SEMI-PARAMETRIC MODELS FOR RESPONSE TIMES AND RESPONSE ACCURACY IN COMPUTERIZED TESTING

BY

CHUN WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang
Professor Jeff Douglas
Professor Lawrence J. Hubert
Professor Carolyn Anderson
Professor Jinming Zhang

Abstract

In computer-administered tests, response times can be recorded conjointly with the corresponding responses. This broadens the scope of potential modeling approaches because response times can be analyzed in addition to analyzing the responses themselves. Current models for response times, however, mainly focus on parametric models that have the advantage of conciseness, but may suffer from a reduced flexibility to fit real data. This thesis presents two types of semi-parametric models that combine the flexibility of nonparametric modeling and the brevity as well as interpretability of the parametric modeling. They are

1. Hierarchical proportional hazard model: This model adopts the hierarchical structure suggested by van der Linden (2007) with the well-known Cox proportional hazard (PH) model in survival analysis. The PH model is comprised of two parts: the non-parametric baseline hazard and the parametric form of the examinee's latent speed. This model acts on the hazard rate, the instantaneous rate at which the event occurs conditioning on the fact that the event has not occurred so far, and it assumes that a unit increase in a latent speed is multiplicative with respect to the hazard rate. The model includes the exponential regression model, Weibull regression model, and many other parametric models as special cases.

2. Hierarchical linear transformation model: This model is a further extension of the Cox PH model. In this model, the response times, after some non-parametric monotone transformation, become a linear model with latent speed as a covariate plus an error term. The distribution of the error term implicitly defines the relationship between the RT and examinees' latent speeds; whereas the non-parametric transformation is able to describe various shapes of RT distributions. The linear transformation model represents a rich family of models that includes the Cox proportional hazard model, the Box-Cox normal model, and many other models as special cases. The linear transformation model is again embedded in a hierarchical framework so that both RTs and responses are modeled simultaneously.

For both new models, we propose two-stage estimation methods. The model checking techniques for both models are provided to help practitioners decide whether the model is appropriate for a real data set.

Finally, the applicability of the new models are demonstrated with simulation studies and applications to actual responses to items.

To my loving and supportive family.

Acknowledgments

This work was made possible through the support of many people. Special thanks to my adviser Professor Hua-Hua Chang for his guidance throughout my graduate study. I have benefited tremendously from his insight and knowledge. He has taught me everything from writing research papers, dealing with hard review comments, drafting proposals, to communicating with people friendly and professionally. He is not only a great mentor, but also one of my closest friends in life. There is no way that I could grow so much without his help and support. Also, thanks to Professor Jeff Douglas, who has provided countless hours of assistance and guidance throughout my dissertation work. His warm personality and kind encouragements have made the journey of my Ph.D. education a pleasure. I would also like to thank Professors Larry Hubert, Carolyn Anderson, and Jinming Zhang, whose lectures and lessons form the cornerstone of my statistical and ethical training as a quantitative psychologist.

Special thanks to my beloved parents and sister, thank you so much for raising me for the past twenty six years, giving me strongest support and walking me through various difficulties in my life. I can never come so close to my dream without you. I also want to express my sincere thanks to my husband, thank you for always taking care of me.

Thanks also goes to my friends, who have helped me a lot during my five-year study. These include Erkao Bao, Ying Guo, Anna Popova, Ping Chen, Haiyan Lin, Yi Zheng, Nathaniel Helwig, Justin Kern, Chris Zwilling, Ehsan Bokhari, Andrej Dietrich, Steve Broomell, Florian Lorenz, and many others. Thank you very much for taking courses with me and giving me so many helps with Latex, Matlab, and English writing.

Lastly, thanks to the National Science Foundation (NSF-MMS 0960822) and the Psychology Department (University of Illinois at Urbana-Champaign), both of which provided essential funding throughout this dissertation.

Table of Contents

Chapter 1

Introduction

Web-based assessment (i.e., on-line testing) is becoming a mainstream form of modern testing due to the internet's flexibility, accessibility, and potential capacities for faster data analysis and reporting. It also makes the collection of examinees' response times straightforward. The analysis of response times (RTs) on tests has recently attracted increase interest. A number of publications have demonstrated the utility of considering RTs on tests. In the realm of personality scales, RTs have been used to measure attitude strength (Bassili, 1996) or detect social desirability (Holden and Kroner, 1992). They have also been used as an additional predictor to enhance criterion validity (Siem, 1996). In the field of achievement tests, RTs have been used to evaluate the speededness of the test, to detect cheating behaviors, and to design a better test (e.g., van der Linden and Guo, 2008; van der Linden, 2009; Bridgeman and Cline, 2004). However, in order to use the full diagnostic potential of RTs, psychometric models are needed to analyze the relationship between the observed RTs and the test takers' latent traits. There are at least three advantages of developing such latent trait models: (1) the latent traits underlying the RTs can be used in addition to latent ability underlying the response accuracy as predictors of future performance, thereby enhancing criterion validity (Siem, 1996); (2) the estimation of ability can be improved by jointly modeling RTs and response accuracy; (3) such models can be used in cognitive psychology for more rigorous cognitive theory development (Rouder, Sun, Speckman, Lu, & Zhou, 2003; Klein Entink, Kuhn, & Fox, 2009b).

In the past couple of decades, researchers tried to formulate models that can maximally explain the variance of RTs as well as the connections among RTs, item characteristics and examinees' behaviors. Most of the models are motivated by the "curve-fitting principle" in the sense that the proposed models are parametric representations of the underlying RT distributions (e.g., Ronder et al., 2003; Schnipke and Scrams, 1997; Klein Entink, van der Linden, & Fox, 2009). The models differ in terms of the assumed response time distributions (e.g., lognormal, exponential, Weibull, and etc), the underlying relations between ability and response speed, and the nature of the items for which the model is designed (Schnipke and Scrams, 2002). Although the parametric models have the advantage of conciseness, they may suffer from a reduced flexibility to fit real data. In addition, with an empirical data set, one often needs to fit each parametric model

separately until a best fitting model is decided based upon some model diagnostic criterion (Schnipke and Scrams, 1997). Even though, the best fitting model may not be the best one for each individual item in the item bank. Ranger and Kuhn (2012) demonstrated that the response time distribution differed dramatically across items within one test. This calls for a flexible model that relaxes the such distributional assumptions.

The idea of proposing a "generalized model" that includes various parametric models as sub-models was first put forward by Ying and Chang (2005), and one example is the Box-Cox normal model (Klein Entink, van der Linden and Fox, 2009), where a power parameter was introduced to represent a number of different transformations. Most recently, Ranger and Kuhn (2012) proposed a generalized linear model with a flexible link function to model discrete response times. Specifically, their model includes a certain parameter (either at item level or test level) that determines the form of the link function, and their model unifies both proportional hazard models and accelerated lognormal failure time models. By fitting the generalized model to a data set, one can immediately pinpoint the most appropriate parametric form for each item from the estimation results.

This dissertation proposes another general modeling approach, namely, the semiparametric approach that reconciles the flexibility of nonparametric modeling and the brevity of the parametric modeling. Two semiparametric models are developed, one originates from the Cox proportional hazard model, and the other is built upon the linear transformation model. This latter model only assumes the existence of a monotone, but otherwise arbitrary transformation of the response times such that the linear model holds. As we will show, it includes the lognormal model, Box-Cox normal model, proportional hazard model and many other models.

Because the response time modeling has a long-term history, much wisdom has been accumulated. Also because this dissertation is motivated by the cutting-edge development in survival analysis, as a prelude for the next two chapters, brief introductions to both the current RT models and the survival analysis techniques are presented below.

1.1 Current Models for Response Time

Response time has been a preferred dependent variable in cognitive psychology since the mid-1950s (Luce, 1986). For relatively uncomplicated cognitive tasks such as Posner's perceptual matching task (Posner and Boies, 1972), response times naturally indicate the processing procedures required by an individual to complete a task. The main idea being that the more (cognitive) steps or processes required to complete a task, the longer the response or reaction time. In testing, Gulliksen (1950) first coined the distinction

between power tests and speed tests. In a pure speed test, the items are easy and the examinees are asked to answer as many items as possible within a limited time period. The goal is to measure how quickly the examinees answer those items. In this sense, the speed test is similar to the simple cognitive tasks. In the pure power test, on the other hand, the items differ in difficulty and there are no time limits. For these tests, examinees' response accuracies are of interest. In practice, although most of the tests (especially achievement tests) are power test, they also contain a speed component in that they are administered with a certain time limit.

Klein Entink et al. (2009a) summarized three different approaches that have been taken in the past to model RT. Here we briefly review each approach with representative examples. Under the first approach, only RT is modeled (Scheiblechner, 1979) such that it is mainly applicable to speed tests that have strict time limits. Within this category, Rouder et al.(2003) proposed a model based on Weibull distribution. In their model, the response time (also called *reaction time* in cognitive psychology) t_{nj} for person n on item j has the density

$$f(t_{nj}) = \frac{\pi_n (t_{nj} - \psi_n)^{\pi_n - 1}}{\sigma_n^{\pi_n}} \exp\left\{ - \left[\frac{t_{nj} - \psi_n}{\sigma_n} \right]^{\pi_n} \right\}, \; t_{nj} > \psi_n, \tag{1.1}$$

where ψ_n, σ_n and π_n are the shift, scale and shape parameters, respectively. Without incorporating any item level parameters, the model in (1.1) treats the RT for a given person as identically distributed across items, that is, characteristics of items do not impact RTs. This assumption is reasonable for the experimental paradigm (Rouder et al., 2003) where every stimuli in each trial requires almost the same cognitive process. The assumption might also hold when analyzing the "addition test" given to the second graders. Because in that test, the test takers can add single digit numbers (say, 100 of them) in 2 or 3 minutes, and every item has quite similar difficulty. When items differ in a test, Scheiblechner (1979) suggested exponential distribution of RT for person n responding to item j with density

$$f(t_{nj}) = (\tau_n + \gamma_j) \exp[-(\tau_n + \gamma_j) t_{nj}]. \tag{1.2}$$

In this model, τ_n is the person speed parameter, γ_j is the item speed parameter. Similar to the linear-logistic test model (LLTM; Fischer, 1973), γ_j can be further decomposed into fine-grained component process as

$$\gamma_j = \sum_{k=1}^{K} a_{jk} \eta_k, \tag{1.3}$$

where η_k indicates the time intensity of component process k, and a_{jk} is the weight with respect to component k within item j. Maris (1993) proposed using a more general gamma distribution but with similar

parameterizations as in (1.2). Although these well-established models are not explicitly characterized in the survival analysis framework, notice that Weibull, exponential and gamma distributions are all common parametric survival time distributions.

The second approach focuses on separate analysis of RTs and response accuracy. For instance, Gorin (2005) regressed log-transformed RTs on decomposed item difficulty parameter. Similar ideas are seen in Embreston (1998) and Primi (2001). Mulholland, Pellegrino, and Glaser (1980), on the other hand, used analysis of variance to predict RTs by item characteristics. Schnipke and Scrams (1997) proposed a lognormal model with a linear composition of its mean parameter into a general-level, person, and item component. That is, the logarithm of the time of n^{th} examinee answer to the j^{th} item is decomposed as

$$\log T_{nj} = \mu + \delta_j + \tau_n + \varepsilon_{nj}, \tag{1.4}$$

where μ is the grand mean response-time for the item bank and the examinee population, τ_n reflects the speed of examinee n, δ_j reflects the time intensity of item j, and $\varepsilon_{nj} \sim \mathcal{N}(0, \sigma^2)$. The same model was used to control differential speededness (van der Linden et al., 1999) and to detect the examinees' aberrant behaviors (van der Linden and van Krimpen-Stoop, 2003). In this approach, RTs and responses are modeled separately assuming these two variables vary independently. However, this assumption may not hold and thus a third approach was proposed.

The third approach advocates joint modeling of both RTs and responses, and such models include those proposed by Thissen (1983), van der Linden (1999), Roskam (1997), Wang and Hanson (2005), just to name a few. A major group of models in this category is motivated by the idea of a speed-accuracy relationship. Cognitive psychologist often focused on the within-person relationship, i.e., whether a person's response accuracy will decrease if he or she chooses to perform a task more quickly? This is termed as "speed-accuracy" tradeoff. The psychometricians, however, are more interested in the across-person relationship between speed and accuracy. For example, one question that psychometricians often explore is whether examinees with higher ability tend to answer the items faster. Both types of speed-accuracy relationships are considered within the model suggested by Verhelst, Verstralen, and Jansen (1997), or Thissen (1983). In their models, the speed-accuracy tradeoff is reflected by letting response accuracy dependent on the time devoted to the item—spending more time on an item increases the probability of a correct response. The speed-accuracy correlation across examinees is reflected by the separate parameters of examinees' ability (or mental power) and speed. Specifically, Verhelst et al.(1997) modeled the probability of a correct response

on item j by examinee n as

$$P_j(\theta_n, \tau_n) = [1 + \exp(\theta_n - \ln \tau_n - b_j)^{-\pi_j}], \tag{1.5}$$

where b_j is the difficulty parameter for item j, θ_n and τ_n are the ability and speed parameter for the nth person, and π_j is an item-dependent shape parameter. For $\pi = 1$, the model reduces to a Rasch type model with $\xi_n = \theta_n - \ln \tau_n$ replacing the traditional ability parameter. The speed-accuracy tradeoff is just reflected through ξ_n. That is, if a person decides to increase the speed τ_n, ξ_n will decrease and so does the correct response probability. Roskam (1997) proposed a similar model that is a Rasch model with an additive parameter structure incorporating logarithm of time as a regressor

$$P_j(\theta_n) = [1 + \exp(\theta_n + \ln t_{nj} - b_j)^{-1}]. \tag{1.6}$$

The model assumes a speed-accuracy tradeoff directly between the ability of the test taker and the actual time spent on a test item; less time on an item results in a higher speed and lower accuracy. Model (1.6) assumes that the actual RT is equivalent to examinees' speed. Though this assumption may be reasonable in the experiment paradigm, it may not hold in testing, in particular when each person takes a different set of items as in adaptive testing. Therefore, it is necessary to make a distinction between the RTs on the items and the speed at which the examinees' operate throughout the test. In this sense, a better way to measure speed is through distinct parameterizations of examinees' speed and items' time intensity. Such a modeling idea is reflected in Thissen (1983)'s model that takes the following form:

$$\ln T_{nj} = \mu + \tau_n + \beta_j - \rho(a_j\theta_n - b_j) + \epsilon_{nj}, \tag{1.7}$$

where $\epsilon_{nj} \sim N(0, \sigma^2)$. The normally distributed error term indicates that the model belongs to the lognormal family. Parameters τ_n and β_j can be interpreted as the speed of the examinee and the amount of time required by the item. The parameter μ is a general intercept parameter, a_j, b_j, and θ_n are the item discrimination, item difficulty and examinee ability parameters respectively. The term $\rho(a_j\theta_n - b_j)$ represents a regression of a two-parameter response model on the logarithm of time with ρ being the regression parameter. The speed accuracy tradeoff is indicated by the term $\rho(a_j\theta_n - b_j)$ when $\rho < 0$. When $\rho > 0$, the speed accuracy relation reverses. A similar idea was adopted by Ferrando and Lorenzo-Seva (2007) in modeling response time data from binary personality items, and the only change is the regression term. Instead of using $(a_j\theta_n - b_j)$, they used a distance measure $\delta_{ij} = \sqrt{a_j^2(\theta_i - b_j)^2}$ based on a distance-difficulty hypothesis in personality theory.

