LIKELIHOOD ESTIMATION FOR JOINTLY ANALYZING ITEM RESPONSES AND RESPONSE TIMES

BY

HYEON-AH KANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang, Chair
Professor Carolyn J. Anderson
Associate Professor Steven A. Culpepper
Professor Jeffrey A. Douglas
Assistant Professor Hans-Friedrich Köhn
Associate Professor Jinming Zhang

Abstract

Response time has become increasingly important for analyzing the relationship between the proficiency and speed of an examinee. In this thesis, statistical estimation procedures are presented for jointly modeling responses and response times in educational and psychological testing. The models under consideration include the three-parameter logistic response model, the lognormal response time model, and the proportional hazards latent trait model. The individual models are conjoined within the hierarchical framework so that parameters in the respective models can be characterized under a unified scheme. The thesis presents estimation methods for each of the combinations of the response and response time models by maximizing the likelihood functions. A series of simulation studies verify that the estimation methods perform appropriately, and the parameters are robustly estimated. The likelihood-based approach provides a practical and efficient alternative to Bayesian estimation procedures, which often comes with high computational intensity and dependence on priors.

Acknowledgments

There are many individuals without whom this doctoral thesis might not have been written and to whom I am greatly indebted.

My deep gratitude goes first to my advisor, Dr. Hua-Hua Chang, to whom I owe a sense of gratitude beyond any possibility of expression in words. I would like to thank him for his valuable guidance, support, and encouragement throughout my graduate education. He has been a constant source of motivation and inspiration for me to grow as a scholar in the fields of psychometrics and educational measurement. Without the help and support from him, it would not have been possible for me to achieve my educational goals. I am also thoroughly grateful to my co-advisor, Dr. Carolyn Anderson, from whom I received so much advice, insightful comments, and encouragement. Her constructive feedback on my work has not only been very influential and essential but also widened my research from various perspectives.

I would like to take this opportunity to record my sincere thanks to my committee as well. I am hugely indebted to Dr. Jeff Douglas for finding out time to answer my questions in his busy schedule and for providing indispensable advice and information regarding the topic of my research. I would also like to express my gratitude to Dr. Steve Culpepper for being so generous as to provide me with sources and datasets. I would like to thank Dr. Frieder Köhn and Dr. Jinming Zhang for their genuine interest in my topic of research, for providing me with material and links, and for their kind words and suggestions.

My appreciation also extends to my fellow laboratory colleagues who have lent their helping hand in this venture. It has been my great privilege and pleasure to work with them for the last five years. In particular, I am grateful to Dr. Yi Zheng for her patient emails and conversations to guide me to the topic of online calibration and Justin Kern for stimulating discussions and sincere advice in my writing during the graduate study.

Last but not the least, I would like to thank my family—my parents, brother, and sister—for their unceasing encouragement and support. My mere expression of thanks would never suffice for their unequivocal support and great patience at all times. I also thank I.J. for being unfailingly supportive and considerate.

Table of Contents

Chapter 1

Introduction

1.1 Background

Given advances in technology and the prevalence of computers in assessments, access to response time data has become readily available. In traditional educational and psychological testing, response scores have been a major source of information for making inferences about unobserved abilities of examinees. Information from response times has been ignored largely due to the difficulty of collecting data in paper-and-pencil tests. As computers are assuming a more prominent role in testing in recent years, studies considering response times as a valuable source of information have begun to effloresce, particularly given the evidence in the psychometric literature for improved measurement.

Applications of response times can be found in many sectors of psychometric testing. Examples include assembling tests (van der Linden, 2011), selecting items adaptively in computerized testing (Fan, Wang, Chang, & Douglas, 2012; van der Linden, 2008), detecting aberrant response behavior (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003), controling test administration time (van der Linden, 2009b; van der Linden, Scrams, & Schnipke, 1999), just to name a few. These studies represent important steps in exploring the role of response times in the context of psychometric testing. One of the fundamental issues to be addressed prior to making inferences in these applications is selection of a proper psychometric model.

Available response time models in the measurement literature fall into three categories (Klein Entink, Kuhn, Hornke, & Fox, 2009). The first approach is to model response times based on a particular parametric distribution, such as exponential (Scheiblechner, 1979, 1985), Weibull (Rouder, Sun, Speckman, Lu, & Zhou, 2003; Tatsuoka & Tatsuoka, 1980), lognormal (van der Linden, 2006), and gamma (Maris, 1993) distributions. This approach is usually applied to speeded tests in which easy items are administered with a strict time limit. Accuracy scores are usually left out of account in these tests because they retain only limited information.

The second approach is to model item responses and response times separately (e.g., Embreston, 1998; Gorin, 2005; Mulholland, Pellegrino, & Glaser, 1980; Primi, 2001). Although this strategy provides information on both accuracy scores and response times, it implicitly assumes that the response accuracy and the pace at which an examinee works during the test are independent. This assumption is unlikely to hold in practice because many operational tests are administered with a time limit, and examinees tend to adopt diverse strategies to complete the tests within the allocated time.

The third approach draws a distinction in this regard by allowing dependence among the parameters associated with response scores and response times within the joint modeling framework. The core principle of this approach lies in a speed-accuracy relationship. In cognitive psychology, the trade-off between the speed and accuracy has been known to exist (Luce, 1986). When working on a test, a subject may choose to work quickly at the expense of low accuracy, or may opt to work slowly to increase accuracy. This notion is reflected in several response time models (e.g., Roskam, 1997; Thissen, 1983; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005). These models let the latent component that describes the probability of a correct response rely on the speed parameter or response time.

van der Linden (2007) claims that in educational testing with a reasonable time constraint, it is possible for a population of examinees to show a positive correlation between the speed and accuracy while each individual examinee may display the negative correlation between the two latent traits. Thus, the speed-accuracy trade-off is a within-person phenomenon. Instead of explicitly modeling the person-level trade-off in a response time model, van der Linden (2007) assumes separate ability and speed parameters as latent traits and allows their covariation at the higher level of the model. The notion of the second level for the latent traits naturally leads to a hierarchical framework, in which responses and response times are nested within an examinee. The latent trait parameters—the ability and speed parameters—are seen as random person effects drawn from a population of examinees. Analogously, the effects of items on the responses and response times are disentangled, and their parameters are allowed to covary at the higher level of the item domain.

While the other joint models are fixed in terms of modeling item responses and response time distributions, the hierarchical framework can be readily applied to other psychometric models of concern. The original hierarchical framework, for instance, was constructed based on the three-parameter normal-ogive model and the lognormal model for relating the person and item effects to the observed responses and response times. The choices of the measurement models, however, do not need to be limited to these models as the only condition for the component models is to have both person and item parameters. The hierarchical framework's promise dwells in this potential for allowing greater flexibility in choices of measurement models while allowing for the dependence of the parameters.

1.2 Statement of the Problem

Despite the high potential, only limited options are available for estimating the hierarchical framework. A majority of the studies are based on Markov chain Monte Carlo (MCMC) algorithm with Gibbs sampling. This is made particularly evident by consideration of a number of recent studies, including Fox, Klein Entink, and van der Linden (2007), Klein Entink, Kuhn, et al. (2009), van der Linden (2006), and van der Linden (2007). A major advantage of a Bayesian approach, especially implemented through MCMC technique, is that it provides a natural and principle way of incorporating prior information in estimation of parameters. Inferences about the parameters are made based on the posterior distribution of estimates without reliance on asymptotic approximation. The MCMC-based estimation, however, often comes at a high computational cost and requires a solid background in computational statistics. The computational overhead increases substantially if data contain large numbers of observations. For this reason, the use of the MCMC has been restrained only to small datasets (Fox et al., 2007).

Perhaps the most attractive alternative to MCMC is likelihood estimation. A likelihood-based approach provides much more efficient estimation routines and avails themselves of statistical properties from the long-standing and rigorously studied large sample theory such as consistency and asymptotic efficiency. Glas and van der Linden (2010) explored the possibility of the marginal likelihood inference for the hierarchical framework; however, the central focus of the study was on fitting the hierarchical framework based on the maximum likelihood estimators. The implementation of the study was confined to the lognormal model, and a procedural technicality was overlooked for the most part. It is clear that work remains to be done to advance understanding of the likelihood approach and to further its application to other promising models within the hierarchical framework.

1.3 Purpose of the Study

The purpose of this study is to examine the possibility of the likelihood estimation approach for fitting the hierarchical framework. This study develops two item calibration methods—marginal maximum likelihood (MML) estimation and marginal maximum a posteriori (MMAP) estimation—for jointly modeling the three-parameter logistic and the log-normal models. Based on the estimated item parameter values, estimators of examinees' latent trait parameters are derived based on maximum a posteriori (MAP) and expected a posteriori (EAP).

While the log-normal model has been conveniently used for modeling response times in many studies (e.g., Thissen, 1983; Schnipke & Scrams, 1997; van der Linden et al., 1999; van der Linden, 2006), it has been shown that log-transformed response times may fail to satisfy the normality assumption, and individual items in a test may manifest different shapes of response time distributions (Klein Entink, van der Linden, & Fox, 2009; Ranger & Kuhn, 2012). This perception has stimulated the development of flexible models that can accommodate various shapes of empirical response time distributions (e.g., Douglas, Kosorok, & Chewing, 1999; Klein Entink, van der Linden, & Fox, 2009; Loeys, Legrand, Schettino, & Pourtois, 2014; Ranger & Ortner, 2012; Ranger & Kuhn, 2012, 2014, 2015; C. Wang, Chang, & Douglas, 2013). A notable feature shared by these models is the adoption of a well-known survival model, the proportional hazards (PH) model (Cox, 1972). The PH model with random effects (Clayton & Cuzick, 1985; Vaupel, Manton, & Stallard, 1979) in particular has shown promise for analyzing data collected from individuals whose response times are correlated due to latent traits. The PH latent trait model (PHLTM) of Ranger and Ortner (2012) is based on the same idea treating the latent trait variable as a random effect.

This thesis is intended to provide a new estimation method for the PHLTM within the hierarchical framework. The proposed method is based on a semiparametric procedure. The semiparametric approach achieves flexibility and simplicity in response time modeling by leaving the baseline hazard functions unspecified. The estimation procedure builds on the penalized partial likelihood function, where the marginal distribution of the latent trait parameters determines the penalty term.

Evaluation of the proposed estimation procedures involves extensive simulation studies under varying factors. Factors considered in this study include the sample size, test length, correlation between parameters, sampling design, response time distributions, prior information and so on. The major research question pursued throughout this thesis is whether the proposed estimation methods perform appropriately under the systematic variation of the factors. The relative performance of the alternative methods also comes within the scope of this study.

1.4 Hypotheses

A number of hypotheses in relation to estimation methods and factors may be summarized as follows.

1. Past research in item response theory has demonstrated that Bayesian procedures typically produce item parameter estimates that are more accurate and consistent than those estimated from maximum likelihood procedures (e.g., Gao & Chen, 2005; Lord, 1986; Mislevy, 1986; Swaminathan & Gifford, 1986). Associated with this is the specification of an informative prior, which leads to shifts of parameter estimates toward the mean of the prior distribution. It is therefore hypothesized that the estimation procedures incorporating accurate prior information about the parameters (e.g., MMAP, MAP, EAP) would outperform the maximum likelihood counterparts.
2. In Bayesian approach, the contribution of the prior distribution to parameter estimation would diminish as the number of observations increases. If the sample size or the test length is large, the posterior probability distribution depends predominantly on the observed data through the likelihood function, and the prior distribution has little effect on the estimates (Baker & Kim, 2004, p. 181). Thus, it can be expected that the Bayesian procedures and the maximum likelihood procedures would perform alike as the number of observations increases. For the same reason, the negative impact of an improper prior is expected to be mitigated along with increase in the data size.
3. van der Linden, Klein Entink, and Fox (2010) suggested that when item responses and response time models are jointly estimated within the hierarchical framework, response times serve as collateral information (Novick & Jackson, 1974) for estimating the response model parameters, and hence, more precise estimates of item parameters can be obtained compared to when only the response model is calibrated. This is possible because information is borrowed from the response time data through the assumption of a common distribution of the parameters at a second level. In like manner, it is

hypothesized that estimates from the proposed methods would gain improvement in statistical accuracy to some extent as a result of incorporating collateral information. Obviously, this tendency would be more pronounced in the Bayesian procedures, where the common distribution for the parameters is explicitly utilized.

4. While the supposition stated above may be pertinent to estimation of the item response model parameters, it is anticipated that the gain in accuracy for estimating the response time parameters would be rather small due to the different nature of observed data. That is, response data are typically observed on the basis of discrete values, whereas response time data are observed on a continuous scale. The disproportion in the amount of information between the two sources of data could therefore lead to the one-sided borrowing of information. Additionally, related to this point, it can be speculated that the parameters of the response time model would be recovered more accurately than those from the response model.
5. Patterns that are commonly seen in estimation practices are also expected to hold true in this study. The increase in the sample size, test length, correlation between the parameters would increase the estimation precision while the use of improper prior would result in declined estimation accuracy.

Chapter 2

Models for Jointly Analyzing Responses and Response Times

The use of response times in educational and psychological assessments has been motivated by the idea that response times can contain vital information about examinees' cognitive processes and item characteristics. For example, response times may provide new insights into the relationship between the latent proficiency and speed. Analysis of response times at an item level may reveal the relationship between the difficulty and time intensity of the item. In this chapter, psychometric models that support joint analysis of item responses and response times are reviewed.

Chapter 1.1 briefly outlined that the earliest response time models had assumed that speed and accuracy measure the same construct. Spearman (1927), for example, argued that an examinee's mental ability can be measured on a scale of accuracy, a scale of speed, or some combination of the two constructs. Example models from this viewpoint include Maris (1993), Rouder et al. (2003), and Scheiblechner (1979). The concept of interchangeability of speed and accuracy may hold for a relatively simple task, where response times can actually indicate the processing capacity of an individual to complete the task. When complex tasks are measured such as in educational testing, these two constructs may act as separate constructs. Tate (1948), for example, investigated the speed and accuracy relationship on number series, arithmetic reasoning, and spatial relations questions. He found that, for a controlled level of accuracy, individual examinees worked at a constant speed. Examinees working at a certain speed did not necessarily demonstrate the same accuracy. Several other studies had remarked that the speed and accruacy are separate constructs (e.g., Baxter, 1941; Bridges, 1985; Foos, 1989; Kennedy, 1930; Myers, 1952).

It was Gulliksen (1950, chap. 17) who made a distinction between a speed test and a power test. He defined a pure speed test as a test with an unlimited number of items easy enough to be answered correctly. The goal of this type of test is to measure how quickly examinees respond to the items. Such tests can be scored as the total time taken to complete a fixed number of items, or as the number of items completed within a fixed time interval. In contrast, a pure power test was defined as a test with no time limit but a fixed number of items that vary in difficulty. The goal of a pure power test is to measure how accurately examinees respond to the items, and hence, the test can be scored by counting the number of correct responses.

In reality, pure speed and pure power tests are rarely employed because they are likely to involve both speed and accuracy to some extent. The question to be addressed therefore boils down to how these two constructs interact within a test. The first two model frameworks presented below explicitly consider the trade-off between the speed and accuracy within a test item. The hierarchical framework that comes next characterizes the two constructs as separate latent traits and links them in a population level. Presented in Chapter 2.4 is a new response time model, namely the proportional hazards latent trait model, which provides more flexibility in modeling response time distributions. Particular attention is devoted to the hierarchical framework and the proportional hazards latent trait model in this chapter as they lay the foundation for the methodologies that will be discussed in later chapters.

2.1 Thissen's Model

Thissen (1983) proposed the response time model that incorporates responses. Rather than specifying response time distributions specific to correct and incorrect responses, response times are directly regressed on the parameter structure for the response model:

$$\log T_{ij} = \mu + \tau_i + \beta_j - \rho(a_j\theta_i - b_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \tag{2.1}$$

where $\log T_{ij}$ denotes the log response time of examinee i on item j; μ is the grand mean for the population of examinees and domains of test items; τ_i and β_j are slowness parameters for the examinee and item; ρ is the slope parameter in the regression of the log response time on the response parameter structure; and ε_{ij} is the error term. The normally distributed ε_{ij} indicates that the model belongs to the lognormal family. Two kinds of trade-offs are present in this model, one between the item difficulty and slowness and the other between the ability and slowness. The regression coefficient, ρ, indicates the direction of the relationships between these two trade-offs (Schnipke & Scrams, 2002).

Several variations and applications of the Thissen's model exist in the literature. Schnipke and Scrams (1997) substituted the two-parameter response model component with that of the three-parameter model in an attempt to clarify the relationship between the speed and accuracy as well as to explore the impact of response time on the estimation of proficiency parameter. Instead of $(a_j\theta_i - b_j)$ in (2.1), they used $\log(c_j + \exp(a_j\theta_i - b_j)) - \log(1 - c_j)$ together with $\varepsilon_{ij} \sim \log N(0, \sigma^2)$. Through the application to the computer-based tests of verbal, quantitative, and reasoning skills, they found that there existed a moderate relationship between examinees' speed and ability as well as between the speed and item difficulty. A similar modification was used in Ferrando and Lorenzo-Seva (2007) for modeling response time data in personality tests. The response model component in this model was replaced by a distance measure, $\sqrt{a_j^2(\theta_i - b_j)^2}$, based on a distance-difficulty hypothesis in personality theory—the response time on an item (i.e., the uncertainty in the decision-making process) increases as the person-item distance decreases.

2.2 Four-Parameter Logistic Response Time Model

Unlike the Thissen's model where accuracy is incorporated into the response time model, Wang and Hanson's (2005) model incoporates response times in the three-parameter logistic model (3PLM; Birnbaum, 1968; Lord, 1980). The item response function is defined as

$$P(U_{ij} = 1 \,|\, \theta_i) = c_j + \frac{1 - c_j}{1 + \exp\big\{ - Da_j \left(\theta_i - d_j \tau_i / t_{ij} - b_j\right) \big\}}, \tag{2.2}$$

where U_{ij} is the response variable; D is a scaling parameter; a_j, b_j, and c_j are the item discrimination, difficulty, and guessing parameters, respectively; d_j is the item slowness parameter; θ_i and τ_i are examinee's accuracy and slowness parameters; and t_{ij} is the observed response time. The model is named as the four-parameter logistic model (4PLM) because it includes an additional item parameter, d_j, in addition to the ususal parameters from the 3PLM. The motivation for adding the slowness parameters is that less time spent on an item has the same effect on the probability of success. These parameters determine the rate of increase in the probability of a correct answer as a function of response time. With increasing time, the probability of a correct response approaches that of the regular 3PLM. Wang (2006) later extended the 4PLM to jointly model the responses and response times allowing dependence between the ability and speed parameters; however, it was found that the model did not show much improvement from the regular item response theory models in terms of parameter recovery.

Roskam's (1987) model and Verhelst, Verstralen, and Jansen's (1997) model resemble the (2.2) except for the term, $d_j \tau_i$. Roskam modified the regular Rasch model by replacing the ability parameter with an effective ability parameter defined as the product of mental speed and processing time. (Roskam used the traditional notation of ability to denote speed as well.) This is realized as the sum on an exponential scale:

$$P(U_{ij} = 1 \,|\, \theta_i) = \Big[1 + \exp\big\{ - D \left(\theta_i - \log t_{ij} - b_j\right) \big\}\Big]^{-1}.$$

Verhelst et al. (1997) replaced $\log t_{ij}$ by a separate speed parameter, τ_i:

$$P(U_{ij} = 1 \,|\, \theta_i) = \Big[1 + \exp\big\{ - D \left(\theta_i - \tau_i - b_j\right) \big\}\Big]^{-\pi_j},$$

where π_j is an item-dependent shape parameter. Both models capture the speed-accuracy trade-off in that an increase in time or speed implies an increase in the probability of success on the item.

2.3 Hierarchical Framework

The hierarchical framework (van der Linden, 2007) consists of two levels. The first level defines the measurement models, one for the item responses and the other for the response times. These two models are nested under the second level in which the relations between the first-level parameters are represented. While the framework allows alternative choices of measurement models through the "plug-and-play"approach, the current chapter focuses on the 3PLM and the lognormal model (van der Linden, 2006) for modeling item responses and response times for their wide applications in educational testing.

The 3PLM models the probability of a correct response to item j for an examinee with latent proficiency θ_i as

$$P_j(\theta_i) = P(U_{ij} = 1 \,|\, \theta_i;\, a_j,\, b_j,\, c_j) = c_j + \frac{1 - c_j}{1 + \exp\big\{ - Da_j(\theta_i - b_j)\big\}}, \tag{2.3}$$

where U_{ij} is a random variable denoting the binary response score; $a_j \in \mathbb{R}^+$, $b_j \in \mathbb{R}$, and $c_j \in [0,\, 1)$ are the discrimination, difficulty, and lower-asymptote parameters for item j, respectively. The D is a scaling constant that approximates a normal ogive model and is typically set as 1.702. If c_j is set to zero in (2.3), the model specializes to the two-parameter logistic model (2PLM). If the a_j is set to one across all items in addition to the zero guessing, the model reduces down to the Rasch model.

The distribution of respose time, $T_{ij} \in \mathbb{R}^+$, for examinee i on item j is assumed to be log-normally distributed as

$$f(T_{ij} = t_{ij} \,|\, \tau_i;\, \alpha_j,\, \beta_j) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left[-\frac{\alpha_j^2}{2}\Big\{ \log t_{ij} - (\beta_j - \tau_i)\Big\}^2\right], \tag{2.4}$$

where τ_i is the speed at which the examinee i performs on the test; $\alpha_j \in \mathbb{R}^+$ and $\beta_j \in \mathbb{R}$ are the time discriminating and the time intensity parameters for item j. The log-transformed response times, $\log T_{ij}$, follows a normal distribution with the mean and variance of

$$E\big[\log T_{ij}\big] = \beta_j - \tau_i \quad \text{and} \quad \text{Var}\big[\log T_{ij}\big] = \alpha_j^{-2}. \tag{2.5}$$

Notice that α_j is the reciprocal of the standard deviation of the log response time distribution. Thus, the larger value of α_j can be interpreted as the less dispersion of the log response times on item j across the examinees. The larger value of β_j indicates that the item j systematically requires examinees more time to solve the item. Because the response times have the positive support with a natural lower-bound at zero, the response time model does not require estimation of any lower-asymptote parameter.

The second-level of the hierarchical framework describes the distributions of the person and item parameters in the population and the item domain, respectively. The population domain assumes that examinees' latent parameters are independent and identically distributed (*i.i.d.*) samples drawn from a bivariate normal distribution

$$(\theta_i,\, \tau_i)' \sim N_2(\boldsymbol{\mu}_P,\, \boldsymbol{\Sigma}_P)^1, \tag{2.6}$$

with mean vector

$$\boldsymbol{\mu}_P = (\mu_\theta,\, \mu_\tau)'$$

and covariance matrix

$$\boldsymbol{\Sigma}_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}.$$

In like manner, the item domain assumes that the item parameters are *i.i.d.* samples from

[1] The subscript P of the $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ denotes the population domain.

a multivariate normal distribution

$$\boldsymbol{\xi}_j = (a_j,\, b_j,\, c_j,\, \alpha_j,\, \beta_j)' \sim N_5(\boldsymbol{\mu}_I,\, \boldsymbol{\Sigma}_I)^2, \tag{2.7}$$

with mean vector

$$\boldsymbol{\mu}_I = (\mu_a,\, \mu_b,\, \mu_c,\, \mu_\alpha,\, \mu_\beta)'$$

and covariance matrix

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}.$$

To establish the identifiability, constraints are imposed such that $\boldsymbol{\mu}_P = 0$ and $\sigma_\theta^2 = \sigma_\tau^2 = 1$. The restriction that $\mu_\theta = 0$ is analogous to the restriction that is usually imposed in the standard maximum likelihood estimation (e.g., Bock & Aitkin, 1981). The restriction that $\mu_\tau = 0$ removes the trade-off between $\beta_j - \tau_i$. Although σ_τ^2 needs not be fixed to a known constant for the purpose of identifiability, the present study assumes that the scale of response times can be standardized to have a unit variance.

A number of extensions of the hierarchical framework are made in the literatue. Klein Entink, Fox, van der Linden, and Fox (2009) extended the framework to a multivariate multilevel regression structure to allow the incorporation of covariates in explaining the variance in the speed and accuracy between individuals who may be nested within groups. Klein Entink, Kuhn, Hornke, and Fox (2009) proposed a variant of the hierarchical framework to address the cognitive procesees required by individual items. Molenaar, Tuerlinckx, and van der Maas (2015) fit the hierarchical framework within a generalized linear factor model by restricting the hierarchical crossed random effects to random person effects only. Instead

[2] The subscript I of the $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$ denotes the item domain.

of the log transformation of the response times, Klein Entink, van der Linden, and Fox (2009) considered a broader class of Box-Cox transformation (Box & Cox, 1964) to address the different shapes of response time distributions.

2.4 Proportional Hazards Latent Trait Model

In psychometric testing, response time is the time elapsed from the onset of an item until an examinee answers the item. If the examinee's responding to an item is viewed as an event, response times have the same meaning as the survival times in biostatistics, and hence, they can be analyzed through survival data techniques (Wang, Fan, Chang, & Douglas, 2013). One of the well-known models for the analysis of event times in survival data is the proportional hazards (PH) model popularized by Cox (1972). The PH model is a regression-like model that accounts for individual differences in the hazard function to predict characteristics of the subjects in survival times.

One of the basic assumptions of the PH model is the independence of event times given the current time and observed values of covariates. Oftentimes, this assumption is not probable because of the unobserved covariates and shared properties in data. For example, in clinical settings groups of patients may have unobserved genetic or environmental determinants in common. Furthermore, if several events are observed for the same person, a within-individual correlation may be present between the events. Ignoring such dependence in the analysis adversely affects the estimation of the relationship between hazards (e.g., Hougaard, 1991; Wei, Lin, & Weissfeld, 1989). In this regard, the PH model with random effects (Ripatti & Palmgren, 2000; Vaida & Xu, 2000)—also known as the frailty model (Clayton & Cuzick, 1985; Vaupel et al., 1979)—can take into account the within-cluster dependencies. The model assumes that event times are independent conditional on unobserved random effects. In psychometric testing, examinees' speed parameters can be seen as random effects for observing response times. Hence, when response time data are analyzed through the PH model with random effects, it is expected that there be no covariation left between response times on different items conditioning on the random speed parameter. Ranger and Ortner's (2012) model is based on this assumption and attempts to explain the variation in response time distributions through the distribution of hazard functions.

The hazard function (or equivalently, the hazard rate) models the probability that an event will occur in the next instant given that the event has not yet occurred. Let f be the probability density function of the response time, T, with corresponding cumulative distribution, $F(t) = P(T \leq t) = \int_0^t f(u)\,du$, and survival function, $S(t) = 1 - F(t) = P(T > t)$. The hazard rate is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,\big|\, T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.8}$$

In survival analysis, the hazard ratio between two individuals (e.g., the hazard of a treated subject over the hazard of a control subject) can never be negative. Therefore, the PH model defines the hazard rate using an exponential function as a link function to describe the impact of covariates.

Let τ_i denote the latent speed parameter of an examinee i ($i = 1, \ldots, N$). The hazard rate of item j ($j = 1, \ldots, J$) for the examinee with τ_i is defined as (Ranger & Ortner, 2012)

$$h_{ij}(t;\, \tau_i) = h_{0j}(t) \exp\left(\gamma_j \tau_i\right), \tag{2.9}$$

where $h_{0j}(\cdot)$ is the baseline hazard rate, and γ_j is the regression coefficient implying the influence of τ_i on the hazard rate. The sign of γ_j is constrained to be positive so that it can be interpreted as a discrimination parameter. The larger the τ_i, the smaller the t's. The latent speed parameter, τ_i, acts multiplicatively on the hazard rate and introduces the unobserved heterogeneity into the model. Similar to the random effect in the generalized linear mixed model, τ_i is assumed to follow a normal distribution with a zero mean. The baseline hazard rate, $h_{0j}(\cdot)$, corresponds to the hazard rate of an examinee with τ_i equal to 0. The value is common to all examinees, yet it can vary for different items. The functional form of $h_{0j}(\cdot)$ can be either assumed to follow a particular parametric distribution or to be completely unknown, each of which leads to the parametric and semiparametric PH latent trait models.

Applying the relationship between the hazard rate and the survival function in (2.8), the conditional density of respons time for examinee i to item j can be obtained as

$$f(t_{ij} \,|\, \tau_i;\, \gamma_j,\, h_{0j}) = h_{0j}(t_{ij}) \exp\left(\gamma_j \tau_i\right) \exp\big\{ -\exp\left(\gamma_j \tau_i\right) H_{0j}(t_{ij}) \big\}, \tag{2.10}$$

where $H_{0j}(t_{ij})$ is the cumulative baseline hazard rate calculated as $\displaystyle\int_0^{t_{ij}} h_{0j}(u)\, du$.

