2 response time models (c, d).*

Correlation between response time-related parameters and estimates

In order to examine the response time related parameters between the response time models, the Pearson product moment correlations for the item and examinee parameters were calculated. The means of the correlations from parameter estimates are summarized in Table 29. These results show that the estimated item parameters from the two response time models are highly correlated. The examinee true ability estimates from the two response time models show a perfect correlation. However, correlations for response time related parameter estimates ($\hat{\beta}, \hat{\tau}$) indicate there are no or weak relationships between two response time models.

Table 29.
Correlations between item and examinee parameter estimates from the 2 response time models

		$r_{\hat{a}\hat{a}}$	$r_{\hat{b}\hat{b}}$	$r_{\hat{c}\hat{c}}$	$r_{\hat{\beta}\hat{\beta}}$	$r_{\hat{\tau}\hat{\tau}}$	$r_{\hat{\theta}\hat{\theta}}$
Examinees	100	0.951	0.981	0.915	0.246	-0.194	1.000
	500	0.960	0.982	0.917	0.305	-0.122	1.000
	1000	0.958	0.984	0.890	0.319	-0.090	1.000
	2000	0.965	0.984	0.879	0.328	-0.059	1.000
Items	30	0.946	0.982	0.892	0.206	-0.023	1.000
	60	0.972	0.984	0.909	0.392	-0.210	1.000
Rho	-0.9	0.880	0.967	0.845	0.249	0.052	1.000
	-0.6	0.943	0.973	0.885	0.149	0.063	1.000
	-0.3	0.970	0.988	0.909	-0.029	0.096	1.000
	0.0	0.977	0.987	0.917	0.047	-0.030	1.000
	0.3	0.981	0.992	0.918	0.526	-0.254	1.000
	0.6	0.980	0.982	0.915	0.570	-0.348	1.000
	0.9	0.980	0.989	0.914	0.582	-0.392	1.000
Total		0.959	0.983	0.900	0.299	-0.116	1.000

In order to examine the relationships among parameters from the item response model and the response time models, correlations between the item difficulty parameter (b) and the item speededness parameter (β) estimates are examined. Mean correlation between two parameter estimates from hierarchical framework is 0.018, which is indicating almost no relationship between item difficulty and item speededness overall. 4PL RT model, however, showed somewhat positive relationship between two parameter estimates ($r_{\hat{b}\hat{\beta}} = .234$). When it was further examined along with the relationship between response time and the IRT structure, the differences from the two models were clearly manifested. The correlations between item difficulty and item speediness ($r_{\hat{b}\hat{\beta}}$) in 4PL RT models showed positive relationships($r_{\hat{b}\hat{\beta}} = .120$) in the $\rho = -0.3$ condition as well as in the $\rho = 0.0$ condition ($r_{\hat{b}\hat{\beta}} = .360$). The hierarchical framework showed the following results: no relationship ($r_{\hat{b}\hat{\beta}} = .022$) in the $\rho = 0.0$ condition, positive relationships ($r_{\hat{b}\hat{\beta}} = .901; r_{\hat{b}\hat{\beta}} = .910; r_{\hat{b}\hat{\beta}} = .921$) in the $\rho = 0.3, 0.6, 0.9$ conditions, and negative relationships ($r_{\hat{b}\hat{\beta}} = -.900; r_{\hat{b}\hat{\beta}} = -.898; r_{\hat{b}\hat{\beta}} = -.821$) in $\rho = -0.3, -0.6, -0.9$ conditions.

Correlations between the examinees true ability (θ) and the examinees speededness parameter (τ) also showed a similar pattern across the two models. Correlations from the 4PL RT model indicated somewhat negative relationship ($r_{\hat{\theta}\hat{\tau}} = -.025$) while the hierarchical framework showed almost no relationship ($r_{\hat{\theta}\hat{\tau}} = .003$). Item discrimination ($a$) and response time discrimination parameters (α) in hierarchical framework showed negative correlations throughout the condition; as the rho parameter increases, the strength of the correlations between item discrimination and response time discrimination also increases. These results are displayed in

Figure 14 and summarized in Table 30.

Table 30.
Correlations between response time related parameter estimates from the 2 response time models

		4PL RT		Hierarchical Framework		
		$r_{\hat{b}\hat{\beta}}$	$r_{\hat{\theta}\hat{\tau}}$	$r_{\hat{b}\hat{\beta}}$	$r_{\hat{\theta}\hat{\tau}}$	$r_{\hat{a}\hat{\alpha}}$
Examinees	100	.411	-.069	.059	.011	-.222
	500	.234	-.023	.005	-.001	-.355
	1000	.173	-.008	.008	.000	-.443
	2000	.119	.000	.000	.001	-.471
Items	30	.235	-.007	.007	.001	-.334
	60	.233	-.043	.029	.004	-.411
Rho	-0.9	-.313	-.013	-.821	.000	-.525
	-0.6	-.197	-.009	-.898	-.002	-.475
	-0.3	.012	-.026	-.900	.006	-.312
	0.0	.360	-.020	.022	.007	-.012
	0.3	.545	-.029	.901	.002	-.294
	0.6	.600	-.036	.910	.007	-.472
	0.9	.632	-.042	.911	-.002	-.518
Total		.234	-.025	.018	.003	-.373

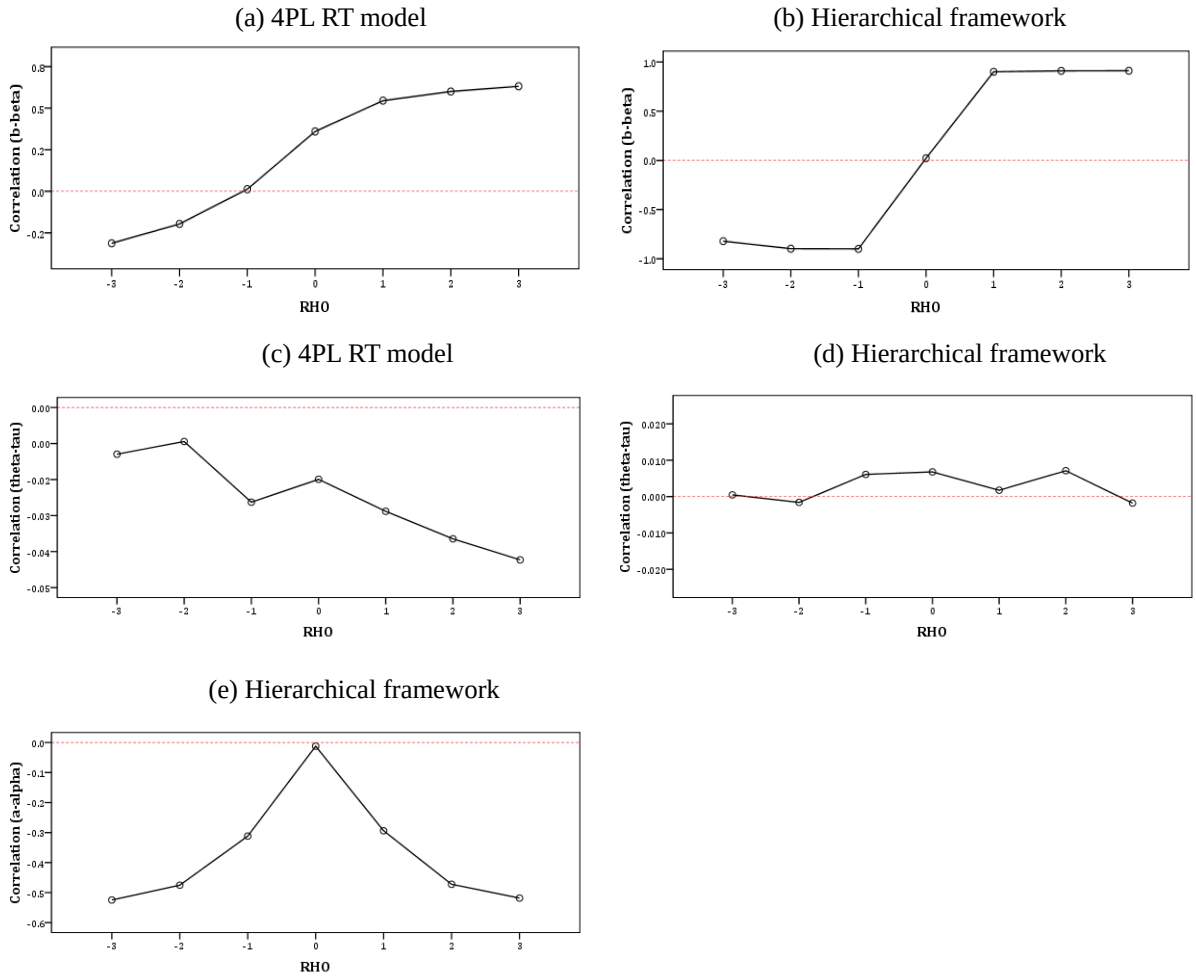


Figure 14. *Correlation between the item speed and item difficulty parameters (a), (b); correlation between the examinee speed and examinee ability parameters (c), (d);correlation between the response time discrimination and item discrimination parameters (e).*

Chapter 5. Discussion

The purpose of this study was to examine two different scoring models using response time data in conjunction with item response models. Two distinctive response time models, incorporated in IRT, were compared, and the relationships both between response time and item characteristics and response time and examinee ability were examined using the real and simulated data. The research questions addressed in this study are as follows:

1. Among the 4PL RT model, hierarchical framework, and Thissen's model, which model is the best method for scoring examinees' item responses when response time data are available in real data?

2. What are the relationships between the response time-related parameters (examinee and item speed, time intensity and time discrimination parameters) from different models that explain the speed-accuracy trade-off among item characteristics and examinee ability in item responses?

3. Which model is better to use for scoring examinees' responses with response time data under several conditions such as various numbers of examinees, different numbers of items, and different strengths of relationships among item characteristics and examinee ability?

These research questions will be discussed in order of overall results from the comparison of the response time models, the relationships in item and examinee parameter estimates, and the relationships between the response speed-related parameter estimates among the models. The discussion on these topics will be followed by the limitations of the study. As a conclusion, the implication of the study for educational practice and further research will be presented.

Model comparison in the real data study

Overall results of analysis

Study 1 examined the six response time models and the relationships between the response time models to the real data. All of the six models showed the evidence of the model convergence in the MCMC estimation method. Graphical diagnostics as well as the Gelman-Rubin ratios were applied and it was confirmed that the models were converged after thousands of burn-in iterations. The estimated item parameters and the examinee true ability parameter showed high correlations across the models. When the estimates from the response time models were compared to the ones from typical IRT methods, they did not show much difference. Overall, hierarchical framework showed the best model goodness of fit to the given data by showing the lowest DIC values among the six response time models. The hierarchical framework also showed the highest correlations both between the item speed and item difficulty parameter estimates and the examinee speed and examinee true ability parameter estimates.

The 4PL RT and Thissen's models also showed high correlations in the item and the examinee true ability parameter estimates with the 2PL and the 3PL IRT estimates. However, somewhat vague estimates for the response time related parameters were examined. The 4PL RT model showed a different direction from the results of hierarchical framework and Thissen's model in both of the item speed and the examinee speed parameters. Item response time-related parameters were further examined through investigating the relationship between response times and item parameters. It was obvious that those two models did not clearly reflect both relationships between response time and item difficulty and response time and examinee ability in the models. Thissen's model showed better results than the 4PL RT model. These results might be

related to the sample size of the items; the real data had only 33 items and there were other compounding source. For example, the items were relatively easy; overall 62% of the examinees responded correctly to the given items, and the mean values of the item difficulty parameter estimates from the 2PL and 3PL IRT models were $-.856$ and $-.007$ respectively.

Response time-related parameter estimation

Different results were shown in the relationships among the response time-related parameter estimates across the models. The hierarchical framework indicated negative relationships in both the item speed and item difficulty and the examinee speed and examinee ability. Both 4PL RT and Thissen's models showed that there was a positive relationship in the item difficulty and item speed. Thissen's model and the hierarchical framework showed the same direction of the relationship between examinee ability and speed; however, the 4PL RT model indicated the opposite direction.

There were two sources of compounding results from the analysis of these parameter estimates in the response time models. First, the response time models in this study did not have the same response speed-related parameters in the models. As discussed in the previous chapter, all of the response time models in this study have the item and examinee response speed parameters that explain both relationships between item difficulty and item speed and examinee true ability and examinee response speed. However, the 4PL RT and Thissen's models have slowness parameters while the hierarchical framework has a time intensity and an examinee speed parameter. The interpretations of these parameters are different depending on the model unless the indicators of the parameters were changed initially by the researcher. Some of the researchers

modified the original models to evaluate on the same ground. Because of these different interpretations of the parameters in the model, the relationships between response time-related parameters and the item and examinee parameters also differ across the models.