Van der Linden (2007) argued that although the speed-accuracy tradeoff is prevalent in reaction-time

research, on a test with a reasonable time limit, there is no need to incorporate a tradeoff in a RT model for a fixed person and a fixed set of test items (van der Linden, 2007). In other words, the tradeoff is a within-person constraint only, and it does not provide information to predict the speed or accuracy of one person from another. Therefore, the speed at which the test taker operates on the items should be assumed as a latent trait, and the response accuracy should only be determined by the examinees' abilities. In fact, as early as 1930, Kennedy (1930) found that individuals tended to perform at a consistent rate of work across a variety of cognitive tasks, even after partialing out the intelligence difference (Schnipke and Scrams, 2002). This conclusion is supported by Tate (1948), who investigated the speed accuracy relationship on number series, arithmetic reasoning, and spatial relations questions. He found that when accuracy was controlled, the fastest examinees were not the most accurate but fast subjects were consistently fast and slow subjects were consistently slow. These results illuminate that we need to model accuracy exclusively dependent on ability, and response time exclusively dependent on examinees' latent speeds. But on the second level of the model, the speed and ability can be correlated. The correlation may differ depending upon the test context and content (Schnipke and Scrams, 2002).

Following this argument, van der Linden (2007) proposed a hierarchical framework, in which RT and responses are modeled separately at the measurement model level; and at a higher level, a population model for the person parameters (speed and ability) is constructed to account for the correlation between them. This model distinguishes the speed-accuracy tradeoff within a person from the speed-accuracy correlation in the population. The formulation of the model is as follows. At the first level, two models for the responses and RTs are specified separately. Responses are assumed to follow a three-parameter logistic (3PL) model:

$$P_j(\theta_n) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_n - b_j)]}{1 + \exp[a_j(\theta_n - b_j)]}, \tag{1.8}$$

where a_j, b_j, and c_j represent item discrimination, difficulty and guessing parameters. For the RTs, a lognormal model with separate person and item parameters was adopted (van der Linden, 2006),

$$T_{nj} \sim f(t_{nj}; \tau_n, \alpha_j, \beta_j) \equiv \frac{\alpha_j}{t_{nj}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j(\ln t_{nj} - (\beta_j - \tau_n))\right]^2\right\}, \tag{1.9}$$

where τ_n, β_j and α_j are the speed parameter for examinee n, the time intensity and discriminating power of item j, respectively. At the second level, $\boldsymbol{\xi}_n = (\theta_n, \tau_n)$ is assumed to be randomly drawn from a bivariate normal distribution, with mean vector $\boldsymbol{\mu}_p = (\mu_\theta, \mu_\tau)$, and covariance matrix $\boldsymbol{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}$. Analogously, the item parameter vector $\boldsymbol{\psi}_j = (a_j, b_j, c_j, \alpha_j, \beta_j)$ is also assumed to follow a multivariate normal

distribution with mean vector $\boldsymbol{\mu}_J = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta)'$, and covariance matrix

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

Several recent attempts have been made to extend the above hierarchical model for more complicated applications. For example, instead of only considering log transformation of RT, Klein Entink, van der Linden and Fox (2009) considered a broader class of Box-Cox transformations (Box and Cox, 1964). This generalization leads to a normal model for the transformed RTs,

$$T^{(v)} = \begin{cases} \frac{t_{nj}^v - 1}{v} \sim N(\beta_j - \tau_n, \alpha_j^{-2}) & v \neq 0 \\ \log t_{nj} \sim N(\beta_j - \tau_n, \alpha_j^{-2}) & v = 0 \end{cases}$$

Here T denotes the original time and $T^{(v)}$ denotes the Box-Cox transformed time. It is apparent that the lognormal distribution belongs to the Box-Cox transformation. Further, Klein Entink, Fox and van der Linden (2009) proposed a multivariate multilevel model for mixed response variables (binary responses and continuous RTs). Their model allows for the incorporation of explanatory variables to identify factors that explain variation in speed and accuracy between individuals who may be nested within groups. Another attempt was made by Klein Entink, Kuhn, Hornke and Fox (2009). They proposed a joint modeling approach by use of responses and RTs to evaluate cognitive theory. Their model takes a similar structure as van der Linden's (2007) hierarchical model, and the innovation is to decompose each item parameter based on the detailed cognitive process as required by the item in order to support the cognitive theory that underlies the item design (Klein Entink et al., 2009b).

1.2 Mixed (Multilevel) and Conditional Regression Perspective

The construction and evolution of response time and response accuracy modeling can also be summarized from mixed and conditional regression perspectives. When analyzing the experimental data from cognitive psychology, there is much literature on the speed-accuracy tradeoff (Kahane and Loftus, 1999) and on mathematical processing models for response speed and accuracy (Luce, 1986; Ratcliff, 1988). These models, however, do not distinguish between person and item parameters, and they are only applicable to within-

subject analysis of a series of replications of the same stimuli in a psychophysical discrimination or detection task (van Breukelen, 2005). On the contrary, in testing or in cognitive tests that typically include tasks like analogical reasoning or series completion, the items vary by difficulty and there is only one replication per person per item. A better approach for this application is, therefore, an extension of item response theory models into the simultaneous modeling of both speed and accuracy as functions of the person and item parameters, such as the models that will be proposed in this dissertation.

The traditional IRT modeling and fixed effects logistic regression can be combined into what is known as mixed or conditional logistic regression. For instance, let the log-odds of a correct response by person i on item j as

$$\log\left[\frac{p_{ij}}{(1-p_{ij})}\right] = \beta_{0i} + \beta_{1i}X_{1ij} + ... + \beta_{pi}X_{pij}, \tag{1.10}$$

where p_{ij} is the correct response probability by person i on item j, and X_1 to X_p represents observed covariates that could either be between-subject (person level) variables such as age or gender, or within-subject (item level) variables like the number of cognitive steps needed to solve an item, or interaction of both. If the covariate is the response time spent on the item, this model inherently models the speed-accuracy tradeoff. The parameter β_{0i} is person dependent intercept, β_{1i} through β_{pi} are person dependent regression weights. The well-known Rasch model can be viewed as one special case of model (1.10) (van Breukelen, 2005).

Similarly, response time could also be modeled via linear mixed modeling approach (Verbeke and Molenberghs, 2000). Because response times are frequently assumed as following a lognormal distribution within persons and items, the mixed model could be

$$\log(t_{ij}) = \gamma_{0i} + \gamma_{1i}X_{1ij} + ... + \gamma_{pi}X_{pij} + e_{ij}, \tag{1.11}$$

where e_{ij} is normally distributed error term. As in (1.10), γ_{0i} through γ_{pi} are person level intercept and slopes. The above two mixed regression models treated persons as random but items as fixed. However, as both subjects (persons) and items can be regarded as random samples from a population of people and a population of items, one can define random residuals for both subjects and items. When subjects and items are in a non-hierarchical relationship, such a model is referred to as a crossed random effect model (Raudenbush, 1993). For instance, Baayen, Davidson, and Bates (2008) proposed a mixed effects modeling approach with crossed random effects for subjects and items, their model can be expressed as

$$T_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2j} + \tau_i + \alpha_{1j} + \varepsilon_{ij}, \tag{1.12}$$

where T_{ij} represents response time for subject i on item j, $\tau_i \sim \mathcal{N}(0, \sigma_\tau^2)$ and $\alpha_{1j} \sim \mathcal{N}(0, \sigma_{\alpha_1}^2)$ denote the random intercepts for subject (i.e., examinees' latent speed) and item, respectively, where as X_{1i} and X_{2j} denote the fixed effects. In this model, there is no by-subject or by-item random slopes for simplicity. ε_{ij} is the by-observation error term. Here the response time itself is modeled directly, but sometimes, certain transformation of response time, say, log transformation, could be imposed first. Following the similar argument, Jaeger (2008) proposed a similar model for response accuracy as

$$\ln\left(\frac{p_{ij}}{(1-p_{ij})}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2j} + \theta_i + \alpha_{2j} + \varepsilon_{ij}, \tag{1.13}$$

with $\theta_i \sim \mathcal{N}(0, \sigma_\theta^2)$ and $\alpha_{2j} \sim \mathcal{N}(0, \sigma_{\alpha_2}^2)$. Large values of θ correspond to higher participant error rate (i.e., lower ability) and large values of α_2 indicates the items have higher difficulty.

The estimation of response time and response accuracy models are separately discussed in Baayen et al. (2008) and Jaeger (2008). Loeys, Rosseel and Baten (2011) recently proposed a joint model by imposing a joint multivariate distribution on the vector of all random effects for subject and item, as follows

$$\boldsymbol{\Sigma}_s = \begin{pmatrix} \sigma_\tau^2 & \rho_{\theta\tau}\sigma_\theta\sigma_\tau \\ \rho_{\theta\tau}\sigma_\theta\sigma_\tau & \sigma_\theta^2 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_s = \begin{pmatrix} \sigma_{\alpha_1}^2 & \rho_{\alpha_1\alpha_2}\sigma_{\alpha_1}\sigma_{\alpha_2} \\ \rho_{\alpha_1\alpha_2}\sigma_{\alpha_1}\sigma_{\alpha_2} & \sigma_{\alpha_2}^2 \end{pmatrix}.$$

The parameter $\rho_{\theta\tau}$ measures the correlation between speed and ability at subject level, whereas $\rho_{\alpha_1\alpha_2}$ measures the correlation between time intensity and difficulty at item level. This modeling approach resembles van der Linden (2007)'s hierarchical framework.

1.3 Survival Analysis and Educational Measurement

Survival analysis is a branch of statistics that concerns the analysis of time-to-event data. Some of the questions survival analysis try to tackle are: what is the proportion of a population that will survive beyond a particular time; among the survivors, at what hazard rate will they die or fail; what factors and how will the factors affect the survival probability or hazard rate of a population. The primary objective of interest is the survival function, specifying the probability that the occurrence (death) of an event is later than some particular time. Survival function is often defined as $S(t) = P(T > t)$, where t is some time,

and T is a random variable denoting the time of death. The survival function must be non-increasing, i.e., $S(t_1) \leq S(t_2)$ if $t_1 \leq t_2$. This reflects the idea that survival to some later time requires survival at all earlier times as well.

Survival analysis has benefited medical scientists in the study of mortality due to chronic diseases, and has helped industrial statisticians to model the longevity of machinery and parts in manufacturing processes. In principle, survival analysis techniques could be used in any science in which outcomes are measured as the time until an awaited event (Douglas, Kosorok and Chewning, 1999). Psychology is also a specific area that survival analysis can shed light on. For instance, Douglas et al. (1999) proposed a discrete version of proportional hazard frailty model to explain the substance abuse of youths. In particular, they modeled the ages at which youths first try alcohol, cigarettes, marijuana, and inhalants, as a function of their latent psychological abilities to abstain from substance abuse. Another example is by Singer and Willett (1993) who used discrete-time survival analysis to study whether and, if so, when the public school teachers stopped teaching between their first year of teaching and the year when the data collection ended. The discrete-time hazard model proposed in their paper not only answers these descriptive questions but also models the relationship between event occurrence and predictors as well. The specific area in educational measurement that survival analysis come into play is response time analysis. Response time (RT) is the time period from the onset of an item until examinee provides an answer to the item. If viewing "giving a response" as an event, RT shares the same meaning as the survival time in biostatistics, and therefore RT can be modeled directly through the survival function $S(t)$.

The most common way to estimate $S(t)$ is through the now ubiquitous Kaplan-Meier estimator (named after Edward L. Kaplan and Paul Meier), also known as the product-limit estimator. It is a non-parametric estimator derived from counting process. The specific construction of Kaplan-Meier estimator is as follows. For the j^{th} item, let the observed response time for the N examinees answering this item be $t_1 \leq t_2 \leq \cdots \leq t_N$. Corresponding to each t_i is the number of examinees n_i whose RTs are longer than t_i ("at risk" set) and the number of examinees d_i who give response to item j at t_i. The Kaplan-Meier maximum likelihood estimator $\hat{S}(t)$ is a product

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}). \tag{1.14}$$

$\hat{S}(t)$ is a non-decreasing step-function, with steps at $t_i, 1 \leq i \leq N$. The original Kaplan and Meier paper that appeared in 1958 is one of the most heavily cited papers in all of the sciences (Hubert & Wainer, unpublished). The importance of Kaplan-Meier curve is apparent in medical/pharmaceutical areas. In educational measurement, Kaplan-Meier curve also provides an overall description of the response time

distribution of an item. For example, Kaplan-Meier curve indicates on average what the probability equals that an examinee answers item j within a time period t_i. Oftentimes, harder items require longer time to answer, and therefore, the Kaplan-Meier curves of the difficult items are often flatter than those of easy items. Another advantage of the Kaplan-Meier estimator is that it has a closed form variance estimator (e.g., the Greenwood formula), and therefore the confidence band around the Kaplan-Meier curve is easily specified.

1.3.1 Regression Models

The Kaplan-Meier estimator only provides a marginal view of the RT distribution for an item, and it does not consider the item-person interaction, that is, the same item may have a different RT distribution for different examinees. To incorporate examinees' parameters (that can be viewed as covariates) into the analysis of RT, we need to resort to regression models. In survival analysis, the covariates can enter into the model in a similar way as in common regression models. In the following, we will separately review two groups of regression models, parametric models and semi-parametric models.

The parametric models assume a fully parametric form of the survival function, or equivalently, the hazard function of response time. The hazard function (usually denoted as $h(t)$) is the instantaneous rate at which events occur. It is defined as

$$h(t) = \lim_{\delta t \to 0} \frac{P[t \leq T < t + \delta t | T \geq t]}{\delta t}.$$

In psychological terms, the hazard rate can be viewed as the processing capacity of an individual, the individual's relative ability to perform mental work in a unit of time. Individuals with a high hazard rate (high conditional probability of finishing the task in the next moment) have a high processing capacity and work more intensely (Wenger & Gibson, 2004). The hazard rate relates to the survival function through $S(t) = \exp[-H(t)]$, where $H(t) = \int_{s=0}^{t} h(s)ds$ is the cumulative hazard function. Taking *exponential model* as an example, the hazard function at time t for an item with covariate $\boldsymbol{Z}$ can be written as

$$h(t|\boldsymbol{Z} = \boldsymbol{z}) = \lambda c(\boldsymbol{z}).$$

In this model, the hazard rate for a given $\boldsymbol{Z}$ is a constant characterizing the exponential distribution. The covariate can be any observed or latent variables, such as the examinees' demographic information, their latent ability, and the like. The function $c(\cdot)$ may be parameterized in a number of ways, oftentimes, the

effects of the covariates are reflected through a linear function, $\boldsymbol{Z}'\boldsymbol{\beta}$, and therefore

$$h(t|\boldsymbol{Z}=\boldsymbol{z}) = \lambda c(\boldsymbol{z}'\boldsymbol{\beta}).$$

Here $\beta' = (\beta_1, ..., \beta_p)$ are regression parameters and c is a specified functional form. The choice of c depends on the particular data being considered, and three common forms have been used in the past (e.g., Feigl & Zelen, 1965): (1) $c(\boldsymbol{Z}) = 1 + \boldsymbol{Z}$, (2) $c(\boldsymbol{Z}) = (1 + \boldsymbol{Z})^{-1}$, and (3) $c(\boldsymbol{Z}) = \exp(\boldsymbol{Z})$. The first two forms correspond to (1) the hazard rate and (2) mean survival time, being linear functions of $\boldsymbol{Z}$. The last form that is also the most widely used form, assumes that a unit increase in a covariate is multiplicative with respect to the hazard rate.

Now consider the model with hazard function

$$h(t|\boldsymbol{Z}=\boldsymbol{z}) = \lambda \exp(\boldsymbol{z}'\boldsymbol{\beta}). \tag{1.15}$$

Taking log transformation on both sides yield a model specifying that the log hazard rate is a linear function of the covariate $\boldsymbol{Z}$. In terms of the log survival time, $Y = \log T$, the above model can be reparameterized as

$$Y = \alpha - \boldsymbol{Z}'\boldsymbol{\beta} + W, \tag{1.16}$$

where $\alpha = -\log\lambda$ and W follows the extreme value distribution. The model in (1.16) can be viewed as a log-linear model, and it is a linear model for Y with the error term W having an extreme value distribution.