It is germane to note that the first attempt to adopt the PH modeling framework to the field of psychometrics is made in Douglas et al. (1999). They presented a discrete version of the frailty model to model waiting times for which items differ in terms of the extent that the speed parameter influences response times. The method of discretizing response times in estimation of parameters was evaluated in comparison with profile likelihood estimation and estimation based on the rank correlation matrix in Ranger and Ortner (2012) and Ranger and Ortner (2013). Several variants of the Ranger and Ortner's model also exist in the literature. Wang et al. (2013) presented an MCMC framework for jointly analyzing response times and response scores using the Ranger and Ortner's model. Loeys et al. (2014) proposed the PH model with crossed random effects by considering both subjects and items to be random. Ranger and Kuhn (2014) considered the PH model for modeling two accumulators, the acquisition of knowledge and the tendency to discontinue. Ranger and Kuhn (2015) proposed a mixture PH latent trait model assuming different subgroups of examinees differ in their way of responding.

Chapter 3

Hierarchical Framework Estimation

This chapter introduces a likelihood-based approach for estimating the item and person parameters in the hierarchical framework of van der Linden (2007). Two methods are developed for estimating the item parameters: marginal maximum likelihood (MML) estimation and marginal maximum a posteriori (MMAP) estimation. For making marginalized inferences about the item parameters, the expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm is employed. Once item parameters are estimated with enough accuracy, person parameters can be estimated by treating the estimated item parameter values as known. The present chapter discusses three likelihood-based methods for the person parameter estimation: maximum likelihood (ML) estimation, maximum a posteriori (MAP) estimation, and expected a posteriori (EAP) estimation. The ML estimator is subsumed under the MAP estimator as a special case of having a uniform prior, and therefore, the chapter gives closer attention to the MAP and EAP estimation methods.

Throughout the chapter, parameter estimation is implemented under several key assumptions. First, it is assumed that observations from an individual examinee are independent conditioned on the examinee's latent trait parameters. This assumption entails three types of conditional independence.

(i) Independence between responses given θ_i. That is, the joint distribution of item responses is equal to the product of the marginal distributions (Lord & Novick, 1968, p. 361):

$$f(\boldsymbol{u}_i \,|\, \theta_i) = \prod_{j=1}^{J} f(u_{ij} \,|\, \theta_i),$$

where $\boldsymbol{u}_i = \{u_{ij}\}_{1\leq j\leq J}$, and $f(u_{ij} \,|\, \theta_i)$ denotes the Bernoulli distribution of response u_{ij} for a fixed person with θ_i.

(ii) Independence between response times given τ_i, which is defined as (van der Linden, 2006)

$$f(\boldsymbol{t}_i \,|\, \tau_i) = \prod_{j=1}^{J} f(t_{ij} \,|\, \tau_i),$$

where $\boldsymbol{t}_i = \{t_{ij}\}_{1 \leq j \leq J}$, and $f(t_{ij} \,|\, \tau_i)$ is the density of response time t_{ij} given τ_i.

(iii) Independence between responses and response times given θ_i and τ_i (van der Linden & Glas, 2010):

$$f(u_{ij},\, t_{ij} \,|\, \theta_i,\, \tau_i) = f(u_{ij} \,|\, \theta_i) f(t_{ij} \,|\, \tau_i)$$

for all $i = 1, \ldots, N$.

Second, to allow for person parameter estimation and item calibration one at a time, it is assumed that examinees are independent, items are independent, and examinees and items are independent. The respective parameters of an examinee or an item are, however, allowed to covary through the second level of the hierarchical framework.

Third, it is assumed that hyperparameters of the item and population domains are either known or estimated with enough precision. This assumption allows to make proper inferences about unknown parameters based on the posterior distribution, which is proportional to the product of the likelihood function and the prior distribution.

In the following derivations, hyperparameters for the person and item parameters are denoted as $\boldsymbol{\Omega} = (\boldsymbol{\mu}_P,\, \boldsymbol{\Sigma}_P)$ and $\boldsymbol{\Psi} = (\boldsymbol{\mu}_I,\, \boldsymbol{\Sigma}_I)$. Examinees represent random samples from a population where latent traits are distributed according to $f(\theta,\, \tau \,|\, \boldsymbol{\Omega})$, while items have the prior distribution of the same form, $f(\boldsymbol{\xi} \,|\, \boldsymbol{\Psi})$. Additionally, $P_j(\theta_i)$ and $Q_j(\theta_i) = 1 - P_j(\theta_i)$ are denoted as P_{ij} and Q_{ij}, respectively, for notational simplicity.

3.1 Item Parameter Estimation

When calibrating items, examinees' latent trait variables remain random and unknown, and hence, a procedure to free the item calibration from its dependence on the person parameters is needed. The seminal work in this regard is Bock and Lieberman (1970) in which an

MML method was developed for estimating the item parameters. Bock and Aitkin (1981) subsequently reformulated the MML estimation approach by employing the EM algorithm to provide a computationally feasible alternative to the Bock and Lieberman's approach. The present chapter follows the solution of Bock and Aitkin, making necessary modifications to adapt to the hierarchical framework.

Let $\mathbf{U} = \{\boldsymbol{u}_i\}_{1 \leq i \leq N}$ and $\mathbf{T} = \{\boldsymbol{t}_i\}_{1 \leq i \leq N}$ denote the observed response matrix and response time matrix for all examinees and items. In terms of estimating item parameters within the hierarchical framework, ($\mathbf{U}$, $\mathbf{T}$, θ, τ) is unobserved complete data, and ($\mathbf{U}$, $\mathbf{T}$) is observed incomplete data. Item parameters are considered structural parameters, the size of which is fixed by the test length. Person parameters are considered incidental parameters because their size depends on the observed sample. Neyman and Scott (1948) (also Little & Rubin, 1983) suggested that when structural parameters are estimated simultaneously with incidental parameters, the ML estimates of the structural parameters would not be consistent as the sample size increases. Bock and Aitkin's approach to this problem is to marginalize the likelihood function of the structural parameters with respect to the incidental parameters and use the iterative procedure to increase the expected complete-data log-likelihood. The procedure removes the dependence on the unknown person parameters through marginalization and ensures that ML estimates of the item parameters are consistent for tests of finite length.

The MML estimation assumes a distinct population distribution so that examinees' latent trait variables can be integrated out of the likelihood function. In this sense, the MML estimation capitalizes on the information from the population domain of the hierarchical framework. The MMAP estimation, on the other hand, takes advantage of the information from both the population and item domains. Thus, the distinction between the MML and MMAP estimation methods can be made by the degree of utilization of the prior information. Although the present study assumes that the prior distributions of the parameters are well defined, impact of wrong prior should not be overlooked. Later in this chapter, robustness of the estimation methods against inappropriate priors is investigated through simulation studies.

Presented below are some derivatives for each item parameter that are useful for deriving the MML and MMAP estimation procedures.

$$\frac{\partial P_{ij}}{\partial a_i} = DQ_{ij}(\theta_i - b_j)\frac{(P_{ij} - c_j)}{(1 - c_j)}, \qquad \frac{\partial}{\partial a_j}\left[\frac{P_{ij} - c_j}{P_{ij}(1 - c_j)}\right] = Dc_j(\theta_i - b_j)\frac{Q_{ij}(P_{ij} - c_j)}{P_{ij}^2(1 - c_j)^2},$$
$$\frac{\partial P_{ij}}{\partial b_j} = -Da_j\frac{Q_{ij}(P_{ij} - c_j)}{(1 - c_j)}, \qquad \frac{\partial}{\partial b_j}\left[\frac{P_{ij} - c_j}{P_{ij}(1 - c_j)}\right] = -Da_jc_j\frac{Q_{ij}(P_{ij} - c_j)}{P_{ij}^2(1 - c_j)^2},$$
$$\frac{\partial P_{ij}}{\partial c_j} = \frac{Q_{ij}}{1 - c_j}, \qquad \frac{\partial}{\partial c_j}\left[\frac{P_{ij} - c_j}{P_{ij}(1 - c_j)}\right] = -\frac{Q_{ij}(P_{ij} - c_j)}{P_{ij}^2(1 - c_j)^2}.$$

3.1.1 Marginal Maximum Likelihood Estimation

Let $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_j\}_{1 \le j \le J}$ denote the item parameter matrix for all J items. The marginal likelihood of observing $(\boldsymbol{u}_i,\, \boldsymbol{t}_i)$ for an examinee is

$$\iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i,\, \boldsymbol{\Xi})\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i.$$

The logarithm of the marginal likelihood function for all examinees is

$$l = \log L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T}) = \sum_{i=1}^{N} \log \iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i,\, \boldsymbol{\Xi})\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i. \tag{3.1}$$

Since items are assumed to be independent, cross second-derivatives of the different items are zero, and the maximization of the marginal log-likelihood can be carried out for each item singly. Thus, the MML estimator of $\boldsymbol{\xi}_j$ can be found as a solution to a set of five equations:

$$\frac{\partial l}{\partial \boldsymbol{\xi}_j} = \left(\frac{\partial l}{\partial a_j},\, \frac{\partial l}{\partial b_j},\, \frac{\partial l}{\partial c_j},\, \frac{\partial l}{\partial \alpha_j},\, \frac{\partial l}{\partial \beta_j}\right)' = \mathbf{0}, \tag{3.2}$$

where

$$
\begin{aligned}
\frac{\partial l}{\partial a_j} &= D \sum_{i=1}^{N} \iint (\theta_i - b_j) \, \frac{(t_{ij} - P_{ij})\,(P_{ij} - c_j)}{P_{ij}\,(1 - c_j)} \, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i \, d\tau_i, \\
\frac{\partial l}{\partial b_j} &= -D a_j \sum_{i=1}^{N} \iint \frac{(u_{ij} - P_{ij})\,(P_{ij} - c_j)}{P_{ij}\,(1 - c_j)} \, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i \, d\tau_i, \\
\frac{\partial l}{\partial c_j} &= \sum_{i=1}^{N} \iint \frac{(u_{ij} - P_{ij})}{P_{ij}\,(1 - c_j)} \, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i \, d\tau_i, \\
\frac{\partial l}{\partial \alpha_j} &= \sum_{i=1}^{N} \iint \Big[\alpha_j^{-1} - \alpha_j \big\{ \log t_{ij} - (\beta_j - \tau_i) \big\}^2 \Big] \, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i \, d\tau_i, \\
\frac{\partial l}{\partial \beta_j} &= \alpha_j^2 \sum_{i=1}^{N} \iint \big\{ \log t_{ij} - (\beta_j - \tau_i) \big\} \, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i \, d\tau_i.
\end{aligned}
$$

The derivatives of the log-likelihood function above involve unknown quantities resulting from the person parameters (e.g., P_{ij} and τ_i). To deal with the dependence on the unobserved latent variables, the current study employs the EM algorithm in conjunction with a numerical iteration technique. A following subsection provides a step-by-step explanation of the procedure for implementing the MML estimation with the EM algorithm (MMLE/EM).

Computational Methods for MMLE/EM

The EM algorithm is an iterative procedure for finding ML or MAP estimates of parameters of probability models in the presence of unobserved latent variables. In the present context, $\boldsymbol{\xi}$ is considered a parameter of interest (i.e., structrual parameter) and $(\theta,\, \tau)$ is considered an unobservable random variable (i.e., incidental parameter). Let $\hat{\boldsymbol{\xi}}_j^{(t)}$ denote the estimated parameter vector for item j at t-th cycle of the EM algorithm[1]. For removing the random noise associated with the unobserved person parameters, the expected value of the complete-data log-likelihood is calculated using the current estimate $\hat{\boldsymbol{\xi}}_j^{(t)}$. Given the observed data $(\mathbf{U},\, \mathbf{T})$, our best knowledge about $(\theta_i,\, \tau_i)$ is summarized by the posterior distribution,

[1] The superscript t within the parentheses implies the number of iteration, whereas the t in the ordinary script represents the response time.

$f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)})$. Thus, the expected log-likelihood function conditioning on the $(\mathbf{U},\, \mathbf{T})$ and $\hat{\boldsymbol{\xi}}_j^{(t)}$ is obtained as

$$E\left[\log f(\mathbf{U},\, \mathbf{T},\, \theta,\, \tau \,\middle|\, \boldsymbol{\xi}_j) \,\middle|\, \mathbf{U},\, \mathbf{T},\, \hat{\boldsymbol{\xi}}_j^{(t)}\right] = \iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \theta_i,\, \tau_i \,|\, \boldsymbol{\xi}_j)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)})\, d\theta_i\, d\tau_i. \tag{3.3}$$

This procedure is called the expectation (E) step because the expected values are substituted for the unknown quantities in (3.3).

The multiple integrals in the above expression can be evaluated using the Gauss-Hermite quadrature (Abramowitz & Stegun, 1972, p. 890). Determine Q^2 nodes X_k $(k = 1,\, \ldots,\, Q)$ and Y_l $(l = 1,\, \ldots,\, Q)$ at the midpoint of each rectangle on the θ- and τ-scale. Using the weight function, $A(X_k,\, Y_l)$, representing the height of the density, the posterior probability that the i-th examinee's latent trait parameters equal $(X_k,\, Y_l)$ is computed as

$$f(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)},\, \boldsymbol{\Omega}) = \frac{L(X_k,\, Y_l)\, A(X_k,\, Y_l)}{\displaystyle\sum_{k=1}^{Q}\sum_{l=1}^{Q} L(X_k,\, Y_l)\, A(X_k,\, Y_l)},$$

where

$$L(X_k,\, Y_l) = L(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)}) = f(\boldsymbol{u}_i \,|\, X_k,\, \hat{\boldsymbol{\xi}}_j^{(t)})\, f(\boldsymbol{t}_i \,|\, Y_l,\, \hat{\boldsymbol{\xi}}_j^{(t)}).$$

The $L(X_k,\, Y_l)$ represents the likelihood of the examinee's item scores and response times at the quadrature node $(X_k,\, Y_l)$. The expected values associated with the person parameters can be obtained as:

$$\bar{\kappa}_{jkl} = \sum_{i=1}^{N} f(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)},\, \boldsymbol{\Omega}) = \sum_{i=1}^{N} \frac{L(X_k,\, Y_l)\, A(X_k,\, Y_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{Q} L(X_k,\, Y_l)\, A(X_k,\, Y_l)},$$

$$\bar{\iota}_{jkl} = \sum_{i=1}^{N} u_{ij}\, f(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)},\, \boldsymbol{\Omega}) = \sum_{i=1}^{N} \frac{u_{ij}\, L(X_k,\, Y_l)\, A(X_k,\, Y_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{Q} L(X_k,\, Y_l)\, A(X_k,\, Y_l)},$$

$$\bar{\lambda}_{jkl} = \sum_{i=1}^{N} \log t_{ij}\, f(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)},\, \boldsymbol{\Omega}) = \sum_{i=1}^{N} \frac{\log t_{ij}\, L(X_k,\, Y_l)\, A(X_k,\, Y_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{Q} L(X_k,\, Y_l)\, A(X_k,\, Y_l)},$$
$$\bar{\varsigma}_{jkl} = \sum_{i=1}^{N} (\log t_{ij})^2 f(X_k,\, Y_l \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \hat{\boldsymbol{\xi}}_j^{(t)},\, \boldsymbol{\Omega}) = \sum_{i=1}^{N} \frac{(\log t_{ij})^2 L(X_k,\, Y_l)\, A(X_k,\, Y_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{Q} L(X_k,\, Y_l)\, A(X_k,\, Y_l)}.$$

The expected values above are called the artificial data (Baker & Kim, 2004, p. 168) because they are artificially created during the estimation. The values of the artificial data still depend on the values of the unknown item parameter (e.g., $P_j(X_k)$ and $f(t_{ij} \,|\, Y_l)$), and hence, an iterative procedure based on an approximation technique is needed. This process is done by the maximization (M) step of the EM algorithm with the Newton-Raphson (NR) iteration based on a Taylor series.

Let g be an objective function to be approximated. In the present case, the g is the first-order derivatives of the marginal log-likelihood function with respect to each item parameter. The first-order Taylor approximation of g at $\hat{\boldsymbol{\xi}}_j^{(t)} = \left(\hat{a}_j^{(t)},\, \hat{b}_j^{(t)},\, \hat{c}_j^{(t)},\, \hat{\alpha}_j^{(t)},\, \hat{\beta}_j^{(t)}\right)'$ is

$$g\left(\boldsymbol{\xi}_j\right) \approx g\left(\hat{\boldsymbol{\xi}}_j^{(t)}\right) + \nabla g\left(\hat{\boldsymbol{\xi}}_j^{(t)}\right)\left(\boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_j^{(t)}\right), \tag{3.4}$$

where $\nabla g\left(\hat{\boldsymbol{\xi}}_j^{(t)}\right)$ is the gradient of g evaluated at $\hat{\boldsymbol{\xi}}_j^{(t)}$. Plugging in each item parameter component into (3.4) leads to

$$g\left(\boldsymbol{\xi}_j\right) \approx g\left(\hat{\boldsymbol{\xi}}_j^{(t)}\right) + \Delta\hat{a}_j^{(t)} \cdot \left.\frac{\partial g}{\partial a_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{b}_j^{(t)} \cdot \left.\frac{\partial g}{\partial b_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{c}_j^{(t)} \cdot \left.\frac{\partial g}{\partial c_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{\alpha}_j^{(t)} \cdot \left.\frac{\partial g}{\partial \alpha_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{\beta}_j^{(t)} \cdot \left.\frac{\partial g}{\partial \beta_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}},$$

where $\Delta\hat{a}_j^{(t)} = a_j - \hat{a}_j^{(t)}$, $\Delta\hat{b}_j^{(t)} = b_j - \hat{b}_j^{(t)}$, $\Delta\hat{c}_j^{(t)} = c_j - \hat{c}_j^{(t)}$, $\Delta\hat{\alpha}_j^{(t)} = \alpha_j - \hat{\alpha}_j^{(t)}$, $\Delta\hat{\beta}_j^{(t)} = \beta_j - \hat{\beta}_j^{(t)}$. If, for example, $g = \partial l / \partial a_j$, the objective function to be solved is

$$\begin{aligned} 0 = \left.\frac{\partial l}{\partial a_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{a}_j^{(t)} \cdot \left.\frac{\partial^2 l}{\partial a_j^2}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{b}_j^{(t)} \cdot \left.\frac{\partial^2 l}{\partial a_j \partial b_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{c}_j^{(t)} \cdot \left.\frac{\partial^2 l}{\partial a_j \partial c_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \cdots \\ \cdots + \Delta\hat{\alpha}_j^{(t)} \cdot \left.\frac{\partial^2 l}{\partial a_j \partial \alpha_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}} + \Delta\hat{\beta}_j^{(t)} \cdot \left.\frac{\partial^2 l}{\partial a_j \partial \beta_j}\right|_{\hat{\boldsymbol{\xi}}_j^{(t)}}. \end{aligned}$$

Applying to all item parameters analogously, a matrix form is obtained as

$$\Delta\hat{\boldsymbol{\xi}}_j^{(t)} = \begin{pmatrix} \Delta\hat{a}_j^{(t)} \\ \Delta\hat{b}_j^{(t)} \\ \Delta\hat{c}_j^{(t)} \\ \Delta\hat{\alpha}_j^{(t)} \\ \Delta\hat{\beta}_j^{(t)} \end{pmatrix} = - \begin{pmatrix} l_{11} & l_{12} & l_{13} & l_{14} & l_{15} \\ l_{21} & l_{22} & l_{23} & l_{24} & l_{25} \\ l_{31} & l_{32} & l_{33} & l_{34} & l_{35} \\ l_{41} & l_{42} & l_{43} & l_{44} & l_{45} \\ l_{51} & l_{52} & l_{53} & l_{54} & l_{55} \end{pmatrix}^{-1} \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \end{pmatrix} = - \left[\mathbf{H}_j^{(t)}\right]^{-1} \boldsymbol{\Lambda}_j^{(t)}$$

where $\mathbf{H}_j^{(t)} = \{l_{mm'}\}_{1\leq m, m' \leq 5}$ is the 5×5 Hessian matrix evaluated at $\hat{\boldsymbol{\xi}}_j^{(t)}$, and $\boldsymbol{\Lambda}_j^{(t)} = \{l_m\}_{1\leq m \leq 5}$ denotes the gradient of the marginal log-likelihood with respect to $\hat{\boldsymbol{\xi}}_j^{(t)}$. Elements of $\boldsymbol{\Lambda}_j^{(t)}$, l_m $(1 \leq m \leq 5)$, are calculated from the derivatives presented in (3.2). Let $P_{jk} = P_j(X_k)$ and $Q_{jk} = 1 - P_j(X_k)$. Using the quadrature nodes for the integral and the artificial data from the E-step, individual l_m's are obtained as

$$\begin{aligned}
l_1 &= \frac{\partial l}{\partial a_j} \approx D \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)\left(\bar{\iota}_{jkl} - P_{jk}\bar{\kappa}_{jkl}\right) \frac{(P_{jk} - c_j)}{P_{jk}(1 - c_j)}, \\
l_2 &= \frac{\partial l}{\partial b_j} \approx -Da_j \sum_{k=1}^{Q}\sum_{l=1}^{Q} \left(\bar{\iota}_{jkl} - P_{jk}\bar{\kappa}_{jkl}\right) \frac{(P_{jk} - c_j)}{P_{jk}(1 - c_j)}, \\
l_3 &= \frac{\partial l}{\partial c_j} \approx \sum_{k=1}^{Q}\sum_{l=1}^{Q} \frac{\left(\bar{\iota}_{jkl} - P_{jk}\bar{\kappa}_{jkl}\right)}{P_{jk}(1 - c_j)}, \\
l_4 &= \frac{\partial l}{\partial \alpha_j} \approx \sum_{k=1}^{Q}\sum_{l=1}^{Q} \left[\alpha_j^{-1}\bar{\kappa}_{jkl} - \alpha_j\left\{\bar{\varsigma}_{jkl} - 2(\beta_j - Y_l)\bar{\lambda}_{jkl} + (\beta_j - Y_l)^2\bar{\kappa}_{jkl}\right\}\right], \\
l_5 &= \frac{\partial l}{\partial \beta_j} \approx \alpha_j^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} \left[\bar{\iota}_{jkl} - (\beta_j - Y_l)\bar{\kappa}_{jkl}\right].
\end{aligned}$$

Equating these five equations to zero simultaneously provides an item parameter estimate that maximizes the marginalized log-likelihood in (3.3). Similarly, elements of $\mathbf{H}_j^{(t)}$ are obtained as follows.

$$
\begin{aligned}
l_{11} &= \frac{\partial^2 l}{\partial a_j^2} \approx -D^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)^2 \left(P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}\right) \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2}, \\
l_{22} &= \frac{\partial^2 l}{\partial b_j^2} \approx -D^2 a_j^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} \left(P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}\right) \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2}, \\
l_{33} &= \frac{\partial^2 l}{\partial c_j^2} \approx -\sum_{k=1}^{Q}\sum_{l=1}^{Q} \frac{(1 - 2P_{jk})\bar{\iota}_{jkl} + P_{jk}^2 \bar{\kappa}_{jkl}}{P_{jk}^2 (1 - c_j)^2}, \\
l_{44} &= \frac{\partial^2 l}{\partial \alpha_j^2} \approx -\sum_{k=1}^{Q}\sum_{l=1}^{Q} \alpha_j^{-2} \bar{\kappa}_{jkl} + \bar{\varsigma}_{jkl} - 2(\beta_j - Y_l)\bar{\lambda}_{jkl} + (\beta_j - Y_l)^2 \bar{\kappa}_{jkl}, \\
l_{55} &= \frac{\partial^2 l}{\partial \beta_j^2} \approx -\alpha_j^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\kappa}_{jkl}, \\
l_{12} &= \frac{\partial^2 l}{\partial a_j \partial b_j} \approx D \sum_{k=1}^{Q}\sum_{l=1}^{Q} \frac{(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2} \times \cdots \\
&\qquad \cdots \times \Big\{ P_{jk}(P_{jk}\bar{\kappa}_{jkl} - \bar{\iota}_{jkl})(1 - c_j) + D a_j Q_{jk}(X_k - b_j)(P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}) \Big\}, \\
l_{13} &= \frac{\partial^2 l}{\partial a_j \partial c_j} \approx -D \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)\, \bar{\iota}_{jkl} \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2}, \\
l_{23} &= \frac{\partial^2 l}{\partial b_j \partial c_j} \approx D a_i \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\iota}_{jkl} \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2}, \\
l_{45} &= \frac{\partial^2 l}{\partial \alpha_j \partial \beta_j} \approx 2\alpha_j \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\lambda}_{jkl} - (\beta_j - Y_l)\bar{\kappa}_{jkl}, \\
l_{14} &= l_{15} = l_{24} = l_{25} = l_{34} = l_{35} = 0.
\end{aligned}
$$

An updated item parameter vector at $(t+1)$-th cycle $\hat{\boldsymbol{\xi}}_j^{(t+1)}$ is obtained as

$$
\hat{\boldsymbol{\xi}}_j^{(t+1)} = \hat{\boldsymbol{\xi}}_j^{(t)} + \Delta\hat{\boldsymbol{\xi}}_j^{(t)} = \hat{\boldsymbol{\xi}}_j^{(t)} - \left[\mathbf{H}_j^{(t)}\right]^{-1} \boldsymbol{\Lambda}_j^{(t)}. \tag{3.5}
$$

Successive approximations are implemented repeatedly until the elements of $\Delta\hat{\boldsymbol{\xi}}_j^{(t)}$ become sufficiently small. The iterative algorithm based on (3.5) may fail to converge if the initial value of $\hat{\boldsymbol{\xi}}_j$ is not in the neighborhood of the true maximum. In such cases, convergence can be ensured through the use of the Fisher's scoring, in which $\mathbf{H}_j^{(t)}$ is replaced by $E\big[\mathbf{H}_j^{(t)}\big]$. For a large sample size, the Fisher's scoring usually converges to a solution faster than does the NR procedure (Kale, 1962). To compute the expected Hessian matrix, expected values are needed for the observed data:

$$E[U_{ij}] = P_{ij}, \qquad E[\log T_{ij}] = \beta_j - \tau_i, \quad \text{and} \qquad E[(\log T_{ij})^2] = \alpha_j^{-2} + (\beta_j - \tau_i)^2.$$

The expected values for the artificial data are correspondingly obtained as:

$$E[\bar{\iota}_{jkl}] = P_{jk}\, \bar{\kappa}_{jkl}, \quad E[\bar{\lambda}_{jkl}] = (\beta_j - Y_l)\, \bar{\kappa}_{jkl}, \quad \text{and} \quad E[\bar{\varsigma}_{jkl}] = \big\{\alpha_j^{-2} + (\beta_j - Y_l)^2\big\}\; \bar{\kappa}_{jkl}.$$

The expected values of the elements of the Hessian matrix are then given by

$$E[l_{11}] \approx -D^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)^2\, \bar{\kappa}_{jkl}\, \frac{Q_{jk}}{P_{jk}} \left[\frac{P_{jk} - c_j}{1 - c_j}\right]^2,$$

$$E[l_{22}] \approx -D^2 a_j^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\kappa}_{jkl}\, \frac{Q_{jk}}{P_{jk}} \left[\frac{P_{jk} - c_j}{1 - c_j}\right]^2,$$

$$E[l_{33}] \approx -\sum_{k=1}^{Q}\sum_{l=1}^{Q} \frac{Q_{jk}}{P_{jk}} \frac{\bar{\kappa}_{jkl}}{(1 - c_j)^2},$$

$$E[l_{44}] \approx -2\alpha_j^{-2} \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\kappa}_{jkl},$$

$$E[l_{55}] \approx -\alpha_j^2 \sum_{k=1}^{Q}\sum_{l=1}^{Q} \bar{\kappa}_{jkl},$$

$$E[l_{12}] \approx D^2 a_j \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)\, \bar{\kappa}_{jkl}\, \frac{Q_{jk}}{P_{jk}} \left[\frac{P_{jk} - c_j}{1 - c_j}\right]^2,$$

$$E[l_{13}] \approx -D \sum_{k=1}^{Q}\sum_{l=1}^{Q} (X_k - b_j)\, \bar{\kappa}_{jkl}\, \frac{Q_{jk}}{P_{jk}} \frac{(P_{jk} - c_j)}{(1 - c_j)^2},$$

$$E[l_{23}] \approx D a_i \sum_{k=1}^{q}\sum_{l=1}^{q} \bar{\kappa}_{jkl}\, \frac{Q_{jk}}{P_{jk}} \frac{(P_{jk} - c_i)}{(1 - c_j)^2},$$

$$E[l_{14}] = E[l_{15}] = E[l_{24}] = E[l_{25}] = E[l_{34}] = E[l_{35}] = E[l_{45}] = 0.$$

Similar to the NR procedure, the iteration process is repeated until the convergence criterion is satisfied. ML estimates asymptotically have a multivariate normal distribution with covariance matrix whose inverse is given by the information matrix. Thus, standard errors of the parameter estimates can be obtained by inverting the diagonal elements of the expected Hessian matrix.