The other compounding source of the analysis is the rho parameter (ρ), a general indicator found in Thissen's model. It is a regression coefficient of the lognormal response time on the 2PL IRT structure $(a_j(\theta_i - b_j))$. If it is a positive value, it implies that response latency is increased as item difficulty increases or examinee ability decrease. A negative value implies a reversal relationship between IRT structure and response latency. Thissen's model showed the complicated results because this model had two different measures of the response speed-related parameters as well as the ρ parameter in the same model. An examinee slowness parameter (τ_i) and an item slowness parameter (β_j) also reflect the relationships among response time-related parameters. Therefore, the relationships captured by correlations between the IRT parameter estimates ($\hat{b}$ and $\hat{\theta}$) and the response speed parameter estimates ($\hat{\beta}$ and $\hat{\tau}$) may not always indicate the same direction as the overall relationship ($\hat{\rho}$) indicates. In Thissen's model, the examinee and item slowness parameter estimates should be interpreted as ones after taking into account the overall relationship of the response time and IRT structure.

Model comparison in the simulated data study

Overall results of analysis

In Study 2, the 4PL RT model and hierarchical framework were applied to the simulated data. The factors of the study included four sample sizes (100, 500, 1,000, and 2,000 examinees), two test lengths (30 and 60 items), and seven different conditions of the relationship

($\rho = -0.9, -0.6, -0.3, 0.0, 0.3, 0.6, \text{and } 0.9$) between the 2PL IRT structure and response time from Thissen's model. In order to avoid potential compounding effects from the response time model, both correlations between response time and examine slowness and response time and item slowness in Thissen's model are set as 0.0. Therefore the ρ parameter in Thissen's model is the only source to regulate the relationships between item difficulty and item speed and examinee true ability and examinee response speed in the generated response time.

Obtained DIC values were compared to examine the model goodness of fit between two response time models. Overall, the hierarchical framework showed lower values than the 4PL RT model consistently throughout marginal conditions. When DIC values for the response data were focused, the 4PL RT model showed comparable results. DIC values for the response time data in the 4PL RT models were much higher than DIC values in the hierarchical framework. Considering model specification procedures in the 4PL RT model, it is natural to have such results. Because the 4PL RT model does not have a response time distribution in it, a lognormal distribution was assigned for the model specification purpose. Implementation of the model in WinBUGS program was impossible without assuming a response time data distribution in the model specification procedure. Therefore, comparisons in DIC values for selecting the better model need to be focused on the response model alone. Although the hierarchical framework showed lower values in overall, the 4PL RT showed comparable results in the response model.

Item and examinee true ability parameters recovery

The item and examinee parameter estimates from the two response time models were compared through examining bias, RMSE and relative efficiency values. The estimates from a

typical 3PL IRT model were also compared with those of the two response time models to measure improvement from the estimation without considering response time data. In the analysis of the recovery of the item parameters, the 4PL RT model and hierarchical framework showed better results: lower bias in absolute terms and lower RMSE values across marginal conditions. Relative efficiency values also indicated that the two response time models were more efficient than the 3PL IRT model. The differences in bias and RMSE values from the 3PL IRT model showed statistical significance through a three-way MANOVA and a post hoc comparison. Thus, it can be stated that recoveries of the item parameters were better when the response times were considered in the estimation procedures.

In the examinee true ability parameter recovery analysis, the hierarchical framework showed better mean bias and lower mean RMSE values than the 4PL RT and 3PL IRT models. However, when relative efficiency values were applied, the 3PL IRT model was shown more efficient than the other two response time models. Results from the three-way MANOVA on the three measured criteria (bias, RMSE, and MSE) confirmed that the hierarchical framework was the best recovered model for the examinee true ability parameter. The differences between the hierarchical framework and the 3PL IRT model in the three measured criteria were statistically significant.

The examinee parameter recovery was further analyzed through an examinee ability group investigation. Examinees were categorized by 10 groups based on the examinee true ability parameter. Graphical analyses indicated that the 3PL IRT model showed better results in bias and RMSE values. Lower and higher ability groups indicated smaller RMSE values for the 3PL IRT model. The hierarchical framework showed lower RMSE values in the middle ability groups; however, the 4PL RT model showed comparable results. The RMSE values in 10 θ groups were

examined through a three-way MANOVA and a post hoc comparison. The 3PL IRT model showed significant differences against the hierarchical framework in $\theta 9$ and $\theta 10$ groups. Against the 4PL RT model, it showed significant differences in $\theta 1$ through $\theta 3$ as well as $\theta 9$ and $\theta 10$ groups. The hierarchical framework showed significant difference against the 3PL IRT model only in the $\theta 4$ group. The 4PL RT model did not show statistically significant difference against the other models in the examinee ability group analysis. Considering the relatively smaller numbers of the examinees in the lower and higher ability groups, it might be related to the difference of the estimation methods. The examinee true ability parameter in the 3PL IRT model was estimated by the Bayes expected a posteriori (EAP) method in BILOG-MG program. The EAP estimation procedures of the examinee true ability are explained in Baker and Kim (2004) in detail. The results of the examinee true ability parameter recovery in this study are not unique from the previous studies. Baker and Kim (2004) also mentioned that the EAP estimation of the examinee true ability consistently yielded lower RMSE and better bias values than Gibbs sampler.

The Pearson product-moment correlation coefficients were examined to compare parameter estimates in the two response time models. The mean correlations between the item parameters and estimates in various conditions across 30 replications showed that overall correlations among the 3 models were high. It was also noted that the 3PL IRT model showed lower correlations in the 100 examinees condition comparing to the response time models. However, in other examinees conditions the results were comparable to those of the response time models. As described in the results of the examinee true ability parameter recovery, the comparable results were due to the difference of estimation methods.

Correlation between response time-related parameters

In order to examine the response time-related parameters between the response time models, the Pearson product moment correlations for the item and examinee parameters were examined. The means of the correlations from parameter estimates showed that the estimated item parameters from the two response time models are highly correlated. However, correlations for response time-related parameter estimates ($\hat{\beta}, \hat{\tau}$) indicated that there were no or weak relationships between two response time models. When correlations between the item difficulty parameter (b) and the item speed parameter (β) estimates were examined, two response time models showed contrasting results. Mean correlation between item difficulty and item speed in the hierarchical framework was 0.018, while a somewhat positive relationship was shown in the 4PL RT model ($r_{\hat{b}\hat{\beta}} = .234$). When the relationship between response time and the IRT structure was considered, the differences from the two models were clearly manifested. The 4PL RT models showed positive correlations even when there were no or negative relationships in true conditions. Hierarchical framework showed clear distinctions based on the direction of the ρ parameter. However, the magnitude of the ρ parameter was not reflected in the correlations between the item difficulty and item speed from the hierarchical framework, while differential magnitude was detected in the 4PL RT model.

Correlations between the examinees true ability (θ) and examinees speed parameter (τ) also showed a similar pattern across the two models. Overall, the mean correlation from the 4PL RT model indicated a somewhat negative relationship ($r_{\hat{\theta}\hat{\tau}} = -.025$) while the hierarchical framework showed almost no relationship ($r_{\hat{\theta}\hat{\tau}} = .003$). When the ρ parameter was considered, the 4PL RT model showed there were differences in the relationship between examinee true ability

and speed; however, the hierarchical framework consistently showed almost no relationships.

Another interesting result was observed in the relationship between the item discrimination (*a*) and response time discrimination (α) parameters in the hierarchical framework. The response time discrimination parameter is a unique among the response time models in this study and the correlation between the item discrimination indicated a negative relationship. As the ρ parameter increases from 0.0 to any direction, the strength of the correlations between the item discrimination and item response time discrimination decreases. Considering that the item discrimination is always affected whenever the item characteristics or the examinee true ability are impacted by other compounding sources of the test (e.g., speededness, different pacing, or change of test taking strategies), this relationship is quite reasonable in practical situations.

Relationship between response time models

The response time models examined in this study showed similar results for the item and examinee true ability parameter estimates. However, there were also several differences in speed-related parameter estimates and the direction of the relationships that parameters captured in the models. Overall, the 4PL RT and Thissen's models showed inconsistent results in Study 1. Thissen's model showed somewhat equivocal results in explaining both relationships between the item difficulty and item speed and the examinee ability and examinee speed. More elaborate explanations are needed in interpreting these relationships, because there are two sources that explain the relationships among the related parameters. One solution to resolve this complication is explaining item and examinee slowness parameters as unique speed parameters after taking into account overall correlation among all the related parameters. Although it is a possible solution in

conceptual term, however, there is still an unresolved problem of the interpretation for practical testing situations. Thissen (1983) also noticed there was an unresolved ambiguity in the relationship of the related parameters. He suggested that the analysis of the two-dimensional response space was required to relocate these complex relationships in this model.

The 4PL RT model showed the opposite directions in explaining the relationships between the examinee ability and examinee speed parameter estimates as well as the item difficulty and item speed in the real data in Study 1. The 4PL RT model also showed somewhat different mechanisms in reflecting item and examinee speed in the model. In Study 2, the examinee ability parameter estimates ($\hat{\theta}$) from the 4PL RT model were shown to be affected by the direction and magnitude of the ρ parameter. The correlation between the examinee true ability parameter and the estimates from the 4PL RT model indicated that the ρ parameter affected the estimation of the examinee true ability. The same patterns were shown in both correlations ($r_{\hat{b}\hat{\beta}}$ and $r_{\hat{\theta}\hat{\tau}}$) between the IRT parameter and the speed parameter estimates.

It is obvious that the hierarchical framework showed clearer relationship with Thissen's model. The generated response time data were analyzed almost precisely in the hierarchical framework. When the relationships of the IRT parameter and the speed-related parameter estimates were examined in the levels of the ρ parameter, the hierarchical framework clearly differentiated the direction of the relationships. However, the magnitude of the ρ parameter was not reflected in the correlations, while differential magnitudes were detected in the 4PL RT model. It seemed that the ρ parameter in Thissen's model affected only the direction of the overall relationship in the hierarchical framework. If true item slowness (β) or true examinee slowness (τ) parameters were considered when generating response time data, it might have shown clearer

magnitude in the correlation analyses.

Limitations of the study and further research questions

There are several limitations and a number of issues for future studies. First, the 4PL RT model assumed that the item response time was independent of the examinee true ability parameter in the model. It was a necessary assumption for the EM algorithm to calibrate the item parameters in Wang and Hanson's (2005) study. Therefore, the assignment of the lognormal distribution for the response time data in the model specification was a somewhat arbitrary decision in this study. However, the model specification in WinBUGS was not available without giving a distribution information in the response time data; it was also a necessary step in this study. It is recommended to compare alternative response time scoring models that have response time modeling. For example, Ingrisone (2008a) introduced a joint distribution of Rasch model and a response time modeling with a 2PL Weibull distribution. This model showed improvement from the 4PL RT model in Wang and Hanson (2005). A marginal maximum likelihood estimation (MMLE) and a maximum a posteriori (MAP) procedures showed that item and examinee parameters recovered quite well in this model.

Second, DIC for the model fit index in this study has shown inconsistent results throughout the different conditions. Especially, it is reported that DIC tended to select a more complex model in the model fit studies (Kang & Cohen, 2007; Li, Cohen, Kim & Cho, 2009). It is recommended to use several other model fit indices to select the best model such as Akaike's information criteria (AIC), Bayesian information criteria (BIC), pseudo Bayes factor (PsBF), posterior model checks (PPMC) and cross validation loglikelihood (CVLL). Some model fit studies have shown the

application of these indices to various item response models, however, these indices have not applied to the item response time. Therefore, the applications of these indices in the response time models need to be further studied.

Third, this study did not consider the content area of the test subject. It was impossible to consider the content area in this study because the item content was not accessible to the researcher. The analysis of the relationships between cognitive complexity and the response time data has been stressed in test validity studies (e.g., Zenisky & Baldwin, 2006). Therefore, further research on the relationship between the difference of the cognitive area coverage and the response time through the analysis of the contents of test items is also recommended. Application of multidimensional item response theory (MIRT) is deemed a well suited method for this area.

Conclusion

In this study, Bayesian estimation using the MCMC method was applied to compare the response time models in the real data as well as the simulated data. Of the response time models investigated in the current study, the hierarchical framework yielded the best result among the response time models. Different response time models were examined through investigating the relationships between the item response theory parameters and the speed related parameters across various different conditions. Although there were several practical issues to the current study, there has been no comparison study among the response time models in the real data as well as the simulated one. Thus, the estimation and the comparison among the response time models in this study makes a unique contribution to the field of educational measurement, especially in the computer based tests utilizing the item response times. It is hoped that the response time models

are explored further and they will contribute to examining human behavior in test taking situations in more detail. The test validation and fairness issues also can be addressed through further examinations of the response time models in the area.

References

Agresti, A. (2002). *Categorical data analysis.* New York, NY: John Wiley .