Another well-known parametric regression model is based on Weibull distribution, with the covariate entered into the model in essentially the same way. Specifically, the conditional hazard is expressed as

$$h(t|\boldsymbol{Z}=\boldsymbol{z}) = \gamma\lambda(\lambda t)^{\gamma-1}\exp(\boldsymbol{z}'\boldsymbol{\beta}). \tag{1.17}$$

Due to the exponential link function (i.e., $c(\boldsymbol{Z}) = \exp(\boldsymbol{Z})$), the effect of the covariates is again acting multiplicatively on the Weibull hazard. By the same token, let $Y = \log T$, the model (1.17) can be expressed in the linear form as

$$Y = \alpha + \boldsymbol{Z}'\boldsymbol{\beta}^* + \sigma W, \tag{1.18}$$

where $\alpha = -\log\lambda$, $\sigma = \gamma^{-1}$ and $\beta^* = -\sigma\beta$. The error term W again follows standard extreme value distribution.

The exponential and Weibull regression models suggest two different generalizations. On one hand, the

covariates in both models (1.15) and (1.17) act multiplicatively on the hazard function. This generalization, as will be shown below, suggests a typical semi-parametric model called the *relative risk model* or *Cox model*. On the other hand, both of these models can be expressed in a log-linear form; that is, the covariates act additively on the log transformed time Y, and this general class of log-linear models is called the *accelerated failure time model*.

In the above models, the covariates are assumed to be observed. Biostatisticians first recognized the usefulness of latent variables to model survival times that are correlated due to either repeated measurements taken on a single subject, or measurements of a common variable taken on genetically associated subjects. These needs gave rise to *frailty models*, in which a latent *frailty* random variable is included in the model to account for possible correlations in failure time distributions (Clayton, 1991; Clayton and Cuzick, 1985). The frailty variables may be viewed as random effects and usually only the influence of the explanatory covariates on failure time is the primary concern. Douglas et al.(1999) used the frailty model in psychology, and their model is a similar version of the conditional proportional hazard model (Clayton and Cuzick, 1985), in which the hazard function for each failure time is a product of the baseline hazard, frailty random variable, and covariate effects. The unique feature of their model is that it has an item-level parameter that measures the influence of the latent variable on the failure time. Thus separate items differ with respect to the extent that the latent variable influences responses. In fact, the Cox PH model represents a standard approach in survival time analysis, it makes only very mild distributional assumptions and, is a flexible semiparametric model. However, it is only very recently that the Cox model has been introduced in the filed of measurement to analyze response times (Ranger and Ortner, 2011). In Chapter 2 of this dissertation, we propose to use Cox model for RT analysis, and that chapter is complementary to Ranger and Ortner's earlier research.

A third widely used parametric regression model comes from log-normal distribution of time (van Breukelen, 2005; van der Linden, 2007). The lognormal regression model takes the following form in general

$$\log(t|\boldsymbol{Z} = \boldsymbol{z}) = \boldsymbol{z}'\boldsymbol{\beta} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then the hazard function can be written as

$$h(t) = \frac{\frac{1}{t\sigma}\phi\left(\frac{\log t - \boldsymbol{z}'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{-\log t + \boldsymbol{z}'\boldsymbol{\beta}}{\sigma}\right)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative density function of the standard normal distribution, respectively. Apparently, the covariate is not multiplicatively related to the hazard

function, and the effect of the covariates cannot be written in an exponential link function as in exponential or Weibull regression models. Because of these reasons, the lognormal regression model is not a special case of the proportional hazard model. However, it does belong to the *linear transformation model* family, and this drives the development of the second new model in the dissertation. In Chapter 3, we will review the linear transformation model, and show how it subsumes both Cox model and lognormal model as special cases.

A semi-parametric extension of the above parametric models with exponential link function leads to the *Cox proportional hazard model*. The Cox model combines the parametric regression term with a non-parametrically defined baseline hazard function. Therefore, the hazard function of the Cox model is expressed as (Cox, 1972)

$$h(t|\theta) = h_0(t)\exp(\boldsymbol{z}'\boldsymbol{\beta}).$$

Here the non-parametric baseline hazard $h_0(t)$ reflects the flexibility of the model to accommodate a variety of different shapes of RT distributions, whereas the regression term succinctly summarizes how RT changes with the covariates. When the baseline hazard is a constant, this model becomes the exponential regression model; when the baseline hazard takes the form of $h_0(t) = \gamma(\lambda t)^{\gamma-1}$, the model becomes Weibull regression model. In light of this, the Cox model is flexible enough to represent different RT distributions, and thus it serves as a good candidate for modeling RTs. The new model proposed in Chapter 2 derives from the Cox model with a frailty term, and the frailty represents the examinee's latent speed.

1.3.2 Model Diagnostic Techniques

Model fit and adequacy checking play an important role in survival analysis. In the classical survival analysis framework, an assessment of the hazard function is usually done through so-called *hazard plot*. Specifically, a parametric specification for the hazard, $h(t)$, can be checked using an empirical estimate of $h(t)$, say, $\hat{h}(t)$. Plots of $\hat{h}(t)$ versus t(or $\log(t)$) are compared with plots assuming the parametric model. Another approach is based on residual analysis, in which a theoretical or empirical Q-Q plot (see, e.g., Cox and Oakes, 1984; Therneau et al., 1990) is developed for certain exponential or martingale based residuals. In this section, we will introduce the second approach in detail for the Cox model under both frequentist and Bayesian paradigms.

In frequentist paradigm, the Cox-Snell (1968) residual can be used to assess the overall fit of model. For item j, denote the observations as $(\boldsymbol{T}_i, \delta_i, \boldsymbol{Z}_i)$, $i = 1, 2, ..., n$. For simplicity, we assume $\boldsymbol{Z}_i$s are time independent; and also due to the fact that all RTs in tests are observed, we drop the censoring indicator δ_i in the following. Suppose the proportional hazard model $h(t|Z_i) = h_0(t)\exp(\boldsymbol{Z}_i'\boldsymbol{\beta}_j)$ has been fit to the model. If

the model is correct, then the true cumulative hazard function conditional on Z, $H(T|\boldsymbol{Z})$, has an exponential distribution with hazard rate equal to 1. This conclusion holds because $H(T|\boldsymbol{Z}) = -\ln[1 - F(T|\boldsymbol{Z})]$ where $F(T|\boldsymbol{Z})$ is the cumulative distribution function and it follows uniform distribution. So the cumulative distribution function of $H(T|\boldsymbol{Z})$ is $1-\exp(-H)$ that is exactly the c.d.f of the unit exponential distribution.

Now, let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood (it is actually the partial likelihood that will be introduced in Chapter 2) estimate of $\boldsymbol{\beta}$, and let $\hat{H}_0$ denotes the estimator of the baseline hazard rate. Then the Cox-Snell residual for the ith examinee and jth item is defined as

$$r_{ij} = \hat{H}_0(t_i)\exp(\boldsymbol{Z}_i'\hat{\boldsymbol{\beta}}_j). \tag{1.19}$$

It is easy to notice that r_{ij} will be approximately exponentially distributed with hazard rate 1, given that the proportional hazards model is correct and $\hat{H}_0$ is close to H_0 and $\hat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}$. To check whether the r_{ij}s behaves as a sample from a unit exponential, the Nelson-Aalen estimator of the cumulative hazard rate of the $r_{ij}, i = 1, ..., n$ can be computed for each item separately. If r_{ij}s are from a unit exponential distribution, then this estimator should be approximately equal to the cumulative hazard rate of the unit exponential $H_E(t) = t$. Thus, a plot of the estimated cumulative hazard rate of the r_i, $\hat{H}_r(r_{ij})$, versus r_{ij} should be a straight line through the origin with a slope of 1.

Here we only review the Cox-Snell residual for the simplest case where the covariates do not depend on time. This assumption is reasonable in response time research where the covariates often include examinees' abilities, speed, or other demographic variables such as age, social economic status, *etc.* Assuming fixed ability and fixed speed for a person during the test, stationarity assumptions leads to standard item response modeling (van der Linden 2007). However, if time-dependent effects such as fatigue or practice are modeled, more complex models need to be constructed and the Cox-Snell residual can be modified accordingly.

For the parametric regression model, the Cox-Snell residual is redefined to incorporate the specific parametric form of the baseline hazard rate. For example, the Cox-Snell residuals for the exponential and Weibull regression models are $r_i = \hat{\lambda} t_i \exp(\boldsymbol{Z}_i'\hat{\boldsymbol{\beta}}_j)$ and $r_i = \hat{\lambda}\exp(\boldsymbol{Z}_i'\hat{\boldsymbol{\beta}}_j)t_i^{\hat{\alpha}}$, respectively. In fact, examination of model fit with the Cox-Snell residual is equivalent to that done using the standardized residual based on the log-linear model representation. To be specific, we define the standardized residual by analogy with those used in normal regression theory as

$$r_i = \frac{\log T_i - \hat{\alpha} - \hat{\boldsymbol{Z}}_i'\boldsymbol{\beta}_j}{\hat{\sigma}}. \tag{1.20}$$

If a Weibull model holds, then the r_i's should be a sample from standard extreme value distribution; if the

log normal distribution holds, these residuals should follow a standard normal distribution.

Chapter 2

The Hierarchical Proportional Hazard Model

This chapter introduces a new hierarchical proportional hazard model to model RTs and response accuracy simultaneously. A critical feature of the model is that examinees' abilities are distinguished from their latent speed and separate latent traits are assigned to both of them. This leads to the key assumption of the current model: a test taker operates at a fixed level of speed during the course of the tests. This stationarity assumption excludes changes in behavior during the test due to fatigue, learning, strategy shifts and other factors. The hierarchical framework proposed by van der Linden (2007) is adopted here. Measurement models at the first level separate the variability in the observed responses and RTs into item and person effects. At a higher level, we assume the examinee's ability θ and latent speed τ are from a bivariate normal distribution. The specific formulation of the model is as follows.

First-Level Model. At the first level, two models for the responses and RTs are specified separately. For the item response model, any appropriate parametric model may be used, but we focus on the three-parameter logistic model:

$$P_j(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \tag{2.1}$$

with a_j, b_j, and c_j representing item discrimination, difficulty and guessing parameters. For the response times, the Cox PH model is chosen and the hazard function for RTs is

$$h_{ij}(t|\tau_i) = h_{0j}(t)\exp(\beta_j\tau_i) \tag{2.2}$$

where the survival function is

$$p(t_{ij} \geq t|\tau_i) = S_{ij}(t) = \exp\left[-\int_0^t h_{0j}(s)\exp(\beta_j\tau_i)ds\right], \tag{2.3}$$

where $\tau_i \in \mathcal{R}$ is the speed parameter for test taker i. The subscript j in h_{0j} implies that different shapes of the RT distributions are possible for different items, and β_j is the regression parameter (i.e., slope). When β_j is positive, the higher the τ_i, the shorter the RT will tend to be. When β_j is negative, it means examinees

with higher answering speed tend to take a longer time to finish the item. This measurement model resembles the one proposed by Ranger and Ortner (2011). The β_j coefficient also determines the influence of the latent speed on the hazard rate. In a psychological sense, it controls the increase in processing capacity that is due to a unit increase in the latent speed. Notice that in traditional survival analysis, the regression parameter is interpreted in a relative sense. But in educational measurement, it is important to be able to make inference about examinees and items, such as how much time is required for each examinee on a particular item on average. Therefore, both the regression parameter and baseline hazard have to be estimated accurately. This point is re-emphasized in the model estimation section below. As every constant multiplier can be absorbed in the baseline hazard rate, the linear predictor $\beta_j\tau_i$ does not include an intercept term. The item time intensity is reflected in the baseline hazard, and more clearly via equation (2.3). In general, items with lower cumulative hazard $H_0(t)$ tend to be more time consuming.

Second-Level Model. This part of the model captures the joint distribution of the person parameters in a population. The values of $\boldsymbol{\xi}_i = (\theta_i, \tau_i)'$ are assumed to be randomly drawn from a bivariate normal distribution, i.e.,

$$\boldsymbol{\xi}_i \sim f(\boldsymbol{\xi}_i; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \equiv \frac{|\boldsymbol{\Sigma}_p^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_i - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\mu}_p)\right], \tag{2.4}$$

with mean vector

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

and covariance matrix

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

Identifiability. To establish identifiability, we suggest the constraints $\mu_\theta = 0, \sigma_\theta^2 = 1, \mu_\tau = 0, \sigma_\tau^2 = 1$. Here, the first two constraints are standard in IRT parameter estimation, when item parameters are unknown. The last two constraints fix the scale of τ to remove the tradeoff between β_j and τ_i and they also fix the scale of h_0.

Model Assumption. Following van der Linden (2007), this model has three independence assumptions. They are

- Independence between responses given θ, that is

$$f(y_{i1}, \cdots, y_{iJ}|\theta_i) = \prod_{j=1}^{J} f(y_{ij}|\theta_i) \tag{2.5}$$

- Independence between response times given τ . This assumption is defined as

$$f(t_{i1}, \cdots, t_{iJ}|\tau_i) = \prod_{j=1}^{J} f(t_{ij}|\tau_i), \tag{2.6}$$

- Independence between responses and response times given θ and τ. That is

$$f(u_{i1}, \cdots, y_{iJ}; t_{i1}, \cdots, t_{iJ}|\theta_i, \tau_i) = \prod_{j=1}^{J} f(y_{ij}|\theta_i) f(t_{ij}|\tau_i) \tag{2.7}$$

In van der Linden (2007)'s model, he imposed a covariance structure on item parameters, whereas we assume the item parameters are independent of one another. There are three reasons. First, according to the results in van der Linden (2007), only the correlation between item time intensity and item difficulty is non-zero (with posterior mean 0.3), all the rest correlations are either very close to 0 or have posterior confidence interval covering 0. Second, the item time intensity information in the new model is reflected in the non-parametric baseline hazard $h_{0j}(t)$, whose correlation with the item difficulty b_j is not easily modeled. Only when the parametric form of $h_{0j}(t)$ is known, for instance, in exponential model, $h_{0j}(t) = \lambda_j$, that one can model the correlation between λ and b. In this case, because $S_{ij}(t) = \exp[-\exp(\beta_j \tau_i)(\lambda_j t)]$, λ and b should be negatively correlated, indicating that more difficult times are more likely to be time consuming. Third, as shown in our simulation study below, even when the correlation between item time intensity and item difficulty is ignored, the estimation accuracy will not be significantly affected.

2.1 Model Estimation

The goal of our investigation is to accurately estimate θ and τ, as well as the regression parameter β in response time model of (2.3) and item parameters in (2.1). In many Cox frailty model applications, only the regression parameter β and frailty τ need to be estimated, but in our case, in order to make inference about the examinees and items, the non-parametric cumulative baseline hazard H_0 also needs to be estimated. Several approaches were proposed in the past to estimate both parametric and non-parametric parts of the Cox model, such as estimation based on a spline approximation of the baseline hazard rate (Cai, Hyndman, & Wand, 2002) or estimation based on piecewise exponential models (Friedman, 1982). Two approaches that have advantages (Ranger and Ortner 2011) are (1) estimation by treating response time as discrete variable (McCullagh, 1980), such that the Cox model can be viewed within the generalized linear model framework and standard software can be used for model estimation; (2) estimation based on partial likelihood;

this approach does require categorization of the response times and thus it is more efficient (Ranger and Ortner 2011). Ranger and Ortner (2011) explored both methods, and they employed a divide-and-conquer approach by estimating the parametric part first and non-parametric part secondly. We proposed a two-stage estimation method that shares the same principle, but instead of using partial likelihood within the marginalized maximum likelihood framework, we used it within the Markov chain Monte Carlo (MCMC) framework.