3.1.2 Marginal Maximum a Posteriori Estimation

Although the MMLE/EM resolves the problem of inconsistent item parameter estimates, it may display undesirable qualities in some situations. First, the MML estimation lacks a means of handling unusual item response patterns such as all correct or all incorrect responses. Second, it can result in item parameter estimates that are substantially deviant from the true values because no information is given on the range of item parameters. Third, without a strong prior on the item parameters, a lack of data at the lower end of the proficiency continuum may lead to convergence problems for the 3PLM. A vehicle for preventing these instances from occurring is the use of prior information on the item parameters. Following the Bayesian approach, a posterior distribution is maximized instead of the likelihood function to draw inferences about the item parameters. The mode of the posterior distribution is known as an MMAP estimator, or equivalently, a marginal Bayesian modal estimator (Mislevy & Stocking, 1989).

The MMAP estimation makes full use of the information at the second level of the hierarchical framework. Let $f(\boldsymbol{\Xi} \,|\, \boldsymbol{\Psi})$ denote the prior density of item parameters conditioned on the hyperparameters $\boldsymbol{\Psi}$. The posterior distribution of $\boldsymbol{\Xi}$ is

$$f(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T},\, \boldsymbol{\Psi}) = \frac{L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T})\, f(\boldsymbol{\Xi} \,|\, \boldsymbol{\Psi})}{f(\mathbf{U},\, \mathbf{T})},$$

where $L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T})$ is the likelihood function, and $f(\mathbf{U},\, \mathbf{T})$ is the marginal probability of observing $(\mathbf{U},\, \mathbf{T})$. The denominator is a constant that does not depend on $\boldsymbol{\Xi}$. Therefore, one can achieve the same solution with the MMAP by maximizing

$$\begin{aligned}\log f(\boldsymbol{\Xi} \,|\, \mathbf{U}, \mathbf{T},\, \boldsymbol{\Psi}) &\propto \log L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T}) + \log f(\boldsymbol{\Xi} \,|\, \boldsymbol{\Psi}) \\ &= \sum_{i=1}^{N} \log \iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i,\, \boldsymbol{\Xi})\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i + \sum_{j=1}^{J} \log f(\boldsymbol{\xi}_j \,|\, \boldsymbol{\Psi}) \qquad (3.6) \\ &= p.\end{aligned}$$

The first component in (3.6) corresponds to the log of the marginal likelihood function given in (3.1). The second term denotes the log-likelihood for the individual item parameters. The hierarchical framework postulates the multivariate normal distribution for the joint relations among the item parameters:

$$f(\boldsymbol{\xi}_j \,|\, \boldsymbol{\Psi}) = (2\pi)^{-\frac{5}{2}} \left|\boldsymbol{\Sigma}_I\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_I)'\boldsymbol{\Sigma}_I^{-1}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_I)\right\}.$$

Some item parameters are however bounded such that $a_j \in \mathbb{R}^+$, $c_j \in [0,\, 1)$, and $\alpha_j \in \mathbb{R}^+$. To place the item parameters on the proper domains, the log, logit, and log transformations are considered for each a_j, c_j, and α_j. These are common transformations in the literature for adequately incorporating prior distributions of item parameters (e.g., Patz & Junker, 1999; van der Linden, 2007; van der Linden & Ren, 2014). Let $\boldsymbol{\xi}_j^* = (a_j^*,\, b_j,\, c_j^*,\, \alpha_j^*,\, \beta_j)$ stand for the vector of the transformed item parameters. The prior density of $\boldsymbol{\xi}_j^*$ can be rewritten as

$$\boldsymbol{\xi}_j^* = (\log a_j,\, b_j,\, \operatorname{logit} c_j,\, \log \alpha_j,\, \beta_j) \sim \boldsymbol{N}_5(\boldsymbol{\mu}_I^*,\, \boldsymbol{\Sigma}_I^*), \qquad (3.7)$$

where $\boldsymbol{\mu}_I^*$ and $\boldsymbol{\Sigma}_I^*$ are the mean vector and the covariance matrix for the transformed item parameters. The $\log a_j$ has a normal prior distribution with a mean μ_{a^*} and a variance $\sigma_{a^*}^2$, which translates into a normal distribution for a_j with a mean

$$\mu_a = \exp\left(\mu_{a^*} + \frac{\sigma_{a^*}^2}{2}\right)$$

and a variance

$$\sigma_a^2 = \exp\left(2\mu_{a^*} + \sigma_{a^*}^2\right)\left(\exp(\sigma_{a^*}^2) - 1\right).$$

Analogous argument holds for α_j and α_j^*. The logit transformation of c_j does not have closed form solutions for the mean and variance; instead, they can be obtained empirically from the data from which prior information is drawn.

The covariance between the individual item parameters can be calculated based on Stein's lemma (Stein, 1981). The covariance between a_j and b_j, for instance, is obtained as

$$\mathrm{Cov}(a_j,\, b_j) = \mu_a \mathrm{Cov}(a_j^*,\, b_j).$$

Likewise, the covariance between a_j and α_j is calculated as

$$\mathrm{Cov}(a_j,\, \alpha_j) = \mu_a \mu_\alpha \mathrm{Cov}(a_j^*,\, \alpha_j^*).$$

The MMAP estimator for the j-th item parameters is obtained by simultaneously solving the equation

$$\frac{\partial p}{\partial \boldsymbol{\xi}_j^*} = \left(\frac{\partial p}{\partial a_j^*},\, \frac{\partial p}{\partial b_j},\, \frac{\partial p}{\partial c_j^*},\, \frac{\partial p}{\partial \alpha_j^*},\, \frac{\partial p}{\partial \beta_j}\right)' = \mathbf{0}. \tag{3.8}$$

Equation (3.8) is solved for separate items due to the independence among items. Explicit expressions for each element are given as follows.

$$\begin{aligned}
\frac{\partial p}{\partial a_j^*} &= Da_j\left[\sum_{i=1}^{N} \iint (u_{ij} - P_{ij})(\theta_i - b_j)\frac{(P_{ij} - c_j)}{P_{ij}(1 - c_j)} f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i\right] - \boldsymbol{v}_1(\boldsymbol{\xi}_j^* - \boldsymbol{\mu}_I^*),\\
\frac{\partial p}{\partial b_j} &= -Da_j\left[\sum_{i=1}^{N} \iint (u_{ij} - P_{ij})\frac{(P_{ij} - c_j)}{P_{ij}\,(1 - c_j)}\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i\right] - \boldsymbol{v}_2\,(\boldsymbol{\xi}_j^* - \boldsymbol{\mu}_I^*),\\
\frac{\partial p}{\partial c_j^*} &= c_j\left[\sum_{i=1}^{N} \iint \frac{(u_{ij} - P_{ij})}{P_{ij}} p(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i\right] - \boldsymbol{v}_3\,(\boldsymbol{\xi}_j^* - \boldsymbol{\mu}_I^*),\\
\frac{\partial p}{\partial \alpha_j^*} &= \left[\sum_{i=1}^{N} \iint \left[1 - \alpha_j^2\left\{\log t_{ij} - (\beta_j - \tau_i)\right\}^2\right]\, p(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i\right] - \boldsymbol{v}_4(\boldsymbol{\xi}_j^* - \boldsymbol{\mu}_I^*),
\end{aligned}$$

$$\frac{\partial p}{\partial \beta_j} = \alpha_j^2 \left[\sum_{i=1}^{N} \iint \left\{ \log t_{ij} - (\beta_j - \tau_i) \right\} p(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega})\, d\theta_i\, d\tau_i \right] - \boldsymbol{\upsilon}_5 \, (\boldsymbol{\xi}_j^* - \boldsymbol{\mu}_I^*),$$

where $\boldsymbol{\upsilon}_m = \{\upsilon_{mm'}\}_{1 \le m' \le 5}$ is the m-th column vector of $\left[\boldsymbol{\Sigma}_I^*\right]^{-1}$. As with the MML estimation, the equations above involve unknown quantities resulting from the incidental parameters, and hence, they must be solved through the EM algorithm accompanied by an iterative procedure such as the NR or Fisher's scoring. The ensuing subsection presents the detailed procedure for implementing the MMAPE/EM procedure.

Computational Methods for MMAPE/EM

The MMAPE/EM algorithm is implemented in an analogous manner with the MMLE/EM. Several corrections are made to adjust the transformation of item parameters as well as to incorporate the prior density. Let $\hat{\boldsymbol{\xi}}_j^{*(t)}$ denote the t-th approximation to the true value of $\boldsymbol{\xi}_j^*$ that maximizes $\log p$. A better approximation $\hat{\boldsymbol{\xi}}_j^{*(t+1)}$ is obtained as

$$\hat{\boldsymbol{\xi}}_j^{*(t+1)} = \hat{\boldsymbol{\xi}}_j^{*(t)} - \left[\mathbf{H}_j^{*(t)}\right]^{-1} \boldsymbol{\Lambda}_j^{*(t)}, \tag{3.9}$$

where $\mathbf{H}_j^{*(t)}$ is the 5×5 Hessian matrix evaluated at $\hat{\boldsymbol{\xi}}_j^{*(t)}$, and $\boldsymbol{\Lambda}_j^{*(t)}$ denotes the first-order derivatives of the posterior distribution with respect to $\hat{\boldsymbol{\xi}}_j^{*(t)}$. Based on the expected values from the E-step—$\bar{\kappa}_{jkl}$, $\bar{\iota}_{jkl}$, $\bar{\lambda}_{jkl}$, and $\bar{\varsigma}_{jkl}$—, elements of $\boldsymbol{\Lambda}_j^{*(t)}$, l_m^* $(1 \le m \le 5)$, are obtained as follows.

$$\begin{aligned}
l_1^* &= \frac{\partial p}{\partial a_j^*} \approx D a_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j) \left(\bar{\iota}_{jkl} - P_{jk} \bar{\kappa}_{jkl} \right) \frac{(P_{jk} - c_j)}{P_{jk}(1 - c_j)} \right] - \boldsymbol{\upsilon}_1 (\boldsymbol{\xi}_j^{*(t)} - \boldsymbol{\mu}_I^*), \\
l_2^* &= \frac{\partial p}{\partial b_j} \approx -D a_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \left(\bar{\iota}_{jkl} - P_{jk} \bar{\kappa}_{jkl} \right) \frac{(P_{jk} - c_j)}{P_{jk}(1 - c_j)} \right] - \boldsymbol{\upsilon}_2 (\boldsymbol{\xi}_j^{*(t)} - \boldsymbol{\mu}_I^*), \\
l_3^* &= \frac{\partial p}{\partial c_j^*} \approx c_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \frac{\left(\bar{\iota}_{jkl} - P_{jk} \bar{\kappa}_{jkl} \right)}{P_{jk}} \right] - \boldsymbol{\upsilon}_3 (\boldsymbol{\xi}_j^{*(t)} - \boldsymbol{\mu}_I^*),
\end{aligned}$$

$$l_4^* = \frac{\partial p}{\partial \beta_j} \approx \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} - \alpha_j^2 \{ \bar{\varsigma}_{jkl} - 2(\beta_j - Y_l) \bar{\lambda}_{jkl} + (\beta_j - Y_l)^2 \bar{\kappa}_{jkl} \} \right] - \boldsymbol{\upsilon}_4 (\boldsymbol{\xi}_j^{*(t)} - \boldsymbol{\mu}_I^*),$$

$$l_5^* = \frac{\partial p}{\partial \alpha_j^*} \approx \alpha_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\iota}_{jkl} - (\beta_j - Y_l) \bar{\kappa}_{jkl} \right] - \boldsymbol{\upsilon}_5 (\boldsymbol{\xi}_j^{*(t)} - \boldsymbol{\mu}_I^*).$$

Let $\upsilon_{mm'}$ denote the $(m,\, m')$-th entry of the matrix $\left[\boldsymbol{\Sigma}_I\right]^{-1}$. Elements of $\mathbf{H}_j^{*(t)}$, $l_{mm'}^*$ $(1 \le m,\, m' \le 5)$, are obtained as follows.

$$l_{11}^* = \frac{\partial^2 p}{\partial a_j^{*2}} \approx -D^2 a_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j)^2 \, (P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}) \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2} \right] - \upsilon_{11},$$

$$l_{22}^* = \frac{\partial^2 p}{\partial b_j^2} \approx -D^2 a_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}) \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2} \right] - \upsilon_{22},$$

$$l_{33}^* = \frac{\partial^2 p}{\partial c_j^{*2}} \approx -c_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \frac{(1 - 2P_{jk}) \bar{\iota}_{jkl} + P_{jk}^2 \bar{\kappa}_{jkl}}{P_{ik}^2} \right] - \upsilon_{33},$$

$$l_{44}^* = \frac{\partial^2 p}{\partial \alpha_j^{*2}} \approx -2\alpha_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\varsigma}_{jkl} - 2(\beta_j - Y_l) \bar{\lambda}_{jkl} + (\beta_j - Y_l)^2 \bar{\kappa}_{jkl} \right] - \upsilon_{44},$$

$$l_{55}^* = \frac{\partial^2 p}{\partial \beta_j^2} \approx -\alpha_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} \right] - \upsilon_{55},$$

$$l_{12}^* = \frac{\partial^2 p}{\partial a_j^* \partial b_j} \approx D a_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \frac{(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)^2} \times \right.$$

$$\left. \cdots \times \left\{ P_{jk} (P_{jk} \bar{\kappa}_{jkl} - \bar{\iota}_{jkl})(1 - c_j) + D a_j Q_{jk} (X_k - b_j)(P_{jk}^2 \bar{\kappa}_{jkl} - c_j \bar{\iota}_{jkl}) \right\} \right] - \upsilon_{12},$$

$$l_{13}^* = \frac{\partial^2 p}{\partial a_j^* \partial c_j^*} \approx -D a_j c_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j) \, \bar{\iota}_{jkl} \, \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)} \right] - \upsilon_{13},$$

$$l_{23}^* = \frac{\partial^2 p}{\partial b_j \partial c_j^*} \approx D a_j c_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\iota}_{jkl} \frac{Q_{jk}(P_{jk} - c_j)}{P_{jk}^2 (1 - c_j)} \right] - \upsilon_{23},$$

$$l_{45}^* = \frac{\partial^2 p}{\partial \alpha_j^* \partial \beta_j} \approx 2\alpha_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\lambda}_{jkl} - (\beta_j - Y_l) \, \bar{\kappa}_{jkl} \right] - \upsilon_{45}$$

otherwise, $l_{mm'}^* = -\upsilon_{mm'}$.

In case of the Fisher's scoring method, following expected values are used instead of the above expressions.

$$
\begin{aligned}
E[l^*_{11}] &\approx -D^2 a_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j)^2 \, \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \, \frac{(P_{jk} - c_j)^2}{(1 - c_j)^2} \right] - \upsilon_{11}, \\
E[l^*_{22}] &\approx -D^2 a_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \, \frac{(P_{jk} - c_j)^2}{(1 - c_j)^2} \right] - \upsilon_{22}, \\
E[l^*_{33}] &\approx -c_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \right] - \upsilon_{33}, \\
E[l^*_{44}] &\approx -2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{ikl} \right] - \upsilon_{44}, \\
E[l^*_{55}] &\approx -\alpha_i^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} \right] - \upsilon_{55}, \\
E[l^*_{12}] &\approx D^2 a_j^2 \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j) \, \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \, \frac{(P_{jk} - c_j)^2}{(1 - c_j)^2} \right] - \upsilon_{12}, \\
E[l^*_{13}] &\approx -D a_j c_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} (X_k - b_j) \, \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \, \frac{(P_{jk} - c_j)}{(1 - c_j)} \right] - \upsilon_{13}, \\
E[l^*_{23}] &\approx D a_j c_j \left[\sum_{k=1}^{Q} \sum_{l=1}^{Q} \bar{\kappa}_{jkl} \, \frac{Q_{jk}}{P_{jk}} \, \frac{(P_{jk} - c_j)}{(1 - c_j)} \right] - \upsilon_{23},
\end{aligned}
$$

otherwise, $E[l^*_{mm'}] = -\upsilon_{mm'}$.

3.2 Person Parameter Estimation

Once item parameters are estimated accurately, the estimated values can be treated as known such that person parameters can be estimated with the known item parameter values. In this section, examinees' latent trait parameters are jointly estimated under the known item parameter values. The posterior distribution of (θ_i, τ_i) is calculated as

$$
f(\theta_i, \tau_i \,|\, \boldsymbol{u}_i, \boldsymbol{t}_i, \boldsymbol{\Omega}) = \frac{f(\boldsymbol{u}_i, \boldsymbol{t}_i \,|\, \theta_i, \tau_i) \, f(\theta_i, \tau_i \,|\, \boldsymbol{\Omega})}{f(\boldsymbol{u}_i, \boldsymbol{t}_i)}, \tag{3.10}
$$

where

$$f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega}) = (2\pi)^{-1}\, |\boldsymbol{\Sigma}_P|^{-1/2}\, \exp\left\{ -\frac{1}{2}\big((\theta_i,\, \tau_i)' - \boldsymbol{\mu}_P\big)' \,\boldsymbol{\Sigma}_P^{-1}\, \big((\theta_i,\, \tau_i)' - \boldsymbol{\mu}_P\big) \right\}.$$

The marginal density $f(\boldsymbol{u}_i,\, \boldsymbol{t}_i)$ does not depend on $(\theta_i,\, \tau_i)$, and hence, the person parameters can be estimated by maximizing the numerator only. Assuming the conditional independence of responses and response times given the person parameters (van der Linden & Glas, 2010), the posterior distribution of $(\theta_i,\, \tau_i)$ is rewritten as

$$f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega}) \propto f(\boldsymbol{u}_i \,|\, \theta_i)\, f(\boldsymbol{t}_i \,|\, \tau_i)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega}). \tag{3.11}$$

Point estimates of the person parameters are found as the mean or the mode of the posterior distribution, each of which leads to the EAP and the MAP estimators. Alternatively, one can maximize the likelihood component by fixing the term pertaining to the prior density equal to one, wherein the ML estimates are obtained for the latent trait parameters. Because the MAP estimator subsumes the ML estimator, our attention will be restricted to attainment of MAP and EAP estimates.

Substituting each corresponding component in (3.11), it can be found that the log of posterior density is proportional to

$$\begin{aligned} \log\, f = &\sum_{j=1}^{J} u_{ij} \log P_{ij} + (1 - u_{ij}) \log Q_{ij} + \log \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} - \frac{\alpha_j^2}{2}\Big[\log t_{ij} - (\beta_j - \tau_i)\Big]^2 \cdots \\ &\qquad - \frac{\sigma_\tau^2(\theta - \mu_\theta)^2 - 2\sigma_{\theta\tau}(\theta - \mu_\theta)(\tau - \mu_\tau) + \sigma_\theta^2(\tau - \mu_\tau)^2}{2(\sigma_\theta^2\sigma_\tau^2 - \sigma_{\theta\tau}^2)}. \end{aligned}$$

Let $(\hat{\theta}_i^{(t)},\, \hat{\tau}_i^{(t)})$ denote the t-th approximation to the true values of $(\theta_i,\, \tau_i)$. An updated approximation is obtained as

$$\begin{pmatrix} \hat{\theta}_i^{(t+1)} \\ \hat{\tau}_i^{(t+1)} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_i^{(t)} \\ \hat{\tau}_i^{(t)} \end{pmatrix} - \left\{ E\left[\begin{pmatrix} \dfrac{\partial^2 \log f}{\partial \theta_i^2} & \dfrac{\partial^2 \log f}{\partial \theta_i \partial \tau_i} \\ \dfrac{\partial^2 \log f}{\partial \tau_i \partial \theta_i} & \dfrac{\partial^2 \log f}{\partial \tau_i^2} \end{pmatrix} \right] \right\}^{-1} \begin{pmatrix} \dfrac{\partial \log f}{\partial \theta_i} \\ \dfrac{\partial \log f}{\partial \tau_i} \end{pmatrix}, \tag{3.12}$$

where

$$\begin{aligned}
\frac{\partial \log f}{\partial \theta_i} &= \left[\sum_{j=1}^{J} \frac{D a_j (u_{ij} - P_{ij})(P_{ij} - c_j)}{P_{ij}(1 - c_j)} \right] - \frac{\sigma_\tau^2(\theta_i - \mu_\theta) - \sigma_{\theta\tau}(\tau_i - \mu_\tau)}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\
\frac{\partial \log f}{\partial \tau_i} &= -\left[\sum_{j=1}^{J} \alpha_i^2 \{ \log t_{ij} - (\beta_j - \tau_i) \} \right] - \frac{\sigma_\theta^2(\tau_i - \mu_\tau) - \sigma_{\theta\tau}(\theta_i - \mu_\theta)}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\
\frac{\partial^2 \log f}{\partial \theta_i^2} &= \left[\sum_{j=1}^{J} \frac{D^2 a_j^2 Q_{ij} (P_{ij} - c_j)(c_j u_{ij} - P_{ij}^2)}{P_{ij}^2 (1 - c_j)^2} \right] - \frac{\sigma_\tau^2}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\
\frac{\partial^2 \log f}{\partial \tau_i^2} &= -\left[\sum_{j=1}^{J} \alpha_j^2 \right] - \frac{\sigma_\theta^2}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}, \\
\frac{\partial^2 \log f}{\partial \theta_i \partial \tau_i} &= \frac{\sigma_{\theta\tau}}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}.
\end{aligned}$$

The iterative process continues until the changes between the two successive approximations become sufficiently small. The procedure presented in (3.12) is based on the concept of Fisher's scoring in which the Hessian matrix is replaced with its expected values. This procedure ensures convergence when initial values of $(\hat{\theta}_i, \hat{\tau}_i)$ are not in the neighborhood of the true maximum. Intead of taking the expectation of the Hessian matrix, one can obtain the estimates based on the NR method as well. In either cases, standard errors of $(\hat{\theta}_i, \hat{\tau}_i)$ can be approximately computed as the square root of the diagonal elements of the inverse of the negative of the Hessian evaluated at the MAP estimates.

Unlike the MAP, the EAP estimation does not require an iterative procedure; it requires an approximation of the integral instead. The EAP estimate of the latent traits is obtained as

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\tau}_i \end{pmatrix} = \frac{\iint (\theta_i,\, \tau_i)' f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i d\tau_i}{\iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i d\tau_i}, \tag{3.13}$$

which can be reasonably approximated using the Gauss-Hermite quadrature nodes as follows.

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\tau}_i \end{pmatrix} \approx \frac{\sum_k \sum_l (X_k,\, Y_l)' f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, X_k,\, Y_l)\, \omega(X_k,\, Y_l)}{\sum_k \sum_l f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, X_k,\, Y_l)\, \omega(X_k,\, Y_l)},$$

where X_k and Y_l denote the finite quadrature nodes, and $\omega(X_k,\, Y_l)$ is the weight corresponding to the bivariate normal distribution with prior of $\boldsymbol{\Omega}$.

3.3 Simulation Study

A series of simulation studies are conducted to evaluate the performance of the estimation procedures under the systematic variation of design factors. The design variables of interest include 1) the calibration sample size (N), 2) the correlation between the item parameters $\big(\rho_I = \frac{\sigma_{mm'}}{\sigma_m \sigma_{m'}}\ (1 \leq m,\, m' \leq 5)\big)$, and 3) the correlation between the person parameters $\big(\rho_P = \frac{\sigma_{\theta\tau}}{\sigma_\theta \sigma_\tau}\big)$. The first study evaluates the performance of the MMLE/EM and MMAPE/EM procedures in recovering the item parameters within the hierarchical framework. The second study examines the appropriateness of MAP and EAP estimators for recovering the examinees' latent traits. The third study evaluates the robustness of the estimation procedures under an improper prior.

3.3.1 Item Parameter Estimation

The present subsection examines the performance of the likelihood-based methods—MMLE/EM and MMAPE/EM—for estimating the item parameters within the hierarchical framework. For reference purposes, the estimation methods were applied to calibrate items for the 3PLM, and results were compared between the two modeling frameworks. Recovery of the item parameters was evaluated with respect to 1) estimation convergence, 2) mean squared error (MSE), 3) bias, and 4) retrievability of correlation between the item parameters. The convergence rate was calculated as the proportion of successfully converged items out of J items in each individual test. The MSE and bias criteria were calculated as

$$\text{MSE} = \frac{1}{J}\sum_{j=1}^{J}\left(\xi_{jm} - \hat{\xi}_{jm}\right)^2, \quad \text{and} \qquad \text{Bias} = \frac{1}{J}\sum_{j=1}^{J}\left(\xi_{jm} - \hat{\xi}_{jm}\right),$$

where ξ_{jm} is the true value of the m-th parameter of item j, and $\hat{\xi}_{jm}$ is its estimated value. Finally, the recovery of the correlation between the item parameters was evaluated by the Pearson correlation between the estimated item parameter values.

The study drawed on simulated data. To mimic large-scale educational testing so far as possible, item parameters were assumed to follow a multivariate normal distribution with means, (-0.043, 0, -1.386, -0.043, 0), and variances, (0.086, 1, 0.040, 0.086, 1), for $\log a$, b, logit$c$, $\log \alpha$, and β, respectively. The means and variances of the item parameters on the original scale correspond to (1, 0, 0.2, 1, 0) and (0.09, 1, 0.001, 0.09, 1). Three levels of dependencies were considered by varying the correlation between the item parameters (ρ_I) as 0.0, 0.4, and 0.8. Within each condition, the item parameters shared the same value of ρ_I to get a clear picture of the impact of different levels of dependencies. To place the item parameter values on the reasonable domains, upper and lower bounds were set such that a, $\alpha \in [0.3, 2]$; b, $\beta \in [-3.5, 3.5]$; and $c \in [0.1, 0.3]$. The test length (J) was fixed at 30.

Examinees' latent trait variables were randomly sampled from a bivariate normal distribution with zero means and unit variances. To investigate the impact of dependency between the person parameters, three levels of correlations were considered: $\rho_P = 0$, 0.3, and 0.6. The values chosen for ρ_P were motivated by the review of van der Linden (2009), where empirical estimates of ρ_P from large-scale operational assessments were found to have values between -0.65 and 0.30. Because the sign of ρ_P has no impact on the amount of information in estimating the parameters, only the positive values were considered. The size of calibration samples (N) were differed as 1000 and 2000. Crossing each condition yielded 72 simulation scenarios—two models $\times$ two estimation methods $\times$ two N's $\times$ three ρ_I's $\times$ three ρ_P's. Within each scenario, 100 replications were made to eliminate the random sampling error.

When estimating the item parameters, convergence criteria are needed to stop the iterative procedure. The EM algorithm and the NR procedure were terminated when the difference between the estimated values from two consecutive iterations was small enough ($< 10^{-3}$). Convergence referred to as a situation where all item parameter values estimated fell in the proper domains within the EM cycles no greater than the maximum (100). Initialization of the EM algorithm was made based on following approximations.

(i) $\hat{a}_j^{(1)} = \sqrt{\dfrac{\rho_{\theta U}^2}{1-\rho_{\theta U}^2}}$, where $\rho_{\theta U}$ is the biserial correlation between ability θ and item response variable U.

(ii) $\hat{b}_j^{(1)} = F^{-1}\left(1-\bar{p}_j \,|\, \mu_P,\, \sigma_P\right)$, where $\bar{p}_j$ is the proportion correct across the examinee sample, and $1-\bar{p}_j = F\left(\hat{b}_j^{(1)} \,|\, \mu_P,\, \sigma_P\right) = \dfrac{1}{\sigma_P\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\hat{b}_j^{(1)}} \exp\left\{-\dfrac{(t-\mu_P)^2}{2\sigma_P^2}\right\} dt.$

(iii) $\hat{c}_j^{(1)} = 0.2.$

(iv) $\hat{\alpha}_j^{(1)} = \left(\mathrm{Var}\left[\log T_j\right]\right)^{-\frac{1}{2}}.$

(v) $\hat{\beta}_j^{(1)} = E\left[\log T_j\right].$

The initial values for $\hat{a}_j$ and $\hat{b}_j$ were obtained based on the relation found from the normal-ogive model (Richardson, 1936; Tucker, 1946). The initival values of $\hat{c}_j$ were fixed to a known constant (e.g., the reciprocal of the number of answer choices). The values for $\hat{\alpha}_j^{(1)}$ and $\hat{\beta}_j^{(1)}$ were obtained using the relation in (2.5).