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

Albert, J., & Ghosh, M. (2000). Item response modeling. In D. K. Dey, S. Ghosh, & B. Mallick (Eds.), *Generalized linear models: A Bayesian perspective* (pp. 173-193). New York: Marcel-Dekker.

Anastasi, A**. (**1976). *Psychological testing*. New York, NY: Macmillan.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26*, 323-336.

Baker, F. B., & Kim, S-H. (2004). Item *response theory: Parameter estimation techniques (2nd ed.).* New York: Marcel Dekker.

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. *Journal of Educational Psychology, 32,* 285-296.

Bejar, I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bergstrom, B., Gershon, R. & Lunn, M. (1994) *Computerized adaptive testing exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Best, N., Cowles, M. K., & Vines, K. (1996). *CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.30*. Cambridge, UK: MRC Biostatistics Unit.

Bontempo, B. D., & Julian, E. R. (1997, March). *Assessing speededness in variable-length computer adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Bridgeman, B. (2000). *Fairness in computer based testing: what we know and what we need to know.* (The GRE FAME Report). Princeton, NJ: Educational Testing Service.

Bridgeman, B. (2004). *Speededness as a threat to construct validity*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 137-148.

Bridgeman, B., Cline, F. & Hessinger, J. (2003). *Effect of extra time on GRE Quantitative and verbal scores.* (GRE No. 00-03P). Princeton, NJ: Educational Testing Service.

Bridges, K. R. (1985). Test-completion speed: Its relationship to performance on three course based objective examination, *Educational Psychological Measurement, 45,* 29-35.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281 -302.

Chang, S. (2006). *Computerized adaptive test item response times for correct and incorrect pretest and operational items: Testing fairness and test-taking strategies.* Unpublished doctoral dissertation, University of Nebraska, Lincoln.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A Practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30,* 295-311.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9,* 123-131.

Foos, P. W. (1989). Completion time and performance on multiple-choice and essay tests. *Bulletin of the Psychometric Society, 27,* 179-180.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software, 20(7),*1-14.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis (2nd ed.).*New York: Chapman & Hall/CRC.

Gewke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (vol. 4, pp. 169-193). Oxford, UK: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* New York: Chapman & Hall/CRC

Gitomer, D. H., Curtis, M. E., Glaser, R., & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. *Journal of Educational Psychology, 79,* 212-219.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217-233.

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associatiates.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12,* 253-261.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hoijtink, H. (2001). Conditional independence and differential item functioning in the twoparameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory.* New York, NY: Springer-Verlag.

Hoijtink, H., & Molenar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-189.

Hornke, L. (2000). Item response time in computerized adaptive testing. *Psicológica*, *21,* 175-189.

Ingrisone, J. N., II. (2008a). *Modeling the joint distribution of response accuracy and response time.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Ingrisone, S. J. (2008b). *An extended item response theory model incorporating item response time.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31,* 331-358.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163-176.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26,* 38-51.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33,* 353-373.

Lu, Y. & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26,* 29–97.

Luce, R. D. (1986). *Response times: their role in inferring elementary mental organization.* NY: Oxford University Press.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York, NY: Springer.

Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17,* 347-352.

Messick, S. (1981). Evidence and Ethics in the Evaluation of Tests. *Educational Researcher, 10,* 9-20.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27,* 289-298.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200-219.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342-366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Plake, B. (1999). *A new breed of CATS: Innovations in computerized adaptive testing*. Paper published by the University of Nebraska, Lincoln.

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

Raftery, A. E., & Lewis, S. M. (1996). Implementing MCMC. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 115-130). London: Chapman & Hall.

Rammsayer, T. (2004). Response times as a function of correct and incorrect answers in two psychophysical discrimination tasks. Retrieved October 30, 2004, from http://www.twk.tuebingen.mpg.de/twk02/Pkognitiv.html

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16,* 261-270.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151-171). Amsterdam: North-Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New

York: Springer.

Rubin, D. B., & Stern, H. S. (1994), Testing in latent class models using a posterior predictive check distribution. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 420-438). Thousand Oaks, CA: Sage Publications.

Schaeffer, G. , Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-bases GRE general test.* (Report No. ETS-RR-93-07). Princeton, NJ: Educational Testing Service.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19,* 18-38.

Scheines, R., Boomsma, A., & Hoijtink, H. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64*, 37-52.

Schmidt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on SAT, *Journal of Educational Measurement, 27,* 67-81.

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times.* Unpublished doctoral dissertation, Johns Hopkins University.

Schnipke, D. L., & Pashely, P. J. (1997). *Assessing subgroup differences in item response times.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Schnipke, D. L., & Scrams, D. J. (1997). *Representing response time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (1999). *Response-time feedback on computer administered tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Scrams, D. J., & Schnipke, D. L. (1997). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B, 13*, 238-241.

Sinharay, S. (2005). *Bayesian item fit analysis for unidimensional item response models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models*. (Report No. RR-03-28). Princeton, NJ: Educational Testing Service.

Sinharay, S., Johnson, M.S., & Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298-321.

Smith, R. (2000). *An exploratory analysis of item parameters and characteristics that influence item response time.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Smith, B. J. (2007). boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software, 21,* 1-37.

Spearman, C. (1927). *The abilities of man.* New York, NY: Macmillan.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS 1.4* user manual* [Computer program]. Cambridge, UK: MRC Biostatistics Unit.

Swygert, K. A. (1998). *An examination of item response times on the GRE-CAT.* Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236-256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 73,* 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of*

Educational Measurement, 46. In press.

van Onna, M. J. H. (2003). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67,* 519-538.

Verhelst, N. D., Verstraalen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York: Springer.

Wang, T. & Hanson, B. A (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29,* 323-339.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Zenisky, A. L., & Baldwin, P. (2007). *Using response time data in test development and validation: Research with beginning computer users.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International.

Appendix A

Table A1.

Generated item parameters in the 30 items and the 60 items

	Item Parameters (30 items)			Item Parameters (60 items)		
Item	a	b	c	a	b	C
1	0.886	0.062	0.155	0.886	0.062	0.155
2	0.528	1.226	0.214	0.528	1.226	0.214
3	0.685	0.509	0.181	0.685	0.509	0.181
4	1.146	-0.644	0.260	1.146	-0.644	0.260
5	0.752	0.229	0.290	0.752	0.229	0.290
6	2.057	0.687	0.218	2.057	0.687	0.218
7	1.219	-0.901	0.195	1.219	-0.901	0.195
8	1.400	1.613	0.133	1.400	1.613	0.133
9	0.954	0.879	0.209	0.954	0.879	0.209
10	0.877	-1.074	0.234	0.877	-1.074	0.234
11	0.979	-0.175	0.407	0.979	-0.175	0.407
12	0.696	-0.592	0.281	0.696	-0.592	0.281
13	0.719	0.470	0.161	0.719	0.470	0.161
14	1.339	1.012	0.088	1.339	1.012	0.088
15	1.182	0.560	0.143	1.182	0.560	0.143
16	1.731	-1.510	0.265	1.731	-1.510	0.265
17	0.902	-0.573	0.155	0.902	-0.573	0.155
18	1.754	1.116	0.114	1.754	1.116	0.114
19	1.029	1.587	0.290	1.029	1.587	0.290
20	1.399	-0.397	0.099	1.399	-0.397	0.099
21	1.021	0.859	0.277	1.021	0.859	0.277
22	0.628	0.339	0.199	0.628	0.339	0.199
23	0.698	-0.642	0.306	0.698	-0.642	0.306
24	0.564	-0.028	0.133	0.564	-0.028	0.133
25	1.256	-1.607	0.069	1.256	-1.607	0.069
26	0.780	-0.198	0.160	0.780	-0.198	0.160
27	1.016	0.549	0.270	1.016	0.549	0.270
28	0.999	-0.144	0.150	0.999	-0.144	0.150
29	0.995	0.447	0.091	0.995	0.447	0.091
30	0.788	0.354	0.286	0.788	0.354	0.286
31				2.129	1.420	0.125
32				1.463	-2.289	0.112
33				1.555	-0.762	0.216
34				1.594	0.491	0.306
35				2.341	-2.120	0.246
36				1.786	1.052	0.350
37				1.496	-0.255	0.244
38				0.906	1.248	0.345
39				0.686	-0.294	0.186
40				1.122	0.835	0.261
41				0.931	0.340	0.225
42				0.963	1.156	0.184
43				2.031	1.287	0.388
44				1.099	-1.063	0.189
45				0.679	0.970	0.244
46				2.409	-0.609	0.313
47				0.886	0.308	0.349
48				0.718	-0.091	0.063
49				1.476	0.166	0.306
50				0.654	0.097	0.258
51				1.213	-1.278	0.409
52				1.296	0.169	0.128
53				0.878	-0.632	0.062
54				0.701	-2.218	0.250
55				0.871	-1.864	0.210
56				2.332	-0.480	0.172
57				1.173	0.997	0.401
58				0.676	-1.603	0.341
59				1.126	0.211	0.171
60				1.181	-0.137	0.202
Mean	1.033	0.134	0.201	1.156	-1.016	0.221

Table A2.

Average Gelman-Rubin diagnostics for item parameter estimates in 4PL RT models

	4PL RT (2PL)			4PL RT (3PL)			
Item	a	b	beta	a	b	c	beta
1	1.012	0.994	0.983	0.993	0.992	1.001	1.006
2	0.999	0.994	1.009	0.992	0.960	1.002	0.993
3	0.978	0.981	1.009	0.994	1.070	1.000	0.986
4	1.006	1.004	0.977	0.994	1.025	1.020	1.005
5	1.015	1.008	1.001	1.006	0.967	0.991	1.009
6	0.992	1.003	1.000	1.005	1.000	0.990	1.002
7	1.001	1.014	1.002	1.014	1.029	1.062	1.007
8	1.001	1.006	1.001	1.067	1.051	1.031	1.000
9	0.992	0.996	0.988	0.996	0.990	1.000	0.991
10	0.995	1.001	0.999	0.973	0.920	1.019	0.995
11	0.997	0.993	0.986	1.004	1.078	1.014	0.994
12	1.003	0.997	1.006	1.001	0.993	1.064	1.008
13	0.983	0.989	1.021	0.990	0.930	0.999	1.011
14	1.009	1.013	0.993	0.992	1.070	0.998	0.988
15	1.006	0.994	1.019	1.007	1.043	1.016	0.998
16	1.023	1.005	0.999	0.999	0.983	0.986	0.998
17	1.001	1.010	1.005	1.002	1.048	1.010	1.008
18	1.002	0.988	1.015	1.021	1.058	1.022	1.011
19	0.996	1.002	1.017	0.970	0.987	0.984	1.004
20	0.996	0.999	0.998	0.982	1.042	1.014	0.987
21	1.004	0.998	1.002	1.012	1.107	1.035	1.002
22	1.002	0.992	0.997	1.003	0.966	1.010	1.009
23	1.019	0.990	0.987	0.988	1.016	1.004	0.992
24	0.985	1.018	0.993	1.008	1.018	1.017	0.944
25	1.006	0.990	0.999	0.998	1.028	1.006	1.018
26	1.014	1.021	1.024	1.003	1.147	1.026	0.992
27	1.024	1.003	0.970	1.008	1.045	0.992	0.925
28	0.987	0.982	0.998	1.000	0.909	1.026	0.991
29	0.999	1.009	1.001	1.008	1.005	1.003	1.013
30	0.993	0.983	1.021	0.992	1.068	0.997	0.992
31	1.007	0.998	1.002	1.007	1.074	1.013	0.996
32	0.997	0.980	1.014	1.016	0.949	1.003	1.000
33	1.007	1.013	1.003	1.051	1.149	1.070	1.029
Total	1.002	0.999	1.001	1.003	1.022	1.013	0.997

Note. The Gelman-Rubin ratios are calculated from the 2,000 post burn-in iterations after the 8,000 iterations of burn-in were discarded.

Table A3.