Marginal likelihood inference involving latent variables is usually challenging because of the integrals that are sometimes numerically intractable. One approach that avoids such difficulties is to use the MCMC method to obtain draws from a distribution that has a density proportional to the joint posterior distribution of the item and person parameters. Another motivation for using the MCMC method is that in computerized adaptive testing (CAT), every test taker is given different items, based on his or her adaptively estimated θ level. So the random sampling of θ (or τ due to the possible correlation between them) from a common distribution can not be assumed. We wish for our estimation technique to allow for data obtained by CAT. Consequently, the usual marginal likelihood approaches used in latent variable modeling are no longer appropriate.

Estimating Cox's PH frailty model with MCMC is not entirely new. Clayton (1991) used Gibbs sampling to fit frailty models to clustered failure data. He sampled iteratively from the full conditional distribution of H_{0j} and all parameters with H_{0j} as an independent increment gamma process (Kalbfleisch, 1978). Gray (1994) used a piecewise constant baseline hazard and also included it as a parameter to be updated in the MCMC scheme. Similarly, Douglas et al.(1999) modeled the discrete failure time, and treated the baseline hazard as a constant at each time point, which again was incorporated in the MCMC algorithm. Most recently, Henschel, Engel, Holzel, and Mansmann (2009) treated the baseline hazard with a stepwise constant function as well as a cubic spline. Sharef, Strawderman, Ruppert, Cowen, and Halasyamani (2010) argued that treating baseline hazard as piecewise constant is somewhat too restrictive because it depends on some discretization of time. Instead, they proposed to model the baseline hazard as a penalized mixture of B-splines. Their approach is even more general in that they allowed the frailty distribution to be unspecified, and modeled it as a penalized mixture of normalized B-splines. As a result, their model estimation method continues to apply to the proportional hazard frailty model, while permits shrinkage towards a specific parametric hazard function or frailty distribution.

Although the Sharef et al. (2010)'s method is flexible and promising, it does not lend itself directly to our case because of two reasons: (1) in our model, we assign an item level regression coefficient β_j in front of the frailty term τ_i whereas in their model, the effect of the frailty is the same across different items; (2) in

our model, we impose a covariance structure on the frailty term, which introduces extra difficulty in model estimation. Due to these reasons, we propose a two-stage estimation method. In the first stage, we avoid the difficulty of modeling and sampling from H_{0j} by using the Cox partial likelihood (Cox, 1975), and in the second stage, we estimate the infinite-dimensional parameter H_{0j} through either non-parametric estimator or B-splines. The use of partial likelihood in the Bayesian context for the frailty model estimation has been demonstrated in Gustafson (1997) and Sargent (1998). The justification for using partial likelihood will be briefly described in section 2.1.2.

2.1.1 Partial Likelihood

For the jth item, suppose that there are no ties between the response times. Let $t_{(1j)} < t_{(2j)} < \cdots < t_{(Nj)}$ denote the ordered RTs and τ_i be the latent trait associated with the individual whose response time is $t_{(ij)}$. Define the risk set $R(t_{(pj)})$ at time $t_{(pj)}$, $1 \leq p \leq N$, as the set of all individuals who have not answered the question yet, i.e., $R(t_{(pj)}) = \{t_{((p+1)j)}, \cdots, t_{(Nj)}\}$. The partial likelihood function for the jth item given τ is specified as:

$$\begin{aligned} L(\beta_j|\boldsymbol{\tau}) &= \prod_{i=1}^{N} \frac{\exp[\beta_j \tau_i]}{\sum_{t_{pj} \in R(t_{pj})} \exp[\beta_j \tau_p]} \\ &= \prod_{i=1}^{N} \frac{\exp[\beta_j \tau_i]}{\sum_{p \geq i}^{N} \exp[\beta_j \tau_p]} \end{aligned} \tag{2.8}$$

The partial likelihood for the vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)'$ is then defined as

$$L(\beta|\boldsymbol{\tau}) = \prod_{j=1}^{J} L(\beta_j|\boldsymbol{\tau}). \tag{2.9}$$

The log of the partial likelihood is $LL(\beta_j|\boldsymbol{\tau}) = \ln[L(\beta_j|\boldsymbol{\tau})]$, and we write $LL(\beta_j|\boldsymbol{\tau})$ as

$$LL(\beta_j|\boldsymbol{\tau}) = \sum_{i=1}^{N} \beta_j \tau_{(i)} - \sum_{i=1}^{N} \ln[\sum_{p \geq i}^{N} \exp(\beta_j \tau_p)]. \tag{2.10}$$

Taking derivatives with respect to β we find the score, $U(\beta_j|\boldsymbol{\tau}) = \partial LL(\beta_j|\boldsymbol{\tau})/\partial \beta_j$, equals to

$$U(\beta_j|\boldsymbol{\tau}) = \sum_{i=1}^{N} \left[\tau_{(i)} - \frac{\sum_{p \geq i}^{N} \tau_{(p)} \exp[\beta_j \tau_p]}{\sum_{p \geq i}^{N} \exp[\beta_j \tau_p]} \right]. \tag{2.11}$$

Kalbfleisch and Prentice (1973) demonstrated that the partial likelihood is a marginal likelihood for β arising out of the distribution of the rank vector associated with the failure times (or response times). The use of

the partial likelihood for inference on β has been justified from both the frequentist viewpoint (Anderson and Gill, 1982) and the Bayesian viewpoint (Kalbfleisch, 1978).

2.1.2 Parameter Estimation: Markov chain Monte Carlo

Suppose the items are indexed by $j = 1, \cdots, J$, and the examinees by $i = 1, \cdots, N$. For the ith test taker, his or her responses and response times are denoted by $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{iJ})'$, and $\boldsymbol{T}_i = (T_{i1}, \cdots, T_{iJ})'$, respectively. We model the jth item's hazard function by (2.2) and specify the partial likelihood function by (2.8). We assume a three-parameter IRT model (2.1) for the response variable $\boldsymbol{Y}_i$, then the likelihood function for the ith subject's ability θ_i can be specified as

$$\text{IRT}(\theta_i) = \prod_{j=1}^{J} P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}}. \tag{2.12}$$

To estimate the parameters $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)'$, note that in CAT it would generally be the case that different examinees take different items, and the items they take are closely associated with their ability level θ (so is related with τ as well). This means some off-the-shelf methods for marginal likelihood estimation or a frailty model procedure will not work. So in our investigation, a Bayesian MCMC method (Metropolis-Hastings algorithm) is used instead. The MCMC method generates samples from $\pi(\omega)$ by creating a Markov chain on the state space of ω that has its equilibrium distribution $\pi(\omega)$. Theoretical details on MCMC methods can be found in Tierney (1994).

Our objective is utilizing the RT information to estimate the non-parametric baseline hazard h_0, regression parameter β, item parameters a, b, c, examinees' speed parameter τ, and also obtain more information for the estimation of θ. Notice that in this model, θ does not play a direct role in RT modeling, but RT still provides additional information for θ estimation through the higher-order relationship between θ and τ. During the estimation, we need to sequentially draw parameters a, b, c,$\sigma_{\theta\tau}$(or $\rho_{\theta\tau}$), θ, τ and β.

Prior Specification

A bivariate normal prior is chosen for the latent parameters (θ, τ), i.e., $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\mu}_p = (0, 0)$ and $\boldsymbol{\Sigma}_p = \begin{pmatrix} 1 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & 1 \end{pmatrix}$. The correlation term $\rho_{\theta\tau}$ is chosen to have a vague prior as in Klein Entink, Fox, and van der Linden (2009), specifically, a truncated normal prior is chosen as $\rho_{\theta\tau} \sim \mathcal{N}_{[-1,1]}(0, 10)$ truncated on the interval $[-1, 1]$. A normal prior is chosen for each regression parameter β_j with means equal to 0 and

variance chosen to be 10. Here we purposely selected a large variance to make the prior less informative. For item parameters, we specify independent priors. This treatment was employed in Patz and Junker (1999) and is assumed to be consistent with some test conventions, such as National Assessment of Educational Progress (NEAP). Specifically, we assume a common beta prior for the guessing parameter as

$$c_j \sim \text{beta}(\gamma, \delta), j = 1, ..., J$$

and assume normal and lognormal priors for a and b parameters separately as

$$p(b_j) \sim \mathcal{N}(0, \sigma_b^2)$$
$$p(a_j) \sim \text{lognormal}(0, \sigma_a^2).$$

Justification of the Partial Likelihood

The partial likelihood may not be seen as a likelihood in a strict sense, yet Kalbfleisch (1978) provides rigorous justification of using partial likelihood in a Bayesian context. Specifically, he showed that marginalizing with respect to an independent-increment gamma process prior on a baseline cumulative hazard led to a posterior density of β that is proportional to the partial likelihood. In the usual Cox model with covariates (denoted as τ's) observed, when integrating out H_{0j} with respect to a diffuse gamma process prior on the cumulative hazard, the posterior marginal density of the regression parameter β is verified to be

$$\pi(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) \propto L(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) p(\beta_j|\mu_\beta, \sigma_\beta^2),$$

where $L(\cdot)$ is the partial likelihood in Eq.(2.8), and $p(\cdot)$ denotes the prior density. This result provides rationale for using partial likelihood in updating the Markov chain. When τ is a latent covariate, Gustafson (1997) used

$$\pi(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) \propto L(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) p(\boldsymbol{\tau}|\mu_\tau, \sigma_\tau) p(\beta_j|\mu_\beta, \sigma_\beta^2),$$

and similarly, we could use

$$\pi(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) \propto L(\beta_j|\boldsymbol{t}, \boldsymbol{\tau}) p(\boldsymbol{\tau}|\boldsymbol{\theta}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\beta_j|\mu_\beta, \sigma_\beta^2)$$

for updating the chain of β. The second level model on person parameters is reflected via the term $p(\boldsymbol{\tau}|\boldsymbol{\theta}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, and it will be cancel out in the Metropolis-Hastings updating algorithm. When updating the

person parameter (θ_i, τ_i) in the Markov chain, we have

$$\pi(\theta_i, \tau_i|\boldsymbol{t}, \boldsymbol{y}_i) \propto p(\tau_i|\boldsymbol{\beta}, \boldsymbol{t})p(\theta_i|\boldsymbol{y}_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})p(\theta_i, \tau_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p),$$

as a result of local independence assumption, where $p(\tau_i|\boldsymbol{\beta}, \boldsymbol{t})$ is calculated from the partial likelihood $L(\boldsymbol{\beta}|\tau_i, \boldsymbol{t})$. We need to show that $L(\boldsymbol{\beta}|\tau_i, \boldsymbol{t})p(\theta_i|\boldsymbol{y}_i, a, b, c)p(\theta_i, \tau_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ yields a proper posterior, that is, it has a bounded integral. Because both $0 < L(\beta_j|\boldsymbol{\tau}, \boldsymbol{t}) < 1$ and $0 < p(\theta_i|\boldsymbol{y}_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) < 1$ are bounded likelihoods, and $p(\theta_i, \tau_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ is a proper prior, implying that, $\int L(\boldsymbol{\beta}|\tau_i, \boldsymbol{t})p(\theta_i|\boldsymbol{y}_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})p(\theta_i, \tau_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)d\theta d\tau < \infty$.

Detailed MCMC Algorithm

To perform the sampling for parameters with support on the entire real line, we use normal proposal distributions with mean equal to the current estimation and variance chosen to give a Metropolis acceptance rate of between 25 and 40 percent. For parameters with support not on the real line, we either transform them to the real line and then sample them from normal proposal distribution, or chose some special proposal distribution.

Step 1: Denote the initial values for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\tau}$ by $\hat{\boldsymbol{\beta}}_0 \equiv (\hat{\beta}_{01}, ..., \hat{\beta}_{0J})'$, $\hat{\boldsymbol{\theta}}_0 \equiv (\hat{\theta}_{01}, ..., \hat{\theta}_{0N})'$ and $\hat{\boldsymbol{\tau}}_0 \equiv (\hat{\tau}_{01}, ..., \hat{\tau}_{0N})'$, respectively. Arbitrary initials are chosen for a and c parameters, such as 1 for all a-parameters, and 0.1 for all c-parameters. Initial values for b-parameters are selected in a slightly more informative way. That is, we rank order the items based on their correct response probability, and assign the corresponding percentile with respect to the standard normal distribution as their initials. Initial values of θ are obtained in a similar fashion. The initial values of $\hat{\boldsymbol{\tau}}$ are obtained differently depending upon the specific test conditions. If all examinees answer the same set of items (often the case in a computer based linear test), we can rank order the examinees' total RT on all the items, and set $\hat{\tau}_{0i} \approx \Phi^{-1}(p_i), i = 1, ..., N$. Here, Φ is the standard normal distribution function, p_i is the percentile of the i^{th} examinee's total RT. In the adaptive test setting, because each examinee answers different sets of items, total RT is no longer comparable. In this case, we generate $\hat{\boldsymbol{\tau}}$ from the normal distribution conditional on $\hat{\boldsymbol{\theta}}_0$. To be specific, arbitrarily choose $\sigma_{\theta\tau}^{(0)}$ as the initial value of the covariance between θ and τ, oftentimes this value is set to be 0.5. Because $\sigma_\tau^2 = 1$ and $\sigma_\theta^2 = 1$ for the sake of identifiability, $\hat{\tau}_{0i} \sim \mathcal{N}(\sigma_{\theta\tau}^{(0)}\hat{\theta}_{0i}, 1 - [\sigma_{\theta\tau}^{(0)}]^2), i = 1, ..., N$. Conditioning on $\hat{\boldsymbol{\tau}}_0$, $\hat{\boldsymbol{\beta}}_0$ is obtained by maximizing the partial likelihood function defined in (2.8). Set the iteration counter iter=1.

Step 2: At rth step, denote the previous positions $\boldsymbol{\beta}^{(r-1)} \equiv (\beta_1^{(r-1)}, ..., \beta_J^{(r-1)})'$, $\boldsymbol{a}^{(r-1)} \equiv (a_1^{(r-1)}, ..., a_J^{(r-1)})'$, $\boldsymbol{b}^{(r-1)} \equiv (b_1^{(r-1)}, ..., b_J^{(r-1)})'$, $\boldsymbol{c}^{(r-1)} \equiv (c_1^{(r-1)}, ..., c_J^{(r-1)})'$, $\boldsymbol{\theta}^{(r-1)} \equiv (\theta_1^{(r-1)}, ..., \theta_N^{(r-1)})'$, $\boldsymbol{\tau}^{(r-1)} \equiv (\tau_1^{(r-1)}, ..., \tau_N^{(r-1)})'$, and $\sigma_{\theta\tau}^{(r-1)}$. Sample each parameter sequentially as follows.