Results

Convergence Rates. Table 3.1 reports convergence rates of the estimation methods under the simulated conditions. The values presented in Table 3.1 are averaged ones across the replications. From Table 3.1, it is clear that MMAP estimation was much more successful

Table 3.1: Estimation Convergence Rates

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	.444	.552	.414	.540	.433	.539
		3PLM	.645	.744	.607	.721	.635	.750
	0.3	HF	.457	.554	.440	.536	.447	.547
		3PLM	.647	.744	.619	.725	.636	.744
	0.6	HF	.469	.546	.436	.537	.449	.554
		3PLM	.642	.744	.606	.730	.626	.737
MMAP	0.0	HF	.999	.999	.999	.999	1	1
		3PLM	1	.999	.999	1	1	.999
	0.3	HF	.999	.998	1	.999	.999	1
		3PLM	1	.999	1	.999	.999	.999
	0.6	HF	.999	1	.999	.999	1	1
		3PLM	1	1	1	.999	1	.999

Note: ρ_I = correlation between item parameters. N = calibration sample size. ρ_P = correlation between person parameters. HF = hierarchical framework. 3PLM = three-parameter logistic model.

compared to MML estimation in terms of convergence. MMLE/EM showed subpar convergence rates across all simulation scenarios. It should be noted that the convergence criterion defined in this study was stringent in that all estimated values were required to be within the pre-specified domains. Most convergence problems in the MML estimation were in fact attributed to the guessing parameter estimates falling outside the pre-specified interval [0.1, 0.3]. Incorporating the prior distribution in the estimation procedure did appear improved convergence performance in estimation of guessing. Overall, increasing the N, ρ_P, or ρ_I resulted in more successful convergence in MML estimation. Compared to the hieararchical framework, estimating the 3PLM alone resulted in higher convergence rates mainly due to the smaller number of parameters to be estimated and converged.

MSE. Presented in Tables 3.2 and 3.3 are MSEs of $\hat{a}$ and $\hat{b}$ for items successfully converged. Overall, MSEs of these estimates were reasonably small, indicating that the true parameter values were recovered well. The b parameters were more accurately recovered than a param-

Table 3.2: Mean Squared Error of Discrimination Parameter Estimates

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	.029	.016	.029	.016	.030	.016
		3PLM	.038	.022	.043	.025	.042	.024
	0.3	HF	.026	.015	.033	.015	.028	.015
		3PLM	.037	.019	.040	.023	.043	.022
	0.6	HF	.027	.014	.032	.014	.031	.014
		3PLM	.041	.022	.044	.025	.043	.023
MMAP	0.0	HF	.018	.012	.018	.012	.012	.008
		3PLM	.018	.011	.019	.012	.015	.010
	0.3	HF	.018	.011	.019	.012	.012	.008
		3PLM	.018	.011	.020	.012	.015	.010
	0.6	HF	.017	.012	.018	.012	.012	.008
		3PLM	.017	.012	.019	.013	.015	.010

eters as shown by smaller MSEs in Table 3.3. Results for $\hat{c}$, $\hat{\alpha}$, and $\hat{\beta}$ were not tabulated because MSEs observed for these estimates were too small to attach any practical meaning. (The maximal MSEs observed for each estimator were 0.003, 0.001, and 0.001.)

The MSE results with respect to the design variables were consistent with expectations in several aspects. Increasing the N or ρ_I improved the MSE statistics of both $\hat{a}$ and $\hat{b}$. Incorporating an informative prior into the item calibration process led to smaller MSEs and thus, more accurate estimates. The differing levels of ρ_P appeared to have minor influence on the estimation of item parameters. A probable cause for this result is the marginalization of the latent trait distributions in item calibration.

Closer examination of Table 3.2 reveals that calibrating the response model jointly with the response time model led to substantial decrease in MSEs for $\hat{a}$. Compared to when the 3PLM was estimated alone, MML estimation of the hierarchical framework resulted in 28.81%, 26.11%, 30.31% reduction in MSEs of $\hat{a}$ under N = 1000, and 27.15%, 38.04%; and 34.12% reduction under N = 2000, along with increasing ρ_I. This tendency was less evident when the parameters were estimated via MMAP. When the item parameters were

Table 3.3: Mean Squared Error of Difficulty Parameter Estimates

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	.021	.013	.019	.014	.018	.015
		3PLM	.021	.013	.019	.014	.018	.013
	0.3	HF	.020	.015	.021	.014	.019	.013
		3PLM	.020	.013	.021	.013	.018	.013
	0.6	HF	.021	.013	.020	.015	.018	.014
		3PLM	.022	.014	.022	.014	.018	.013
MMAP	0.0	HF	.014	.009	.013	.007	.010	.006
		3PLM	.014	.009	.014	.008	.012	.008
	0.3	HF	.013	.008	.013	.007	.010	.006
		3PLM	.013	.008	.014	.008	.012	.007
	0.6	HF	.014	.009	.012	.008	.009	.006
		3PLM	.014	.009	.013	.008	.011	.007

uncorrelated (i.e., $\rho_I = 0.0$), MMAP estimation of the hierarchical framework resulted in slight increase in MSEs. As the item parameters were correlated with $\rho_I = 0.4$ or 0.8, calibrating the hierarchical framework with MMAP showed consistent decrease in MSEs. While the proportional reduction in error provides a sense of relative performances between the two conditions, it should not be considered as an absolute measure of comparing the two cases. Results for MMAP estimator in Table 3.2, for example, had so small MSE values on the original scale that making inference based on the proportional reductions may falsely amplify the relative performances.

The similar argument can be followed for MSEs of $\hat{b}$'s in Table 3.3. The original MSE values observed for $\hat{b}$ were very small. Estimating the hierarchical framework via MMAP resulted in marginally smaller MSEs than estimating the 3PLM separately. When the MML estimator was used, estimating the 3PLM alone led to slightly smaller MSEs in $\hat{b}$'s. The differences between the two conditions were on average less than 0.002. Overall, the two tables seemed to suggest that improvements in MSEs as a result of calibrating the hierarchical framework could be made only when the item parameters have nonzero correlations and they are estimated through the MMAP.

Although not presented in the tables, stability of the error statistics across the replications deserves comments. Standard deviations (SDs) of the MSE values for the MML estimator of a ranged from 0.016 to 0.021 when $N = 1000$, and from 0.009 to 0.012 when $N = 2000$. The MMAP estimator in the meantime showed SDs between 0.005 and 0.006 when $N = 1000$, and between 0.003 and 0.005 when $N = 2000$. The result indicates that the estimation method was the most prominent factor influencing the MSE performances, followed by the sample size. The MSE results for $\hat{b}$ displayed the similar patterns. When b parameters were estimated via MML, the MSE values had SDs between 0.008 and 0.012 ($N = 1000$) or between 0.006 and 0.007 ($N = 2000$). When the MMAP estimation was carried out, the values of SDs were between 0.004 and 0.006 ($N = 1000$) or between 0.003 and 0.004 ($N = 2000$), generally supporting the findings from the overall MSEs for $\hat{a}$.

Bias. Tables 3.4 and 3.5 report biases of $\hat{a}$ and $\hat{b}$. Results for $\hat{c}$, $\hat{\alpha}$, and $\hat{\beta}$ were omitted due to small errors and minor impact of the design variables. (The largest biases in absolute value were less than 0.005, 0.002, and 0.002 for each of $\hat{c}$, $\hat{\alpha}$, and $\hat{\beta}$.) The tables reported for $\hat{a}$ and $\hat{b}$ suggest that the parameter estimates were biased only to a very small degree. The estimates obtained from MMAP appeared almost unbiased as shown by the values occurred at the third decimal place. The MML estimator resulted in slightly larger biasedness; yet, the maximal bias observed in both tables was never greater than 0.04.

In Table 3.4, the overall biasedness of the MML estimator of a was found as 0.023 when $N = 1000$, and 0.011 when $N = 2000$, suggesting the decrease in bias along with the increase in the sample size. Under the same conditions, the MMAP estimator of a produced the overall bias of 0.004 and 0.002 as N increased from 1000 to 2000. The two estimators yielded slightly smaller biases for estimating b. The MML estimator showed the overall bias of 0.010 and 0.008 along with increasing N, whereas the MMAP estimator showed the bias of 0.003 and 0.002 at the same time. Across the bias results, no consistent patterns could be detected in regards to changes in the models being estimated, ρ_I, and ρ_P.

Table 3.4: Bias of Discrimination Parameter Estimates

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	-.021	-.005	-.016	-.006	-.024	-.009
		3PLM	-.034	-.017	-.022	-.014	-.028	-.014
	0.3	HF	-.013	-.008	-.014	-.004	-.028	-.011
		3PLM	-.025	-.015	-.026	-.013	-.017	-.016
	0.6	HF	-.016	-.009	-.025	-.009	-.014	-.005
		3PLM	-.036	-.019	-.030	-.019	-.016	-.011
MMAP	0.0	HF	-.002	-.002	-.006	-.006	-.002	.000
		3PLM	-.002	-.002	-.005	-.005	-.004	-.001
	0.3	HF	.001	-.001	-.008	-.003	.001	-.001
		3PLM	.001	.000	-.007	-.002	.000	-.002
	0.6	HF	-.003	-.003	-.010	-.006	.001	.001
		3PLM	-.002	-.002	-.008	-.006	.001	.000

Table 3.5: Bias of Difficulty Parameter Estimates

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	-.010	.000	-.008	-.007	-.013	-.015
		3PLM	-.003	.000	-.006	-.006	-.018	-.013
	0.3	HF	-.001	-.001	-.008	-.009	-.017	-.013
		3PLM	-.005	-.002	-.013	-.010	-.018	-.014
	0.6	HF	-.004	-.004	-.011	-.008	-.011	-.011
		3PLM	-.009	-.003	-.015	-.007	-.016	-.014
MMAP	0.0	HF	.006	.003	.001	.002	-.001	.000
		3PLM	.006	.003	.001	.002	-.003	-.001
	0.3	HF	.004	.003	.002	.001	-.001	-.001
		3PLM	.004	.003	.002	.001	-.003	-.002
	0.6	HF	.005	.002	.000	.000	-.001	-.001
		3PLM	.005	.002	.000	.001	-.002	-.002

The overall SD results for the biases indicated the similar trends with those for the MSEs. SDs of the biases resulting from the MML estimates of a were on average 0.042 and 0.026 under each $N = 1000$ and 2000 condition. The MMAP estimator of a resulted in overall SDs of 0.021 and 0.015 along with the increase in N. The results reported for $\hat{b}$ had the average SDs of 0.032 and 0.025 when the MML estimator was used, and SDs of 0.015 and 0.012 when the MMAP estimator was used, as the N increased from 1000 to 2000.

Correlation between Item Parameters. Correlations between the estimated item parameters are provided in Table 3.6. The results were obtained by averaging the correlations between each pair of item parameters. That is, when the 3PLM was calibrated, three correlation values—ρ_{ab}, ρ_{bc}, and ρ_{ac}—were averaged; when the hierarchical framework was calibrated, ten correlation values obtained from pairs of five item parameters were averaged.

Table 3.6 suggests that MML estimation of the item parameters was generally unsuccessful in recovering the true values of correlations. Although the correlations computed from the estimated item parameters increased as ρ_I increased and became closer to the true values as N increased, the overall recovery of the true correlation level was inferior to that of the MMAP estimation. Incorporation of the prior density in the item calibration indeed resulted in the better recovery of the true correlation levels among the item parameters. Some clear differences related to N were likewise apparent; the larger the N, the more accurate recovery of ρ_I. The impact of different choices of models was manifested by SDs of the observed correlations. Calibrating the hierarchical framework showed more consistent performances in predicting the ρ_I compared to the 3PLM calibration. The MML estimation of the hierarchical framework, for instance, showed SDs of 0.105 and 0.098 across the increasing N while the MML estimation of the 3PLM showed SDs of 0.136 and 0.137 under the same settings. When the MMAP estimator was used, calibration of the hierarchical framework showed SDs of 0.052 and 0.054; and estimating of the 3PLM showed SDs of 0.075 and 0.081 as N increased from 1000 to 2000.

Table 3.6: Correlation between Estimated Item Parameters

			$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
			N		N		N	
Method	ρ_P	Model	1000	2000	1000	2000	1000	2000
MML	0.0	HF	.104	.106	.223	.266	.383	.433
		3PLM	.087	.047	.195	.223	.392	.454
	0.3	HF	.099	.103	.231	.256	.390	.435
		3PLM	.060	.043	.203	.217	.384	.443
	0.6	HF	.092	.105	.233	.242	.388	.435
		3PLM	.060	.047	.181	.233	.383	.441
MMAP	0.0	HF	.085	.094	.530	.516	.859	.847
		3PLM	-.119	-.084	.504	.492	.894	.874
	0.3	HF	.079	.088	.535	.517	.856	.848
		3PLM	-.119	-.097	.533	.490	.891	.875
	0.6	HF	.088	.091	.531	.516	.857	.846
		3PLM	-.113	-.093	.517	.495	.894	.871

Note: The averages of true correlations under each ρ_I = 0.0, 0.4, and 0.8 condition were 0.000, 0.384, and 0.774.

3.3.2 Person Parameter Estimation

Recovering examinees' true latent trait levels from a test is always a major concern. In this subsection, the accuracy of MAP and EAP estimators is evaluated treating the item parameter estimates obtained from the previous study as known values. Because MML estimators had difficulty in converging, only the results from the MMAP estimation were considered to make inferences about the person parameters. In addition, since the impact of the calibration sample size is made only through the item parameter estimates (rather than directly on the estimation of person parameters), the size of the examinee samples was fixed at $N = 2000$. Evaluation of the estimators was made with respect to 1) MSE, 2) bias, and 3) the recovery of correlation level between the person parameters. The MSE criterion for each trait dimension was calculated as

$$\mathrm{MSE}_\theta = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2, \qquad \text{and} \qquad \mathrm{MSE}_\tau = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i - \tau_i\right)^2.$$

Likewise, the bias was computed for each dimension as follows.

$$\text{Bias}_\theta = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right), \qquad \text{and} \qquad \text{Bias}_\tau = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i - \tau_i\right).$$

Finally, Pearson correlation between θ and τ was used to evaluate how well the population-level correlation is recovered from the estimated person parameters.

Results

MSE. Table 3.7 reports MSEs of the trait estimators under the evaluated scenarios. The results generally supported findings from the item parameter estimation. Estimating θ jointly with τ showed improvement in MSEs over estimating θ alone; the magnitude of the improvement in MSEs increased as ρ_P increased. MAP estimator, for example, showed 0.03%, 1.66%, and 8.15% smaller MSEs in the joint estimation condition as ρ_P increased from 0.0 to 0.3 and 0.6. EAP displayed 0.02%, 1.55%, and 7.76% reduction in MSEs under the same conditions. In general, EAP estimator performed marginally better than the MAP estimator. The largest difference between the two methods observed was less than 0.004. Holding the other factors constant, increasing ρ_P resulted in smaller MSEs due to the increased amount of information exchanged between the two latent trait dimensions. Across the varying design factors, SDs of the MSE statistics remained stable. The overall SDs observed were between 0.007 and 0.011.

Bias. Table 3.8 reports biases of the person parameter estimates. The overall bias levels were reasonably small across the simulation scenarios under consideration. Comparison between the two estimators revealed that the MAP estimator produced slightly more biased estimates than the EAP estimator. The level of ρ_P seemed to have no significant impact on the bias results as shown by small differences in the biases and the absence of clear patterns. Across all simulation conditions, the estimates of τ appeared essentially unbiased irrespective of the design variables. Consistent with the results for MSEs, SDs of the bias

Table 3.7: Mean Squared Error of Person Parameter Estimates

			$\rho_P = 0.0$		$\rho_P = 0.3$		$\rho_P = 0.6$	
Par	ρ_I	Model	MAP	EAP	MAP	EAP	MAP	EAP
θ	0.0	HF	.140	.137	.137	.135	.129	.127
		3PLM	.140	.137	.140	.137	.140	.137
	0.4	HF	.145	.142	.143	.140	.133	.131
		3PLM	.145	.142	.146	.142	.144	.141
	0.8	HF	.150	.146	.147	.143	.136	.133
		3PLM	.150	.146	.150	.146	.149	.145
τ	0.0	HF	.031	.031	.031	.031	.030	.030
	0.4	HF	.030	.030	.030	.030	.030	.030
	0.8	HF	.031	.031	.031	.031	.030	.030

Table 3.8: Bias of Person Parameter Estimates

			$\rho_P = 0.0$		$\rho_P = 0.3$		$\rho_P = 0.6$	
Par	ρ_I	Model	MAP	EAP	MAP	EAP	MAP	EAP
θ	0.0	HF	-.027	.001	-.027	.001	-.023	.002
		3PLM	-.026	.002	-.027	.001	-.026	.002
	0.4	HF	-.033	.000	-.032	.000	-.027	.001
		3PLM	-.033	.000	-.032	.001	-.031	.001
	0.8	HF	-.041	.000	-.041	-.001	-.034	.001
		3PLM	-.041	.000	-.042	-.001	-.040	.001
τ	0.0	HF	.000	.000	-.002	-.002	.003	.004
	0.4	HF	.000	.000	-.002	-.002	.002	.003
	0.8	HF	.000	.000	-.003	-.002	.002	.003

statistics remained stable across the simulated conditions. The MAP and EAP estimators of θ showed the overall SDs of 0.013 and 0.008, respectively. When τ parameters were estimated, both estimators showed the average SD of 0.025.

Correlation between Person Parameters. Presented in Table 3.9 are correlations between $\hat{\theta}$'s and $\hat{\tau}$'s. The table suggests that the true values of ρ_P were overall well recovered. MAP estimator showed better recovery of the true ρ_P's than the EAP estimator; the differences between the two methods were however generally negligible. The two estimators showed very similar SDs in predicting the correlations. As the ρ_P increased from 0.0 to 0.3 and 0.6, the SDs observed decreased from 0.023 to 0.019 and 0.013 irrespective of the estimation methods.

Table 3.9: Correlation between Estimated Person Parameters

	$\rho_P = 0.0$		$\rho_P = 0.3$		$\rho_P = 0.6$	
ρ_I	MAP	EAP	MAP	EAP	MAP	EAP
0.0	-.001	-.001	.326	.328	.643	.647
0.4	-.001	-.001	.326	.328	.645	.648
0.8	.000	.000	.327	.330	.645	.649

3.3.3 Robustness

The preceding simulation studies suggest that the use of the well-defined prior distribution of the parameters can improve the estimation precision. A question then arises about the robustness of the estimation methods against the ill-defined priors. In this subsection, impact of inappropriate prior information on the estimation precision is examined. Maintaining the same simulation settings with the previous studies, prior densities of the item and person parameters were deliberately misspecified. That is, the datasets generated under the $\rho_I = 0.4$ condition were used to calibrate items under the improper priors, $\rho_I = 0.0$ and 0.8. Likewise, the datasets generated under the $\rho_P = 0.3$ condition were used to estimate the person parameters assuming the ρ_P as 0.0 and 0.6. The studies fixed the ρ_P at 0.3 when the item calibration was implemented and the ρ_I at 0.4 when the person parameters were estimated. To examine the impact of wrong priors as a function of the size of observations, the size of samples for item calibration was varied as 1000 and 2000, and the test length for person parameter estimation was varied as 20 and 30.

Results

MSE. Table 3.10 presents MSEs when the items were calibrated using improper priors. MSEs tended to slightly increase when items were calibrated using the under- and over-predicted ρ_I's. Increases in the MSEs were generally small and became smaller as N increased. The MSEs of $\hat{a}$'s increased approximately from 0.001 to 0.003 when ρ_I was underpredicted and from 0.003 to 0.007 when ρ_I was overpredicted. The MSEs of $\hat{b}$'s showed the smaller deviation from the values observed under the true specification. Increases in the MSE values of $\hat{b}$'s were overall between 0.001 and 0.002. No differences were observed for $\hat{c}$, $\hat{\alpha}$, and $\hat{\beta}$ at the three significant decimal points. SDs of the MSE values followed the similar patterns. When the ρ_I was misspecified, the SDs of $\hat{a}$ and $\hat{b}$ increased less than 0.003.

Table 3.10: Mean Squared Error of Item Parameter Estimates under Improper Priors

		$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
		N		N		N	
Par	Model	1000	2000	1000	2000	1000	2000
a	HF	.022	.014	.019	.012	.023	.015
	3PLM	.021	.014	.020	.012	.027	.017
b	HF	.015	.009	.013	.007	.015	.009
	3PLM	.016	.009	.014	.008	.015	.009
c	HF	.001	.001	.001	.001	.001	.001
	3PLM	.001	.001	.001	.001	.001	.001
α	HF	.001	.000	.001	.000	.001	.000
β	HF	.001	.001	.001	.001	.001	.001

Note: Item parameters generated under $\rho_I = 0.4$ were estimated using improper priors $\rho_I = 0.0$ and $\rho_I = 0.8$, each of which corresponded to under- and over-prediction of the strength between the item parameters.

Presented in Table 3.11 are MSEs of $\hat{\theta}$ and $\hat{\tau}$ obtained under the improper priors. The characters of the results were very similar to those for item calibration. MSEs of $\hat{\theta}$ slightly increased as the ρ_P was incorrectly specified. The magnitude of the increase was generally smaller for the longer test, indicating that the adverse impact of the wrong prior on the estimation can be mitigated with more observarions. Overall, the estimation of the person parameters within the hierarchical framework displayed larger increase in MSEs of $\hat{\theta}$ compared to when the θ was estimated within the response model only. This is mainly because the information from ρ_P was not used in estimation of θ under the 3PLM, and thus, no negative influence was present from the wrong prior information.

Correlation between Parameters. In reviewing the trends in MSEs, it was apparent that the MMAP estimator was fairly robust against the misspecification of the priors. To further examine its effects on the recovery of the correlation between the parameters, Pearson correlations computed from the parameter estimates under the wrong priors are summarized in Tables 3.12 and 3.13. Table 3.12 suggests that the item parameter estimates showed weaker correlations than the true ones when the ρ_I was underpredicted, while they showed stronger correlations when the ρ_I was overpredicted. The extent of the deviation from the true correlation values became smaller as N increased within the hierarchical framework.

Table 3.11: Mean Squared Error of Person Parameter Estimates under Improper Priors

			$\rho_P = 0.0$		$\rho_P = 0.3$		$\rho_P = 0.6$	
			J		J		J	
Par	Method	Model	20	30	20	30	20	30
θ	MAP	HF	.201	.146	.197	.143	.208	.150
		3PLM	.201	.146	.201	.146	.201	.146
	EAP	HF	.197	.142	.193	.140	.204	.147
		3PLM	.197	.142	.197	.142	.197	.142
τ	MAP	HF	.045	.030	.045	.030	.045	.030
	EAP	HF	.045	.030	.045	.030	.045	.030

Note: J = test length. Person parameters generated under $\rho_P = 0.3$ were estimated using improper priors $\rho_P = 0.0$ and $\rho_P = 0.6$.

Table 3.12: Correlation between Estimated Item Parameters under under Improper Priors

	$\rho_I = 0.0$		$\rho_I = 0.4$		$\rho_I = 0.8$	
	N		N		N	
Model	1000	2000	1000	2000	1000	2000
HF	.299	.332	.535	.517	.633	.602
3PLM	.085	.139	.533	.490	.786	.720

Table 3.13: Correlation between Estimated Person Parameters under under Improper Priors

	$\rho_P = 0.0$		$\rho_P = 0.3$		$\rho_P = 0.6$	
	J		J		J	
Method	20	30	20	30	20	30
MAP	.265	.275	.337	.326	.452	.409
EAP	.265	.275	.341	.328	.459	.414

Similar outcomes were observed for the person parameter estimates. Despite the inappropriate prior information, the observed correlation levels from the estimators were quite close to the true correlation values. Overall, the correlations recovered from the hierarchical framework showed less divergence from the true ones compared to those from the 3PLM.

3.4 Summary

Thus far, the likelihood-based procedures have been presented for estimating the item and person parameters within the hierarchical framework. The primary estimation setting was based on linear tests. Results from the simulation studies suggest that the proposed estimators performed well. The MMAP estimator was generally preferred to the MML estimator in terms of the convergence and estimation accuracy. The performances of the MAP and EAP estimators were very comparable each other. Despite the dependence on the prior information, the Bayesian procedures showed robust performances against the wrong priors. In the next chapter, the performances of the likelihood-based estimation methods are examined in the adaptive testing settings where items administered vary depending on the examinees' proficiency levels. Building on the estimation methods presented here, strategies for adaptively selecting calibration samples are discussed.

Chapter 4

Calibrating Hierarchical Framework Online in Computerized Adaptive Testing

Computerized adaptive testing (CAT) is a test delivery mode that adapts questions to examinees' proficiency levels. Because of its efficient and accurate estimation routines, it has become increasingly popular in educational, psychological, and clinical assessments. For administering CAT continuously, an item pool needs to be routinely replenished by replacing over-exposed, obsolete, or flawed items. One efficient way to replenish an item pool is to calibrate new items on the fly during the test administrations. Unlike the conventional field-testing practice in which linking is needed to place the scale of new item parameters on a common metric, online calibration (Wainer & Mislevy, 1990) automatically places parameter estimates of the new items on the same scale as the operational items by taking advantage of fixed operational item parameters.

The present study is designed to propose online calibration strategies for the hierarchical framework and evaluate their performances in CAT. The traditional online calibration methods employed a randomized sampling design (e.g., Ban, Hanson, Wang, Yi, & Harris, 2001; Ban, Hanson, Yi, & Harris, 2002) or an examinee-centered sampling design (e.g., Chen, Xin, Wang, & Chang, 2012). These two methods, however, may beget sample data that contain very little information on the item parameters because neither is optimized for item calibration. In this study, an adaptive selection of samples is proposed that can maximize the information for calibrating both the 3PLM and the lognormal response time model. Such design can lead to a reduction of the sample size needed for item calibration, and thus, a reduction of the cost of field-testing.

4.1 Fisher Information Matrix

In online calibration, field-test items are calibrated instantly after response data are collected. To whom a field-test item should be assigned is thus of critical importance. A common approach for optimal sampling is to select examinees who can lead to the greatest reduction in sampling variances of item parameter estimates. This is equivalently achieved by administering the field-test item to examinees who can provide the largest Fisher information. The Fisher information matrix of an item is calculated as the negative expectation of the second derivatives of the log-likelihood with respect to item parameters $\boldsymbol{\xi}_j$ (Kendall & Stuart, 1979, p. 54-55). The log-likelihood function in the presence of known person parameters is given by

$$l = \log L = \sum_{i=1}^{N} \log f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i,\, \boldsymbol{\Xi}) = \sum_{i=1}^{N} \log f(\boldsymbol{u}_i \,|\, \theta_i,\, \boldsymbol{\Xi}) + \sum_{i=1}^{N} \log f(\boldsymbol{t}_i \,|\, \tau_i,\, \boldsymbol{\Xi}), \tag{4.1}$$

where

$$f(\boldsymbol{u}_i \,|\, \theta_i,\, \boldsymbol{\Xi}) = \prod_{j=1}^{J} f(u_{ij} \,|\, \theta_i;\, a_j,\, b_j,\, c_j) = \prod_{j=1}^{J} P_{ij}^{u_{ij}}\, Q_{ij}^{1-u_{ij}},$$

and

$$f(\boldsymbol{t}_i \,|\, \tau_i,\, \boldsymbol{\Xi}) = \prod_{j=1}^{J} f(t_{ij} \,|\, \tau_i;\, \alpha_j,\, \beta_j) = \prod_{j=1}^{J} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left[-\frac{\alpha_j^2}{2}\big\{ \log t_{ij} - (\beta_j - \tau_i)\big\}^2\right]$$

under the usual local independence assumptions of u_{ij} given θ_i, and of t_{ij} given τ_i.