Average Gelman-Rubin diagnostics for item parameter estimates in Hierarchical framework

	Hierarchical Framework (2PL)				Hierarchical Framework (3PL)				
Item	a	b	alpha	beta	a	b	c	alpha	beta
1	0.995	0.973	1.004	1.014	1.005	1.009	0.995	0.979	1.015
2	0.989	0.980	0.999	0.994	0.996	1.023	0.973	0.980	0.995
3	0.995	0.998	0.980	0.993	0.997	1.051	1.087	1.011	0.976
4	1.030	1.021	1.010	1.011	1.040	1.070	1.092	1.001	1.002
5	0.982	0.981	1.010	0.988	0.991	1.007	1.012	0.999	0.993
6	0.977	1.015	1.004	0.989	0.998	1.106	1.014	0.999	0.989
7	1.009	0.985	0.981	0.999	0.993	1.026	0.999	1.031	0.999
8	1.020	0.994	0.992	0.973	1.002	0.952	0.975	1.000	0.970
9	0.981	1.011	0.994	1.010	0.995	1.009	0.999	0.994	0.993
10	0.992	0.977	0.989	0.991	1.010	0.991	0.998	1.006	0.995
11	1.001	1.005	0.990	0.981	0.961	0.950	1.008	1.019	0.991
12	0.990	1.003	0.990	1.014	1.037	1.073	1.045	0.997	0.983
13	1.013	1.002	1.011	0.984	1.031	1.005	1.008	1.004	0.995
14	0.985	1.016	1.001	1.012	1.060	0.999	1.044	1.011	0.981
15	0.986	0.985	1.000	1.019	1.003	1.020	1.014	0.996	0.995
16	1.019	1.017	0.988	0.996	1.016	1.012	1.039	1.012	0.986
17	1.006	1.010	0.999	1.002	1.007	1.075	0.989	0.998	0.997
18	1.021	0.991	1.003	0.987	1.017	1.026	1.095	1.010	1.015
19	1.020	1.038	1.014	1.007	1.013	1.059	1.003	1.008	1.001
20	1.013	1.001	0.994	0.981	1.050	1.016	1.030	1.002	1.006
21	1.011	0.992	1.005	1.005	1.013	1.037	1.034	1.002	0.998
22	0.984	0.997	1.006	0.980	1.037	1.005	1.018	0.995	0.985
23	1.004	0.998	0.989	1.008	1.051	1.122	1.179	0.989	0.993
24	0.981	0.997	0.998	1.009	0.984	0.996	0.971	1.010	1.008
25	0.992	0.998	1.009	1.020	1.006	0.939	1.006	1.001	0.991
26	1.034	1.028	0.988	1.022	1.041	1.028	0.980	0.997	1.001
27	0.994	1.006	0.996	0.988	1.004	1.044	0.994	0.990	1.008
28	0.991	0.993	0.991	1.012	0.986	1.108	0.995	0.996	0.992
29	1.007	1.024	0.990	1.000	1.028	1.067	1.016	1.006	0.999
30	1.015	1.004	1.006	0.987	1.002	0.982	1.011	0.990	0.997
31	1.060	1.020	0.996	0.997	1.042	1.037	1.015	1.007	1.013
32	1.020	1.011	1.012	0.999	1.030	0.983	1.014	1.020	1.002
33	0.996	0.991	1.008	0.996	1.006	0.953	1.008	1.009	1.005
Total	1.003	1.002	0.998	0.999	1.014	1.024	1.020	1.002	0.996

Note. The Gelman-Rubin ratios are calculated from the 2,000 post burn-in iterations after the 8,000 iterations of burn-in were discarded.

Table A4.

Average Gelman-Rubin diagnostics for item parameter estimates in Thissen's lognormal response time models

	Thissen's lognormal RT (2PL)			Thissen's lognormal RT (3PL)			
Item	a	b	beta	a	b	c	beta
1	1.025	1.005	0.993	1.053	0.996	0.997	1.022
2	1.004	0.997	0.991	1.018	0.965	1.039	1.000
3	0.990	0.997	1.003	1.051	1.085	1.108	1.025
4	1.011	1.013	0.985	0.984	0.985	0.996	0.999
5	1.001	1.005	1.007	1.012	1.046	1.025	0.989
6	1.083	1.049	0.994	1.201	1.292	1.221	1.021
7	1.020	0.997	1.003	1.076	1.001	1.050	0.991
8	1.012	0.985	1.001	0.988	1.015	1.002	0.998
9	1.003	0.996	0.990	0.998	0.966	1.041	1.008
10	1.020	0.983	1.012	1.004	1.040	1.005	0.999
11	0.972	1.005	0.992	1.051	1.067	1.141	0.999
12	0.988	1.016	0.995	1.087	0.978	1.108	0.991
13	1.009	1.005	0.989	0.988	1.109	0.990	1.000
14	1.023	1.011	0.994	0.999	1.016	1.041	0.988
15	0.998	0.995	0.986	1.073	1.178	1.203	1.015
16	0.994	1.013	0.999	1.032	1.094	1.070	1.003
17	0.997	0.998	0.998	0.991	0.923	1.001	0.993
18	0.996	1.022	1.006	0.992	1.055	1.038	1.021
19	1.018	1.036	1.000	0.977	1.063	0.969	0.997
20	0.999	1.000	1.007	1.037	1.121	0.995	1.009
21	1.004	0.997	1.009	0.994	1.148	1.065	1.003
22	0.980	1.012	1.006	1.023	1.067	1.016	0.992
23	1.018	0.981	0.986	1.006	1.112	1.033	1.023
24	1.015	1.013	0.994	1.021	1.076	1.046	1.016
25	1.016	0.994	0.996	1.029	0.997	1.010	0.997
26	0.988	1.012	0.986	1.029	1.027	1.026	1.010
27	0.981	1.008	0.978	1.005	0.936	0.972	0.990
28	1.010	0.988	0.999	1.129	1.198	1.337	1.005
29	1.006	0.992	1.007	1.223	1.332	1.117	1.022
30	1.007	0.995	0.991	0.993	1.052	1.041	0.986
31	1.000	0.995	1.012	1.027	1.006	0.957	1.003
32	1.019	1.046	1.007	0.999	0.955	0.997	1.007
33	1.019	0.990	1.001	1.014	1.025	0.982	0.994
Total	1.007	1.005	0.997	1.033	1.058	1.050	1.004

Note. The Gelman-Rubin ratios are calculated from the 2,000 post burn-in iterations after the 8,000 iterations of burn-in were discarded.

Table A5.

Item parameter estimates in 4PL RT models

	4PL RT (2PL)			4PL RT (3PL)			
Item	a	b	beta	a	b	c	beta
1	0.703	-2.264	1.278	0.414	-1.689	0.262	1.770
2	0.486	0.715	1.970	0.537	1.367	0.215	2.610
3	0.605	-1.967	0.948	0.439	-0.929	0.306	1.453
4	0.410	-1.609	1.465	0.322	-0.153	0.287	1.958
5	0.832	-0.622	5.057	0.648	-0.184	0.177	6.151
6	0.451	-2.595	6.611	0.295	-1.279	0.306	6.692
7	0.461	-0.728	2.549	0.423	0.392	0.253	3.444
8	0.401	0.617	1.180	0.535	1.494	0.277	4.287
9	0.821	-0.334	6.935	0.606	0.058	0.142	7.419
10	1.025	-1.731	2.826	0.693	-1.298	0.274	7.422
11	0.733	-2.006	1.903	0.550	-0.987	0.390	7.325
12	0.310	0.223	6.448	0.319	1.292	0.207	6.506
13	1.166	-2.242	7.476	0.721	-1.943	0.187	8.272
14	0.431	1.360	1.169	0.591	1.787	0.234	4.497
15	0.708	-1.219	3.232	0.625	-0.338	0.297	6.315
16	0.538	-0.865	1.440	0.424	0.100	0.253	3.886
17	0.835	-0.072	2.293	0.724	0.362	0.201	6.356
18	0.426	-1.683	6.098	0.309	-0.597	0.227	7.257
19	0.753	0.196	2.442	0.607	0.521	0.146	3.724
20	1.167	-1.577	4.457	0.700	-1.301	0.186	4.091
21	0.945	-0.471	5.055	2.691	0.457	0.357	1.440
22	0.816	0.252	0.996	1.863	0.756	0.280	2.129
23	0.817	-2.105	5.412	0.545	-1.537	0.257	7.290
24	0.453	-0.272	0.225	1.208	1.002	0.391	0.296
25	0.481	0.137	7.584	0.389	0.774	0.158	7.358
26	0.850	-2.835	4.770	0.528	-2.283	0.280	5.667
27	0.936	0.585	0.379	0.740	0.847	0.126	0.974
28	0.690	-1.029	1.202	0.584	-0.192	0.264	1.539
29	1.171	-1.874	5.156	0.724	-1.589	0.181	5.572
30	0.856	-2.112	7.544	0.507	-1.862	0.148	7.518
31	1.108	-1.214	1.665	0.845	-0.742	0.294	6.745
32	0.650	-2.112	3.200	0.450	-1.203	0.327	6.275
33	0.689	0.069	0.495	0.620	0.686	0.207	0.702
Mean	0.719	-0.951	3.378	0.672	-0.249	0.245	4.695

Table A6.

Item parameter estimates in Hierarchical framework

	Hierarchical Framework (2PL)				Hierarchical Framework (3PL)				
Item	a	b	alpha	beta	a	b	c	alpha	beta
1	0.728	-2.173	2.222	3.842	0.421	-1.633	0.262	2.231	3.840
2	0.499	0.743	1.864	3.808	0.576	1.436	0.225	1.866	3.810
3	0.591	-1.986	2.279	3.747	0.437	-0.898	0.307	2.282	3.750
4	0.446	-1.336	0.690	2.761	0.357	-0.046	0.270	0.692	2.761
5	0.857	-0.451	1.810	3.497	0.647	0.014	0.175	1.806	3.498
6	0.422	-2.547	2.058	3.726	0.282	-0.880	0.349	2.052	3.724
7	0.475	-0.642	1.934	3.822	0.433	0.496	0.259	1.930	3.821
8	0.434	0.671	0.776	2.911	0.537	1.575	0.251	0.771	2.920
9	0.839	-0.133	1.469	3.843	0.673	0.339	0.176	1.471	3.845
10	1.072	-1.547	1.627	3.629	0.655	-1.178	0.224	1.623	3.631
11	0.766	-1.852	1.399	3.687	0.519	-1.037	0.303	1.400	3.689
12	0.335	0.496	1.314	3.425	0.479	1.698	0.277	1.316	3.428
13	1.239	-1.716	1.333	3.444	0.759	-1.376	0.238	1.336	3.446
14	0.451	1.437	0.564	2.837	0.617	1.973	0.224	0.562	2.841
15	0.735	-1.074	1.755	3.583	0.567	-0.326	0.246	1.743	3.588
16	0.546	-0.787	1.622	3.555	0.421	0.115	0.231	1.629	3.555
17	0.848	-0.006	1.827	3.667	0.653	0.406	0.155	1.823	3.669
18	0.430	-1.358	1.256	3.311	0.332	-0.022	0.266	1.263	3.314
19	0.822	0.451	0.756	2.576	0.656	0.796	0.138	0.753	2.575
20	1.206	-1.394	1.746	3.592	0.740	-1.077	0.203	1.748	3.597
21	0.943	-0.335	2.474	3.717	2.485	0.520	0.363	2.467	3.719
22	0.805	0.274	2.152	3.843	1.731	0.830	0.282	2.150	3.845
23	0.858	-1.806	1.779	3.570	0.547	-1.277	0.242	1.773	3.570
24	0.458	-0.237	0.608	2.701	1.137	1.062	0.389	0.606	2.701
25	0.486	0.340	1.870	3.817	0.429	1.041	0.184	1.872	3.820
26	0.860	-2.572	1.766	3.487	0.538	-1.942	0.305	1.765	3.489
27	0.917	0.595	1.566	3.405	0.729	0.863	0.119	1.565	3.408
28	0.733	-0.785	0.689	2.522	0.615	-0.017	0.252	0.691	2.523
29	1.276	-1.513	1.565	3.415	0.810	-1.175	0.218	1.570	3.418
30	0.945	-1.581	1.435	3.326	0.611	-1.125	0.217	1.437	3.326
31	1.150	-1.090	1.253	3.524	0.826	-0.622	0.229	1.250	3.525
32	0.696	-1.790	1.116	3.373	0.482	-0.964	0.278	1.119	3.372
33	0.720	0.111	0.663	2.953	0.673	0.752	0.219	0.665	2.955
Mean	0.745	-0.776	1.492	3.422	0.678	-0.051	0.245	1.492	3.423

Table A7.