1. c: To enable the Gibbs sampling of guessing parameter, the data augmentation involves the definition of a latent variable W_{ij}, which is equal to 1 when person i knows the correct answer to item j, and 0 otherwise. Sampling c from its posterior distribution includes sampling $W_{ij}|Y_{ij}, \theta_i, a_j, b_j, c_j$. Specifically, we have

   (a) if $y_{ij} = 0$ then $w_{ij}^{(r)} = 0$;

   (b) if $y_{ij} = 1$ then $w_{ij}^{(r)} = 1$ with probability $\frac{\phi(\theta_i)}{c_j+(1-c_j)\phi(\theta_i)}$ where $\phi(\theta_i) = \frac{1}{1+\exp(-a_j(\theta_i-b_j))}$

   To sample $c_j|W, Y$, define $T_j^{(r)} = \sum_{i=1}^{N} I(w_{ij}^{(r)} = 0)$ as the number of persons who do not know the correct response to item j, and define $M_j^{(r)} = \sum_{i=1}^{N} I(w_{ij}^{(r)} = 0)I(y_{ij} = 1)$ as the number of persons who do not know the correct response to item j but correctly answer the item. Apparently, $M_j^{(t)}$ follows binomial distribution with parameters $T_j^{(r)}$ and c_j. Because c has a beta prior, then

$$c_j^{(r)} \sim Beta(M_j^{(t)} + \gamma, T_j^{(r)} - M_j^{(r)} + \beta)$$

2. a and b: Draw a_j^* ~lognormal$(\log(a_j^{(r-1)}), c_a^2)$ and $b_j^* \sim \mathcal{N}(b_j^{(r-1)}, c_b^2)$ independently for each $j = 1, 2, ...J$. Following Patz and Junker (1999), we can update a and b simultaneously. The acceptance probability is calculated as

$$\alpha((a_j^{(r-1)}, b_j^{(r-1)}), (a_j^*, b_j^*)) = \min\left\{1, R_{ab}\right\},$$

   where

$$R_{ab} = \frac{L(\boldsymbol{Y}_j|\boldsymbol{\theta}^{(r-1)}, a_j^*, b_j^*, c_j^{(r)})\mathrm{T}(a_j^*, a_j^{(r-1)})p(a_j^*)p(b_j^*)}{L(\boldsymbol{Y}_j|\boldsymbol{\theta}^{(r-1)}, a_j^{(r-1)}, b_j^{(r-1)}, c_j^{(r)})\mathrm{T}(a_j^{(r-1)}, a_j^*)p(a_j^{(r-1)})p(b_j^{(r-1)})},$$

   where $p(\cdot)$ denotes the prior density. $T(\cdot)$ denotes the transition kernel (Patz and Junker, 1999). This term is not canceled out because of the lack of symmetry in the lognormal proposal density.

3. $\rho_{\theta\tau}$: Sample correlation between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$: Because $-1 \leq \rho_{\theta\tau} \leq 1$, we need to first transform $\rho_{\theta\tau}$ to the real line, the transformation we adopt is $\rho_{\theta\tau} = -1 + 2\frac{e^{\varphi}}{1+e^{\varphi}}$. Then draw φ^* from $\mathcal{N}(\varphi^{r-1}, 1)$ with acceptance probability

$$\alpha(\varphi^{r-1}, \varphi^*) \equiv \min\left\{1, \frac{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2r}, \varphi^*, \mu_\theta^{(r-1)})\pi_\rho(\rho_{\theta\tau}^*)J(\varphi^*)}{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2r}, \varphi^{(r-1)}, \mu_\theta^{(r-1)})\pi_\rho(\rho_{\theta\tau}^{(r-1)})J(\varphi^{(r-1)})}\right\} \tag{2.13}$$

   where

   $p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2r}, \varphi^*, \mu_\theta^{(r-1)}) = \prod_{i=1}^{N} p(\theta_i^{(r-1)}, \tau_i^{(r-1)}|\sigma_\theta^{2r}, \varphi^*, \mu_\theta^{(r-1)})$

$\sim \prod_{i=1}^{N} \frac{|\mathbf{\Sigma}_2^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}\boldsymbol{\xi}_i^T \mathbf{\Sigma}_2^{-1} \boldsymbol{\xi}_i\right]$, where the variance-covariance matrix is $\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & \sigma^*_{\theta\tau} \\ \sigma^*_{\theta\tau} & 1 \end{pmatrix}$. with $\sigma^*_{\theta\tau} = \rho^*_{\theta\tau} = -1 + 2\frac{\exp(\varphi^*)}{1+\exp(\varphi^*)}$ and $\pi_\rho(\rho^*_{\theta\tau})$ is the normal prior density of the correlation term. The $J(\cdot)$ is the Jacobian term expressed as $J(\varphi^*) = \frac{2\exp(\varphi^*)}{(1+\exp(\varphi^*))^2}$.

4. θ and τ: Sample examinees' ability and speed parameter: For the i^{th} pair $(\theta_i, \tau_i), 1 \leq i \leq N$, draw (θ_i^*, τ_i^*) from a bivariate normal distribution with mean $(\theta_i^{(r-1)}, \tau_i^{(r-1)})$. The acceptance probability is

$$\alpha(\theta_i^{(r-1)}, \tau_i^{(r-1)}, \theta_i^*, \tau_i^*) \equiv \min\left\{1, \frac{\text{IRT}(\theta_i^*) L(\boldsymbol{\beta^{(r-1)}}|\boldsymbol{\tau}^*)\pi(\theta_i^*, \tau_i^*)}{\text{IRT}(\theta_i^{(r-1)}) L(\boldsymbol{\beta^{(r-1)}}|\boldsymbol{\tau}^{(r-1)})\pi(\theta_i^{(r-1)}, \tau_i^{(r-1)})}\right\} \tag{2.14}$$

where $\boldsymbol{\tau}^* = (\tau_1^{(r-1)}, ..., \tau_i^*, ..., \tau_N^{(r-1)})'$. $\pi(\theta_i^r, \tau_i^{(r-1)})$ is a bivariate normal with mean $(0, 0)$ and variance-covariance matrix $\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & \sigma^{(r)}_{\theta\tau} \\ \sigma^{(r)}_{\theta\tau} & 1 \end{pmatrix}$. IRT$(\cdot)$ is calculated from equation (2.12) and $L(\cdot)$ is defined by (2.9), respectively.

5. β: Sample survival regression parameter: For jth item, draw β_j^* from a normal distribution $\mathcal{N}(\beta_j^{r-1}, 1)$ with the acceptance probability defined as

$$\alpha(\beta_j^{(r-1)}, \beta_j^*) \equiv \min\left\{1, \frac{L(\beta_j^*|\boldsymbol{\tau}^r)p(\beta_j^*)}{L(\beta_j^{(r-1)}|\boldsymbol{\tau}^r)p(\beta_j^{(r-1)})}\right\} \tag{2.15}$$

where $L(\cdot)$ is defined in equation (2.8).

Step 4: At the end of the chain, compute the posterior mean of each parameter. A burn-in period of the initial K iterations is often required to allow the chain to reach equilibrium.

2.1.3 Estimation of the Cumulative Baseline Hazard

The nonparametric cumulative baseline hazard can be estimated via the Breslow estimator (Breslow, 1972). For the jth item, in order to estimate h_{0j}, we express the complete likelihood as

$$\begin{aligned} L(\beta_j, h_{0j}(t)) &= \prod_{i=1}^{N} f(t_{ij}|\tau_i) = \prod_{i=1}^{N} -\frac{dS(t_{ij}|\tau_i)}{dt_{ij}} \\ &= \prod_{i=1}^{N} h_{0j}(t_{ij}) \exp(\beta_j \tau_i) \exp[-H_{0j}(t_{ij}) \exp(\beta_j \tau_i)]. \end{aligned}$$

Replace β_j by its estimator $\hat{\beta}_j$ from the MCMC estimation and consider maximizing the above likelihood as a function of $h_{0j}(t)$ only. It can be verified that the likelihood is maximized when $h_{0j}(t) = 0$ except for times

at which the events occur. The Breslow estimator for the cumulative baseline hazard takes the following form (Breslow, 1972),

$$\hat{H}_{0j}(t) = \sum_{i=1}^{N} \frac{I_{t_i \leq t}}{\sum_{p \geq i}^{N} \exp[\hat{\beta}_j \hat{\tau}_p]}. \tag{2.16}$$

The non-parametric baseline hazard $h_0(t)$, though flexible, is somewhat inconvenient in that the whole hazard function has to be stored for each item to be able to recover the entire response time distribution. To fix this, we propose retaining much of the flexibility of the new models, but directly fitting the cumulative hazard $H_0(t)$ estimated from Brewslow estimator with B-splines, such that the entire RT distribution, conditional on τ, can be expressed without a great many parameters. In mathematics, a spline is a special function defined piecewise by polynomials. B-splines refers to a linear basis for the piecewise polynomials, and offers spline functions that have minimal support with respect to a given degree, smoothness, and domain partition. We chose B-splines here because it can describe a variety of shapes with a minimal number of parameters, while avoiding computational problems.

Specifically, we will adopt a cubic B-spline basis. When the knots and boundary points are specified, the basis functions are determined recursively from the following formula:

$$\begin{aligned} B_{i,0} &= I_{(u_i \leq t \leq u_{i+1})} \\ B_{i,p} &= \frac{t - u_i}{u_{i+p} - u_i} B_{i,p-1}(t) + \frac{u_{i+p+1} - t}{u_{i+p+1} - u_{i+1}} B_{i+1,p-1}(t). \end{aligned}$$

Here u_is are the ordered knot points (including boundary points), p is the degree of the B-spline basis and the number of basis functions equals to $p + m + 1$ with m being the number of inner knots. The knots are often chosen to the equally spanned along the range of the data. For example, if the number of knots is 3, then the three knots are the 25th, 50th, and 75th percentile of the whole range of the data. Usually, increasing the number of knots or increasing the degree will lead to a better fit. But oftentimes, the degree is chosen to be 3, indicating a cubic basis function. Once the B-spline bases are specified, we treat them as predictors and fit linear regression model to the Breslow estimated baseline hazard. In this way, we obtain the regression coefficient for each basis. For details about B-spline, please refer to de Boor (1978) and He and Shi (1998). An apparent advantage of the B-spline idea here is that only the knots, boundary points and regression parameters are needed to recover the whole baseline hazard.

2.2 Model Diagnosis

Model fit checking is an important step in any model development. In this section, we propose to use two approaches of evaluating model fit: (1) posterior predictive checks (Gelman, Carlin, Stern, & Rubin, 1995) and (2) a survival analysis specific residual method.

Posterior predictive checks. Given the posterior distribution of the model parameters, one can calculate the predicted response time for test taker i and item j, denoted as $\tilde{t}_{ij}$. For each observation, t_{ij}, we can calculate the left-sided probability of exceedance of the observation under its predictive density,

$$Pr\{\tilde{t}_{ij} < t_{ij}\}, i = 1, ..., N, j = 1, ..., J. \tag{2.17}$$

The distributions of the above probabilities over all the person item combinations in the sample will be used to evaluate the global fit of the model (van der Linden, Breithaupt, Chauah, & Yang, 2007). If the model fits, the cumulative distributions of these probabilities will follow the identity line (e.g., Casalla and Berger, 1990). This model diagnosis method is appropriate for any kind of model.

Residual checks. The Cox model can be rewritten as $S(t) = [S_0(t)]^{\exp(\tau'\beta)}$. It follows that

$$\log\{-\log[S(t)]\} = \log\{-\log[S_0(t)]\} + \tau\beta.$$

We can further rewrite the equation as

$$\log\{-\log[S(t)]\} = T(t) + \tau\beta, \tag{2.18}$$

where $S(\cdot)$ is the survival function of T given τ. $T(t) = \log\{-\log[S_0(t)]\} = \log[\int_0^t h_0(s)ds]$ is an unspecified strictly monotone function (because of the unknown form of the nonnegative function $h_0(t)$), which maps the positive half-line onto the whole real line. Now it is clear to see that (2.18) is equivalent to the so-called linear transformation model (Cuzick, 1988) as $T(t) = -Z'\beta + \varepsilon$ where ε follows the extreme value distribution $F = 1 - g^{-1} = 1 - \exp\{-\exp(s)\}$. Following this argument, we can calculate the residual for each item-person pair as

$$\varepsilon_{ij} = \log(\hat{H}_{0j}(t)) + \hat{\tau}_i\hat{\beta}_j. \tag{2.19}$$

If the model fits the data well, the ε_{ij} should follow the extreme value distribution closely. In terms of graphical representation, one can draw the distribution plot for $\varepsilon_{ij}, i = 1, ..., N$ against standard extreme value distribution for item j. Departure from the theoretical distribution of ε_{ij} signals the possible model

misfit for the item. In addition to the graphical check, we also proposed to use an index to summarize the goodness of fit statistically. That is, we can calculate the Kullback-Leibler (KL) distance between the empirical density estimated from ε_{ij}'s and its theoretical exponential distribution. One most widely used non-parametric density estimation method is the kernel smoothing method. Thus, the KL distance is calculated between the kernel smoothed density and extreme-value density. Smaller distance indicates better fit. Mallick and Walker (2003) first used KL distance to measure the precision of the density estimation, and we applied the similar idea here. These two diagnostic methods will be used in real data analysis in Chapter 4.

2.3 Simulation Study

2.3.1 Study One: Check the Estimation Accuracy

A simulation study was carried out to check the performance of the proposed MCMC estimation method. As a starting point, we only consider the non-adaptive situation, in which each examinee has taken the same set of items. A total of $2 \times 2 \times 3 = 12$ different test conditions are simulated. The first factor represents test length J, and two levels ($J = 20, 40$) are considered. The second factor represents sample size N, and again two levels (N=250, 500) are considered. The third factor represents three different shapes of baseline hazard functions: exponential, Weibull, and a non-monotone hazard. For the exponential baseline hazard, $h(\cdot) = \lambda$ with λs drawn from a uniform distribution $\lambda \sim U(0.25, 1.5)$; for the Weibull baseline hazard, $h(\cdot) = \lambda\alpha t^{\alpha-1}$ with λs drawn from a uniform distribution $\lambda \sim U(0.25, 1.5)$ and αs drawn from another uniform distribution $\alpha \sim U(1, 3)$. The selection of these values, though arbitrary, yields a baseline hazard function with reasonable mean and variance. We intentionally chose a non-monotone baseline hazard as a third option to show that the proposed model is flexible enough to recover various shapes of the RT distribution, even when the hazard is not monotonically increasing or decreasing. The specific parametric form we chose is $h(\cdot) = 0.5\lambda(x - \alpha)^2$ with λs drawn from a uniform distribution $\lambda \sim U(0.25, 1.5)$ and αs drawn from $\alpha \sim U(1, 3)$. This quadratic form yields a inverse-bell shaped baseline hazard. To show that the parameters chosen here generate reasonable response time distribution, Figure 2.1 illustrates the RT distributions generated from Cox model with different baseline hazard and for certain fixed values of λ, β, and α. Each curve represents the shape of the histogram of the RT distributions. The curves were obtained by averaging over 100 replications. As one will notice later, the curves resemble the RT distributions obtained from real example very closely.

The three-parameter logistic model was used for generating item responses. Item discrimination and

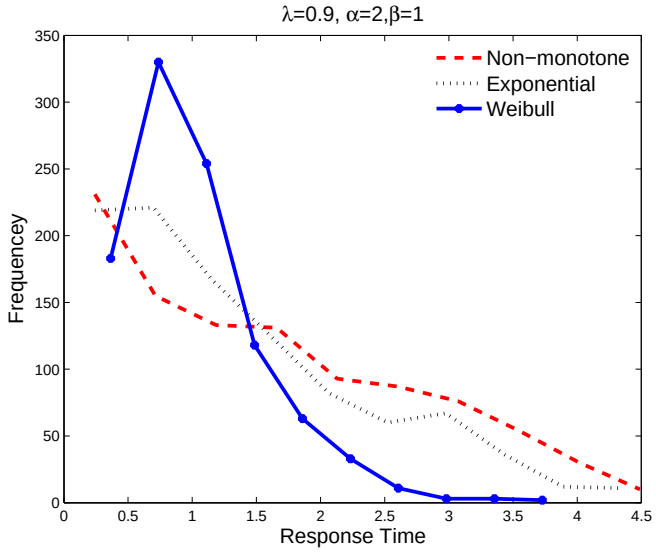


Figure 2.1: Illustration of response time distribution under different shapes of baseline hazard

difficulty parameters were simulated from $a \sim U(1, 2.5)$, $b \sim \mathcal{N}(0, 1)$, item pseudo-guessing parameter is simulated from $c \sim U(0, 0.2)$. Examinees' latent trait (θ, τ) was drawn from a bivariate normal distribution with mean $\mu = [0, 0]$ and covariance matrix $\sigma = [1, 0.5; 0.5, 1]$. The regression parameter was drawn from $\beta \sim U(0.5, 1.5)$. To implement the Bayesian MCMC algorithm, chains of length 4000 with an initial burn-in period 1000 were chosen. There were 10 replications for each simulation condition. Item and examinee parameters for each replication are generated separately.