The information matrix for an item in the joint model is then obtained as

$$\boldsymbol{I}(\boldsymbol{\xi}_j) = -\sum_{i=1}^{N} E\left[\frac{\partial^2 \log l}{\partial \boldsymbol{\xi}_j^2} \,\middle|\, \boldsymbol{u}_j,\, \boldsymbol{t}_j,\, \boldsymbol{\xi}_j\right] = \begin{pmatrix} I_{aaj} & I_{abj} & I_{acj} & 0 & 0 \\ I_{abj} & I_{bbj} & I_{bcj} & 0 & 0 \\ I_{acj} & I_{bcj} & I_{ccj} & 0 & 0 \\ 0 & 0 & 0 & I_{\alpha\alpha j} & 0 \\ 0 & 0 & 0 & 0 & I_{\beta\beta j} \end{pmatrix}, \tag{4.2}$$

where $\boldsymbol{u}_j = \{u_{ij}\}_{1\le i\le N}$, $\boldsymbol{t}_j = \{t_{ij}\}_{1\le i\le N}$, and,

$$I_{aaj} = D^2 \sum_{i=1}^{N} (\theta_i - b_j)^2 \left[\frac{P_{ij} - c_j}{1 - c_j}\right]^2 \frac{Q_{ij}}{P_{ij}}, \qquad I_{bbj} = D^2 a_j^2 \sum_{i=1}^{N} \left[\frac{P_{ij} - c_j}{1 - c_j}\right]^2 \frac{Q_{ij}}{P_{ij}},$$

$$I_{ccj} = \frac{1}{(1 - c_j)^2} \sum_{i=1}^{N} \frac{Q_{ij}}{P_{ij}}, \qquad I_{\alpha\alpha j} = N\alpha_j^{-2},$$

$$I_{\beta\beta j} = N\alpha_j^2, \qquad I_{ab} = -D^2 a_j \sum_{i=1}^{N} (\theta_i - b_j) \left[\frac{P_{ij} - c_j}{1 - c_j}\right]^2 \frac{Q_{ij}}{P_{ij}},$$

$$I_{acj} = D \sum_{i=1}^{N} (\theta_i - b_j) \frac{(P_{ij} - c_j)}{(1 - c_j)^2} \frac{Q_{ij}}{P_{ij}}, \qquad I_{bcj} = -Da_j \sum_{i=1}^{N} \frac{(P_{ij} - c_j)}{(1 - c_j)^2} \frac{Q_{ij}}{P_{ij}}.$$

The notion of the item information matrix reveals a number of interesting features. First, the off-diagonal blocks of the information matrix are zero due to the conditional independence assumption for the two measurement models given the person parameters. Second, the amount of information in the sample for estimating the response time model parameters depends only upon α_j. Hence, when the information matrix in (4.2) is used for sequentially selecting calibration samples, impact of incorporating response times on the optimal sampling is manifested by only the provisional estimates of α_j. Third, the off-diagonal terms within the response time model component are zero, meaning that α_j and β_j are orthogonal parameters. Fourth, the information for estimating the item parameters is additive across examinees. Finally, the information matrix of item parameters does not depend on τ_i, and hence, the information matrix can be expressed as a function of θ_i only. Thus, in the subsequent notations, the information matrix for item parameters will be denoted as $\boldsymbol{I}(\boldsymbol{\xi}_j;\, \theta_i)$ to account for individual contribution of an examinee with proficiency level θ_i.

4.2 Optimal Sampling Design

An important question arises as to how the Fisher information matrix can be translated into a formal criterion of sample selection. Among others (see Silvey, 1980 for other criteria), the use of determinant or trace of the information matrix has been standard practice in experimental design (e.g., Chaloner & Verdinelli, 1995). Each criterion is known as D-optimality and A-optimality. In the online calibration context, D- and A-optimality for selecting an examinee are expressed as follows.

$$D\text{-optimality}: \ \arg\max_{\theta_i} \ \det\left[\boldsymbol{I}(\hat{\boldsymbol{\xi}}_j;\, \theta_i)\right], \tag{4.3}$$

and

$$A\text{-optimality}: \ \arg\min_{\theta_i} \text{trace}\left[\boldsymbol{I}^{-1}(\hat{\boldsymbol{\xi}}_j;\, \theta_i)\right]. \tag{4.4}$$

D-optimality seeks calibration samples for minimizing the volume of the confidence ellipsoid (i.e., the generalized variance) of the item parameter estimates by maximizing the determinant of the Fisher information matrix. A-optimality, on the other hand, selects samples that minimize the average variance of the item parameter estimates by minimizing the trace of the inverse of the information matrix. Within each optimality criterion, the inverse of the prior covariance matrix, $\boldsymbol{\Sigma}_I^{-1}$, can be incorporated to obtain Bayesian versions of D- and A-optimality, each of which leads to the sampling design that minimizes the determinant or the trace of the posterior covariance matrices.

The optimality criteria in (4.3) and (4.4) select examinees who provide the most information for simultaneously estimating all item parameters. If the focus of field-testing is on a subset of item parameters, it would be more desirable to maximize the information for the targeted item parameters. For example, in practice it is likely that the central interest of field-testing is in estimating a_j, b_j, and c_j accurately, and estimating α_j and β_j is of secondary importance. In such cases, selection of calibration samples may be optimized with respect to the intentional item parameters. The distinction between the intentional and nuisance parameters calls for a new optimality criterion such that calibration sample can be selected to maximize the information for an intentional subset of item parameters. The present study proposes D_S- and A_S-optimality (Mulder & van der Linden, 2009; Silvey, 1980) for this purpose:

$$D_S\text{-optimality}: \ \arg\max_{\theta_i} \ \det \left[\boldsymbol{W}^T \boldsymbol{I}^{-1}(\hat{\boldsymbol{\xi}}_j;\, \theta_i) \boldsymbol{W} \right]^{-1}, \tag{4.5}$$

and

$$A_S\text{-optimality}: \ \arg\min_{\theta_i} \ \text{trace} \left[\boldsymbol{W}^T \boldsymbol{I}^{-1}(\hat{\boldsymbol{\xi}}_j;\, \theta_i) \boldsymbol{W} \right], \tag{4.6}$$

where $\boldsymbol{W}^T$ is a 3×5 matrix consisting of $[\boldsymbol{I}_3 \ \boldsymbol{0}]$ with $\boldsymbol{I}_3$ being 3×3 identity matrix and $\boldsymbol{0}$ being 2×2 zero matrix.

From the implementational perspective, it is hardly feasible to have a static pool of examinee samples because CAT is administered continuously or at time intervals. Thus, instead of comparing examinees for an item, comparing the finite set of field-testing items would be more achievable. This strategy selects a field-test item that has the largest optimality statistic of concern from a field-test item pool for a given examinee. That is to say, at a given seeding location in the adaptive testing session, all field-testing items are compared and evaluated at the current value of the estimated ability, and the item with the maximum optimality statistic is administered. It should be further noted that the adaptive selection of calibration samples lacks the independence property among the observations because the samples are sequentially obtained based on the information matrix evaluated at the provisional item parameter estimates. Asymptotic properties and consistency of ML estimates under the sequential designs are well-established through martingale theories (e.g., Chang & Lu, 2010; Wu, 1985a, 1985b; Wynn, 1970; Ying & Wu, 1997).

In reality, computation of the optimality criteria in (4.3)-(4.6) is subject to measurement error because estimated values from the operational CAT are substituted for the unknown θ_is. Nevertheless, it will be assumed that neglecting this error introduces only minor differences in the optimal samples. (A relative efficiency measure presented later in the simulation study serves to examine this assumption.) Because one is chiefly interested in finding optimal samples that enable efficient estimation of item parameters, loss of efficiency will be minimal as long as the samples are obtained using the proficiency estimates close enough to true values (Berger, 1994). A practical way of attaining nearly optimal samples is to assign field-test items toward the end of tests so that the information matrices can be calculated using the proficiency estimates that are accurate as possible.

4.3 Simulation Study

4.3.1 Adaptive Online Calibration of the Hierarchical Framework

Simulation studies were conducted to evaluate the effectiveness of the proposed online calibration strategies for jointly estimating the response and response time models. Calibration samples were adaptively selected according to D-, D_S-, A-, or A_S-optimality criteria using the 5×5 information matrix. For comparison purposes, a condition that considers the calibration of the response model only was also included. When the response model parameters were estimated, the 3×3 information matrix pertaining to a, b, and c was used to obtain D- or A-optimality samples. To ensure convergence in the 3PLM, the MMAPE/EM algorithm was used for calibrating the field-test items.

The hyperparameters for generating items were the same with the simulation study presented in Chapter 3.3.1. The numbers of field-test items and operational items were 15 and 300, respectively. Before entering the online calibration process, every field-test item was given initial parameter estimates obtained from relatively small random samples of 300. The initial parameter values were assigned to select samples adaptively at the early stage of online calibration. In practice, one may avail content experts' crude approximation to obtain these values. After entering the online calibration, 400 examinees were adaptively selected based on the initial estimates. The parameter estimates of the field-test items were not updated during this period because item calibration could be unstable when the sample size is too small. Once the size of the calibration samples exceeded the minimum of 400, the parameters of the field-test items were sequentially estimated and updated after every batch of 10 additional observations. The online calibration process continued until the field-test items reached the maximum sample size of 800.

CAT was continuously administered to 40000 examinees with the test length of 33, consisting of 30 operational items and 3 field-test items. Examinees' latent trait parameters were randomly sampled from a bivariate normal distribution with zero means and unit variances. The correlation between θ and τ was fixed at $\rho_P = 0.3$, which would be considered a moderate level. During the CAT administrations, operational items were adaptively assigned to examinees according to the maximum information criterion. The person parameters were estimated via EAP when the number of operational items administered was no more than five; otherwise, the MAP method was used. Seeding locations for the field-test items were determined randomly at the later stage of CAT—between the 24th and 33rd items—in order to regulate the measurement error in the optimal sampling. The maximum exposure rate was imposed on the operational items at 0.2, meaning that no more than 800 examinees received the same item.

Crossing the three factors—two models, field-test item selection methods (D-, D_S-, A- and A_S-optimality for joint model calibration; D- and A-optimality for the 3PLM calibration), and three levels of ρ_I—yielded 18 online calibration conditions. Within each condition, 100 replications were executed.

For assessing the loss of efficiency caused by measurement error in the optimal sampling design, a relative efficiency measure (Berger, 1991, 1994) was inspected. The relative efficiency of a sampling design with measurement error compared to its counterpart without the error was obtained by comparing the logarithms of the determinants of the information matrices given by

$$\text{Relative Efficiency}_j = \frac{\log\Big(\det\big[\boldsymbol{I}\big(\hat{\boldsymbol{\xi}}_j;\, \hat{\boldsymbol{\theta}}\big)\,\big]\Big)}{\log\Big(\det\big[\boldsymbol{I}\left(\boldsymbol{\xi}_j;\, \boldsymbol{\theta}\right)\,\big]\Big)}, \tag{4.7}$$

where $\boldsymbol{I}(\hat{\boldsymbol{\xi}}_j;\, \hat{\boldsymbol{\theta}})$ denotes the 3×3 or 5×5 item information matrix evaluated at the calibration sample $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\, \hat{\theta}_2,\, \ldots, \hat{\theta}_N)$; and $\boldsymbol{I}(\boldsymbol{\xi}_j;\, \boldsymbol{\theta})$ is the information matrix based on the true values. The relative efficiency statistics show how much efficiency is lost as a result of replacing true proficiency values with estimated ones when selecting the calibration samples. Hence, comparison should be made within each calibration condition (i.e., between with and without measurement error). If the resulting value is less than 1, the sampling design under the measurement error is less efficient than the sampling design without the measurement error; if the value is greater than 1, the sampling design with the measurement error is more efficient than the other.

The recovery of individual item parameters was evaluated by MSE, bias, and correlation between the true parameters and the estimated values. The analysis was made on the field-test items that were successfully converged across all selection methods within the same optimality. This was to make valid comparison across the field-test item selection strategies by controlling for the true item parameter values because the recovery of individual item parameters depends on their true values.

Results

Relative Efficiency. Tables 4.1 and 4.2 report relative efficiency statistics computed for each *D*- and *A*-optimality sampling design. The primary interest in this stage is to examine

Table 4.1: Relative Efficiency of Online Calibration under D-optimality

	$\rho_I = 0.0$			$\rho_I = 0.4$			$\rho_I = 0.8$		
	HF		3PLM	HF		3PLM	HF		3PLM
	D	D_S	D	D	D_S	D	D	D_S	D
Overall	.998	.998	.998	.995	.998	.993	.995	.996	.988
(SD)	(.022)	(.029)	(.025)	(.022)	(.030)	(.024)	(.020)	(.027)	(.024)

Note: HF = hierarchical framework. 3PLM = three-parameter logistic model. D = D-optimality sampling design. D_S = D_S-optimality sampling design.

Table 4.2: Relative Efficiency of Online Calibration under A-optimality

	$\rho_I = 0.0$			$\rho_I = 0.4$			$\rho_I = 0.8$		
	HF		3PLM	HF		3PLM	HF		3PLM
	A	A_S	A	A	A_S	A	A	A_S	A
Overall	1.004	1.002	1.003	1.005	1.003	1.005	1.001	.998	1.002
(SD)	(.028)	(.023)	(.034)	(.040)	(.033)	(.042)	(.032)	(.032)	(.036)

Note: A = A-optimality sampling design. A_S = A_S-optimality sampling design.

the impact of measurement error on the efficiency of optimal sampling, and hence, the indices were averaged across the same sampling design. The values in parentheses provide SDs of the relative efficiency statistics over the replications. Tables 4.1 and 4.2 suggest that the observed relative efficiency statistics were very close to 1, and their SDs were very small. Although slight departures from 1 were observed among the optimal sampling designs, the overall magnitude was negligible enough to consider them as random error. Thus, it can be concluded that replacing the true person parameters with the estimated values had minimal impact on the efficiency of the optimal sampling designs.

MSE. Figure 4.1 plots MSEs of item parameter estimates obtained from the D-optimality sampling design as a function of the sample size. Overall, all online calibration conditions showed improvement in MSEs along with increasing N and ρ_I. As shown by larger decreases in MSE$_a$ and MSE$_b$, the effect of increasing N was most apparent for estimating a and b. This is mainly due to the fact that c, α, and β were already estimated sufficiently well

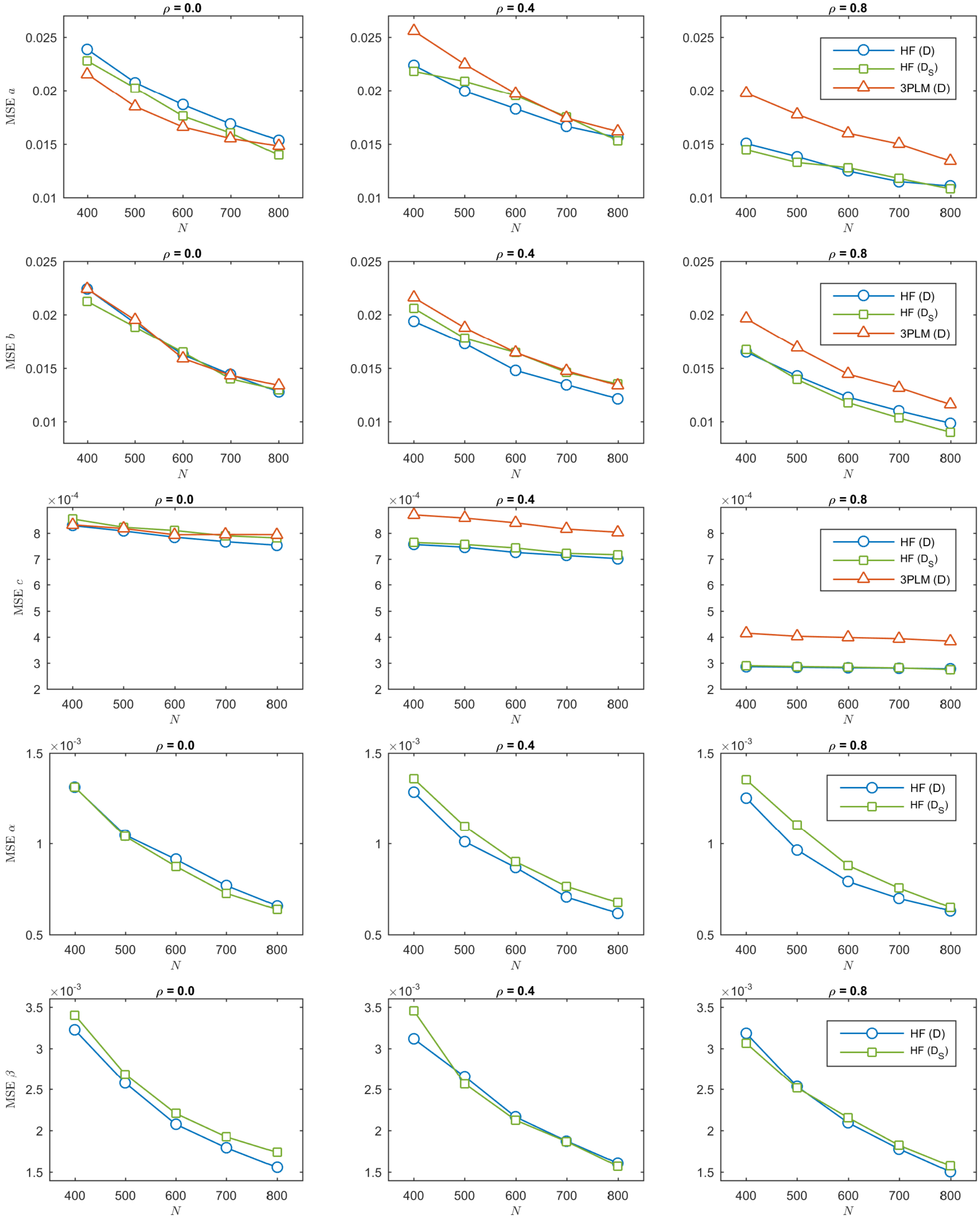


Figure 4.1: MSE of Item Parameter Estimates under *D*-optimality Sampling Design

with the small samples. The online calibration scenarios under evaluation showed very minor differences when the item parameters had the zero correlations. As ρ_I increased to 0.4, slight improvement in MSEs were observed as a result of calibrating the hierarchical framework. The gain in the estimation accuracy became pronounced as ρ_I increased to 0.8. Despite the fact that a and α, and b and β are defined on the same respective domains, MSEs of $\hat{\alpha}$ and $\hat{\beta}$ were much smaller than those of $\hat{a}$ and $\hat{b}$. This is due in large part to the property of response time data. Because response times were observed on the continuous scale, much information could be used for pinpointing the true locations of α and β.

Figure 4.2 provides MSEs of item parameter estimates obtained from the A-optimality sampling design. Compared to Figure 4.1, A-optimality resulted in generally larger MSEs for $\hat{a}$ and $\hat{b}$. The differences, however, occurred mostly at the second decimal place, and the overall MSEs remained small for the A-optimality design. The impact of the differential sampling design on the recovery of c, α, and β seemed minimal as suggested by the small differences in the MSEs between Figures 4.1 and 4.2[1]. The effects of N and ρ_I were consistent with those seen under the D-optimality design. When $\rho_I = 0.0$, calibrating the 3PLM alone produced the smallest MSEs for a and b. As ρ_I increased to 0.4 and 0.8, calibrating the hierarchical framework under A_S-optimality consistently showed the most accurate recovery for all item parameters. The degree to which MSEs reduced as a result of calibrating the hierarchical framework instead of the 3PLM was generally greater under the A-optimality design than under the D-optimality design.

Bias. Tables 4.3 and 4.4 report biases of the final item parameter estimates. When D-optimality was used for adaptive selection of the calibration samples, the parameter estimates showed the biases less than 0.02. Although the A-optimality sampling design produced slightly larger biases, the overall bias level remained modest, resulting in the maximal bias of -0.034 for $\hat{b}$ in the $\rho_I = 0.8$ condition. Under both D-optimality and A-optimality, the

[1] The vertical scales in the two figures were differed to examine the comparative performances of the calibration scenarios conditioning on the same optimality design.

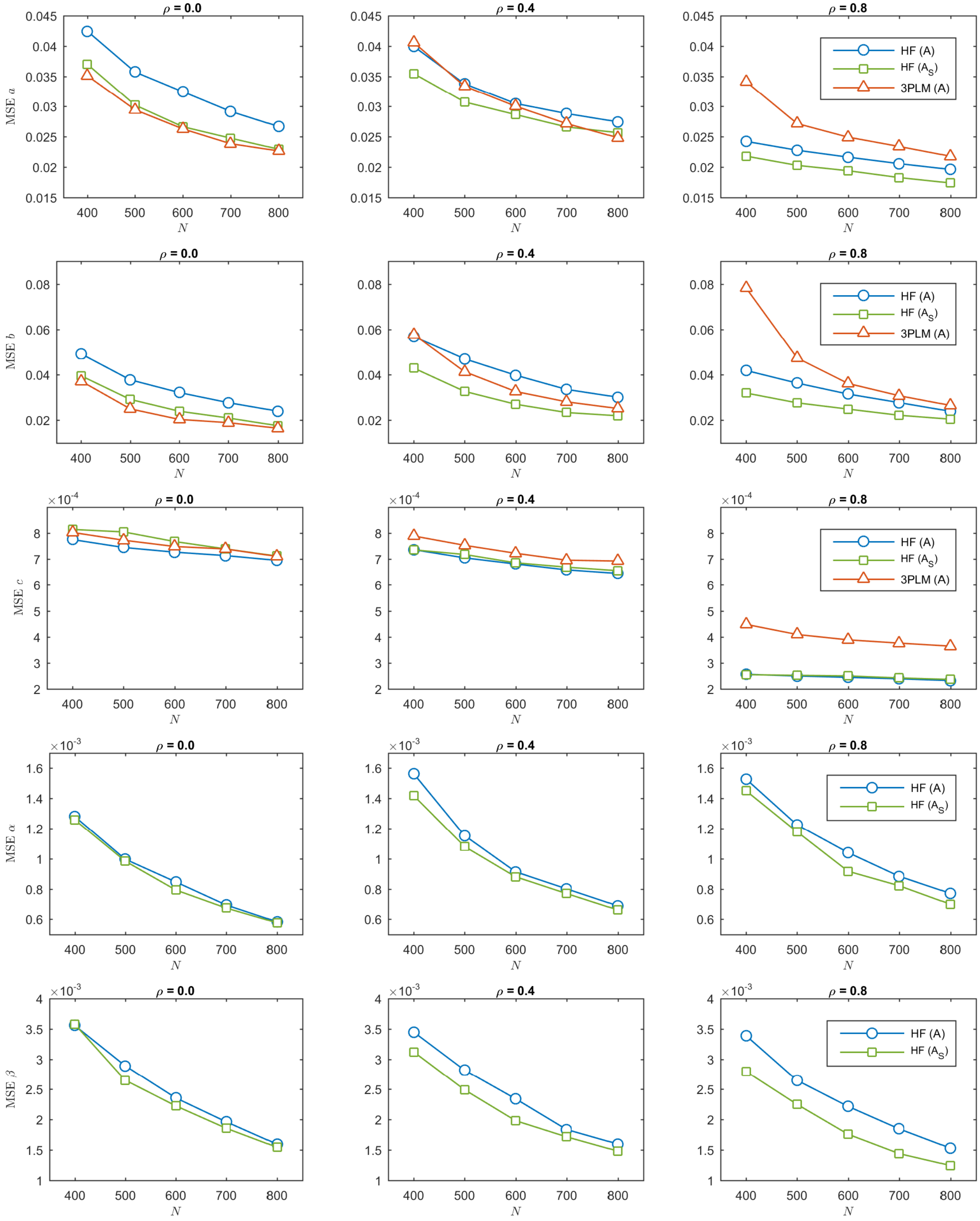


Figure 4.2: MSE of Item Parameter Estimates under A-optimality Sampling Design

Table 4.3: Bias of Item Parameter Estimates under D-optimality Sampling Design

	$\rho_I = 0.0$			$\rho_I = 0.4$			$\rho_I = 0.8$		
	HF		3PLM	HF		3PLM	HF		3PLM
Par	D	D_S	D	D	D_S	D	D	D_S	D
a	-.005	-.009	-.005	-.008	-.012	-.006	-.005	-.011	-.017
b	-.011	-.007	-.002	-.012	-.011	-.012	-.010	-.008	-.013
c	-.002	-.002	-.002	-.004	-.004	-.003	-.001	-.002	-.002
α	.000	.000		.000	-.001		-.001	-.001	
β	.005	.004		.004	.002		.003	.005	

Table 4.4: Bias of Item Parameter Estimates under A-optimality Sampling Design

	$\rho_I = 0.0$			$\rho_I = 0.4$			$\rho_I = 0.8$		
	HF		3PLM	HF		3PLM	HF		3PLM
Par	A	A_S	A	A	A_S	A	A	A_S	A
a	.010	.012	.004	.016	.022	.015	.011	.016	.023
b	-.028	-.013	-.009	-.031	-.007	-.006	-.035	-.002	.013
c	-.002	-.002	-.002	-.002	-.002	.000	-.001	.000	.002
α	-.001	.000		-.001	.000		-.001	.001	
β	.002	.003		.003	.001		.002	.001	

parameter estimates for α and β appeared essentially unbiased. The maximal bias observed across all evaluated conditions was 0.005. Impact of increasing ρ_I was demonstrated by marginally increased biases under each optimal sampling design. The differences across the varying ρ_I levels were generally too small to attach any practical meaning to the relative performance of the different online calibration scenarios.

Correlation. Figure 4.3 plots correlations between the true and estimated item parameters obtained from the D-optimality sampling design. The three online calibration scenarios showed very comparable performances under the zero correlation condition. As ρ_I increased to 0.4, patterns related to model calibration became more pronounced. When $\rho_I = 0.8$, calibrating the hierarchical framework consistently led to higer correlations compared to the 3PLM calibration condition, indicating the borrowing of collateral information from response times in estimation of the response model parameters. Although $\hat{c}$'s had smaller errors

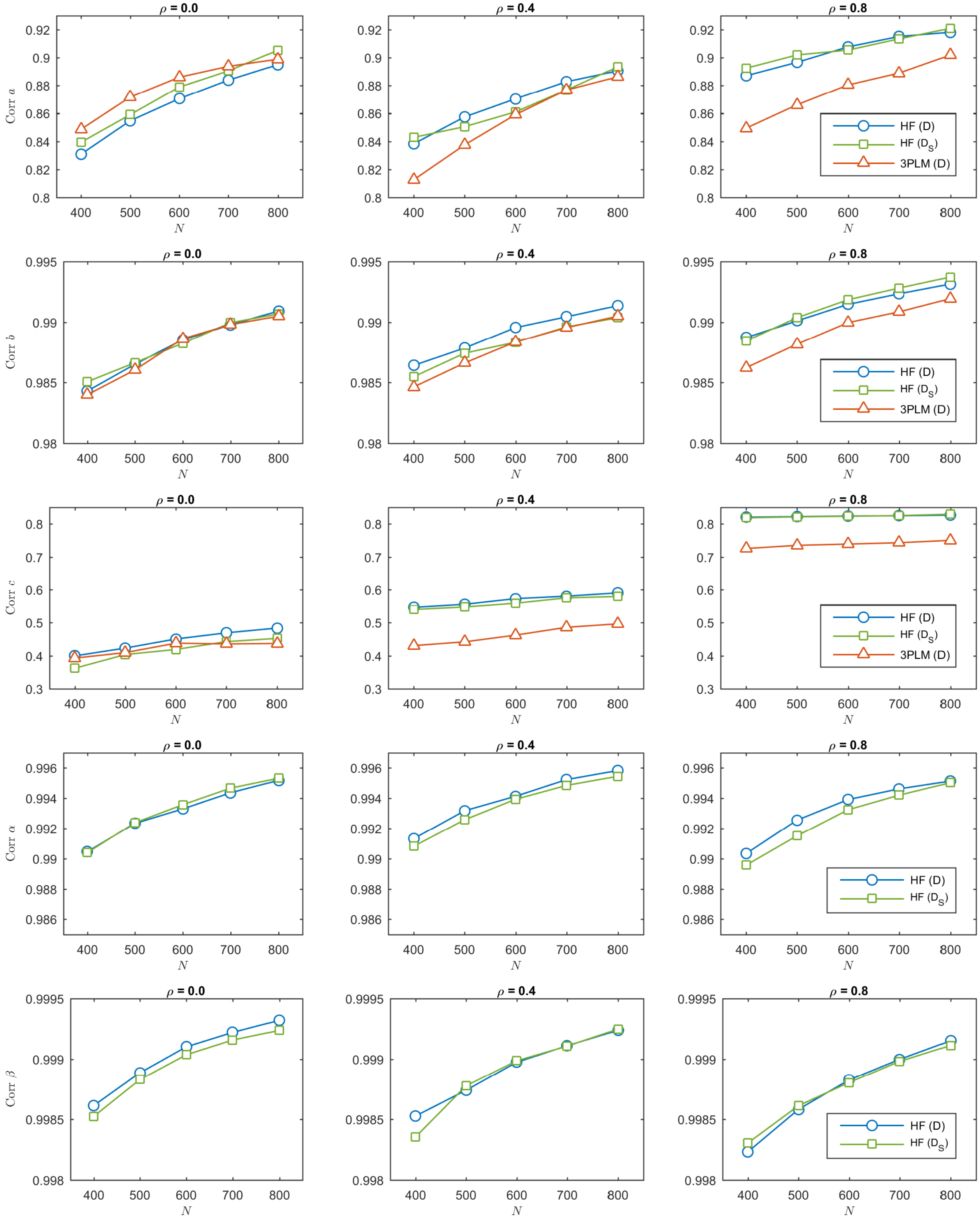


Figure 4.3: Correlation between True and Estimated Item Parametes under *D*-optimality Sampling Design

compared to $\hat{a}$'s and $\hat{b}$'s in the previous results, the correlations between c and $\hat{c}$ were actually lower than those of a and $\hat{a}$ as well as those of b and $\hat{b}$. The high dependency among the item parameters appeared to greatly improve the precision of c estimation. The improvement in the correlation between c and $\hat{c}$ was further increased by jointly calibrating the response and response time models. The overall patterns in regards to design variables were consistent with expections. Correlations between the true and estimated values increased as N and/or ρ_I increased across all item parameters.