Item parameter estimates in Thissen's lognormal response time models

	Thissen's lognormal RT (2PL)			Thissen's lognormal RT (3PL)			
Item	a	b	beta	a	b	c	beta
1	1.013	-1.665	0.797	0.545	-1.369	0.242	0.609
2	0.619	0.621	0.127	0.570	1.303	0.205	0.282
3	0.794	-1.543	0.564	0.541	-0.737	0.310	0.479
4	0.333	-1.740	-0.627	0.308	0.009	0.281	-0.598
5	0.868	-0.443	0.055	0.672	0.010	0.173	0.083
6	0.677	-1.688	0.513	0.373	-0.860	0.311	0.435
7	0.614	-0.506	0.354	0.494	0.500	0.268	0.409
8	0.280	0.965	-0.727	0.396	1.832	0.229	-0.590
9	0.881	-0.127	0.319	0.734	0.352	0.188	0.394
10	1.031	-1.586	0.573	0.716	-1.087	0.232	0.404
11	0.718	-1.934	0.552	0.518	-1.093	0.288	0.439
12	0.392	0.430	-0.187	0.477	1.718	0.282	-0.059
13	0.999	-1.990	0.496	0.738	-1.381	0.245	0.276
14	0.300	2.029	-0.902	0.543	2.298	0.220	-0.728
15	0.819	-0.975	0.270	0.605	-0.334	0.235	0.242
16	0.670	-0.654	0.130	0.464	0.062	0.221	0.158
17	0.866	-0.007	0.109	0.671	0.370	0.148	0.190
18	0.454	-1.266	-0.077	0.332	-0.072	0.259	-0.055
19	0.559	0.616	-1.088	0.542	0.900	0.134	-0.950
20	1.205	-1.385	0.547	0.843	-0.993	0.199	0.370
21	0.943	-0.339	0.256	1.597	0.456	0.330	0.320
22	0.838	0.267	0.214	1.252	0.819	0.258	0.342
23	0.846	-1.819	0.480	0.564	-1.265	0.236	0.327
24	0.236	-0.466	-0.829	0.613	1.307	0.342	-0.729
25	0.570	0.290	0.206	0.449	0.990	0.182	0.327
26	0.992	-2.278	0.620	0.625	-1.740	0.302	0.353
27	0.876	0.618	-0.320	0.722	0.847	0.118	-0.152
28	0.670	-0.850	-0.862	0.627	0.032	0.270	-0.841
29	1.050	-1.718	0.410	0.795	-1.157	0.236	0.225
30	0.846	-1.719	0.215	0.593	-1.143	0.225	0.076
31	0.807	-1.423	0.314	0.685	-0.727	0.225	0.235
32	0.636	-1.922	0.186	0.446	-1.081	0.267	0.097
33	0.392	0.164	-0.628	0.484	0.923	0.212	-0.519
Mean	0.721	-0.729	0.062	0.622	-0.009	0.238	0.056

Table A8.
Correlations between item discrimination parameter (a) estimates among the models

		4PL RT		Hierarchical framework		Thissen's model		IRT	
		2PL	3PL	2PL	3PL	2PL	3PL	2PL	3PL
4PL RT model	2PL								
	3PL	.356							
Hierarchical framework	2PL	.995	.317						
	3PL	.370	.995	.337					
Thissen's model	2PL	.837	.224	.812	.221				
	3PL	.584	.923	.545	.931	.528			
IRT model	2PL	.986	.309	.995	.334	.772	.530		
	3PL	.379	.961	.350	.971	.196	.905	.355	

Table A9.
Correlations between guessing parameter (c) estimates among the models

	4PL RT	Hierarchical framework	Thissen's model	IRT model
4PL RT				
Hierarchical framework	.848			
Thissen's model	.796	.971		
IRT model	.724	.874	.873	

Table A10.
Correlations between item speed parameter (beta) estimates among the models

		4PL RT		Hierarchical framework		Thissen's model	
		2PL	3PL	2PL	3PL	2PL	3PL
4PL RT model	2PL						
	3PL	.717					
Hierarchical framework	2PL	.336	.322				
	3PL	.336	.323	1.000			
Thissen's model	2PL	.413	.431	.882	.882		
	3PL	.391	.391	.962	.961	.977	

Table A11.
Correlations between examinee speed parameter (tau) estimates among the models

		4PL RT		Hierarchical framework		Thissen's model	
		2PL	3PL	2PL	3PL	2PL	3PL
4PL RT model	2PL						
	3PL	.815					
Hierarchical framework	2PL	.064	.182				
	3PL	.066	.184	.999			
Thissen's model	2PL	-.178	-.359	-.769	-.770		
	3PL	-.153	-.315	-.862	-.862	.980	

Table A12.

Bias for the examinee ability parameter based on the groups in the 3 models

		4PL RT										Hierarchical Framework										3PL IRT									
Ability group		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Examinees	100	-0.844	-0.503	-0.385	-0.207	-0.095	-0.002	0.096	0.177	0.294	0.000	-0.701	-0.457	-0.344	-0.170	-0.047	0.041	0.159	0.240	0.355	0.000	-0.517	-0.298	-0.218	-0.123	-0.099	-0.081	-0.060	-0.031	-0.003	0.000
	500	-0.510	-0.332	-0.195	-0.080	-0.017	0.057	0.107	0.170	0.224	0.441	-0.470	-0.305	-0.178	-0.074	-0.009	0.067	0.121	0.194	0.253	0.485	-0.366	-0.222	-0.118	-0.040	-0.014	0.029	0.062	0.098	0.137	0.353
	1000	-0.654	-0.316	-0.185	-0.081	-0.016	0.036	0.070	0.101	0.166	0.296	-0.571	-0.277	-0.166	-0.071	-0.009	0.042	0.074	0.108	0.178	0.316	-0.514	-0.228	-0.130	-0.058	-0.017	0.009	0.028	0.038	0.098	0.218
	2000	-0.571	-0.284	-0.128	-0.039	0.013	0.039	0.062	0.081	0.141	0.313	-0.487	-0.238	-0.105	-0.031	0.013	0.033	0.054	0.071	0.131	0.308	-0.444	-0.199	-0.077	-0.017	0.012	0.016	0.029	0.029	0.082	0.253
Items	30	-0.887	-0.512	-0.322	-0.141	-0.018	0.076	0.145	0.211	0.294	0.362	-0.742	-0.424	-0.269	-0.117	-0.007	0.074	0.138	0.202	0.290	0.354	-0.713	-0.396	-0.252	-0.115	-0.031	0.029	0.066	0.110	0.179	0.307
	60	-0.402	-0.205	-0.125	-0.062	-0.039	-0.011	0.022	0.054	0.118	0.164	-0.372	-0.214	-0.127	-0.056	-0.019	0.017	0.066	0.105	0.168	0.200	-0.208	-0.077	-0.020	-0.004	-0.028	-0.043	-0.037	-0.043	-0.022	0.105
Rho	-0.9	-0.802	-0.443	-0.271	-0.122	-0.034	0.044	0.107	0.162	0.260	0.300	-0.714	-0.403	-0.246	-0.107	-0.018	0.056	0.125	0.183	0.282	0.315										
	-0.6	-0.725	-0.401	-0.249	-0.115	-0.034	0.039	0.094	0.149	0.234	0.287	-0.637	-0.362	-0.224	-0.100	-0.018	0.051	0.112	0.171	0.256	0.302										
	-0.3	-0.650	-0.363	-0.227	-0.106	-0.033	0.034	0.081	0.134	0.209	0.269	-0.562	-0.324	-0.202	-0.090	-0.018	0.047	0.099	0.155	0.232	0.284										
	0.0	-0.599	-0.338	-0.211	-0.096	-0.027	0.029	0.077	0.125	0.193	0.255	-0.511	-0.298	-0.185	-0.080	-0.011	0.042	0.095	0.147	0.216	0.269										
	0.3	-0.584	-0.327	-0.204	-0.094	-0.027	0.025	0.074	0.119	0.185	0.248	-0.497	-0.287	-0.179	-0.079	-0.011	0.038	0.092	0.140	0.207	0.262										
	0.6	-0.577	-0.320	-0.200	-0.090	-0.024	0.031	0.076	0.119	0.184	0.242	-0.490	-0.281	-0.175	-0.075	-0.009	0.044	0.094	0.140	0.206	0.256										
	0.9	-0.577	-0.319	-0.200	-0.090	-0.022	0.028	0.076	0.118	0.181	0.237	-0.489	-0.279	-0.175	-0.074	-0.006	0.041	0.094	0.139	0.203	0.252										
Total		-0.645	-0.359	-0.223	-0.102	-0.029	0.033	0.084	0.132	0.206	0.263	-0.557	-0.319	-0.198	-0.086	-0.013	0.046	0.102	0.153	0.229	0.277	-0.460	-0.237	-0.136	-0.059	-0.030	-0.007	0.014	0.033	0.079	0.206

Table A13.

RMSE for the examinee ability parameter based on the groups in the 3 models

		4PL RT										Hierarchical Framework										3PL IRT									
Ability group		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Examinees	100	0.852	0.621	0.547	0.457	0.389	0.387	0.376	0.421	0.460	0.000	0.708	0.569	0.497	0.412	0.370	0.361	0.374	0.433	0.484	0.000	0.601	0.496	0.485	0.451	0.428	0.431	0.424	0.441	0.398	0.000
	500	0.640	0.533	0.460	0.423	0.420	0.413	0.402	0.413	0.429	0.567	0.592	0.499	0.429	0.398	0.413	0.398	0.387	0.407	0.431	0.595	0.547	0.485	0.442	0.420	0.409	0.394	0.395	0.396	0.396	0.496
	1000	0.793	0.541	0.459	0.430	0.423	0.421	0.409	0.404	0.412	0.457	0.707	0.500	0.435	0.413	0.414	0.412	0.391	0.391	0.403	0.459	0.677	0.494	0.446	0.422	0.409	0.397	0.400	0.392	0.392	0.407
	2000	0.724	0.530	0.460	0.434	0.435	0.439	0.414	0.411	0.417	0.497	0.641	0.486	0.439	0.420	0.426	0.423	0.401	0.397	0.400	0.484	0.617	0.484	0.446	0.425	0.413	0.402	0.406	0.398	0.401	0.452
Items	30	0.958	0.669	0.557	0.490	0.462	0.470	0.471	0.487	0.492	0.461	0.816	0.594	0.513	0.464	0.454	0.455	0.455	0.471	0.481	0.451	0.788	0.581	0.517	0.476	0.458	0.456	0.456	0.462	0.430	0.414
	60	0.547	0.443	0.406	0.382	0.372	0.360	0.330	0.338	0.367	0.299	0.508	0.433	0.387	0.357	0.358	0.342	0.321	0.343	0.378	0.318	0.433	0.399	0.392	0.383	0.371	0.356	0.356	0.351	0.363	0.264
Rho	-0.9	0.906	0.629	0.520	0.462	0.429	0.427	0.418	0.436	0.463	0.412	0.815	0.587	0.489	0.437	0.418	0.410	0.405	0.431	0.463	0.416										
	-0.6	0.830	0.593	0.499	0.445	0.422	0.417	0.407	0.424	0.447	0.402	0.739	0.551	0.468	0.420	0.410	0.401	0.395	0.418	0.447	0.406										
	-0.3	0.755	0.558	0.482	0.434	0.411	0.421	0.399	0.411	0.430	0.386	0.665	0.515	0.450	0.408	0.400	0.405	0.386	0.406	0.430	0.391										
	0.0	0.707	0.536	0.470	0.428	0.417	0.409	0.396	0.406	0.421	0.373	0.616	0.494	0.438	0.403	0.406	0.393	0.384	0.401	0.421	0.377										
	0.3	0.694	0.529	0.467	0.428	0.416	0.403	0.394	0.402	0.415	0.367	0.604	0.487	0.436	0.402	0.405	0.386	0.381	0.397	0.415	0.372										
	0.6	0.687	0.524	0.465	0.428	0.410	0.421	0.396	0.404	0.416	0.363	0.597	0.481	0.434	0.403	0.399	0.405	0.384	0.398	0.416	0.367										
	0.9	0.688	0.523	0.467	0.427	0.413	0.405	0.393	0.403	0.414	0.359	0.597	0.481	0.435	0.402	0.402	0.389	0.381	0.397	0.414	0.363										
Total		0.752	0.556	0.481	0.436	0.417	0.415	0.400	0.412	0.430	0.380	0.662	0.514	0.450	0.411	0.406	0.399	0.388	0.407	0.429	0.385	0.611	0.490	0.455	0.430	0.414	0.406	0.406	0.407	0.396	0.339