2.3.2 Results

The Markov chain for each parameter appeared to reach equilibrium, and had small autocorrelations beyond the first couple of lags. Mean squared error (MSE) and average bias were calculated to check how close the estimated parameters were to their true values. Table 2.1 presents the MSE and bias of θ and τ for the 12 simulation conditions. All values were averaged over all examinees and all replications within a simulation condition. Tables 2.2 tabulates the MSE and average bias for item parameters, including β and a, b, c. Notice that the true value of $\sigma_{\theta\tau}$ across all conditions is 0.5. We report the final estimates of correlation term in Table 2.3, the mean value is calculated from the 10 replications. Please ignore the last two columns of each table for the moment.

For the log hazard ratio regression parameter β, the estimation is quite accurate in general, as indicated by the small MSE in Table 2.2. There is an apparent trend that increasing the population size reduces the MSE of β. The results also show that no matter which shape the baseline hazard takes, the model can always be accurately recovered. Increasing the test length reduces the MSE of τ and θ. Figure 2.2 shows the true and estimated cumulative baseline hazards. Here we only present the results for $J = 20$ and $N = 250$ under one replication because the other conditions are alike. The Breslow estimator appears to reconstruct the baseline cumulative hazard functions well under all three different shapes except at the right boundaries. The possible reason is at the right boundary, the size of the risk set is very small and thus the hazard estimation may be inflated. But considering only a small portion of examinees will have

Table 2.1: MSE and average bias for the θ and τ estimation

		J=20,N=250		J=40,N=250		J=20,N=500		J=40,N=500		$\rho_{b\lambda} = 0.3$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Exponential baseline	$\hat{\theta}$	0.018	0.148	0.038	0.111	-0.007	0.152	0.009	0.091	0.026	0.172
	$\hat{\tau}$	0.017	0.078	0.037	0.052	0.027	0.076	-0.013	0.055	0.011	0.098
Weibull baseline	$\hat{\theta}$	0.034	0.148	0.021	0.095	0.039	0.133	-0.011	0.076		
	$\hat{\tau}$	0.039	0.056	0.049	0.029	0.041	0.067	-0.018	0.03		
Non-monotone baseline	$\hat{\theta}$	0.023	0.166	-0.005	0.108	0.001	0.152	-0.009	0.106		
	$\hat{\theta}$	0.011	0.071	0.033	0.045	-0.011	0.069	0.013	0.051		

Table 2.2: MSE and average bias for the item parameter estimation

		J=20,N=250		J=40,N=250		J=20,N=500		J=40,N=500		$\rho_{b\lambda} = 0.3$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Exponential baseline	a	-0.120	0.194	-0.261	0.178	-0.102	0.157	-0.149	0.102	0.020	0.237
	b	0.009	0.052	0.008	0.050	0.061	0.028	-0.024	0.028	0.027	0.100
	c	0.024	0.004	0.020	0.004	0.019	0.004	0.016	0.003	0.010	0.009
	β	-0.047	0.027	-0.023	0.023	0.058	0.015	-0.049	0.013	-0.044	0.047
Weibull baseline	a	-0.107	0.124	-0.139	0.115	0.076	0.081	0.038	0.085		
	b	0.096	0.050	0.092	0.051	0.093	0.042	0.097	0.039		
	c	-0.011	0.007	-0.013	0.006	-0.031	0.007	-0.033	0.007		
	β	0.012	0.022	0.038	0.019	0.005	0.014	-0.002	0.011		
Non-monotone baseline	a	-0.171	0.167	-0.239	0.208	-0.085	0.178	-0.093	0.163		
	b	0.007	0.040	0.061	0.057	0.045	0.037	0.044	0.040		
	c	0.020	0.004	0.022	0.004	0.019	0.004	0.017	0.004		
	β	-0.072	0.023	-0.011	0.019	-0.049	0.016	0.053	0.015		

Table 2.3: Mean and Standard Deviation for the integrated absolute difference between $H_0(t)$ and Breslow estimator & mean of $\rho_{\theta\tau}$

		J=20,N=250		J=40,N=250		J=20,N=500		J=40,N=500		$\rho_{b\lambda} = 0.3$	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Exponential baseline	d_j	1.321	0.983	1.407	1.051	1.152	0.791	0.910	0.732	1.471	1.182
	$\rho_{\theta\tau}$	0.507		0.491		0.519		0.485		0.521	
Weibull baseline	d_j	1.980	1.69	1.449	1.238	1.471	1.131	1.503	1.219		
	$\rho_{\theta\tau}$	0.475		0.491		0.519		0.513			
Non-monotone baseline	d_j	1.296	0.934	0.994	0.841	0.918	0.643	0.899	0.651		
	$\rho_{\theta\tau}$	0.511		0.466		0.492		0.491			

extreme RTs, this inflation is tolerable. To further quantify the discrepancy between the true and estimated cumulative hazard, we calculate the integrated absolute different between the true $H_{0j}(t)$ and the Breslow estimator for the j^{th} item

$$d_j = \int |H_{0j}(t) - \hat{H}_{0j}(t)|dt. \tag{2.20}$$

The mean and standard deviation of d_j are reported in Table 3. The results are again based on 10 replications.

2.3.3 Study Two: When the Item Parameters are Correlated

This study is designed to show that even if the item parameters have some moderate correlation, especially between item time intensity and item difficulty parameters, the proposed algorithm can still generate satisfactory results, with the item covariance matrix unestimated. As an illustration, we only consider the exponential model, in which the baseline hazard λ_j is negatively correlated with the item difficulty b_j to produce a positive correlation between item time intensity and item difficulty. Specifically, λ's and b's were generated from bivariate normal with mean $[1, 0]$, and covariance matrix $[1, \rho_{\lambda b}; \rho_{\lambda b}, 1]$, with two levels of $\rho_{\lambda b} = 0, 0.3$. All the rest parameters were simulated in the same fashion as in simulation study one. The MSE and average bias for all parameters with $\rho_{\lambda b} = 0.3$ are presented in the last two columns of Tables 2.1 and 2.2. When $\rho_{\lambda b} = 0$, the results are very close to the results from simulation study one, and they are omitted here. As one can tell, with the increased correlation of $\rho_{\lambda b}$, the estimation errors only slightly inflated, but they are still acceptable. Because the item covariance matrix will not influence our conclusion about the data, the second level model on the item covariance matrix can be ignored to simplify the model estimation.

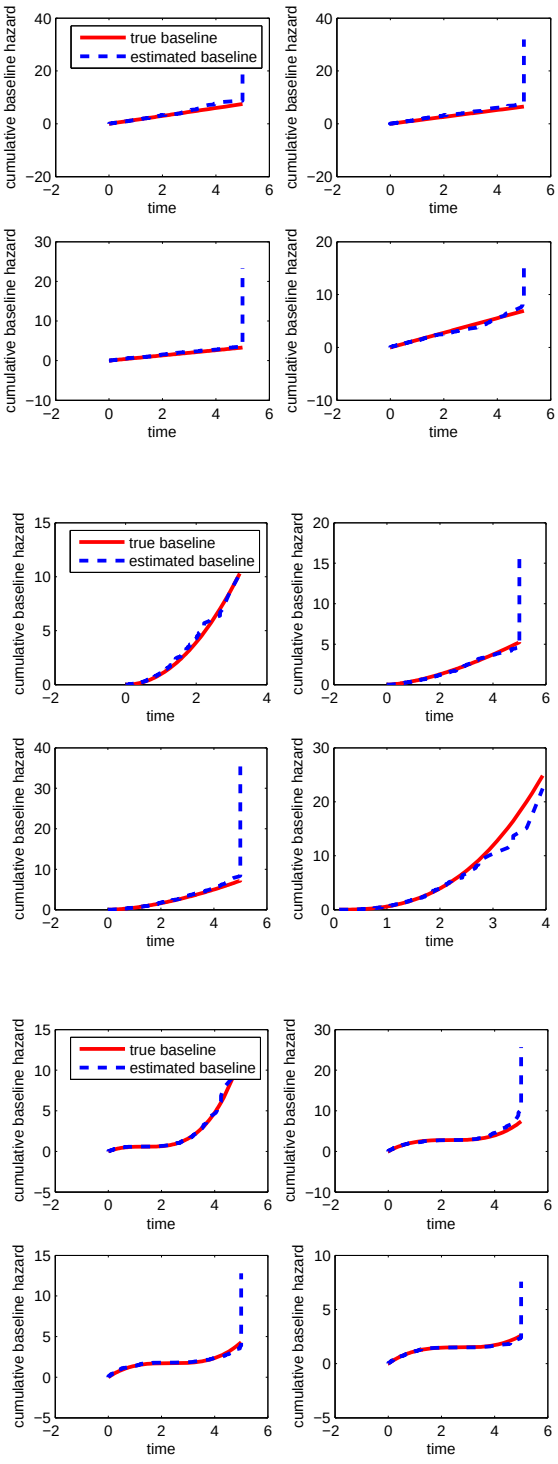


Figure 2.2: True vs. Estimated cumulative baseline hazard for different shapes of baseline hazard

Chapter 3

The Linear Transformation Model with Frailties

[1] The hierarchical PH model is demonstrated to be flexible in that it encompasses various parametric regression models for response times, such as the exponential model or Weibull model. The general expression for the PH model is

$$h(t|Z = z) = h_0(t) \exp(\boldsymbol{Z}'\boldsymbol{\beta}), \tag{3.1}$$

where $\boldsymbol{\beta}' = (\beta_1, ..., \beta_p)$ are regression parameters, and $\boldsymbol{Z}$ is the covariate vector. The effect of the covariate is reflected via a linear form, $\boldsymbol{Z}'\boldsymbol{\beta}$. Since the link function is exponential, which assumes that a unit increase in a covariate is multiplicative with respect to the hazard rate. It is this property that excludes other possible functional relationships between the covariates and RTs. One obvious example is the widely used lognormal regression model. If let $Y = \log T$, then $Y = \alpha + \sigma W + \boldsymbol{Z}'\boldsymbol{\beta}$, where W is a standard normal term. With different location parameters α_1 and α_2, the hazard rates are not proportional to each other. Also because the effect of the covariates can not be written in an exponential link function as in Equation (3.1), the lognormal regression model does not belong to the proportional hazard model. However, it does belong to a more general model—the linear transformation model.

In this chapter, we propose a new semi-parametric model based on the linear transformation model that only assumes the existence of a monotone but otherwise arbitrary transformation of the response times such that the linear model holds. The semi-parametric nature of the model allows considerable generality and applicability but enough structure for useful substantive interpretation. In fact, by allowing the error term to take on different distributions, the linear transformation model includes the lognormal model, the Box-Cox model, the proportional hazard model and many other models as special cases. Due to its flexibility, this model has already been widely used in biostatistics to explore the effects of the covariates on the (cancer) patients' survival times. In those applications, however, the covariates are often observed, such as tumor type, measure of general fitness, and so on (Prentice, 1973; Cheng, Wei, & Ying, 1995). Researchers later recognized the correlations among survival times that are due to either repeated measurements taken on

[1] This chapter was currently accepted as a peer-reviewed paper entitled "The linear transformation model with frailties for the analysis of item response times" by the *British Journal of Mathematical and Statistical Psychology*.

a single subject, or measurements of a common variable taken on genetically associated subjects, and this gave rise to the development of the *frailty models*, in which a latent *frailty* random variable is included in the model to account for possible correlations in survival time distributions (Clayton, 1991; Clayton and Cuzick, 1985). However, the standard frailty model only generalizes the proportional hazard model by incorporating a random effect, such that units within the same group (or response times for all test items within an individual) share the same frailty. The response time for each item within a same individual is assumed to be independent. It is only recently that some researchers introduced the frailty term into the linear transformation model, such as Mallick and Walker (2003) in which the model is used in Veteran's Admission lung cancer trial data. Dunson (2003) proposed a slightly different model called "dynamic latent variable models" for multidimensional longitudinal data. In that model, the dependent variables are assumed from a distribution in an exponential family with canonical parameter, after certain known monotone transformation, being equal to a linear combination of covairates (either observed or latent) plus an error term.

With such a flexible model, the challenge is the model estimation. We propose a two-stage estimation method. First, the linear transformation frailty model is placed as a first-level measurement model in a two-level model framework such that the response times and response accuracy are estimated simultaneously. In the second stage, we propose a two-stage estimation method, incorporating a rank-based marginal likelihood as a key building block. This method offers a way of estimating linear transformation models with latent covariates together with the population covariance matrix at the second level. The new method is also flexible enough to deal with the sparse data, such as the data often collected from computerized adaptive testing.

3.1 The Linear Transformation Model

The independent random variables, $T_1, ..., T_n$, are said to follow a linear transformation model if for some increasing transformation H,

$$H(T_i) = \boldsymbol{Z}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, ..., n. \tag{3.2}$$

Z_i is an observed covariate (it can also be a vector), $\boldsymbol{\beta}$ is the regression parameter, $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. with distribution F. This model indicates that, after some order preserving transformation, the dependent variable is related to $\boldsymbol{Z}$ in a simple linear fashion except for random errors (Cuzick, 1988). Parametric forms for H have been studied extensively in the literature, some examples are (1) $H(t) = (t + c)^\lambda$, (2) $H(t) = sign(t)|t|^\lambda$, and (3) $H(t) = (t^\lambda - 1)/\lambda(\lambda \neq 0), H(t) = \log t, \lambda = 0$ (Box & Cox, 1964). Here t is a

realization of the random variable T. The third example is the well-known Box-Cox power transformation. When the parametric family of H is not specified, equation (3.2) becomes a non-parametric model where H is continuous and monotone, but otherwise arbitrary. Linear transformation models represents a rich family of different models. For example, if ε follows the standard logistic distribution, then (3.2) reduces to the proportional odds model (Pettitt, 1982; Bennett, 1983); if ε follows the standard normal distribution, then (3.2) becomes a semi-parametric extension of the Box-Cox model (Doksum, 1987).

The proportional hazard model proposed in Chapter 2 assumes a constant relative risk compared to the baseline hazard function given the covariates. When the assumption is violated, the proportional odds (PO) model provides an alternative. In the PO model, the log-odds of the response time distribution depends on the linear combination of the covariates. As we have emphasized, the linear transformation model provides a unified approach to include those specific semi-parametric models. For instance, to see that the Cox proportional hazard model is a special case of the linear transformation model, first write the Cox model in Lehmann (1953) form as $S_Z(t) = [S_0(t)]^{\exp(Z'\beta)}$. It follows that

$$\log\{-\log[S_Z(t)]\} = \log\{-\log[S_0(t)]\} + \boldsymbol{Z}_i'\boldsymbol{\beta}.$$

We can further rewrite the equation as

$$\log\{-\log[S_Z(t)]\} = H(t) + \boldsymbol{Z}_i'\boldsymbol{\beta}, \tag{3.3}$$

where $S_Z(\cdot)$ is the survival function of T given Z. $H(t) = \log\{-\log[S_0(t)]\} = \log[\int h_0(t)dt]$ is a unspecified strictly monotone function (because of the unknown form of $h_0(t)$), which maps the positive half-line onto the whole real line. A natural generalization of model (3.3) is

$$g\{S_Z(t)\} = H(t) + \boldsymbol{Z}_i'\boldsymbol{\beta}, \tag{3.4}$$

where $g(\cdot)$ is a known decreasing function. Now it is clear to see that (3.4) is equivalent to the linear transformation model [2] $H(t) = -\boldsymbol{Z}_i'\boldsymbol{\beta} + \varepsilon$ where ε follows extreme value distribution $F = 1 - g^{-1} = 1 - \exp\{-\exp(s)\}$.