Figure 4.4 shows the correlation values obtained from the A-optimality sampling design. When the item parameters were uncorrelated, calibrating the hierarchical framework with the A_S-optimality sampling produced very comparable correlations to those from the calibration of the response model. As the item parameters had nonzero associations, estimation of the hierarchical framework with A_S-optimality consistently produced the highest correlations except for $\hat{c}$ under $\rho = 0.4$. This result provides an interesting point that the sampling design can interact with the objective set of the item parameters. In the D-optimal sampling design, for example, calibrating the hierarchical framework with the consideration of all parameters resulted in best outcomes for the estimation of the response model parameters under the nonzero correlation conditions. When the A-optimal design was chosen for its relative ease in computation, on the other hand, considering only the subset of parameters would result in the most accurate recovery for the parameters of interest.

4.4 Summary

In this chapter, online calibration strategies have been presented for efficiently obtaining item parameter estimates for the hierarchical framework. Results from the simulation study suggest that the MMAP estimator performed well in jointly calibrating the response and response time models during the CAT administrations. With the employment of the optimal sampling design, the estimator achieved the desired estimation accuracy with much smaller

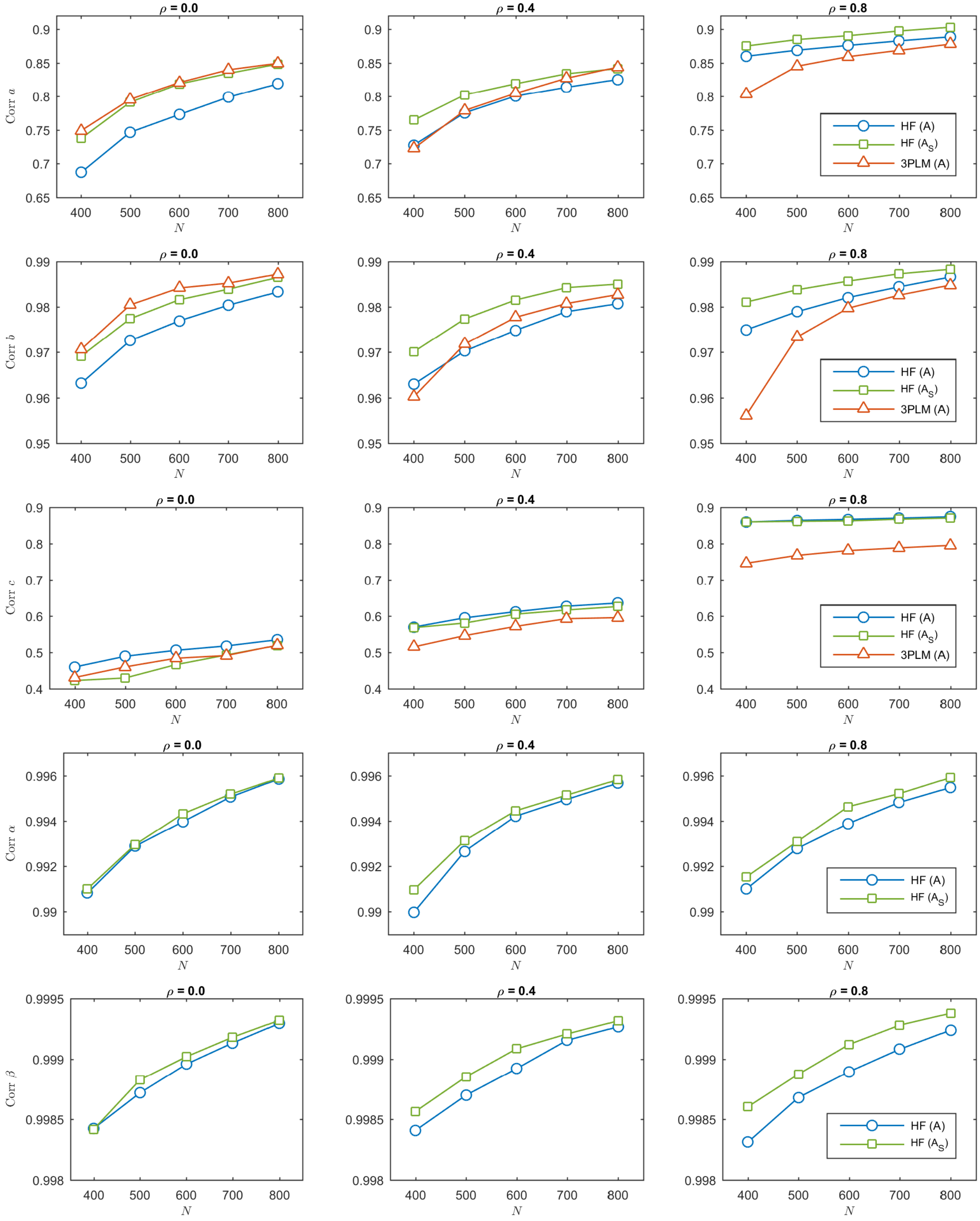


Figure 4.4: Correlation between True and Estimated Item Parameters under A-optimality Sampling Design

calibration samples. Increasing the sample size and the correlation between the item parameters consistently led to improvement in the estimation precision across all online calibration scenarios. The present study provides corroborating evidence that the likelihood-based estimation methods are indeed a viable alternative to the MCMC estimation procedures. Implementation of the simulation study was manageable enough to carry out 100 replications despite the high computational load in sequential estimation.

Chapter 5

Proportional Hazards Model Estimation

The Cox proportional hazards (PH) model (Cox, 1972) with random effects, also known as the frailty model (Clayton & Cuzick, 1985; Vaupel et al., 1979), has shown promise in modeling various shapes of response time distributions. In contrast to parametric models that are bound to certain distributional assumptions (e.g., Maris, 1993; Rouder et al., 2003; Scheiblechner, 1979; van der Linden, 2006), the PH model can accommodate different shapes of response time distributions. Despite the popularity in survival analysis, however, the PH model has seldom been used in psychological or educational measurement because of the difficulty in model estimation. Estimation of the PH model parameters is often complicated by the presence of latent variables and the nonparametric nature of baseline hazard rates. It has been only recently that several attempts have been made to fit the PH latent trait model (PHLTM; Ranger & Ortner, 2012) in the psychometric context (e.g., Douglas et al., 1999; Loeys et al., 2014; Ranger & Ortner, 2012, 2013; Ranger & Kuhn, 2015; Wang, et al., 2013).

Douglas et al. (1999) presented estimation based on discrete response times such that standard estimation methods, such as the marginal maximum likelihood estimation, can be used within the framework of the generalized linear model. However, this approach requires arbitrary decisions about thresholds for categorizing the response times and can result in loss of efficiency as a result of discretizing the continuously defined response times. The profile likelihood (PL) estimation (Ranger & Ortner, 2012) employs the EM algorithm to deal with the unknown latent traits, which, in turn, results in expensive computation and requires further computation for variance estimates (Therneau, Grambsch, & Pankratz, 2003). Later, Ranger and Ortner (2013) proposed a computationally more affordable estimation method based on the rank correlation matrix of Kendall's Tau. The study provided a comparison of the profile likelihood estimator, the rank-based estimator, and the marginal maximum likelihood estimator with discretized response times and concluded that the rank-based estimator and profile likelihood estimator performed better than the estimation based on the discrete response times in terms of accuracy and efficiency. One limitation of the rank-based estimator, however, is that it does not allow joint analysis of item responses and response times. Although one can opt for MCMC estimation for jointly analyzing the accuracy scores and response times (Wang et al., 2013), implementing this procedure requires a solid background in computational statistics and is computationally much more demanding than implementing the likelihood-based approach[1].

The purpose of the present study is to provide a viable alternative for fitting the PHLTM. The new procedure is based on the penalized partial likelihood function where the marginal distribution of the latent speed parameters determines the penalty term. The procedure follows Ripatti and Palmgren (2000), which applied the penalized likelihood theory (Green, 1987) and the Laplace approximation of the likelihood function (Breslow & Clayton, 1993) to estimate the multivariate frailty models. Compared to the existing methods, the new estimation approach is much simple and easy to implement. Furthermore, it allows the joint analysis of response times and accuracy scores.

In what follows, two approaches for fitting the PHTLM are introduced, the PL estimation and the penalized partial likelihood (PPL) estimation. The PL method is presented for the purpose of comparison in simulation studies. Next, a joint analysis approach for responses and response times is proposed using the PHLTM. Finally, the performance of the proposed methods are validated in simulation studies.

[1] Patton (2015) reports that each Markov chain takes from 12 to 18 hours for estimating the three-parameter logistic model and the PHLTM within the hierarchical framework.

5.1 Item Parameter Estimation

To set the stage for parameter estimation, we assume that items are independent of each other and response times are independent conditioning on τ_i. Since examinees are drawn at random from a population of interest, it is also assumed that examinees are independent, and examinees and items are independent. Suppose that, for a given item j, there are no ties in the response times. Thus, the response times can be ordered as $t_{pj} < t_{p'j}$ for all $1 \leq p \neq p' \leq N$. Let $\tau_{(p)}$ denote the p-th latent speed parameter whose ordered response time is t_{pj}. We can therefore define the risk set at time t_{pj}, $R(t_{pj})$, as the set of all individuals who have not answered the item yet, that is, $R(t_{pj}) = \{t_{(p+1)j}, \ldots, t_{Nj}\}$.

Our main focus in this section is on estimation of the PHLTM parameters from the observed response time data. Let $\mathbf{T}$ denote the response time data for all examinees and items. The unknown parameters to be estimated are denoted as vectors—i.e., $\boldsymbol{h}_0 = \{h_{0j}(\cdot)\}_{1\leq j\leq J}$, $\boldsymbol{\gamma} = \{\gamma_j\}_{1\leq j\leq J}$, and $\boldsymbol{\tau} = \{\tau_i\}_{1\leq i\leq N}$. The PL estimator maximizes the complete-data likelihood function, $L(\boldsymbol{\gamma}, \boldsymbol{h}_0)$, by treating the $\boldsymbol{\tau}$ as known and obtains estimates of $\boldsymbol{\gamma}$ (and possibly $\boldsymbol{h}_0$). The newly proposed estimation method that bases on the PPL attempts to maximize the joint density of $\mathbf{T}$ and $\boldsymbol{\tau}$, $L(\mathbf{T}, \boldsymbol{\tau} \,|\, \boldsymbol{\gamma}, \boldsymbol{h}_0)$, by considering the $\boldsymbol{\tau}$ as the observed random variables. In the following subsections, the PL estimation method and the PPL estimation method are presented based on the assumptions and notations made above. For the moment, the primary interest is in estimation of γ. The following section will devote attention to the estimation of the latent traits by using the item parameter estimates obtained from the either methods described in here.

5.1.1 Review of Profile Likelihood Estimation

The PL estimation maximizes the complete-data likelihood function by treating the examinees' latent traits as known. Assuming the randomness of the examinees and the conditional independence of response times given τ_i, the complete-data likelihood can be written as

$$L(\boldsymbol{\gamma},\, \boldsymbol{h}_0) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(t_{ij} \,|\, \tau_i;\, \gamma_j,\, h_{0j}). \tag{5.1}$$

When it comes to estimating $\boldsymbol{\gamma}$, the $\boldsymbol{h}_0$ and $\boldsymbol{\tau}$ are a nuisance parameter and an incidental parameter, respectively. Hence, it may be desirable to write the likelihood function only in terms of $\boldsymbol{\gamma}$ by profiling out the nuisance parameter and then use the EM algorithm to cope with the unobservable variables.

Since items are assumed independent each other, the likelihood function for individual γ_j's can be maximized instead of the likelihood for $\boldsymbol{\gamma}$. The profile likelihood for γ_j is obtained by fixing the γ_j for each item and writing the likelihood as a function of h_{0j} only:

$$L_{\gamma_j}(h_{0j}) = \left[\prod_{i=1}^{N} h_{0j}(t_{ij}) \exp\left(\gamma_j \tau_i\right)\right] \exp\left[-\sum_{i=1}^{N} H_{0j}(t_{ij}) \exp\left(\gamma_j \tau_i\right)\right],$$

which may be written as (Klein & Moeschberger, 2003, p. 258)

$$L_{\gamma_j}(h_{0j}) \propto \prod_{i=1}^{N} h_{0j}(t_{ij}) \exp\left[-h_{0j}(t_{ij}) \sum_{p \in R(t_{ij})} \exp\left(\gamma_j \tau_{(p)}\right)\right].$$

The maximizer of this profile likelihood is

$$\hat{h}_{0j}(t_{ij}) = \left[\sum_{p \in R(t_{ij})} \exp\left(\gamma_j \tau_{(p)}\right)\right]^{-1}.$$

Combining these estimates yields an estimate of $\hat{H}_{0j}(t)$:

$$\hat{H}_{0j}(t) = \sum_{t_{ij} \le t} \left[\sum_{p \in R(t_{ij})} \exp\left(\gamma_j \tau_{(p)}\right)\right]^{-1}, \tag{5.2}$$

also known as the Breslow estimator (Breslow, 1972) of the cumulative baseline hazard rate.

Substituting these estimators into the original complete-data log-likelihood function yields the partial log-likelihood function for the γ_j:

$$l^{pl} = \sum_{i=1}^{N} \left[\gamma_j \tau_i - \log \sum_{p \in R(t_{ij})} \exp\left(\gamma_j \tau_{(p)}\right) \right]. \tag{5.3}$$

A maximizer of (5.3) leads to the profile likelihood solution for γ_j.

The quantity (5.3) cannot be calculated unless the dependence on the unknown latent traits is properly addressed. Ranger and Ortner (2012) seek a solution from the EM algorithm. During the E-step of the EM algorithm, the unobservable latent variables are replaced by the Gauss-Hermite quadrature nodes with corresponding weights. The conditional expectation of the complete log-likelihood function is then determined using the provisional estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{h}_0$. In the M-step, the conditional expectation is maximized over the structural parameters (i.e., $\boldsymbol{\gamma}$ and $\boldsymbol{h}_0$). Since inserting the Breslow estimator into the conditional expectation of the log-likelihood yields the profile likelihood function for the $\boldsymbol{\gamma}$, maximizing the expected complete log-likelihood function will be equivalent to maximizing the profile likelihood.

To compute standard errors of the estimates, it would be necessary to invert the observed information matrix for ($\boldsymbol{\gamma}$, $\boldsymbol{h}_0$). The observed information matrix is obtained as the inverse of the negative Hessian matrix computed from the EM algorithm (Louis, 1982). Inverting the information matrix is a non-trivial task especially when the dimension of the matrix is large. A possible way to tackle this problem is to assume independence among the item parameters—between γ_j and h_{0j}, and between items—and to invert only the relevant information submatrices (Cortiñas & Burzykowski, 2005). The current study follows this approach and obtains the standard error of the final estimate of γ_j as the square root of the diagonal element of the inverted information matrix.

5.1.2 Penalized Partial Likelihood Estimation

The PPL estimator begins with the complete-data likelihood function with $\boldsymbol{\tau}$ being an observed random variable:

$$L(\boldsymbol{\gamma},\, \boldsymbol{h}_0) = L(\mathbf{T},\, \boldsymbol{\tau} \,|\, \boldsymbol{\gamma},\, \boldsymbol{h}_0,\, \boldsymbol{\Omega}) = f(\mathbf{T} \,|\, \boldsymbol{\tau},\, \boldsymbol{\gamma},\, \boldsymbol{h}_0)\, f(\boldsymbol{\tau} \,|\, \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega}$ denotes the hyperparameter for $\boldsymbol{\tau}$. Within the PHLTM framework, τ is considered a random effect, and thus, it is assumed that $\tau \sim N(0,\, \sigma_\tau^2)$; that is, $\boldsymbol{\Omega} = (0,\, \sigma_\tau^2)$. The corresponding marginal likelihood of the observed data is expressed as

$$\prod_{i=1}^{N} \int f(\boldsymbol{t}_i \,|\, \tau_i,\, \boldsymbol{\gamma},\, \boldsymbol{h}_0)\, f(\tau_i \,|\, \sigma_\tau^2)\, d\tau_i. \tag{5.4}$$

Assuming the conditional independence of response times given τ_i and applying the Laplace approximation (Breslow & Clayton, 1993) for the integral in (5.4), one can obtain the approximate marginal log-likelihood function for an examinee as follows (Ripatti & Palmgren, 2000).

$$\log \int f(\boldsymbol{t}_i \,|\, \tau_i,\, \boldsymbol{\gamma},\, \boldsymbol{h}_0)\, f(\tau_i \,|\, \sigma_\tau^2)\, d\tau_i \approx -\frac{1}{2} \log\left(\frac{1}{2\pi\sigma_\tau^2}\right) - \frac{1}{2}\log|\kappa''(\tilde{\tau})| - \kappa(\tilde{\tau}),$$

where

$$\kappa(\tilde{\tau}) = -\left[\sum_{j=1}^{J} \log h_{0j}(t_{ij}) + \gamma_j \tilde{\tau}_i - H_{0j}(t_{ij}) \exp\left(\gamma_j \tilde{\tau}_i\right) - \frac{\tilde{\tau}_i^2}{2\sigma_\tau^2}\right], \tag{5.5}$$

and $\tilde{\tau}$ is the solution to the first-order partial derivative of $\kappa(\tilde{\tau})$ with respect to τ: that is,

$$\kappa'(\tilde{\tau}) = -\left[\sum_{j=1}^{J} \gamma_j - \gamma_j H_{0j}(t_{ij}) \exp\left(\gamma_j \tilde{\tau}_i\right) - \frac{\tilde{\tau}_i}{\sigma_\tau^2}\right] = 0.$$

The $\kappa''(\tilde{\tau})$ in (5.5) denotes the second-order partial derivative of $\kappa(\tilde{\tau})$ with respect to τ:

$$\kappa''(\tilde{\tau}) = \sum_{j=1}^{J} \gamma_j^2 H_{0j}(t_{ij}) \exp(\gamma_j \tilde{\tau}_i) + \frac{1}{\sigma_\tau^2}.$$

Ripatti and Palmgren (2000) suggest that if τ_i is considered a fixed effect, the quantity $\kappa(\tilde{\tau})$ in (5.5) would equal to a penalized log-likelihood (Green, 1987). For a fixed σ_τ^2, the values $\hat{\gamma}_j$ and $\hat{\tau}_i$ that maximize the penalized log-likelihood can maximize the penalized partial log-likelihood. Hence, the objective function to be maximized for all examinees can be obtained as

$$l^{ppl} = \sum_{i=1}^{N} \left[\sum_{j=1}^{J} \left(\gamma_j \tau_i - \log \sum_{p \in R(t_{ij})} \exp(\gamma_j \tau_{(p)}) \right) - \frac{\tau_i^2}{2\sigma_\tau^2} \right]. \tag{5.6}$$

The quantity within the parentheses in (5.6) denotes the partial log-likelihood of the response times treating the random effects as fixed. The following is the penalty term penalizing for extreme values of τ_i. Since the penalty term does not involve the γ_j, the partial derivative of (5.6) with respect to γ_j is equivalent to that of (5.3), that is, $\dfrac{\partial l^{ppl}}{\partial \gamma_j} = \dfrac{\partial l^{pl}}{\partial \gamma_j}$. Hence, the score equation for γ_j in the PPL estimation is identical to those for ordinary PH models treating the random effects term as the offset term.

It follows from (5.6) that the score equation for γ_j can be obtained as

$$\frac{\partial l^{ppl}}{\partial \gamma_j} = \sum_{i=1}^{N} \left[\tau_i - \frac{\displaystyle\sum_{p \in R(t_{ij})} \tau_{(p)} \exp(\gamma_j \tau_{(p)})}{\displaystyle\sum_{p \in R(t_{ij})} \exp(\gamma_j \tau_{(p)})} \right]. \tag{5.7}$$

The term to the right of the minus sign is a weighted average of τ's over the risk set, $R(t_{ij})$, with weights equal to the relative risks, $\exp(\gamma_j \tau_{(p)})$. The score function for τ has the form

$$\frac{\partial l^{ppl}}{\partial \tau_i} = -\sum_{j=1}^{J} \sum_{\{i' : i \in R(t_{i'j})\}} \frac{\gamma_j \exp(\gamma_j \tau_i)}{\sum_{p \in R(t_{i'j})} \exp(\gamma_j \tau_{(p)})} + \sum_{j=1}^{J} \gamma_j - \frac{\tau_i}{\sigma_\tau^2}, \tag{5.8}$$

where $\{i' : i \in R(t_{i'j})\} = \{i' : t_{i'j} \leq t_{ij}\}$. Equation (5.8) is obtained by summing the contribution of examinee i to the score function and those of examinees $i \neq i'$ who include i in their risk sets.

Computational Methods for PPL Estimation

The maximization of l^{ppl} is done by alternating between solving equations for $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$. Conceptually, this is analogous to the two-stage procedure (Birnbaum, 1968, p. 42) in item response theory (IRT). First, an initial value for $\boldsymbol{\tau}$ is guessed. In the first stage of PPL estimation, $\boldsymbol{\gamma}$ is estimated for the fixed value of $\boldsymbol{\tau}$. In the second stage, $\boldsymbol{\tau}$ is estimated assuming that the $\boldsymbol{\gamma}$ estimate obtained from the first stage is true. These two stages comprise one cycle. At the beginning of the next cycle, the $\boldsymbol{\tau}$ estimate from the second stage of the previous cycle is considered known and a new estimate for $\boldsymbol{\gamma}$ is obtained. Each cycle is repeated until a suitable convergence criterion is reached.

Within each stage, parameter estimates are found by successively obtaining better approximations to the true γ_j's. A standard practice to implement this process is to use the NR technique. Let $\hat{\gamma}_j^{(t)}$ represent the current approximation to the true value of γ_j obtained from the t-th iteration. Correspondingly, let $\hat{\gamma}_j^{(t+1)}$ be the updated value at $(t+1)$-th iteration. The NR equation for γ_j to be solved iteratively is written as

$$\hat{\gamma}_j^{(t+1)} = \hat{\gamma}_j^{(t)} - \left[\frac{\partial^2 l^{ppl}}{\partial \gamma_j^2}\right] \left[\frac{\partial l^{ppl}}{\partial \gamma_j}\right], \tag{5.9}$$

where

$$\frac{\partial^2 l^{ppl}}{\partial \gamma_j^2} = \sum_{i=1}^{N} \left[-\frac{\sum\limits_{p \in R(t_{ij})} \tau_{(p)}^2 \exp(\gamma_j \tau_{(p)})}{\sum\limits_{p \in R(t_{ij})} \exp(\gamma_j \tau_{(p)})} + \left(\frac{\sum\limits_{p \in R(t_{ij})} \tau_{(p)} \exp(\gamma_j \tau_{(p)})}{\sum\limits_{p \in R(t_{ij})} \exp(\gamma_j \tau_{(p)})} \right)^2 \right].$$

The first-order partial derivative of l^{ppl} with respect to γ_j is defined in (5.7). In the second stage, the $\boldsymbol{\gamma}$ estimates resulting from this stage are treated as the true regression coefficients, and the NR procedure is implemented for iteratively solving for $\boldsymbol{\tau}$. Analogous to above, let $\hat{\tau}_i^{(t)}$ and $\hat{\tau}_i^{(t+1)}$ denote the current and updated values for τ_i at iteration t and $t+1$, respectively. The NR equation for τ_i has the form

$$\hat{\tau}_i^{(t+1)} = \hat{\tau}_i^{(t)} - \left[\frac{\partial^2 l^{ppl}}{\partial \tau_i^2} \right] \left[\frac{\partial l^{ppl}}{\partial \tau_i} \right], \tag{5.10}$$

where

$$\frac{\partial^2 l^{ppl}}{\partial \tau_i^2} = \left[\sum_{\{i' :\, i \in R(t_{i'j})\}} \sum_{j=1}^{J} \frac{-\gamma_j^2 \exp(\gamma_j \tau_i)}{\sum\limits_{p \in R(t_{i'j})} \exp(\gamma_j \tau_{(p)})} + \left(\frac{\gamma_j \exp(\gamma_j \tau_i)}{\sum\limits_{p \in R(t_{i'j})} \exp(\gamma_j \tau_{(p)})} \right)^2 \right] - \frac{1}{\sigma_\tau^2}.$$

The first-order partial derivative of l^{ppl} with respect to τ_i is specified in (5.8). The notations of l^{ppl} in both (5.9) and (5.10) are implicit in that the penalized partial log-likelihood function is obtained using the current estimate, $\hat{\gamma}_j^{(t)}$ and $\hat{\tau}_i^{(t)}$, respectively. An approximate variance for each of the estimate can be obtained from the inverse of the minus second partial derivatives, as described above.

5.2 Joint Analysis of Responses and Response Times

In this section a method that analyzes examinees' latent traits is discussed based on the 3PLM and the PHLTM. van der Linden's (2007) hierarchical framework is employed for achieving this goal because of its flexibility in choices of measurement models and the capability to model correlation between the parameters.

Analogous to Chapter 2.3, constraints are imposed such that $\boldsymbol{\mu}_P = 0$ and $\sigma_\theta^2 = \sigma_\tau^2 = 1$ to establish the identifiability of the framework. Although σ_τ^2 needs not be fixed to a known constant within the original hierarchical framework, it is done so to remove the trade-off between τ and γ as well as to fix the scale of h_0. Additionally, we assume that the item parameters from the two measurement models are independent of one another. Wang et al. (2013) give three reasons to support this proposition: (i) the weak correlations of the parameters found in the previous literature, (ii) the complication of modeling the correlation between the time intensity and the item difficulty, and (iii) the minor influence on the parameter estimation even in the presence of the dependence. Molenaar et al. (2015) provide another example of the third point, demonstrating that neglecting the item parameter correlation does not result in bias or inefficiency. The independence assumption of the item parameters also lays groundwork for separate calibration of the measurement models, which will be explained further below.

Let $\boldsymbol{\Xi}$ denote the matrix of all item parameters. The likelihood of $\boldsymbol{\Xi}$ given the response matrix $\mathbf{U}$ and the response time matrix $\mathbf{T}$ is obtained as

$$L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T}) = \prod_{i=1}^{N} f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \boldsymbol{\Xi}),$$

where $\boldsymbol{u}_i = \{u_{ij}\}_{1 \le j \le J}$ is the vector of observed response scores for examinee i, and $f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \boldsymbol{\Xi})$ is the joint distribution of $(\boldsymbol{u}_i,\, \boldsymbol{t}_i)$ conditioned on $\boldsymbol{\Xi}$. The log of the marginal likelihood is expressed as

$$l = \log L(\boldsymbol{\Xi} \,|\, \mathbf{U},\, \mathbf{T}) = \sum_{i=1}^{N} \log \iint f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i,\, \boldsymbol{\Xi})\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega})\, d\theta_i d\tau_i, \tag{5.11}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\mu}_P,\, \boldsymbol{\Sigma}_P)$ denotes the hyperparameters for $(\theta_i,\, \tau_i)$.

Recall that the PPL estimation concurrently estimates $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$, which makes it difficult to collate information from the other sources of data. Hence, the latent trait parameters may be conveniently assigned independent priors when the PPL estimation is used for fitting the PHLTM. The prior research as well as the simulation studies presented in Chapter 3.3.1 suggested that the impact of the nonzero correlation between θ and τ on the item parameter estimation is rather small. For example, the simulation study of van der Linden et al. (2010) indicated that the reduction in MSEs due to the use of response times is at most 0.1 for a parameters and 0.05 for b parameters. The largest reduction in the MSEs occurred at the large a values and the extreme b values when the correlation between θ and τ, $\rho_{\theta\tau}$, was as high as 0.9. Because the relationship between θ and τ is only needed for marginalizing the likelihood function and the gain in the collateral information tends to be slight when the $\rho_{\theta\tau}$ is modest (see Ranger, 2013), the current study neglects potential auxiliary information when estimating the item parameters and assigns independent priors for each θ and τ. This strategy will allow us to calibrate the measurement models independently and conduct joint analysis on the latent trait parameters based on the separately estimated item parameters.