Appendix B

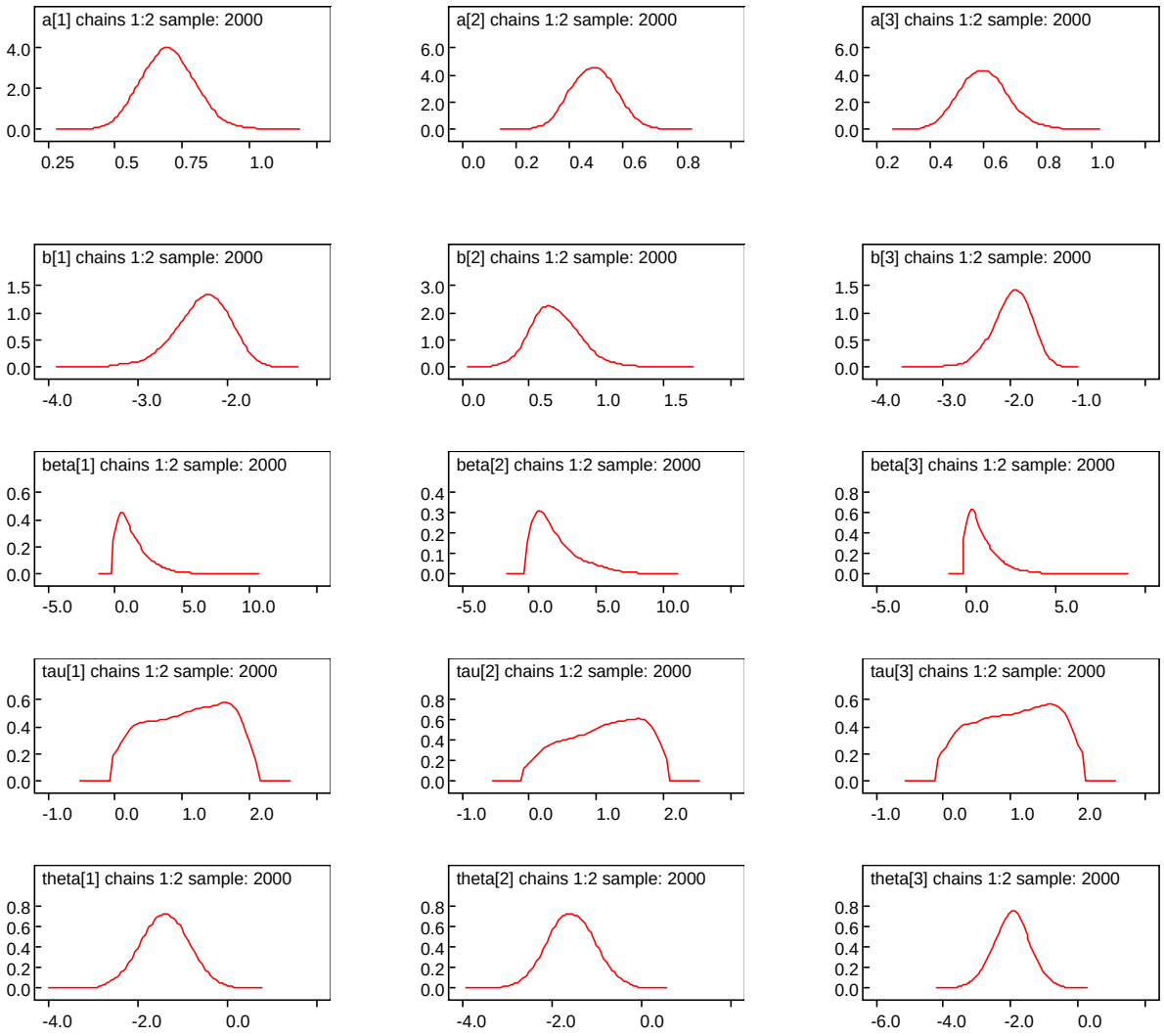


Figure B1. *Some of the representative item and examinee parameter estimates from the 4PL RT (2PL) model*

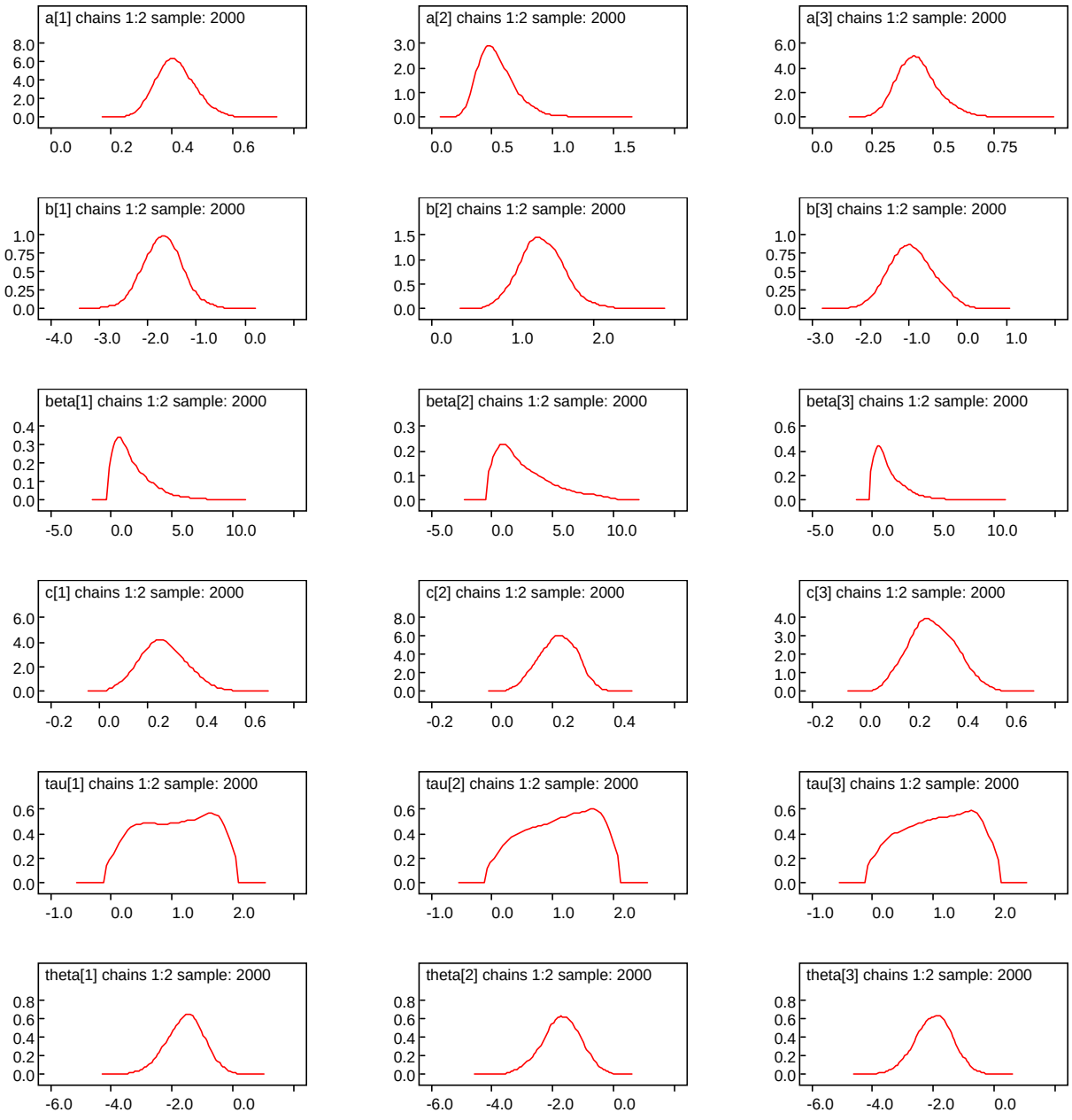


Figure B2. *Some of the representative item and examinee parameter estimates from the 4PL RT (3PL) model*

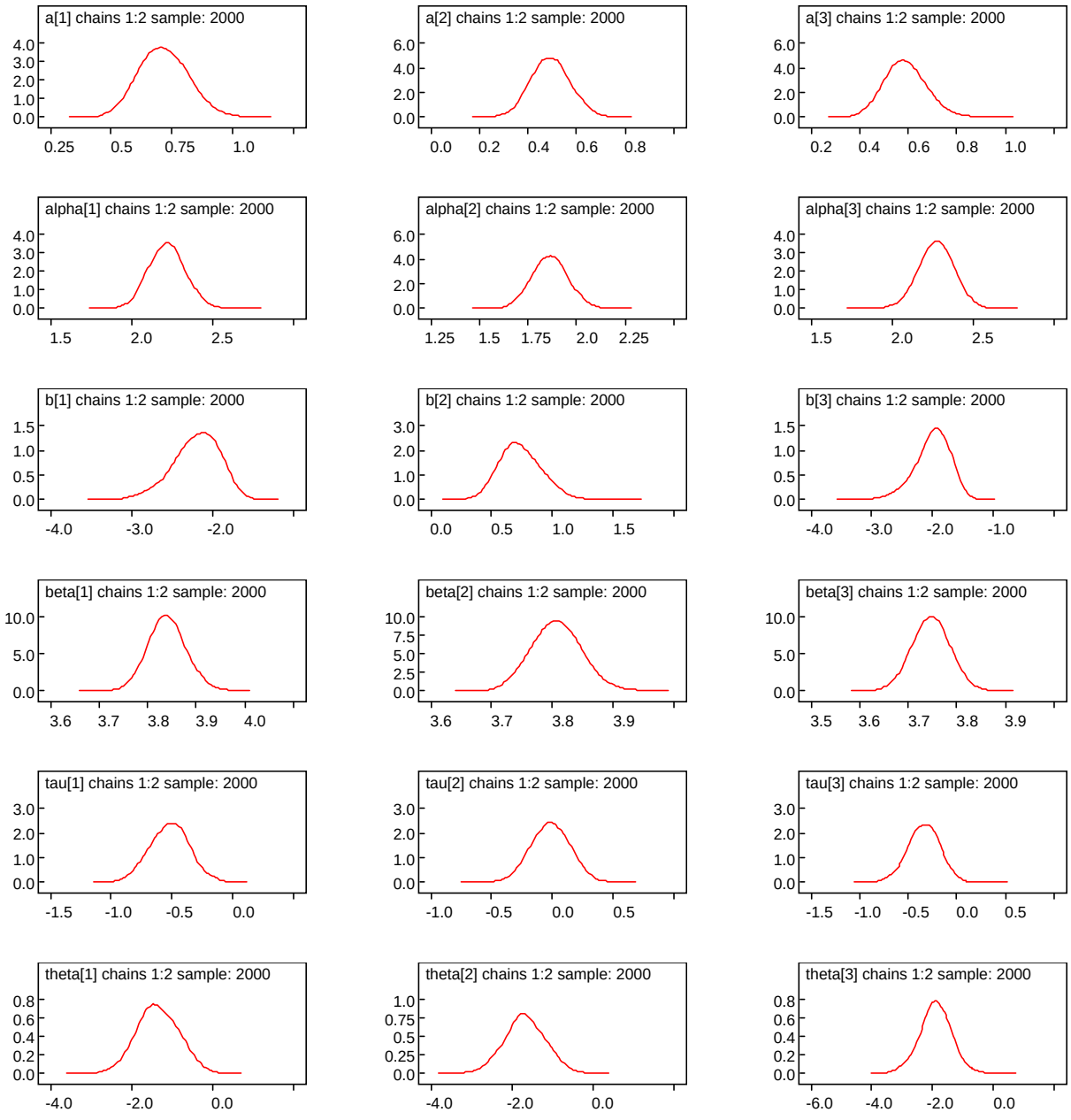


Figure B3. *Some of the representative item and examinee parameter estimates from the hierarchical framework (2PL) model*

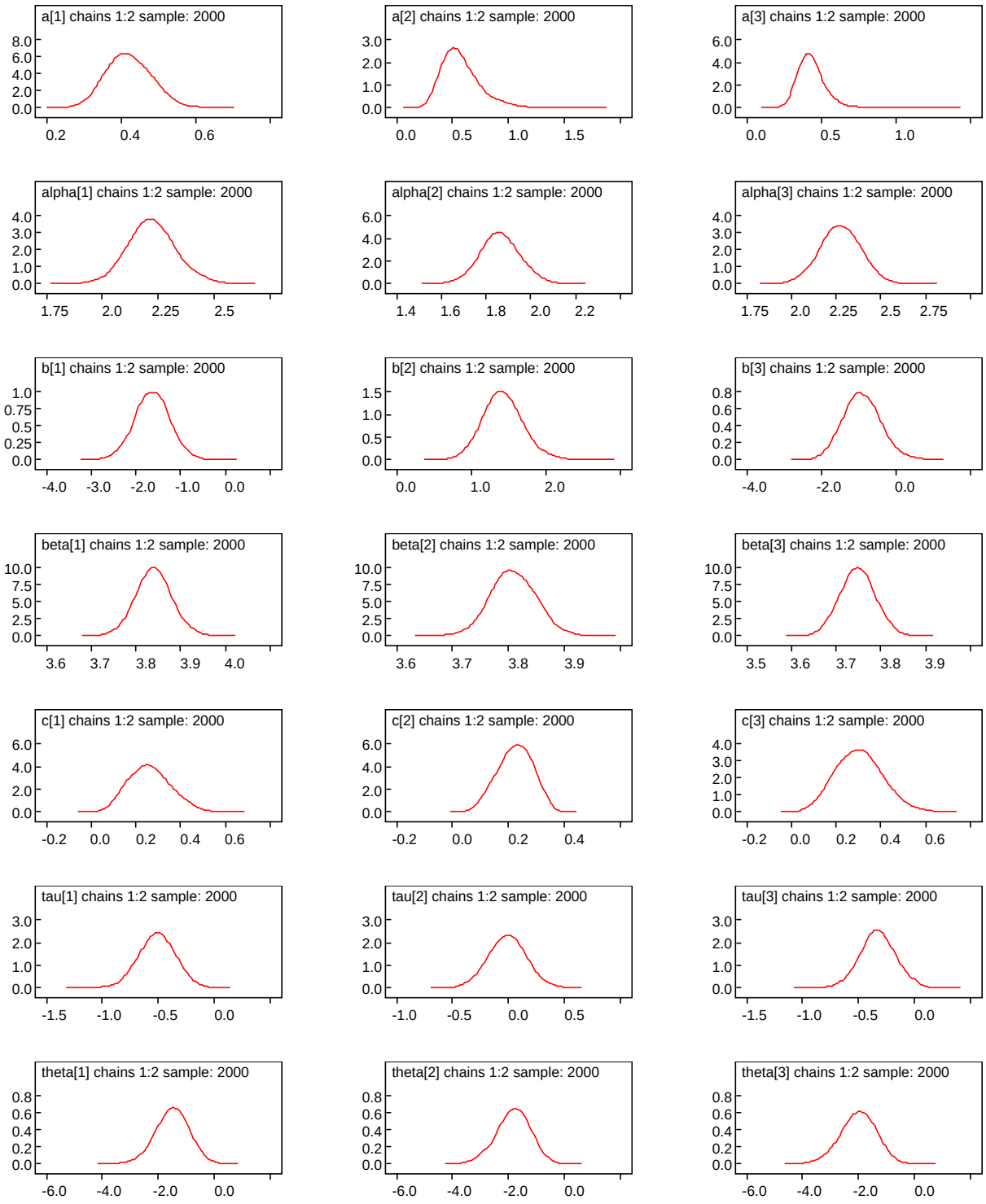


Figure B4. *Some of the representative item and examinee parameter estimates from the hierarchical framework (3PL) model*