[2]Notice that when re-parameterizing the proportional hazard model in the linear transformation model form, and compare this form with Equation (3.2), they are not exactly equal, but there is a negative sign in front of regression parameter $\boldsymbol{\beta}$ in the re-parameterized form. Therefore, one should be careful when interpreting the regression parameters, because the sign of the parameter depends on the specific model parameterization used in estimation.

3.1.1 Hierarchical Linear Transformation IRT Model

In this paper, we adopt van der Linden's (2007) hierarchical framework while replacing the lognormal model with the linear transformation model for response times. In particular, the model we propose is as follows.

First-Level Model. At the first level, two models for the responses and RTs are specified separately. For the item response model, any appropriate parametric model may be used, but we focus on the three-parameter logistic model:

$$P_j(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \tag{3.5}$$

with a_j, b_j, and c_j representing item discrimination, difficulty and guessing parameters. For the RTs, a linear transformation model is adopted as

$$H_j(t_i) = \beta_j \tau_i + \varepsilon_{ij} \tag{3.6}$$

where $\tau_i \in \mathcal{R}$ is the speed parameter for test taker i. The function $H_j(\cdot)$ represents the monotone transformation for item j, and this item-level transformation implies that different types of the RT transformations will be possible for different items. β_j is a discrimination-like parameter. Negative β_j means examinees with higher speed will tend to have shorter RTs. The residuals ε_{ij} are i.i.d. with distribution F and it is independent of τ_i. Because τ_i's are latent variables, they are sometimes referred to as frailties. We assume F is known and the same across different items. The three distributions we consider in the simulations are normal, extreme value and logistic distributions.

Second-Level Model. Similar to the van der Linden (2007) model, this level captures the joint distribution of the person parameters in a population. The values of $\boldsymbol{\xi}_i = (\theta_i, \tau_i)'$ are assumed to be randomly drawn from a multivariate normal distribution, i.e.,

$$\boldsymbol{\xi}_i \sim f(\boldsymbol{\xi}_i; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \equiv \frac{|\boldsymbol{\Sigma}_p^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_i - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\mu}_p)\right], \tag{3.7}$$

with mean vector

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

and covariance matrix

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

Identifiability. To establish identifiability, we suggest the constraints $\mu_\tau = 0, \sigma_\tau^2 = 1$. The mean of τ is

fixed to fix the center of the non-parametric transformation H, and the variance of τ is fixed to remove the tradeoff between β_j and τ_i. The scale of H is determined by the fixed distribution of the error term ε_{ij}. For instance, when the error term ε_{ij} follows a normal distribution, i.e., $N(0, \sigma^2)$, we restrict it to be standard normal with $\sigma^2 = 1$. This is because the σ can be easily absorbed into the monotone transformation and the item parameters β on both sides of Equation (3.6), $\frac{h_j(t_i)}{\sigma} = \frac{\beta_j}{\sigma}\tau_i + \varepsilon_{ij}$ is equivalent to the original equation (3.6) and ε is now from standard normal distribution. The mean and variance for θ are free to estimate in this cases because we assume the item parameters (including discrimination, difficulty and guessing parameters) have already been well calibrated. This is often the case in item banks for which response times are recorded, IRT parameters have been calibrated, but no response time model has been calibrated.

Assumption. Both the response model and response time model described above rely on the assumption of local independence. The responses are locally independent given the examinee's latent ability θ; the RTs are locally independent given the examinee's latent speed τ; and the responses and response times are independent given θ and τ. The latent variables θ and τ are normally distributed, and the distribution for the error term F_ε is fixed and known. Different from van der Linden (2007) and the settings in Chapter 2, the item parameters for the 3PL model are known in our case, thus we do not estimate the correlation structure of the item parameters.

To show that this new model is a generalization of van der Linden's (2007) model, recall that in van der Linden's (2007) model, the response time, after log transformation, follows the normal distribution as

$$\log(t_{ij}) \sim \mathcal{N}(\delta_j - \tau_i, \sigma_j^{-2}). \tag{3.8}$$

Here τ_i is the speed parameter for examinee i, δ_j and α_j are the time intensity and discriminating power of item j. Similar to the difficulty and discrimination parameters in 3PL model, higher δ_j indicates that the items requires longer time to finish, and higher σ_j indicates higher power to differentiate slow examinees from fast ones. Our linear transformation model takes a very similar form. Suppose H_j is a log transformation, and to let different items have distinct transformations, we embed a scale term, i.e., $H_j(t) = \log(\lambda_j t)$, then the linear transformation model is expressed as $\log(\lambda_j t) = \beta_j \tau_i + \varepsilon_{ij}$, and we can further rewrite it as $\frac{1}{\beta_j}\log(t) = -\frac{1}{\beta_j}\log(\lambda_j) + \tau_i + \frac{\varepsilon_{ij}}{\beta_j}$, which implies, $\frac{1}{\beta_j}\log(t) \sim \mathcal{N}(-\frac{1}{\beta_j}\log(\lambda_j) + \tau_i, \beta_j^{-2})$. The item time intensity parameter, in this case, is represented by $-\frac{1}{\beta_j}\log(\lambda_j)$.

3.2 Model Estimation

Several research attempts have been made for estimating linear transformation model with or without fraitly terms. When the covaraites are completely observed, Chen, Jin and Ying (2002) proposed an estimating equation method, in which two separate estimating equations are constructed for estimating β and H respectively. Their estimator for β reduces to the Cox partial likelihood estimator when the error term follows the extreme value distribution. It is easy to compute the estimator through the estimating equation that resembles the Cox partial likelihood score function; and the estimator also has nice asymptotic properties such as closed form variance and asymptotic normality. When there are latent covariates (or fraility) terms, Mallick and Walker (2003) proposed a fully Bayesian Markov chain Monte Carlo (MCMC) method that is able to estimate both the parametric regression parameter β and non-parametric transformation H within separate parallel Markov chains. Their method is even more flexible in that the distribution of the error term F_ε is also unknown and free to be estimated. Specifically, they propose to use a mixtures of incomplete beta functions for $H(\cdot)$ and model F_ε as Polya tree distributions. Dunson (2003) also employed Bayesian MCMC estimation method with nicely specified full conditional distributions for each parameter, but his method is dependent on the fixed and known transformation H.

The goal of our investigation involves accurately estimating the parameters, i.e., $\theta_i, \tau_i,\ i = 1, 2, ...N$; $\beta_j,\ j = 1, 2, ..., J$; $\mu_\theta,\ \sigma^2_\theta, \sigma_{\theta\tau}$ as well as the non-parametric transformations H_j. Especially we wish for our estimation technique to allow for data obtained by computerized adaptive testing, in which every test taker is given different items, based on his or her adaptively estimated θ level. So the random sampling of θ (or τ) from a common distribution can not be assumed. Consequently, the usual marginal likelihood approaches used in latent variable modeling are no longer appropriate. Also because we have latent frailty term and unknown transformation $H(\cdot)$, Chen et al.'s (2002) and Dunson's (2003) methods do not lend themselves directly in our model estimation. Mallick's and Walker's (2003) method seems promising, but our preliminary investigation showed that using mixtures of incomplete beta functions for $H(\cdot)$ involves two tuning part—the number of mixands and the size of the parameters in the beta functions—which need to be adjusted with every single data set to research accurate estimations. In addition, mixture beta functions is not always enough to approximate various shapes of the non-parametric transformation. In this manuscript, we propose a two-stage estimation method. In the first stage, we will focus on the parametric part of the model, and only rely on the ranks of the observations instead of the observations per se so as to avoid the complications introduced by $H(\cdot)$. Specifically, we propose to use the "rank-based likelihood" coupled with the MCMC method for parameter estimation. In the second stage, we will use the estimating equation method proposed in Chen et al. (2002) for estimating $H(\cdot)$ while treating the parametric part of the model

as known.

3.2.1 Rank-based Marginal Likelihood for β

In model (3.6), for fixed β_j, a maximal invariant for τ_i under the group of monotone transformation is the vector

$$\tilde{\tau} \equiv (\tau_{(1)}, \cdots, \tau_{(N)}),$$

where $t_{(1)} < \cdots < t_{(N)}$ are the ordered t_i and $\tau_{(i)}$ is the corresponding covariate, so that $(\tau_{(i)}, t_{(i)})$ $i = 1, ..., N$ is a permutation of (τ_i, t_i). Knowing $\tilde{\tau}$ is equivalent to knowing the ranks of the t_i and $\tau_i, \cdots, \tau_N$. Therefore, it is reasonable to use the marginal likelihood of $\tilde{\tau}$, which does not depend on H_j, to make inference about β_j without any loss of information (Bickel & Ritov, 1998).

Denote the density of ε_{ij} in model (3.6) by $f_\varepsilon(\cdot)$. Consider the group G of increasing differentiable transformations acting on t. If $H \in G$, the density function of t is

$$f(t|\tau) = f_\varepsilon(H(t) - \beta\tau)H'(t). \tag{3.9}$$

Because the inference is based only on the ranks of the t_i's, the general location of the t_i's cannot be estimated, and a constant term in the linear model is not needed. To fix the scale, it is assumed that $\sum_i^n \tau_i = 0$, which resonates with the model assumption described earlier. By the definition of Barnard (1962), the rank vector is marginally sufficient for β and inferences on β can be based on the marginal likelihood generated by the probability function of rank statistics $\boldsymbol{r} = r_1, ..., r_N$ as

$$L(\boldsymbol{r}|\beta) = p(t_{\alpha_1} < t_{\alpha_2} < ... < t_{\alpha_N}|\beta) = \int \prod_1^n f_\varepsilon(t_{\alpha_i} - \beta\tau_i) dt_{\alpha_i} ... dt_{\alpha_N}, \tag{3.10}$$

where α_i is the anti-rank of t_j, i.e., $\alpha_i = j$ if and only if t_j is the ith smallest of $t_1, ..., t_n$ (Cuzick, 1988; Pettitt, 1982). When the covariates τ_i's are observed and fixed, the estimation $\hat{\beta}$ is obtained by solving the estimating equations

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^n \tau_i E_R \Big\{ - \frac{f'_\varepsilon(t_i - \beta\tau_i)}{f_\varepsilon(t_i - \beta\tau_i)} \Big\}, \tag{3.11}$$

where $E_R\{h(t_i)\}$ is the conditional expectation given the ranks R and that the regression parameter equals β_j. Clayton and Cuzick (1985) have proposed a solution to Equation (3.11), and Cuzick (1988) further proposed another solution which has the nice form variance estimator and asymptotic properties. Notice that this estimating equation method depends on the observed covariates, and it is not readily applied in our model estimation. However, we can take one step back and adopt the marginal likelihood based on ranks in

Equation (3.10).

For most choices of the density $f_\varepsilon(\cdot)$, the integral in (3.10) is not tractable, but the extreme value density that yields Cox PH model is an exception. When $f_\varepsilon = \exp(x - e^x)$, the integral in (3.10) becomes exactly the partial likelihood of β as long as the covariates are time-invariant (Kalbfleish & Prentice, 1973; Kalbfleisch, 1978). For other densities of $f_\varepsilon(\cdot)$, the integral in (3.10) can only be obtained through approximation. Let $\zeta = \beta\tau$ for notational simplicity. Assuming $f_\varepsilon(\cdot)$ satisfies regularity condition such that the asymptotic theory for maximum likelihood estimation of ζ holds when a random sample is taken from a distribution with density $f_\varepsilon(t-\zeta)$ and $-\infty < \zeta < \infty$ (Cox, 1974), we can expand $\log f_\varepsilon(t-\zeta)$ via Taylor series as

$$\log f_\varepsilon(t-\zeta) \simeq \log f_\varepsilon(t) + \zeta g(t) - \frac{\zeta^2}{2} g'(t) \tag{3.12}$$

or

$$f_\varepsilon(t-\zeta) \simeq f_\varepsilon(t) \exp\left[\zeta g(t) - \frac{\zeta^2}{2} g'(t)\right],$$

with $g(t) = -f'_\varepsilon(t)/f_\varepsilon(t)$. Substituting $f_\varepsilon(t-\zeta)$ with the above expansion in Eq.(3.10), we have

$$\begin{aligned} L(\boldsymbol{r}|\beta) &\simeq \int \exp\left\{\sum \zeta_i g(t_{\alpha_i}) - \frac{1}{2}\zeta_i^2 g'(t_{\alpha_i})\right\} \prod_{i=1}^{N} f(t_{\alpha_i}) dt_{\alpha_i} ... dt_{\alpha_N} && (3.13) \\ &= (N!)^{-1} E\left[\exp\left\{\sum \zeta_i g(t_{\alpha_i}) - \frac{1}{2}\zeta_i^2 g'(t_{\alpha_i})\right\}\right], && (3.14) \end{aligned}$$

and the key is to obtain approximation for Eq.(3.14). Pettitt (1982) provided a detailed derivation for approximating Eq.(3.14) by

$$f(\boldsymbol{r}|\beta) \simeq (N!)^{-1} \exp\{-\frac{1}{2}\beta Z' \boldsymbol{C}\boldsymbol{\tau}\beta + \beta\boldsymbol{\tau}'\boldsymbol{a}\}, \tag{3.15}$$

where the matrix $\boldsymbol{C}$ and vector $\boldsymbol{a}$ has explicit analytical form for some specific densities $f_\varepsilon(\cdot)$, such as normal distribution, logistic distribution and double exponential distribution (Pettitt, 1982). The quality of the approximation worsens as the density $f_\varepsilon(\cdot)$ departs from normality (Pettitt, 1983). Below we will introduce the specific form of the approximated marginal likelihood for the case of the normal and logistic distributions of ε_{ij} because we show their performance in the simulations.

When $\varepsilon_{ij} \sim \mathcal{N}(0,1)$, let $Z_{\alpha_1} < Z_{\alpha_2} < ... < Z_{\alpha_N}$ be the order statistics of a sample of size N from the standard normal distribution. Then $\boldsymbol{a} = E(\boldsymbol{Z})$ with $\boldsymbol{Z} = (Z_{\alpha_1}, \cdots, Z_{\alpha_N})$, and $\boldsymbol{C} = \boldsymbol{I}_{N\times N} - \boldsymbol{A}$ where $\boldsymbol{A} = \text{var}(\boldsymbol{Z})$. The derivation details are given in Pettitt (1982). Thus, if ξ_k is the mean of the k^{th} order statistic in a random sample of standard normal distribution with size N, and r_i is the rank of the response

time from the ith examinee, then $a_i = \xi_{r_i}$. For the similar token, $\boldsymbol{A}$ is the variance-covariance matrix for the normal order statistics $Z_{\alpha_1}, ..., Z_{\alpha_N}$ from the observed response time. So if ξ_{ik} is the (i,k)th element in the variance covariance matrix for the standard normal order statistics, then $A_{ik} = \xi_{r_i r_k}$ where r_i and r_k are the ranks for the ith and kth observation. Because the response times are continuous, we assume that the ranks are uniquely assigned to each observation without ties. To calculate the mean and covariance matrix of standard normal order statistics, i.e., the $(\xi)_i$ and $(\xi)_{ij}$, one can refer to the tabulation given in Pearson and Hartley (1972). Some numerical algorithms are also proposed to approximate those statistics (see David, 1970), and they are implemented in subroutines for various programming languages (such as MATLAB, FORTRAN) by NAG (Numerical Algorithm Group).

When the error term follows logistic distribution, i.e., $f(y) = e^{-\varepsilon_{ij}}/(1+e^{-\varepsilon_{ij}})^2$ that results in proportional odds model, the approximation to (3.10) is given below. The general form of the approximation in (3.15) stays the same, but

$$\begin{aligned} a_i &= 2r_i/(N+1) - 1 \\ (\boldsymbol{A})_{ik} &= 4r_k(N+1-r_i)/[(N+1)^2(N+2)], r_k \leq r_i \\ (\boldsymbol{B})_{ii} &= 0.5(N+1)\boldsymbol{A}_{jj} \end{aligned}$$

and $\boldsymbol{C} = \boldsymbol{B} - \boldsymbol{A}$ where $\boldsymbol{B}$ is a diagonal matrix (Pettitt, 1982). The marginal rank-based likelihood will be used for both β and τ estimation in the Metropolis-Hastings algorithm.