Assuming the independence between the latent traits, the marginal log-likelihood in (5.11) can be decomposed as

$$l \approx \sum_{i=1}^{N} \log \int f(\boldsymbol{u}_i \,|\, \theta_i;\, \boldsymbol{a},\, \boldsymbol{b},\, \boldsymbol{c})\, f(\theta_i \,|\, \mu_\theta,\, \sigma_\theta^2)\, d\theta_i + \sum_{i=1}^{N} \log \int f(\boldsymbol{t}_i \,|\, \tau_i;\, \boldsymbol{h}_0,\, \boldsymbol{\gamma})\, f(\tau_i \,|\, \mu_\tau,\, \sigma_\tau^2)\, d\tau_i,$$

where $\boldsymbol{a} = \{a_j\}_{1\le j\le J}$, $\boldsymbol{b} = \{b_j\}_{1\le j\le J}$, and $\boldsymbol{c} = \{c_j\}_{1\le j\le J}$. The first component in the above equation is only a function of the 3PLM parameters, and hence, its partial derivatives with respect to h_{0j} or γ_j are zero. For the same reason, the partial derivatives of the second component with respect to a_j, b_j, or c_j will equal zero. Therefore, item parameter estimates for the 3PLM and the PHLTM can be obtained separately using the standard estimation routines. The present study uses the marginal maximum likelihood estimation with Bayesian priors (Bock & Aitkin, 1981; Mislevy & Stocking, 1989) for estimating the 3PLM and either the PPL and PL estimation for fitting the PHLTM.

The joint analysis of response times and accuracy scores is implemented using the item parameter estimates obtained from the previous step. The posterior density of $(\theta_i,\, \tau_i)$ given observed data is

$$f(\theta_i,\, \tau_i \,|\, \boldsymbol{u}_i,\, \boldsymbol{t}_i,\, \boldsymbol{\Omega}) \propto f(\boldsymbol{u}_i,\, \boldsymbol{t}_i \,|\, \theta_i,\, \tau_i)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega}).$$

Conditional on the latent traits, the responses and response times are assumed to be independent (van der Linden & Glas, 2010). Hence, the posterior density can be rewritten as

$$f = f(\boldsymbol{u}_i \,|\, \theta_i)\, f(\boldsymbol{t}_i \,|\, \tau_i)\, f(\theta_i,\, \tau_i \,|\, \boldsymbol{\Omega}). \tag{5.12}$$

As stated in Chapter 3.3.2, point estimates of $(\theta_i,\, \tau_i)$ can be found by MAP or EAP estimators. Instead of maximizing (5.12), one can alternatively maximize the likelihood function by fixing the prior term equal to one, producing ML estimates for the latent traits. As alluded to above, the ML estimator is a special case of MAP estimator, and the EAP estimator can be easily computed using the same formulation presented in 3.13. Hence, the focus will be given to obtainment of MAP estimates in this section.

The NR technique is used to find the MAP estimates iteratively. Let $(\hat{\theta}_i^{(t)},\, \hat{\tau}_i^{(t)})$ represent the t-th approximation to the true values of the i-th latent traits. A better approximation, $(\hat{\theta}_i^{(t+1)},\, \hat{\tau}_i^{(t+1)})$, is obtained as

$$\begin{pmatrix} \hat{\theta}_i^{(t+1)} \\ \hat{\tau}_i^{(t+1)} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_i^{(t)} \\ \hat{\tau}_i^{(t)} \end{pmatrix} - E\left[\begin{pmatrix} \dfrac{\partial^2 \log f}{\partial \theta_i^2} & \dfrac{\partial^2 \log f}{\partial \theta_i \partial \tau_i} \\ \dfrac{\partial^2 \log f}{\partial \tau_i \partial \theta_i} & \dfrac{\partial^2 \log f}{\partial \tau_i^2} \end{pmatrix}\right]^{-1} \begin{pmatrix} \dfrac{\partial \log f}{\partial \theta_i} \\ \dfrac{\partial \log f}{\partial \tau_i} \end{pmatrix}. \tag{5.13}$$

Explicit expressions of the partial derivatives are obtained as follows.

$$\begin{aligned} \frac{\partial^2 \log f}{\partial \theta_i^2} &= \left[\sum_{j=1}^{J} \frac{D^2 a_j^2 (1 - P_{ij})(P_{ij} - c_j)(c_j u_{ij} - P_{ij}^2)}{P_{ij}^2 (1 - c_j)^2}\right] - \frac{\sigma_\tau^2}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2} \\ \frac{\partial^2 \log f}{\partial \theta_i \partial \tau_i} &= \frac{\sigma_{\theta\tau}}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2} \end{aligned}$$

$$\frac{\partial^2 \log f}{\partial \tau_i^2} = -\left[\sum_{j=1}^{J} \gamma_j^2 H_{0j}(t_{ij}) \exp(\gamma_j \tau_i)\right] - \frac{\sigma_\theta^2}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}$$

$$\frac{\partial \log f}{\partial \theta_i} = \left[\sum_{j=1}^{J} \frac{D a_j (u_{ij} - P_{ij})(P_{ij} - c_j)}{P_{ij}(1 - c_j)}\right] - \frac{\sigma_\tau^2(\theta_i - \mu_\theta) - \sigma_{\theta\tau}(\tau_i - \mu_\tau)}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}$$

$$\frac{\partial \log f}{\partial \tau_i} = \left[\sum_{j=1}^{J} \gamma_j - \gamma_j H_{0j}(t_{ij}) \exp(\gamma_j \tau_i)\right] - \frac{\sigma_\theta^2(\tau_i - \mu_\tau) - \sigma_{\theta\tau}(\theta_i - \mu_\theta)}{\sigma_\theta^2 \sigma_\tau^2 - \sigma_{\theta\tau}^2}$$

Equation (5.13) is based on the concept of Fisher's scoring in which the Hessian matrix is replaced with its expected values. For computing the partial derivatives with respect to τ_i, the nonparametric cumulative baseline hazard can be replaced by its Breslow estimator as presented in (5.2). The iterative process continues until the changes between the two successive approximations become sufficiently small. The standard errors of the final estimates can be found by calculating the square roots of the diagonal elements of the inverse of the negative Hessian matrix.

5.3 Simulation Study

5.3.1 Estimation of the Proportional Hazards Latent Trait Model

A simulation study was conducted to examine the performance of the PPL estimator. For comparison purposes, the PL estimator with the EM algorithm was evaluated using the same simulation conditions[2]. The simulation study considered two factors to create different test conditions. The first factor was the sample size, varied in two levels (N=250, 500)[3]. The second factor was the shape of baseline hazard rates. Specifically, two types of hazard functions were considered for generating response times, the exponential and the Weibull

[2] The author is indebted to Jochen Ranger for sharing his sample R code to implement the profile likelihood estimation.

[3] The sample size did not need to be as large as the previous studies because the parameters were estimated accurately with the small samples and the 2PL model was used as the response model.

distributions. The exponential distribution is the simplest parametric model and assumes a constant risk over time. However, this model can be sensitive to even a modest variation because it has only one adjustable parameter. The Weibull distribution overcomes this limitation by allowing flexibility in the shapes of the hazard functions. Characteristics of these distributions and the corresponding formulas for simulating response times (Bender, Augustin, & Blettner, 2005; Klein & Moeschberger, 2003, p. 38) are presented in Table 5.1.

Choices of the hyperparameters were made in accordance with the preceding studies (e.g., Ranger & Ortner, 2012, 2013; Wang et al., 2013). Specifically, for creating the exponential baseline hazard rates, the rate parameters were randomly sampled from a uniform distribution $U(0.25,\ 1.5)$. For the Weibull baseline hazards, the scale parameters were drawn from $U(0.25,\ 1.5)$, and the shape parameters were drawn from $U(1,\ 3)$. Equally spaced values were used for the regression parameters on the interval of [0.5, 1.5) with steps of 0.1, irrespective of the shapes of the hazard functions. The reason for this choice was to examine the estimation precision as a function of the γ values. Test length was fixed at $J = 20$. Examinees' speed parameters were randomly sampled from a standard normal distribution.

Table 5.1: Characteristics of Exponential and Weibull Distributions and Formulas for Generating Response Times

Characteristic	Exponential	Weibull
Parameter	Rate Parameter $\lambda > 0$	Scale parameter $\lambda > 0$ Shape parameter $\nu > 0$
Hazard rate	$h_0(t) = \lambda$	$h_0(t) = \lambda \nu t^{\nu-1}$
Cum. hazard rate	$H_0(t) = \lambda t$	$H_0(t) = \lambda t^{\nu}$
Cond. hazard rate	$h(t\|\tau) = \lambda \exp(\gamma\tau)$	$h(t\|\tau) = \lambda \exp(\gamma\tau)\nu t^{\nu-1}$
Density function	$f_0(t) = \lambda \exp(-\lambda t)$	$f_0(t) = \lambda \nu t^{\nu-1} \exp(-\lambda t^{\nu})$
Survival function	$S_0(t) = \exp(-\lambda t)$	$S_0(t) = \exp(-\lambda t^{\nu})$
Mean	$E[T] = \lambda^{-1}$	$E[T] = \lambda^{-1/\nu}\Gamma(1/\nu + 1)$
Variance	$\mathrm{Var}[T] = \lambda^{-2}$	$\mathrm{Var}[T] =$ $\lambda^{-2/\nu}\left[\Gamma\left(2/\nu + 1\right) - \Gamma^2\left(1/\nu + 1\right)\right]$
Survival time	$T = -\dfrac{\log U}{\lambda \exp(\gamma\tau)}$	$T = \left(-\dfrac{\log U}{\lambda \exp(\gamma\tau)}\right)^{1/\nu}$

Note. U is a random variable following a uniform distribution $U(0,\ 1)$

Within each simulation condition, 100 replications were implemented to regulate the sampling error and to examine the stability of the estimation performance. Datasets for each replication were generated separately.

Results

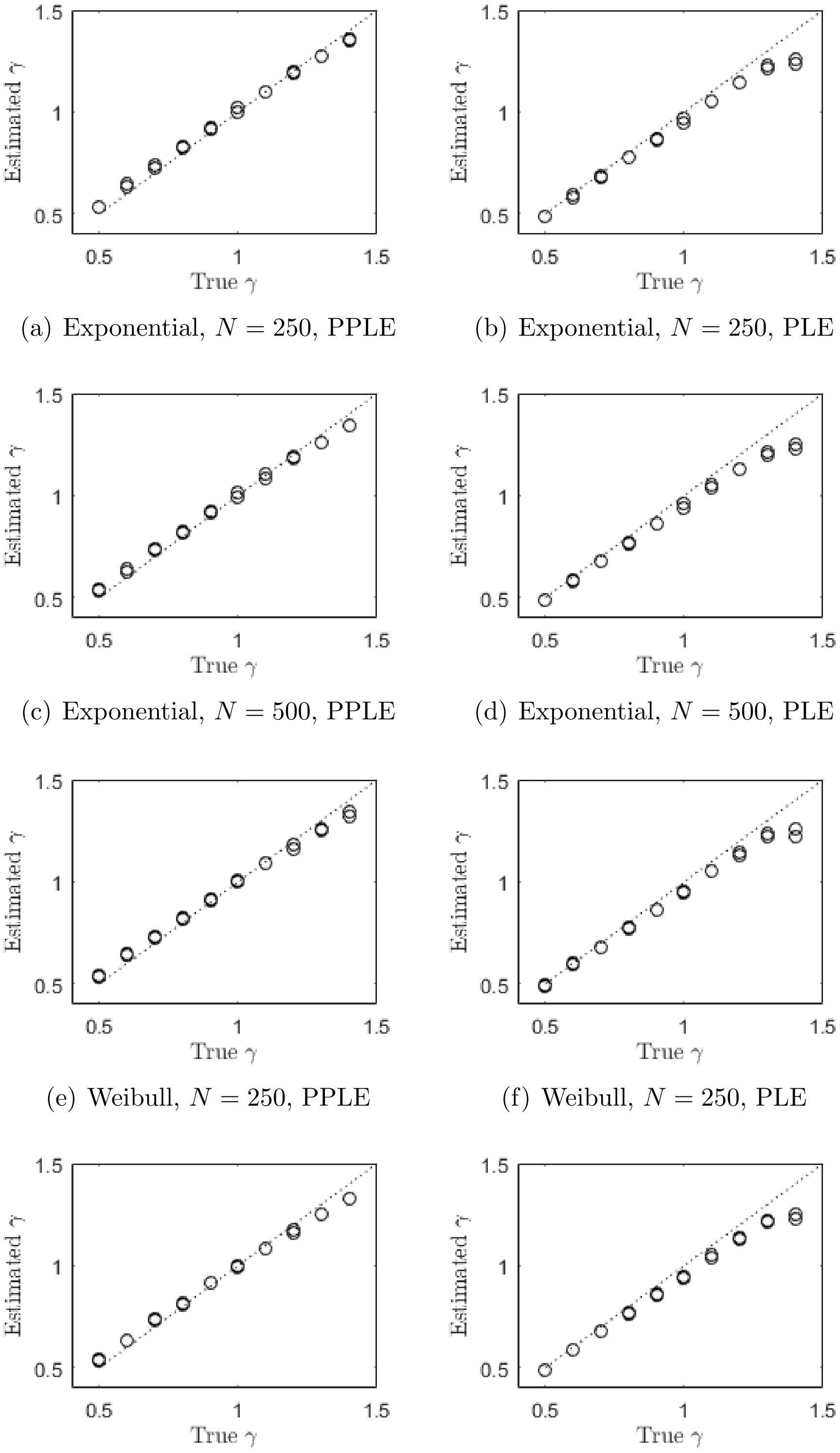
Table 5.2 reports biases and MSEs for recovering the true γ values. The values in parentheses provide SDs, which show the variability of the evaluation criteria across the replications. Table 5.2 suggests that the two estimation approaches performed reasonably well, yielding small estimation errors. The PPL estimation produced slightly smaller biases and MSEs than the PL estimation. While the MSE statistics were quite comparable between the two estimation methods, the differences in the biases seemed substantial. Increasing N generally resulted in smaller estimation errors, except for the biases in the PL estimation. The shapes of the baseline hazard rates appeared to have only minor influence on the recovery of the true γ values. MSEs and biases differed by less than a hundredth of a decimal place across the different distribution conditions.

To further examine the estimation behavior in terms of bias, the γ estimates averaged across the replications were plotted against the true values. Figure 5.1 suggests that the estimates from the PPL maximization had the positive biases for the small γ values and the

Table 5.2: Recovery of Regression Parameters

	Exponential				Weibull			
	$N = 250$		$N = 500$		$N = 250$		$N = 500$	
Criterion	PPLE	PLE	PPLE	PLE	PPLE	PLE	PPLE	PLE
Bias	.009	-.048	.003	-.055	-.001	-.048	-.004	-.055
	(.048)	(.042)	(.036)	(.029)	(.048)	(.043)	(.034)	(.028)
MSE	.010	.012	.006	.009	.013	.013	.008	.009
	(.005)	(.006)	(.003)	(.005)	(.005)	(.006)	(.003)	(.004)

Note. N = sample size. PPLE = penalized partial likelihood estimation. PLE = profile likelihood estimation. Values in the parentheses are standard deviations of the evaluation statistics.



(a) Exponential, $N = 250$, PPLE

(b) Exponential, $N = 250$, PLE

(c) Exponential, $N = 500$, PPLE

(d) Exponential, $N = 500$, PLE

(e) Weibull, $N = 250$, PPLE

(f) Weibull, $N = 250$, PLE

(g) Weibull, $N = 500$, PPLE

(h) Weibull, $N = 500$, PLE

Figure 5.1: True and Estimated Regression Parameters

negative biases for the large γ values. Nevertheless, the degree to which the estimates were biased seemed minimal at most. The estimates from the PL maximization were negatively biased throughout the scale of γ. The magnitude of the biases tended to increase as the true values of γ increased. The reason why the PL estimator had the large negative at the large γ's should be further investigated.

An additional simulation study was conducted to examine the appropriateness of standard error computation. For a fixed sample size $N = 250$, 10 datasets were randomly selected from the above simulation study, and generation of response time data and estimation of the parameters were repeated 100 times given the sets of true parameters. For each dataset selected, the average standard errors were calculated from the Hessian matrices, and the empirical standard deviations were computed across the 100 repetitions. Note that in contrast to PPL estimator, where both the γ and τ values are obtained simultaneously, the PL estimator requires separate estimation of τ using the estimated item parameter values. To obtain the τ estimates under the PL estimation, the equations pertinent to τ in (5.13) were considered, and ML estimates of τ's were calculated. Table 5.3 reports the average values over the examined datasets. The SDs presented in parentheses represent the variability in the observed standard errors across the 200 items under consideration. Table 5.3 suggests that theoretical standard errors of the PPL estimator were very close to the empirical coun-

Table 5.3: Theoretical and Empirical Standard Errors

	Exponential				Weibull			
	PPLE		PLE		PPLE		PLE	
Par	Theoretic	Empirical	Theoretic	Empirical	Theoretic	Empirical	Theoretic	Empirical
γ	.088	.087	.080	.091	.087	.087	.079	.089
	(.010)	(.012)	(.008)	(.013)	(.010)	(.012)	(.007)	(.012)
τ	.220	.217	.181	.190	.223	.220	.182	.192
	(.008)	(.017)	(.006)	(.025)	(.010)	(.018)	(.006)	(.025)

Note. Par = parameter. Theoretic = theoretical standard error. Empirical = empirical standard error. Values in the parentheses are standard deviations of the standard errors.

terparts, indicating that the estimated standard errors appropriately captured the variability in the parameter estimates across the samples. The PL estimator, on the other hand, tended to produce smaller standard errors than those from the empirical standard deviations.

5.3.2 Joint Estimation of Response and Response Time Models

The second simulation study was designed to show the performance of the proposed estimation methods in jointly analyzing response times and accuracy scores. A total of five factors were considered: (i) the item parameter estimation method (PPL and PL maximization), (ii) person parameter estimation method (EAP and MAP), (iii) test length (J =20 and 40), (iv) sample size (N =250 and 500), and (v) the correlation between the latent traits ($\rho_{\theta\tau}$ =0.0, 0.3, and 0.6). The values chosen for the $\rho_{\theta\tau}$ were motivated by the review of van der Linden (2009). Given the fixed $\rho_{\theta\tau}$, examinees' θ and τ parameters were obtained by randomly sampling from a bivariate normal distribution with zero means and unit variances.

Response times were simulated assuming the baseline hazard functions of the exponential or the Weibull distributions, analogous with Study 5.3.1. While the previous study shared the same distributional features within each test, the current study allowed the items to vary in terms of both the hazard rate distributions and the corresponding parameterization. Specifically, when the test length equaled 20, response times for randomly chosen 10 items were generated using the exponential baseline hazard rates, and the remainder were simulated assuming the Weibull baseline hazard rates. Likewise, in case of J =40, each set of 20 items had the exponential and the Weibull distributed baseline hazard functions. The rate and scale parameters were generated using the same hyperparameters with the earlier study. The regression parameters were randomly sampled from $U(0.5,\, 1.5)$.

Item parameters for the 3PLM were generated from a multivariate normal distribution. That is, $\log a \sim N(-0.043,\, 0.086)$ and $b \sim N(0,\, 1)$. The a and b parameters were allowed to covary within the 3PLM with the correlation of 0.4 as suggested in the prior study (e.g., Chan, Qian, & Ying, 2001). The hyperparameters for generating the a parameters corresponded to a mean of 1 and a standard deviation of 0.3 on the original scale. With the small sample sizes conditioned in this study, estimation of the c parameters can be unstable. Therefore, instead of estimating individual c_j's, the c values were fixed at a plausible constant, 0.2 (e.g., the reciprocal of the number of response alternatives of the items). The item parameter estimates for the 3PLM were obtained via marginal maximum likelihood (MML) estimation with the EM algorithm using the known priors.

Results

Table 5.4 reports the results of item parameter recovery analysis. Under the constructed settings, the sample size was the most influential factor in terms of item parameter estimation, and hence, the results were averaged over the J and $\rho_{\theta\tau}$ to get a clear picture of the estimation behavior. Table 5.4 suggests that a and b parameters were recovered quite well, having small biases and MSEs. These estimates tended to have larger errors and more variability in the estimation errors compared to γ estimates due to the property of the main source of data. While the response model parameters were estimated using the dichotomous data, and thereby, limited information was used to locate the true values, the γ parameters were estimated using the continuous response time data, exploiting much more information

Table 5.4: Recovery of Item Parameters Used for Joint Analysis

	$N = 250$				$N = 500$			
	a	b	γ		a	b	γ	
Criterion			PPLE	PLE			PPLE	PLE
Bias	.007	.003	-.002	-.074	.006	.004	-.006	-.079
	(.055)	(.069)	(.045)	(.043)	(.055)	(.053)	(.032)	(.031)
MSE	.067	.051	.011	.017	.067	.037	.006	.013
	(.017)	(.025)	(.004)	(.008)	(.018)	(.020)	(.002)	(.006)
SE	.212	.159	.088	.082	.151	.113	.062	.058
	(.011)	(.010)	(.002)	(.003)	(.008)	(.008)	(.001)	(.002)

Note. Values in the parentheses are standard deviations of the evaluation statistics.

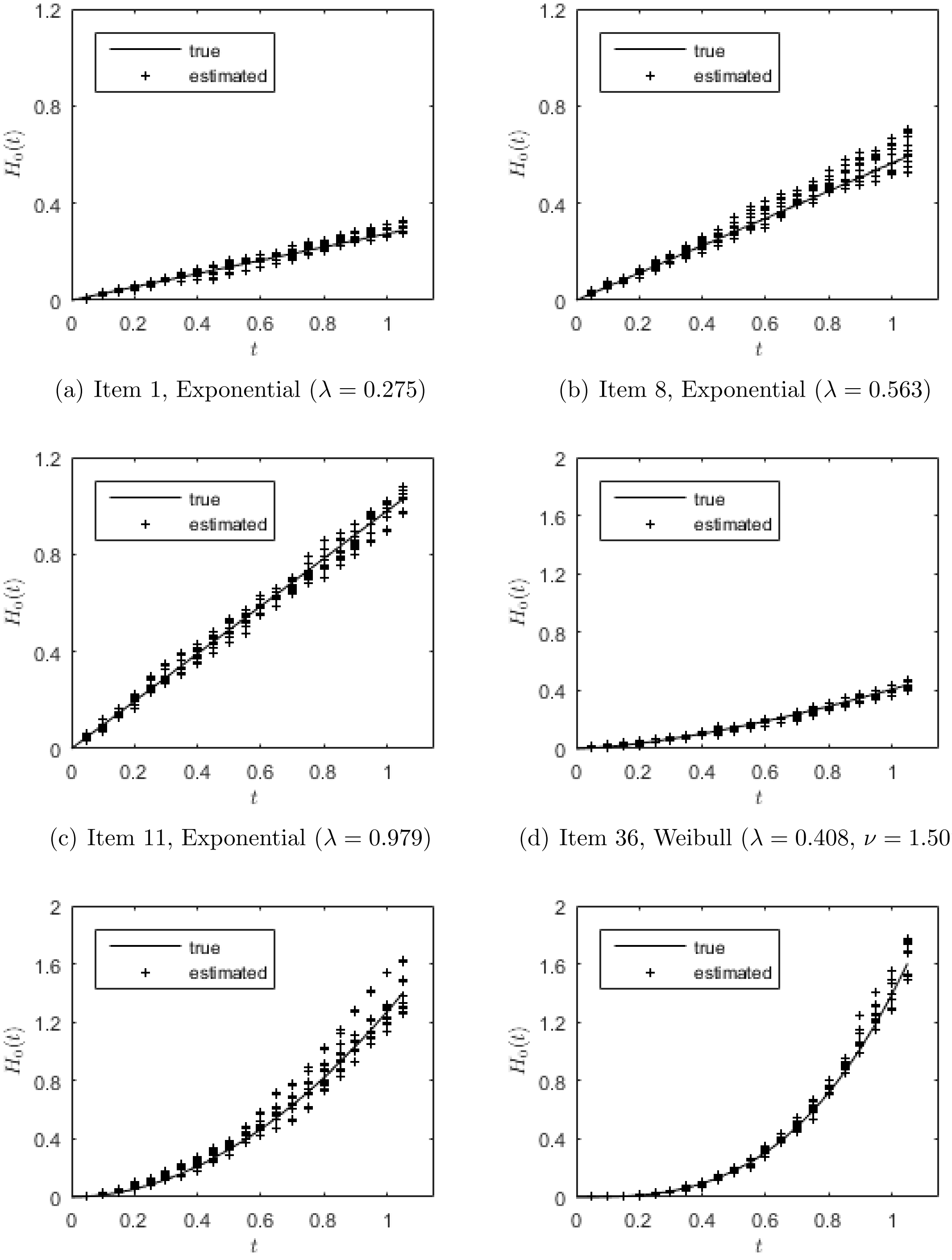
in finding the true locations. Irrespective of the kinds of the parameters, there was a clear tendency for the estimation errors to decrease as N increased. The results of the γ estimation followed the similar patterns as those reported in the previous study. The PPL estimator in general produced smaller biases and MSEs, and larger SEs than did the PL estimator.

Another issue that merits attention is the accuracy of the estimated cumulative baseline hazard rates. Because the H_{0j} is needed to estimate individual τ's, there is a parallel need for the H_{0j} to be accurately estimated. Figure 5.2 provides a good illustration of how well the true H_{0j}'s were approximated in the simulation study. For illustration purposes, we selected three items, varied in the true item parameter values, from a randomly chosen dataset and plotted the $\hat{H}_{0j}$ against the H_{0j}. Because each data set was used for 12 different simulation conditions—2 N's × 3 $\rho_{\theta\tau}$'s × 2 estimation methods—, there were 12 approximated lines for each true H_{0j}. At each observation, the $\hat{H}_{0j}(t_{ij})$ value was obtained using the $\hat{\gamma}_j$ as a proxy for the true value. A piecewise linear interpolation was used to approximate $\hat{H}_{0j}$ within the time grid points (t) of $[0, 1.05]$ at increments of 0.05. In Figure 5.2, the Breslow estimator appeared to recover the true H_{0j}'s quite well under the different shapes. The estimation accuracy conditioned on t tended to decline as the t increased. The possible reason is the diminishing risk sets along with the increasing t. Item 27, for instance, gradually deviated from the true H_{0j} as t increased to 1. Provided that the item had a mean of response times with 0.899, there were fewer individuals left for computing the risk sets when the t was close to 1, thereby rendering the estimation of H_{0j} unstable in this region.

To quantify the overall discrepancy between the true and estimated H_{0j}'s under all conditions simulated, the mean absolute error (MAE) statistics were computed:

$$MAE_j = \frac{1}{N}\sum_{i=1}^{N}\left|H_{0j}(t_{ij}) - \hat{H}_{0j}(t_{ij})\right|.$$

Table 5.5 reports the MAE statistics classified according to the same functional forms of the baseline hazard rates. Since no obvious pattern was found across the levels of $\rho_{\theta\tau}$, the



(a) Item 1, Exponential ($\lambda = 0.275$)

(b) Item 8, Exponential ($\lambda = 0.563$)

(c) Item 11, Exponential ($\lambda = 0.979$)

(d) Item 36, Weibull ($\lambda = 0.408$, $\nu = 1.507$)

(e) Item 27, Weibull ($\lambda = 1.270$, $\nu = 1.957$)

(f) Item 32, Weibull ($\lambda = 1.387$, $\nu = 2.955$)

Figure 5.2: True and Estimated Cumulative Baseline Hazard Rates

MAE statistics were averaged over the $\rho_{\theta\tau}$. Table 5.5 suggests that the two item parameter estimation methods had minor differences in the MAEs. There was no systematic pattern across these factors. Increasing N had a distinct impact on the estimation of the H_{0j}'s. The larger the N, the smaller the MAEs. The larger N also appeared to result in more consistent performances in retrieving the true values of H_{0j}. Increasing J generally led to more accurate estimates of H_{0j}, possibly attributed to the use of more precise τ estimates.

Based on the item parameter estimates obtained above, the latent traits were jointly estimated as described in Chapter 5.2. Table 5.6 summarizes the results averaged over N. (The average was taken because the sample size does not affect the individual parameter recovery.) In Table 5.6, results for θ estimates corresponded to when the 3PLM and the PHLTM item parameters were estimated from the MML estimator and the PPL estimator, respectively. The results for the MML and the PL estimators were omitted from the table due to small differences with those reported in Table 5.6. (The values were the same other than a few occasions under the EAP conditions, in which the differences occurred less than 0.001).