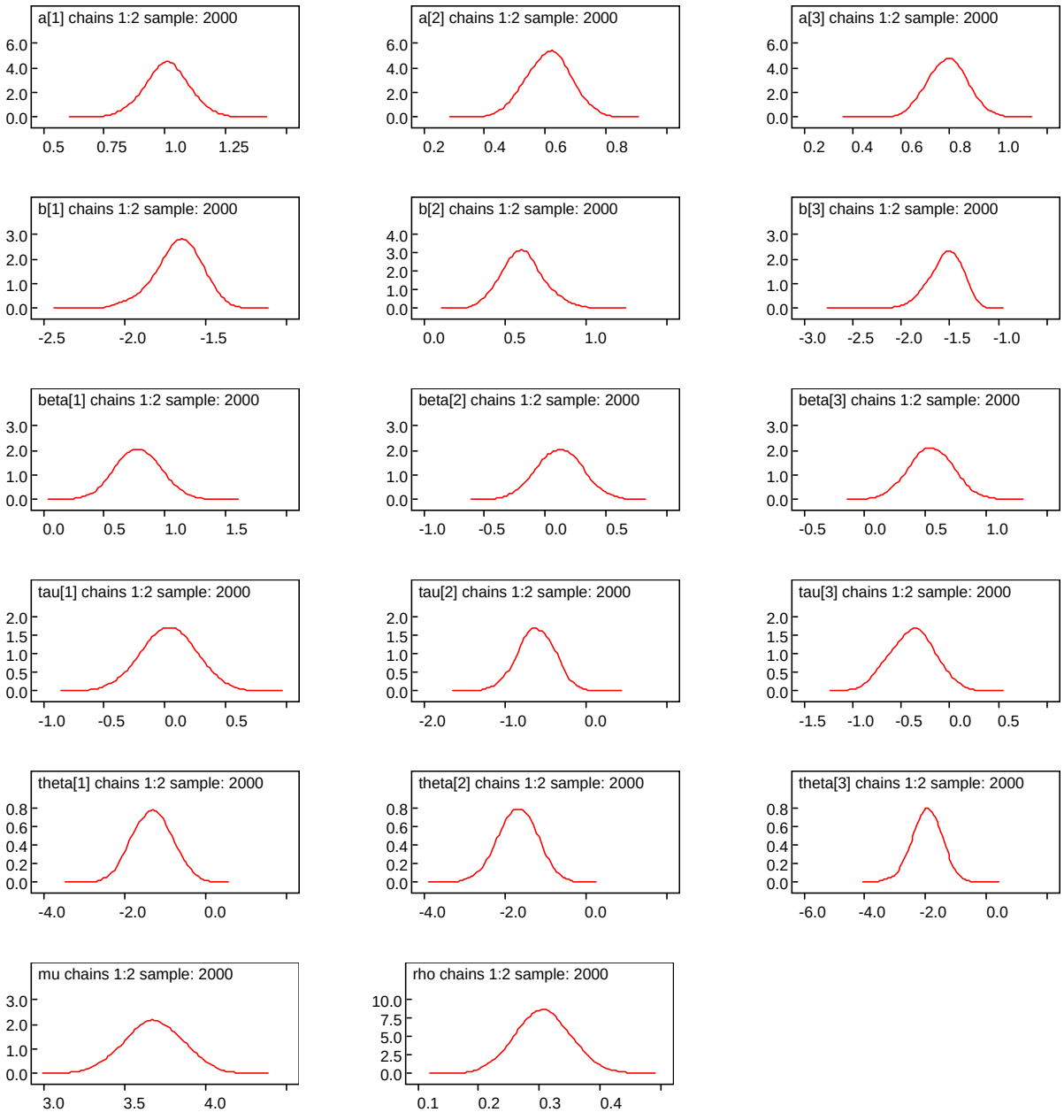


Figure B5. *Some of the representative item and examinee parameter estimates from Thissen's model (2PL)*

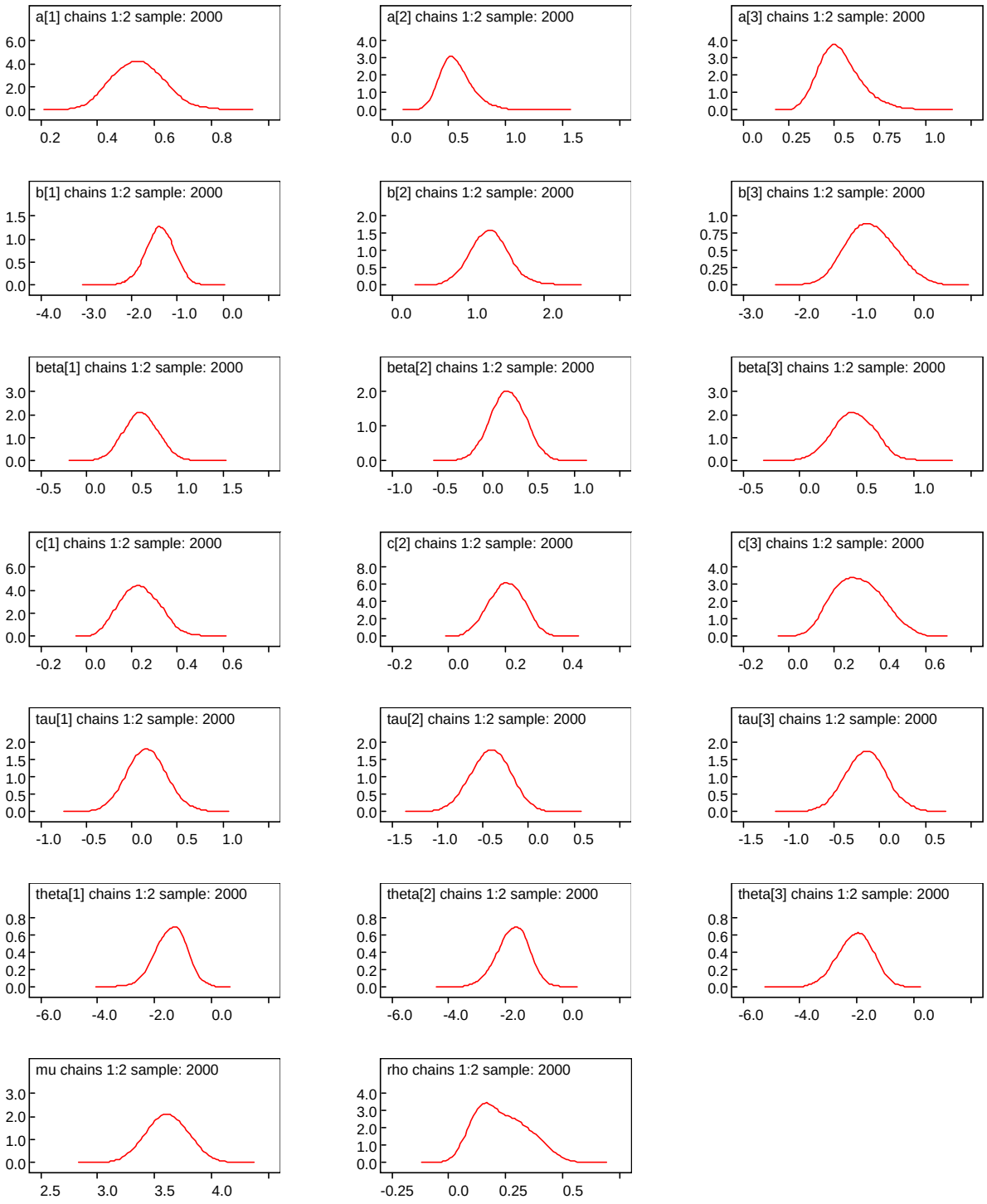
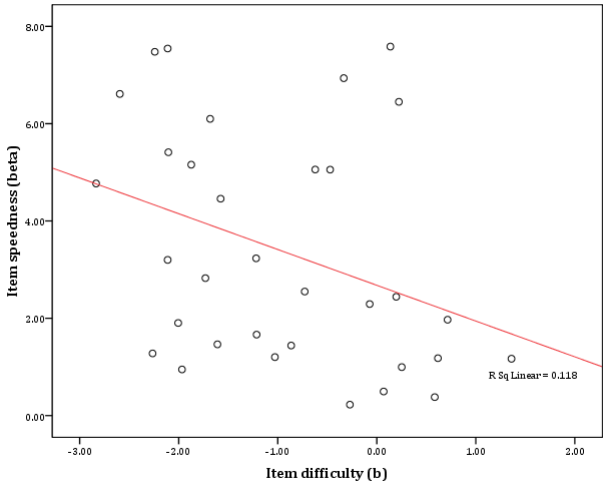
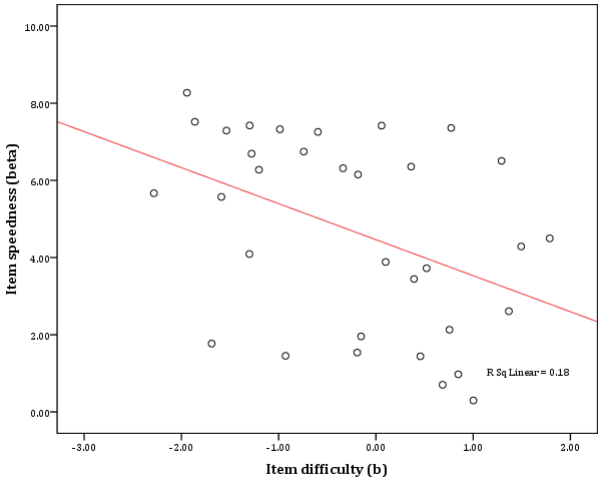


Figure B6. *Some of the representative item and examinee parameter estimates from Thissen's model(3PL)*

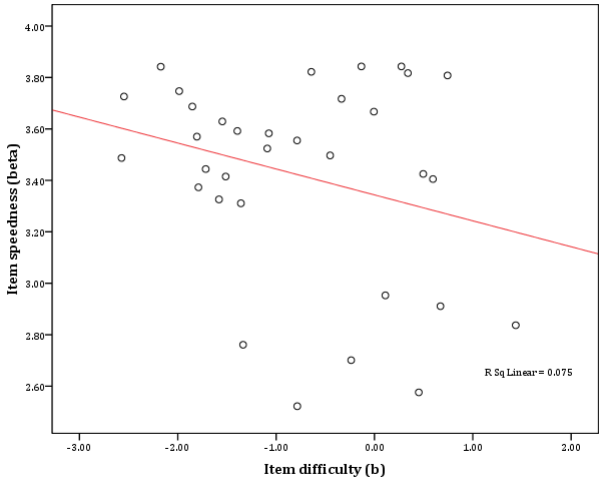
(a) 4PL RT (2PL)



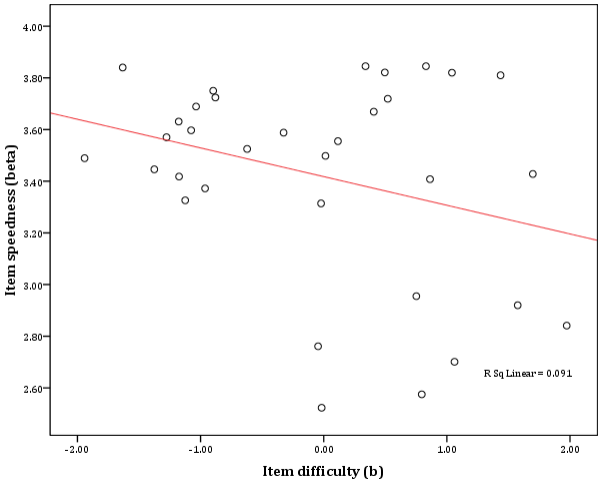
(b) 4PL RT (3PL)