Table 5.6 suggests that the θ estimates had small biases and reasonable MSEs under the conditions established. The impact of increasing J was manifested by smaller biases and MSEs. In general, EAP estimates resulted in smaller biases and MSEs than the MAP

Table 5.5: Mean Absolute Errors of Estimated Cumulative Baseline Hazard Rates

	$N = 250$				$N = 500$			
	Exponential		Weibull		Exponential		Weibull	
J	PPLE	PLE	PPLE	PLE	PPLE	PLE	PPLE	PLE
20	.296	.287	.282	.283	.216	.213	.215	.210
	(.233)	(.213)	(.226)	(.220)	(.159)	(.156)	(.173)	(.158)
40	.283	.278	.282	.284	.208	.209	.208	.210
	(.218)	(.215)	(.229)	(.236)	(.151)	(.155)	(.159)	(.164)

Note. Values in the parentheses are standard deviations of the mean absolute error statistics across replications.

Table 5.6: Recovery of Latent Trait Parameters via Joint Analysis

			$\rho_{\theta\tau} = 0.0$			$\rho_{\theta\tau} = 0.3$			$\rho_{\theta\tau} = 0.6$		
			θ	τ		θ	τ		θ	τ	
Method	J	Criterion		PPLE	PLE		PPLE	PLE		PPLE	PLE
MAP	20	Bias	.043	.022	.026	.034	.017	.022	.033	.020	.024
		MSE	.215	.051	.056	.210	.051	.056	.192	.051	.054
	40	Bias	.030	.014	.017	.020	.010	.013	.021	.012	.015
		MSE	.123	.028	.035	.123	.028	.034	.115	.028	.034
EAP	20	Bias	.001	-.003	.000	-.009	-.008	-.004	-.006	-.006	-.003
		MSE	.209	.051	.055	.206	.051	.055	.188	.051	.054
	40	Bias	.002	.001	.003	-.007	-.003	-.001	-.005	-.001	.001
		MSE	.119	.028	.035	.120	.028	.034	.112	.028	.034

estimates. When the J equaled 40, the differences in the MSEs seemed inconsequential between the two estimation methods. Table 5.6 also suggests that the level of $\rho_{\theta\tau}$ had a distinct impact on the θ estimation. The increasing value of $\rho_{\theta\tau}$ resulted in improved estimation accuracy, confirming the results of van der Linden et al. (2010), where response times were used as a collateral information for estimating θ.

While considering the response times improved the θ estimation, the other way around—that is, considering the responses in τ estimation—did not appear to have much influence on the τ estimation. Overall, no systematic pattern was found across the varying $\rho_{\theta\tau}$, possibly because the τ values were already well estimated using the continuous response time data. The amount of information gain as a result of considering the discrete response data seemed negligible. Comparison between the PPL and PL estimation conditions revealed that the PPL estimator produced more accurate τ estimates. The overall biases and MSEs from the PPL estimator were smaller than those from the PL estimator except for a few occasions in biases under the EAP estimation. Similar to θ estimation, increasing J improved the accuracy of the τ estimates. The τ estimates from the EAP produced smaller biases than those from MAP; MSEs were quite comparable between the two estimation methods.

The accurate recovery of the true level of $\rho_{\theta\tau}$ is also of concern in the present analysis.

Table 5.7 reports the empirical estimates of the correlation between the latent trait estimates. The results were averaged across the MAP and EAP conditions. Despite the fact that item parameters were estimated assuming the independence between the latent traits, the true values of $\rho_{\theta\tau}$ seemed to be recovered quite well. Overall, compared to the PL estimation, the PPL estimation tended to produce closer estimates to the true $\rho_{\theta\tau}$ values. The accuracy of the estimates improved as the J increased and/or N increased.

Table 5.7: Recovery of Correlation between Latent Traits

		$\rho_{\theta\tau} = 0$		$\rho_{\theta\tau} = .3$		$\rho_{\theta\tau} = .6$	
N	J	PPLE	PLE	PPLE	PLE	PPLE	PLE
250	20	.013	.013	.336	.341	.662	.669
		(.064)	(.064)	(.053)	(.053)	(.035)	(.035)
	40	.012	.012	.318	.321	.629	.635
		(.069)	(.069)	(.052)	(.052)	(.036)	(.036)
500	20	.008	.008	.337	.342	.661	.668
		(.045)	(.045)	(.043)	(.043)	(.022)	(.023)
	40	.007	.007	.316	.319	.634	.640
		(.043)	(.043)	(.042)	(.042)	(.024)	(.024)

Note. Values in the parentheses are standard deviations of the correlations.

5.4 Real Data Example

The proposed methods were applied to an empirical dataset[4]. The dataset contained observations from 250 examinees. Each examinee answered 30 items from a spatial rotation test—i.e., the Purdue Spatial Visualization Test: Rotations (PSVT-R) of Guay (1976). The test has been found to have predictive validity for success in science, technology, engineering, and mathematics majors (e.g., Maeda & Yoon, 2013).

[4] The data set was supplied by Professor Steven A. Culpepper. The author is grateful to Dr. Culpepper for his generous support.

5.4.1 Model Diagnosis

Model fit statistics were checked for fitting the 2PLM and the PHLTM. The 2PLM parameters were estimated from the commercial calibration program, PARSCALE (Muraki & Bock, 2003). Using the baseline of 15 intervals, the likelihood-ratio χ^2-statistic for the whole test was found as 215.885 with the degrees of freedom of 242. These values corresponded to the p-value of 0.885, indicating the good fit of the 2PLM. The average p-value for the individual items was 0.563, with the minimum of 0.035[5] and the maximum of 0.98. Overall, the items were found to adequately fit the 2PLM at the significance level 0.01.

To evaluate the global fit of the PHLTM, the posterior predictive probability was computed for each observation t_{ij} as

$$\Pr(\tilde{t}_{ij} < t_{ij}), \quad i = 1, \ldots, N, \ j = 1, \ldots, J.$$

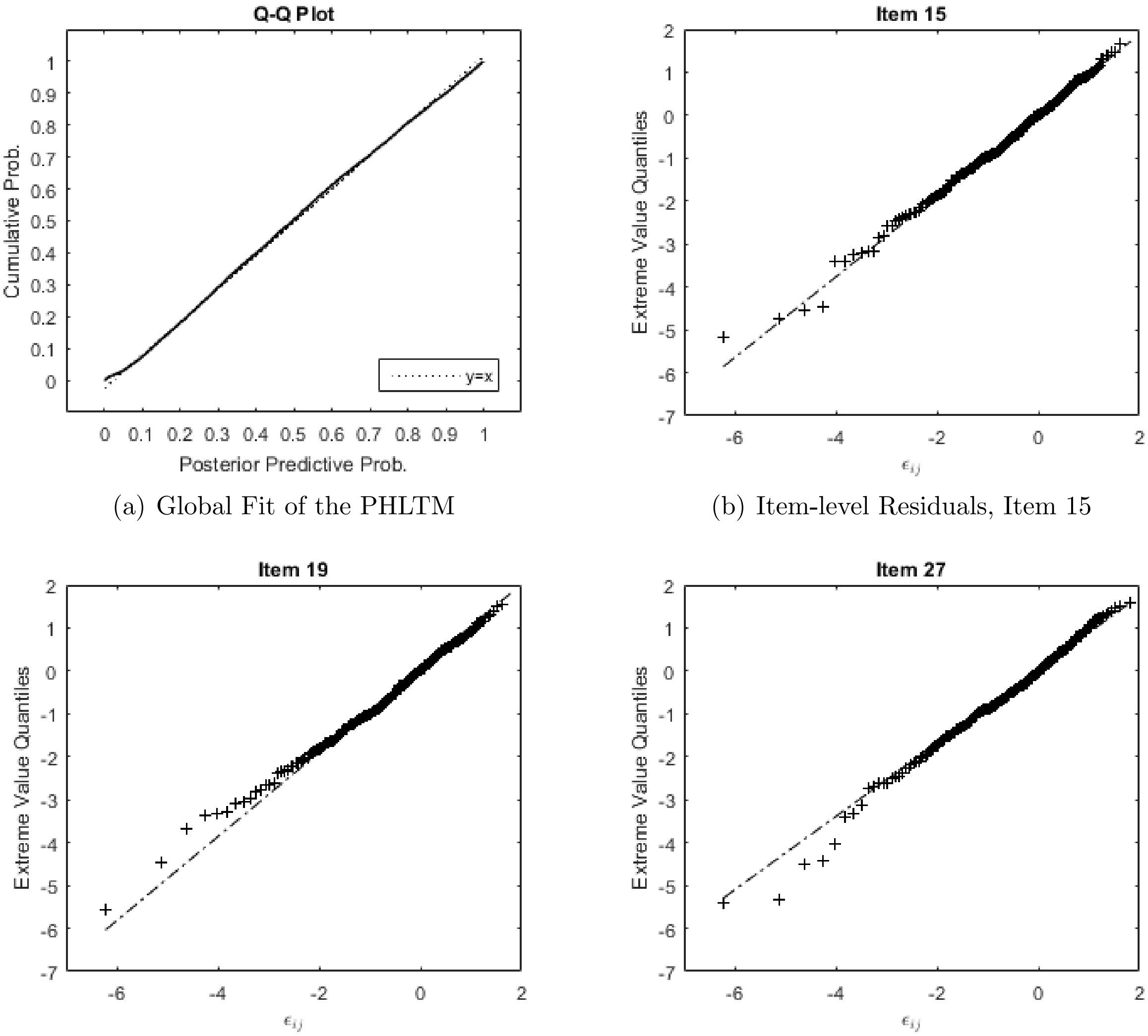
If the model fits, the cumulative distribution of these probabilities over the examinee-item combinations follows the identity line (van der Linden, Breithaupt, Chuah, & Zhang, 2007). Figure 5.3(a) presents the Q-Q plot of the empirical cumulative distribution of the posterior predictive probabilities for all observations. The empirical distribution coincided with the identity line suggests the appropriate fit of the PHLTM to the dataset of concern.

To further examine the item-level fit of the model, residuals under the PHLTM were calculated for each item-examinee pair as follows.

$$\hat{\varepsilon}_{ij} = \log\left[\hat{H}_{0j}(t_{ij})\right] - (-\hat{\gamma}_j\hat{\tau}_i).$$

If the model fits the data well, the N residuals for a given item follow the standard type I extreme value distribution (Wang et al., 2013), which is also known as the Gumbel dis-

[5] Two items showed p-values between 0.01 and 0.05. Note that the χ^2-statistic is a function of the number of intervals specified. When the wider intervals were used (e.g., 10), these two items showed p-values of 0.358 and 0.675.



(a) Global Fit of the PHLTM

(b) Item-level Residuals, Item 15

(c) Item-level Residuals, Item 19

(d) Item-level Residuals, Item 27

Figure 5.3: Diagnostic Plots for PHLTM

tribution for minimums with the location parameter being 0 and the scale parameter being 1. Examples of the distributions of $\hat{\varepsilon}_{ij}$ are plotted against the corresponding Gumbel quantiles in Figures 5.3(b)–5.3(d). Figure 5.3(b) provides an example of items with a good fit, whereas Figures 5.3(c) and 5.3(d) present examples of ill-fitting items. Overall, items under evaluation shared the alike residual plots. Some showed good adherence to the identity line as in Figure 5.3(b); others exhibited slight deviation at the negative residuals as in Figure 5.3(c) or 5.3(d). Although the negative residuals were larger or smaller than expected for some items, the distribution of the residuals closely resembled the Gumbel density when

$\varepsilon_{ij} \geq -4$. Provided that the consistent misfit exhibited at the low end of ε_{ij} did not yield tangible evidence of global misfit in Figure 5.3(a), the misfitting residuals at the item level probably comprised a relatively small proportion of the data.

Finally, a Lagrange multiplier (LM) test was implemented to check the local independence of the observations. A violation of the conditional independence assumption between response times and responses can be evaluated by embedding a plausible parameter in the PHLTM (van der Linden & Glas, 2010; C. Wang, Fan, et al., 2013):

$$h_{ij}(t) = h_{0j}(t)\exp(\gamma_j\tau_i + \eta_j u_{ij}),$$

where η_j is a plausible value that denotes the dependence on the item response u_{ij}. The null hypothesis of the local independence is stated as $\mathcal{H}_0 : \eta_j = 0$. The test statistic is obtained as

$$LM(\eta_j) = \left. \frac{g(\eta_j)^2}{g(\eta_j,\, \eta_j) - \boldsymbol{G}(\boldsymbol{\tau},\, \eta_j)'\boldsymbol{G}(\boldsymbol{\tau},\, \boldsymbol{\tau})^{-1}\boldsymbol{G}(\boldsymbol{\tau},\, \eta_j)} \right|_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}},\, \eta_j=0},$$

where $g(\eta_j)$ is the first-order derivatives of the log-likelihood of the alternative model with respect to η_j; $g(\eta_j,\, \eta_j)$ is the corresponding observed information for η_j; $\boldsymbol{G}(\boldsymbol{\tau},\, \eta_j)$ is the observed information vector with size of N; and $\boldsymbol{G}(\boldsymbol{\tau},\, \boldsymbol{\tau})$ is the $N \times N$ diagonal information matrix for $\boldsymbol{\tau}$. Each quantity is calculated as follows.

$$\begin{aligned}
\boldsymbol{G}^{ii}(\boldsymbol{\tau},\, \boldsymbol{\tau}) &= \gamma_j^2 H_{0j}(t_{ij})\exp(\gamma_j\tau_i + \eta_j u_{ij}) + \sum_{l\neq j}^{J} \gamma_l^2 H_{0l}(t_{il})\exp(\gamma_l\tau_i) \\
g(\eta_j) &= -\sum_{i=1}^{N} u_{ij}\left[1 - H_{0j}(t_{ij})\exp(\gamma_j\tau_i + \eta_j u_{ij})\right] \\
g(\eta_j,\, \eta_j) &= \sum_{i=1}^{N} u_{ij}^2 H_{0j}(t_{ij})\exp(\gamma_j\tau_i + \eta_j u_{ij}) \\
\boldsymbol{G}^{i}(\boldsymbol{\tau},\, \eta_j) &= H_{0j}(t_{ij})\exp(\gamma_j\tau_i + \eta_j u_{ij})\gamma_j u_{ij}
\end{aligned}$$

The LM statistic, $LM(\eta_j)$, is asymptotically χ^2-distributed with one degree of freedom.

When the LM tests were applied to the empirical data, test statistics were computed using the estimates obtained from the PPL estimation. The average p-value was found as 0.4679, with the minimum of 0.005 and the maximum of 0.981. There was one item that showed the significance at the 1% level, and this item was removed from further analysis.

5.4.2 Parameter Estimation

Findings from the above analysis suggest that the 2PLM and the PHLTM fit the observed data satisfactorily. Building on the outcomes of the previous analysis, summative statistics of the empirical item parameter estimates and their corresponding standard errors are examiend in Table 5.8. Overall, the parameter estimates seemed to have small standard errors despite the small calibration sample size. The test items were found moderately discriminating and relatively easy for test takers with zero ability levels. The a and b parameter estimates were found slightly negatively correlated with -0.116. The standard errors of the γ estimates were consistently smaller than those of the item response model. The PPL and PL estimators appeared to produce similar estimates in general. The absolute differences between the PPL and PL estimation conditions were on average 0.107, with the maximum of 0.312 and the minimum of 1.7×10^{-5}. The overall correlation between the two sets of estimates was 0.784.

In a like manner, Table 5.9 summarizes the descriptive statistics for the latent trait estimates. The third line indicates the method used for estimating the PHLTM item parameters.

Table 5.8: Descriptive Statistics of Item Parameter Estimates from Empirical Data

	Parameter Estimates				Standard Errors			
	a	b	γ		a	b	γ	
Criterion			PPLE	PLE			PPLE	PLE
Avg	.831	-.563	.656	.726	.126	.146	.072	.074
SD	.223	.664	.193	.168	.027	.063	.006	.005
Min	.447	-1.807	.273	.309	.086	.082	.062	.065
Max	1.333	1.320	1.069	1.016	.215	.358	.084	.085

Note. Avg=Average. SD=Standard deviation.

The item parameters for the 2PLM were uniformly estimated via MML estimation. Table 5.9 suggests that the latent trait estimates were centered around 0. The SDs of the θ estimates were less than 1, whereas those of the τ estimates were slightly larger than 1. In line with the prior results, the τ estimates generally had smaller standard errors than the θ estimates. The correlation between the two sets of latent trait estimates ranged from -0.411 to -0.412 depending on the item and person parameter estimation methods. The negative values of the correlations indicate that the more capable test takers tended to answer the items slower. The two item parameter estimation methods produced very similar solutions. The θ and τ estimates from the PPL estimation and the PL estimation had the correlation of 1 and 0.998, respectively.

Table 5.9: Descriptive Statistics of Person Parameter Estimates from Empirical Data

		Parameter Estimates				Standard Errors			
		θ		τ		θ		τ	
Method	Criterion	PPLE	PLE	PPLE	PLE	PPLE	PLE	PPLE	PLE
MAP	Avg	.016	.018	.026	.061	.343	.343	.273	.251
	SD	.798	.800	1.058	1.016	.093	.094	.012	.011
EAP	Avg	.041	.043	.002	.039	.349	.349	.275	.253
	SD	.829	.831	1.056	1.014	.101	.101	.013	.011

Note. Avg=Average. SD=Standard deviation.

5.5 Summary

The present chapter has provided an estimation routine for fitting the semiparametric PHLTM. The procedure was based on the PPL estimation, in which latent speed variables are constrained by a penalty function. The implementation of the proposed method was validated via simulation studies. The studies provided comparisons of the PPL estimation method with the PL-based solutions. An application of the estimation methods to a real dataset was also provided.

Chapter 6

Discussion and Future Work

6.1 Discussion

The nature of speededness in operational tests and easy availability of response time data in computerized tests have inspired much research on the response times in educational and psychological testing. In this thesis, likelihood-based approaches to estimating the hierarchical framework (van der Linden, 2007) are developed to efficiently and accurately obtain parameter estimates for the response and response time models.

Two approaches were presented for jointly estimating the 3PLM and the lognormal response time model. One is based on the MML estimation, and the other builds on the MMAP estimation. Both approaches were implemented with the EM algorithm to cope with unknown latent variables in item calibration. Using the estimated item parameter values, examinees' latent trait parameters—the proficiency and speed parameters—were estimated through the MAP and EAP estimators.

The estimation procedures, as implemented in this study, provided highly reliable and accurate results with respect to recovery of both the item and person parameters. Simulation studies presented in Chapters 3.3.1 and 3.3.2 suggest that the overall MSEs and biases of the estimated parameters were small, and the values of correlations between the parameters were recovered well. Although the MML estimation procedure showed occasional convergence problems, they mostly concerned the failure to home in on the pre-specified interval, which was set in a somewhat stringent manner. Despite the strict convergence criteria specified, the MMAP estimation procedure, on the other hand, showed excellent convergence rates as a result of utilizing prior information at the second level of the hierarchical framework. The results relating to the simulation factors were largely consistent with expectations. As a general rule, the larger the samples, the better the quality of parameter estimates.

The higher levels of correlations resulted in better estimation of the parameters than lesser levels as well. Overall, no substantive differences across varying ρ_P's were evident in item calibration, most likely due to marginalization, so too were there no systematic differences in latent trait estimates due to different ρ_I levels.

The other aspect explored in this study was the feasibility of the likelihood-based procedures in calibrating the items in CAT. In Chapter 4.1, Fisher information matrix of item parameters for the hierarchical framework was developed for adaptively selecting calibration samples during the CAT administrations. A total of four optimality sampling designs were proposed that differ in terms of treatment of the information matrix and the purpose of the online calibration. A simulation study presented in Chapter 4.3 suggests that the MMAP estimation accompanied with the EM algorithm performed well despite the relatively small sample sizes ($N = 400 \sim 800$). The overall MSEs observed remained small, and the biases of the estimates were close to zero. Increasing N or ρ_I consistently resulted in reduced estimation errors across the simulation conditions.

With respect to the sampling design, D-optimality was generally found more effective than A-optimality for selecting the calibration samples. The possible reason for this trend is that, while D-optimality takes into account the information from the respective item parameters as well as the joint information between the parameters, A-optimality only considers the information from the individual item parameters, and thus, capitalizing on the limited information in selecting the calibration samples. From the perspective of implementation, the two approaches are comparable in complexity, and hence, the choice of strategy should be driven by better outcomes. The second trend of note concerns the differences relating to the purpose of online calibration. Results from the simulation study suggest that when the purpose of field-testing is centered on accurately estimating the parameters of only the response model, D_S- or A_S-optimality should be preferred to D- or A-optimality. While the D- and D_S-optimal sampling design produced comparable results, A_S optimality clearly outperformed A-optimality by improving the estimation precision of the parameters of interest.

Provided in Chapter 5 was the extension of the hierarchical framework into a more flexible response time model. The PHLTM (Ranger & Ortner, 2012), a modified version of the Cox PH model, was chosen for its increasing popularity in the measurement literature. In this study, the PHLTM was fit in a semiparametric fashion by leaving the baseline rates unknown, whereby the model allows flexibility in modeling response time distributions. The estimation procedure was based on the PPL in which latent speed parameters are constrained by a penalty function. It is computationally similar to other shrinkage methods for penalized regression, such as ridge regression and smoothing splines. Simulation studies presented in Chapter 5.3 suggest that the PPL estimator produced smaller errors than the PL estimator in recovering the true regression parameters and latent speed parameters. While the PL estimator tended to underestimate standard errors of the parameter estimates, the PPL estimator appeared to faithfully capture the true standard errors of the estimates. The application of the proposed estimation method within the hierarchical framework was also provided for jointly analyzing accuracy scores and response times.

6.2 Future Work

There are a number of research areas that seem worthy of pursuing in the future. The first direction of interest concerns the investigation of the calibration-based procedures in detecting compromised items. CAT, for example, administers tests continuously or at frequent time intervals using a pre-constructed item pool. When CAT is administered for a while, some items may be compromised due to cheating or item sharing among test takers. Compromised items typically reveal themselves as aberrances in both responses and response times. Beneficiaries of a leaked item, for instance, are likely to respond to the item correctly spending distinctly short time, wherein the item may become easier and less time intensive across the span of time. Therefore, considering both responses and response times simultaneously as a source of evidence may improve the statistical power of detecting the presence of compromised items.

There have been indeed several approaches for detecting examinees' cheating behaviors through person fit testing (e.g., Marianti et al., 2014; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003); however, not much attention has been given to identification of compromised items with the aid of response times. If a testing program is run based on particular psychometric models that are properly assumed, item parameter estimates for the response and response time models may be used to evaluate changes in statistical properties of an item over time. In so doing, clearer assurance about which item parameters have shifted from the original values as well as to what extent the parameter change has occurred can be obtained. Information from this procedure can also lead to more targeted actions when mitigating the compromised items.

The other important extension of this study would be to explore a fully parametric approach for estimating the PHLTM. Compared to the semiparametric approach, the parametric estimation approach offers much simple and powerful estimation methods for the PHLTM. It capitalizes on the knowledge about the functional forms of baseline hazard rates, and therefore the complication arising from the unknown baseline hazard rates can be removed. In the parametric approach, the estimation procedure can be done in a standard manner that maximizes the log of the likelihood function marginalized with respect to latent speed parameters.

Finally, investigation of linking methods for the PHLTM may be a direction of interest for researchers. Similar to the linear indeterminacy problem in the IRT (Lord, 1980, p. 36-38), the PHLTM bears identification problem due to its exponent component. Thus, in order to ensure valid inference about examinees' test performances across time and different testing occasions, linking procedures are needed for placing parameter estimates on a common scale. While there has been literature for linking parameters of the lognormal response time model (e.g., van der Linden, 2010), no study to date has been done for the PHLTM despite its high potential. It appears desirable for future studies to develop more thorough theoretical grounds for the further application of the PHLTM.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (9th ed.). New York: Dover.

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

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 191-212.

Ban, J.-C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, *39*(3), 207-218.

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

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*, 1713-1723.

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement*, *15*, 293-306.

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, *19*, 43-56.

Birnbaum, A. (1968). Theories of mental test scores. In F. M. Lord & M. R. Novick (Eds.), *Some latent trait models and their use in inferring an examinee's ability* (p. 397-479). MA: Addison-Wesley, Reading.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an em-algorithm. *Psychometrika*, *46*, 443-459.

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

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211–252.

Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B*, *34*, 216217.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear models. *J. Amer. Statist. Assoc*, *88*, 925.

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

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, *10*(3), 273-304.

Chang, H.-H., Qian, J., & Ying, Z. (2001). α-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*, 333-341.

Chang, Y.-c. I., & Lu, H.-Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika*, *75*, 140-157.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77*, 201-222.

Clayton, D. G., & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, *148*, 82117.

Cortiñas, A. J., & Burzykowski, T. (2005). A version of the EM algorithm for proportional hazards model with random effects. *Biometrical J*, *47*, 847862.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 187-220.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Douglas, J., Kosorok, M., & Chewing, B. (1999). A latent variable model for discrete multivariate psychometric waiting times. *Psychometrika*, *64*, 69–82.

Embreston, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380-396.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655-670.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543.

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.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351–380.

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603-626.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension items: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351-373.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression model. *International Statistical Review*, *55*, 245-259.

Guay, R. (1976). *Purdue spatial visualization test.* West Layfette, IN: Purdue University.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hougaard, P. (1991). Modelling heterogeneity in survival data. *Journal of Applied Probability*, *28*, 695–701.

Kale, B. K. (1962). On the solution of likelihood equations by iteration processes: The multiparameter case. *Biometrika*, *49*, 479-486.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics. volume 2. inference and relationship.* New York: Macmillan.

Kennedy, M. (1930). Speed as a personality trait. *Journal of Social Psychology*, *1*, 286–298.

Klein, J., & Moeschberger, M. (2003). *Survival analysis: Techniques for censored and truncated data.* New York: Springer-Verlag.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54–75.

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, *37*, 218-220.

Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, *67*, 304-327.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157–162.

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

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. (Ser. B)*, *44*, 190200.

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

Maeda, Y., & Yoon, S. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT:R). *Educational Psychology Review*, *25*, 69-94.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426-451.

Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445469.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195.

Mislevy, R. J., & Stocking, M. L. (1989). A consumers guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57-75.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197–219.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*(2), 273-296.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*, 252-284.

Muraki, E., & Bock, D. (2003). Parscale (version 4.1): IRT item analysis and test scoring for rating-scale data [Computer software manual]. Chicago, IL.

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

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

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement.* University of Notre Dame. (Unpublished doctoral dissertation)

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

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*, 41-70.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*, 538-544.

Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31-47.

Ranger, J., & Kuhn, J.-T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *67*, 388–407.

Ranger, J., & Kuhn, J.-T. (2015). A mixture proportional hazards model with random effects for response times in tests. *Educational and Psychological Measurement*. doi: 10.1177/0013164415598347

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*, 334-349.

Ranger, J., & Ortner, T. M. (2013). Response time modeling based on the proportional hazards model. *Multivariate Behavioral Research*, *48*, 503–533.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, *1*(2), 33–49.

Ripatti, S., & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, *56*, 10161022.

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. Linden. & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589606.

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

Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embretson (Ed.), *Test design: Developments in psychology and education* (p. 219-244). New York: Academic Press.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method for measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

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.

Silvey, S. D. (1980). *Optimal design.* London: Chapman & Hall.

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

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, *9*(6), 1135–1151.

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

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of diffculty. *Educational and Psychological Measurement*, *8*, 353–374.

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* (p. 236-256).

Therneau, T. M., Grambsch, P. M., & Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, *12*(1), 156-175.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (p. 179-203). New York: Academic Press.

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

Vaida, F., & Xu, R. (2000). Proportional hazards model with random effects. *Statist. Medicine*, *19*, 3309–3324.

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*, *72*, 287-308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5-20.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.

van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 24–41.

van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, *47*, 92–114.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*, 44-60.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120-139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365?384.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Pscyhological Measurement*, *34*(5), 327–347.

van der Linden, W. J., & Ren, H. (2014). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, *80*, 263-288.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, *68*, 251-265.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

Verhelst, N. D., Verstralen, 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-Verlag.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (p. 65-102). Hillsdale, NJ: Erlbaum.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381-417.

Wang, T. (2006). *A model for the joint distribution of item response and response time using a one-parameter Weibull distribution* (Tech. Rep.). Iowa City, IA: University of Iowa. (Center for Advanced Studies in Measurement and Assessment Research Report, no. 20)

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

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, *84*, 1065–1073.

Wu, C. F. J. (1985a). Asymptotic inference from sequential design in nonlinear situation. *Biometrika*, *72*, 553-558.

Wu, C. F. J. (1985b). Efficient sequential designs for binary data. *Journal of the American Statistical Association*, *392*, 974-984.

Wynn, H. P. (1970). The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics*, *41*, 1655-1664.

Ying, Z., & Wu, C. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica*, *7*, 75-91.