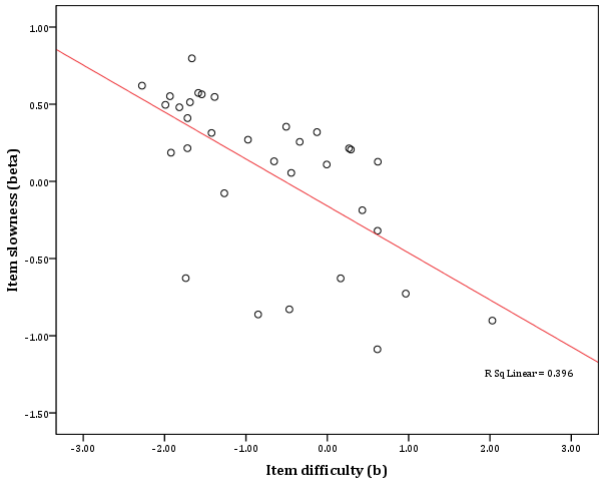
(c) Hierarchical framework (2PL)



(d) Hierarchical framework (3PL)



(e) Thissen's model (2PL)



(f) Thissen's model (3PL)

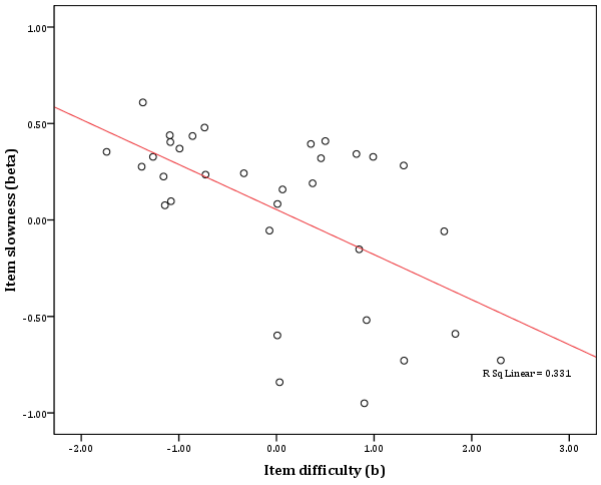
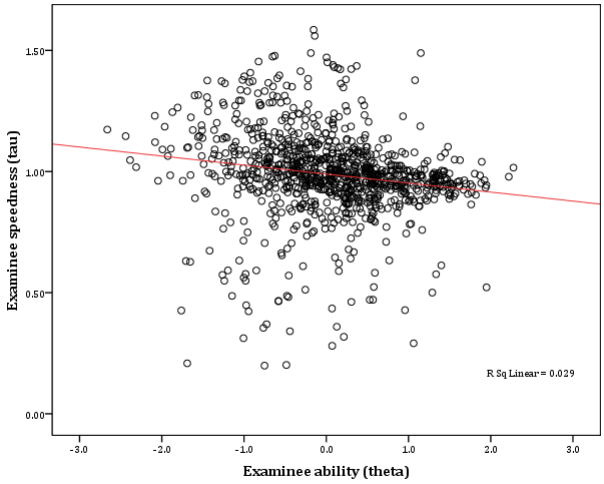
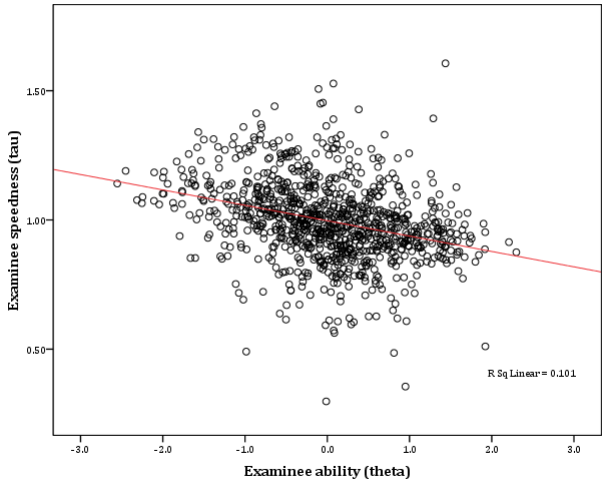


Figure B7. *Scatter plots of item difficulty and item speediness (slowness) parameter estimates.*

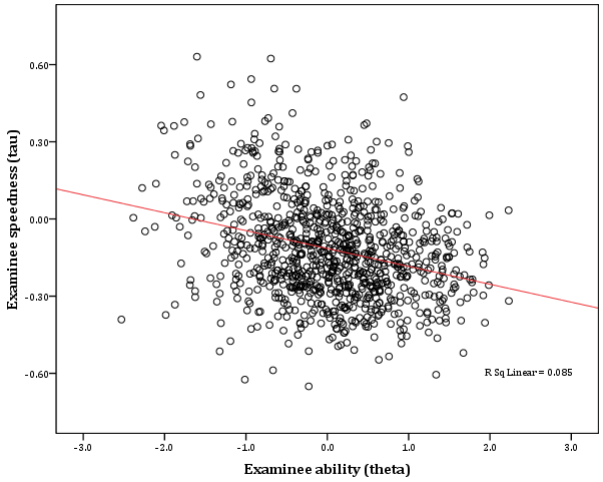
(a) 4PL RT (2PL)



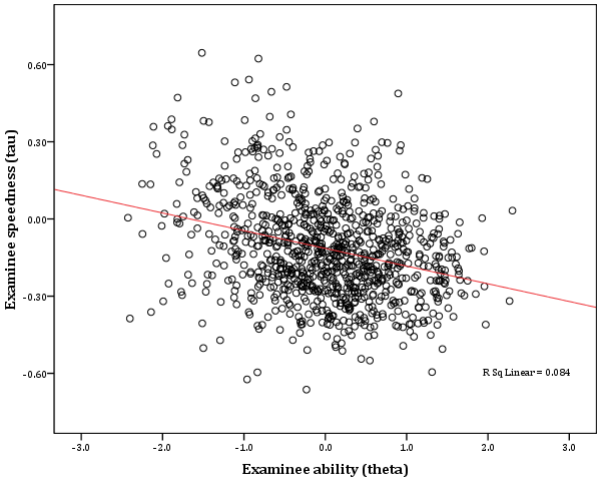
(b) 4PL RT (3PL)



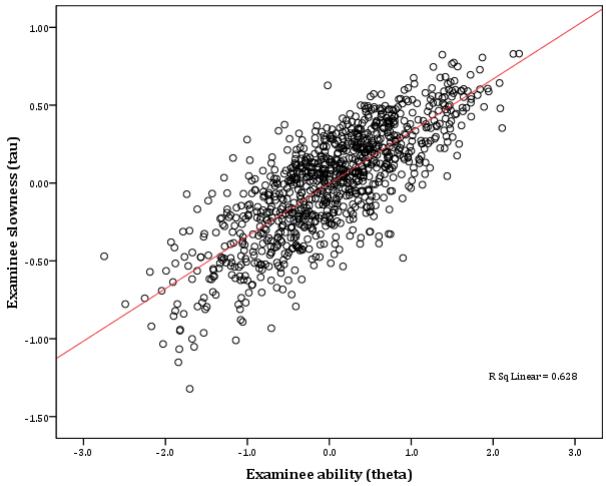
(c) Hierarchical framework (2PL)



(d) Hierarchical framework (3PL)



(e) Thissen's model (2PL)



(f) Thissen's model (3PL)

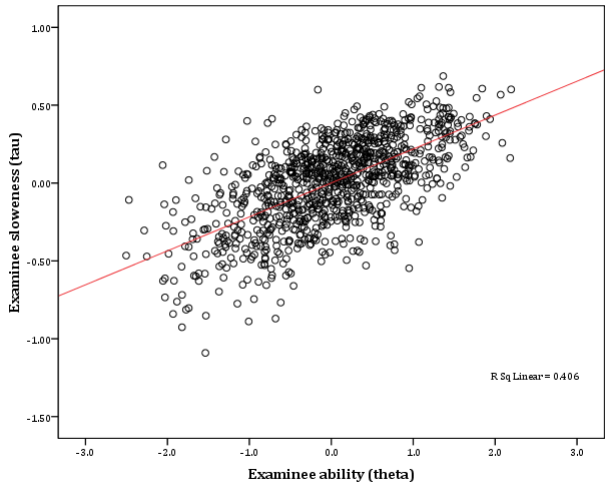


Figure B8. *Scatter plots of examinee ability and examinee speediness (slowness) parameter estimates.*

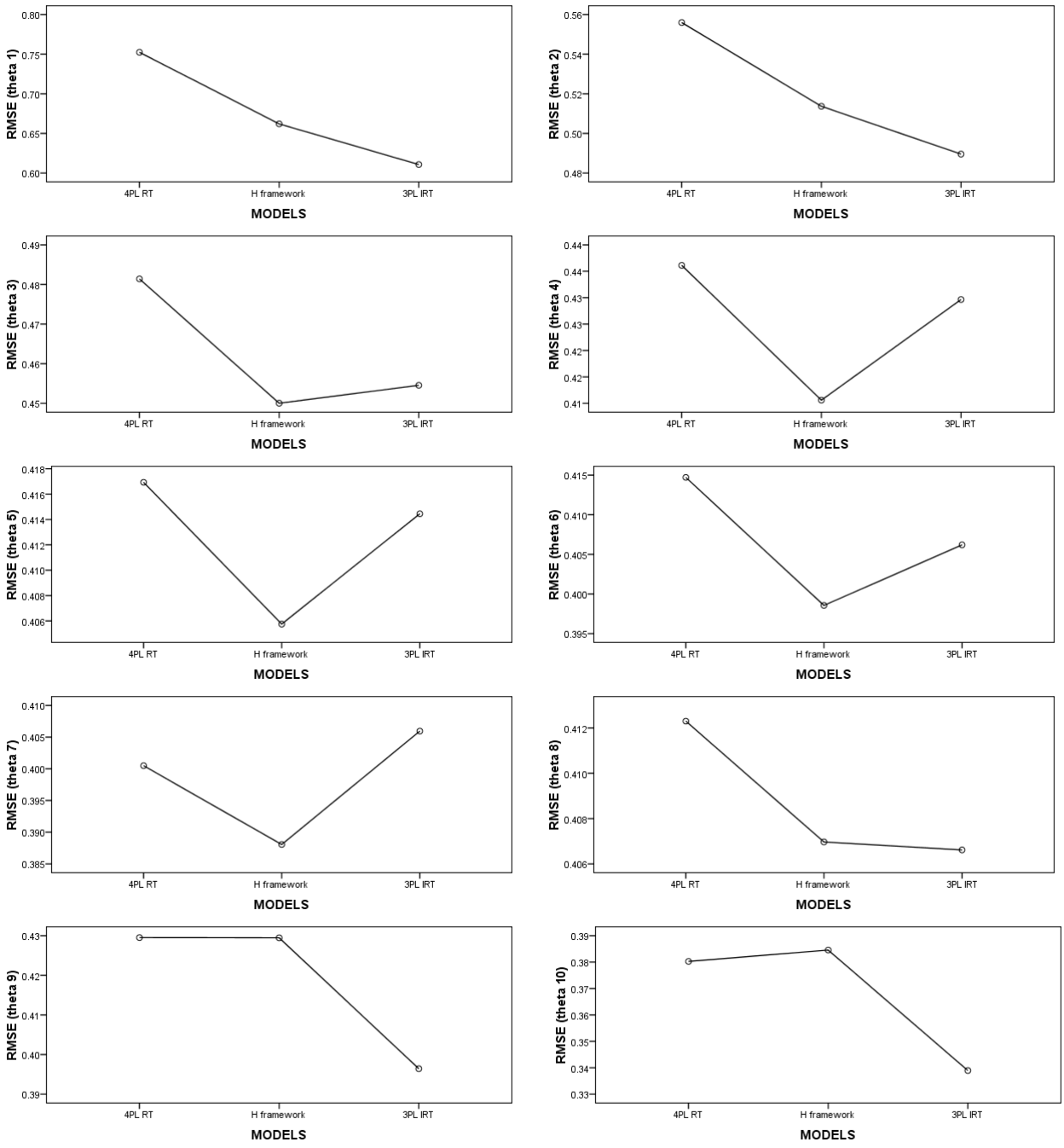


Figure B9. *Mean RMSE values for the examinee true ability parameters based on the examinee ability groups.*