3.2.2 Estimating Equation Method for $\hat{H}(t)$

We use the estimating function developed by Chen et.al.(2002) for estimating the non-parametric monotone transformation $H(t)$. Let $\lambda(\cdot)$, $\Lambda(\cdot)$ be the known hazard and cumulative hazard functions of ε respectively. Let $Y(t) = I(T \geq t), N(t) = I(T \leq t)$ and let $\{Y_i(t), N_i(t)\}$ be the corresponding samples of $\{Y(t), N(t)\}$. For a single item, suppose there are N examinees answering that item, then the estimating equation for $H(t)$ is

$$\sum_{i=1}^{N}[dN_i(t) - Y_i(t)d\Lambda\{\hat{\beta}\tau_i + H(t)\}] = 0(t \geq 0), \tag{3.16}$$

assuming $\hat{\tau}_i$ $(i = 1, ..., N)$ are the estimation in the first step. The solution $\hat{H}$ to (3.16) is the estimation of the unknown monotonic transformation H. Chen et al.(2002) further proposed a numerical algorithm for

obtaining the solution. In particular, the first point estimate $\hat{H}(t_1)$ is obtained by solving

$$\sum_{i=1}^{N} Y_i(t_1)\Lambda\{\hat{\beta}\hat{\tau}_i + H(t_1)\} = 1, \tag{3.17}$$

and the others recursively by

$$\hat{H}(t_k) = \hat{H}(t_{k-1}) + \frac{1}{\sum_{i=1}^{N} Y_i(t_k)\lambda\{\hat{\beta}\hat{\tau}_i + H(t_{k-1})\}}. \tag{3.18}$$

In computation, when the error term follows logistic distribution, then $\lambda(x) = -\frac{d\ln[1-F(x)]}{dx} = (1+e^{-x})^{-1}$ and $\Lambda(x) = -\ln[1-(1+e^{-x})^{-1}]$. For the first time point t_1, $Y_i(t_1) = 1$ for all examinees $i = 1, 2, ..., N$ by definition. The Equation (3.17) is solved by plugging the corresponding term, and the MATLAB function fzero is used to solve the non-linear equation. As to the normal distribution of the error term, $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$ where $\phi(x)$ is the density for standard normal distribution and $\Phi(x)$ is the c.d.f; and $\Lambda(x) = -\ln(1-\Phi(x))$.

A familiar special case occurs when the error term follows the extreme value distribution. In this case, because the linear transformation model is equivalent to the Cox's PH model, we can also use the Breslow estimator (Breslow, 1972) to approximate the non-parametric transformation. Breslow estimator targets at estimating the baseline cumulative hazard ($\int_0^t h_0(s)ds$) in the PH model, and according to the one-to-one connection between the transformation ($H(t)$) and the baseline hazard ($H_0(t) = \log(\int_0^t h_0(s)ds)$), we can estimate $H(t)$ directly from Breslow estimator, and the results will be comparable to the estimation calculated from (3.18).

3.2.3 Parameter Estimation

Prior Specification

A bivariate normal prior is chosen for the latent parameters (θ, τ), i.e., $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\mu}_p = (\mu_\theta, 0)$ and $\boldsymbol{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & 1 \end{pmatrix}$. A normal prior is chosen for each regression parameter β_j with means equal to 0 and variance chosen to be 10. Here we purposely selected a large variance to make the prior less informative. The correlation term $\rho_{\theta\tau}$ is also chosen to have a vague normal prior as in Klein Entink, Fox, and van der Linden (2009). But we restrict the prior to be within the range of $[-1, 1]$, i.e., we employed a truncated normal prior $\rho_{\theta\tau} \sim \mathcal{N}_{[-1,1]}(0, 10)$. This treatment of restricting the prior region of the covariance (in our case the correlation) parameter to the area that supports a positive definite covariance is discussed in Mulder and Fox (2011). For the variance term, we impose a inverse-gamma prior, that is $\sigma_\theta^{-2} \sim \Gamma(g1, g2)$. The Gamma

prior was chosen in Klein Entink et.al. (2009) because it is a conjugate prior for normal distribution with known mean, although the inverse-gamma is no longer a conjugate prior in our case with a logistic response model, we still adopt this prior.

Markov chain Monte Carlo

As mentioned before, we assume a setting in which item parameters are previously calibrated and are taken as known. This will be the case in item banks for which response times are recorded, IRT parameters have been calibrated, but no response time model has been calibrated. However, the estimation method introduced above can still be used when the item 3PL parameters are unknown. If that happens, one just has to add three additional chains for estimating the a, b, and c-parameters (Patz and Junker, 1999) separately. The reason we assume them as know is because we want to emphasize the estimation of the linear transformation model parameters, which are the focus and innovation of the current manuscript. During the estimation, we need to sequentially draw parameters $\sigma_\theta^2, \sigma_{\theta\tau}$(or $\rho_{\theta\tau}$), θ, τ and β. The details of the Metropolis-Hastings algorithm within Gibbs sampler is presented below.

Suppose the items are indexed by $j = 1, \cdots, J$, and the examinees by $i = 1, \cdots, N$. For the ith test taker, his or her responses and response times are denoted by $\boldsymbol{Y_i} = (Y_{1i}, \cdots, Y_{Ji})'$, and $\boldsymbol{T_i} = (T_{1i}, \cdots, T_{Ji})'$, respectively. To perform the sampling for parameters with support on the entire real line, we use normal proposal distributions with mean equal to the current estimation and variance chosen to give a Metropolis acceptance rate of between 25 and 50 percent. For parameters with support not on the real line, we first transform them to the real line and then sample them from normal proposal distribution.

Step 1: Denote the initial values for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\tau}$ by $\hat{\boldsymbol{\beta}}_0 \equiv (\hat{\beta}_{01}, ..., \hat{\beta}_{0J})'$, $\hat{\boldsymbol{\theta}}_0 \equiv (\hat{\theta}_{01}, ..., \hat{\theta}_{0N})'$ and $\hat{\boldsymbol{\tau}}_0 \equiv (\hat{\tau}_{01}, ..., \hat{\tau}_{0N})'$, respectively. $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator (MLE) by maximizing the likelihood function formed by equation (3.5). The initial value $\sigma_\theta^{2(0)}$ is calculated by the sample variance of $\hat{\boldsymbol{\theta}}_0$. The initial value of $\hat{\boldsymbol{\tau}}$ is obtained differently depending upon the specific test conditions. The details are given in Wang, et.al. (under review). Conditioning on $\hat{\boldsymbol{\tau}}_0$, $\hat{\boldsymbol{\beta}}_0$ is obtained by maximizing the approximation of the rank-based likelihood. The initial value of $\mu_\theta^{(0)}$ is the sample mean of the θ's. Set the iteration counter iter=1.

Step 2: At rth step, denote the previous positions $\boldsymbol{\beta}^{(r-1)} \equiv (\beta_1^{(r-1)}, ..., \beta_J^{(r-1)})'$, $\boldsymbol{\theta}^{(r-1)} \equiv (\theta_1^{(r-1)}, ..., \theta_N^{(r-1)})'$, $\boldsymbol{\tau}^{(r-1)} \equiv (\tau_1^{(r-1)}, ..., \tau_N^{(r-1)})'$, $\sigma_{\theta\tau}^{(r-1)}$ and $\sigma_\theta^{2(r-1)}$, $\mu_\theta^{(r-1)}$. Sample each parameter sequentially as follows.

1. σ_θ^2: Sample the variance for $\hat{\boldsymbol{\theta}}$: Because the variance is always non-negative, we draw $\log(\sigma_\theta^{2*})$ from

$\mathcal{N}(\log(\sigma_\theta^{2(r-1)}), 1)$ with acceptance probability as

$$\alpha(\sigma_\theta^{2(r-1)}, \sigma_\theta^{2*}) \equiv \min\left\{1, \frac{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2*}, \sigma_{\theta\tau}^{(r-1)}, \mu_\theta^{(r-1)}) p_{\sigma^2}(\sigma_\theta^{2*})\sigma_\theta^{2*}}{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2(r-1)}, \sigma_{\theta\tau}^{(r-1)}, \mu_\theta^{(r-1)}) p_{\sigma^2}(\sigma_\theta^{2(r-1)})\sigma_\theta^{2(r-1)}}\right\}, \tag{3.19}$$

where

$p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2*}, \sigma_{\theta\tau}^{(r-1)}, \mu_\theta^{(r-1)}) = \prod_{i=1}^N p(\theta_i^{(r-1)}, \tau_i^{(r-1)}|\sigma_\theta^{2*}, \sigma_{\theta\tau}^{(r-1)}, \mu_\theta^{(r-1)}) \sim$
$\prod_{i=1}^N \frac{|\boldsymbol{\Sigma}_2^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\xi}_i - \boldsymbol{\mu})\right]$ with $\boldsymbol{\mu} = (\mu_\theta^{(r-1)}, 0)$, $\boldsymbol{\xi}_i = (\theta_i^{(r-1)}, \tau_i^{(r-1)})$ and variance-covariance matrix $\boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_\theta^{2*} & \sigma_{\theta\tau}^{(r-1)} \\ \sigma_{\theta\tau}^{(r-1)} & 1 \end{pmatrix}$. $p_{\sigma^2}(\cdot)$ is the inverse-Gamma prior for the variance term.

2. $\rho_{\theta\tau}$: Sample correlation between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$: Because $-1 \leq \rho_{\theta\tau} \leq 1$, we need to first transform $\rho_{\theta\tau}$ to the real line, the transformation we adopt is $\rho_{\theta\tau} = -1 + 2\frac{e^\varphi}{1+e^\varphi}$. Then draw φ^* from $\mathcal{N}(\varphi^{r-1}, 1)$, and the acceptance probability of the corresponding $\rho_{\theta\tau}^*$ is

$$\alpha(\rho_{\theta\tau}^{r-1}, \rho_{\theta\tau}^*) \equiv \min\left\{1, \frac{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2r}, \varphi^*, \mu_\theta^{(r-1)})\pi_\rho(\rho_{\theta\tau}^*)J(\varphi^*, \rho_{\theta\tau}^*)}{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2r}, \varphi^{(r-1)}, \mu_\theta^{(r-1)})\pi_\rho(\rho_{\theta\tau}^{(r-1)})J(\varphi^{(r-1)}, \rho_{\theta\tau}^{(r-1)})}\right\} \tag{3.20}$$

where $\pi_\rho(\rho_{\theta\tau}^*)$ is the truncated normal prior density of the correlation term with $\rho_{\theta\tau}^*$ being the one-on-one transformation of φ^*. The Jacobian matrix $J(\varphi^*, \rho_{\theta\tau}^*) = \frac{2\exp(\varphi^*)}{(1+\exp(\varphi^*))^2}$ is involved due to the transformation of $\rho_{\theta\tau}^*$, and one will notice that it is the same as the transition matrix as if drawing $\rho_{\theta\tau}^*$ from a non-symmetric distribution (instead of drawing φ^* from a symmetric normal distribution).

3. μ_θ: Sample the population mean of θ. Draw μ_θ^* from $\mathcal{N}(\mu_\theta^{r-1}, 1)$ with acceptance probability

$$\alpha(\mu_\theta^{r-1}, \mu_\theta^*) \equiv \min\left\{1, \frac{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2(r)}, \sigma_{\theta\tau}^{(r)}, \mu_\theta^*)\pi(\mu_\theta^*)}{p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2(r-1)}, \sigma_{\theta\tau}^{(r)}, \mu_\theta^{(r-1)})\pi(\mu_\theta^{(r)})}\right\} \tag{3.21}$$

$p(\boldsymbol{\theta}, \boldsymbol{\tau}|\sigma_\theta^{2(r)}, \sigma_{\theta\tau}^{(r)}, \mu_\theta^*)$ is again a product of multivariate normal densities, and $\pi(\mu_\theta^*)$ is the normal prior density.

4. θ and τ: Sample examinees' ability and speed parameter: For the i^{th} pair $(\theta_i, \tau_i), 1 \leq i \leq N$, draw (θ_i^*, τ_i^*) from a bivariate normal distribution with mean $(\theta_i^{(r-1)}, \tau_i^{(r-1)})$ and variance-covariance $\boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_\theta^{2(r)} & \sigma_{\theta\tau}^{(r)} \\ \sigma_{\theta\tau}^{(r)} & 1 \end{pmatrix}$. The acceptance probability for is

$$\alpha(\theta_n^{(r-1)}, \tau_n^{(r-1)}, \theta_i^*, \tau_i^*) \equiv \min\left\{1, \frac{\text{IRT}(\theta_i^*)L(\boldsymbol{\beta^{(r-1)}}|\boldsymbol{\tau}^*)\pi(\theta_i^*, \tau_i^*)}{\text{IRT}(\theta_i^{(r-1)})L(\boldsymbol{\beta^{(r-1)}}|\boldsymbol{\tau}^{r-1})\pi(\theta_i^{(r-1)}, \tau_i^{(r-1)})}\right\} \tag{3.22}$$

where $\boldsymbol{\tau}^* = (\tau_1^{(r-1)}, ..., \tau_i^*, ..., \tau_N^{(r-1)})'$. $\pi(\theta_i^r, \tau_i^{(r-1)})$ is a bivariate normal with mean $(\mu_\theta^{(r)}, 0)$ and variance-covariance $\boldsymbol{\Sigma}_2$. IRT$(\cdot)$ is calculated from equation (3.5) and $L(\cdot)$ is defined by (3.15), respectively.

5. β: Sample survival regression parameter: For jth item, draw β_j^* from a normal distribution $\mathcal{N}(\beta_j^{r-1}, 1)$ with the acceptance probability defined as

$$\alpha(\beta_j^{(r-1)}, \beta_j^*) \equiv \min\left\{1, \frac{L(\beta_j^*|\boldsymbol{\tau}^r)p(\beta_j^*)}{L(\beta_j^{(r-1)}|\boldsymbol{\tau}^r)p(\beta_j^{(r-1)})}\right\} \tag{3.23}$$

where $L(\cdot)$ is defined in equation (3.15).

Step 3: Change the iteration counter from r to $r+1$ and return to step 1 until iter=M, where M is a pre-specified number.

Step 4: At the end of the chain, compute the posterior mean of each parameter. A burn-in period of the initial K iterations is often required to allow the chain to reach equilibrium. Once the parameters are well estimated, we move on to the second step of estimating the non-parametric monotone transformation.

When the data is collect from a cognitive task, it is often the case that the items (also known as "trials" in cognitive experiment) are very similar to each other within a block. In this regard, we can treat the items belonging to the same class as having identical item parameters. There are two possible ways to approach this issue. We can either view those items as single items and aggregate examinees' responses/RTs to those items, or we can treat them as conditional independent given the examinees' latent speed and ability but impose an equality constraint (i.e., we will update those same item parameters together in a single chain while pooling information from responses and RTs to all those items to construct the acceptance ratio). In either way, the method described above can be applied in a straightforward fashion.

3.3 Model Diagnosis

To check global fit, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), which is analogous to the AIC but designed specifically for Bayesian hierarchical models, may be used. Lower DIC usually indicates better fit. The DIC is equal to a deviance plus a penalty term for model complexity. The deviance is calculated as

$$\begin{aligned} D(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\beta}, H) &= -2\log p(\boldsymbol{t}|\boldsymbol{\tau}, \boldsymbol{\beta}, H)p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \\ &= -2\sum_{i=1}^{N_j}\sum_{j=1}^{J}\Big[\log f_\varepsilon(H(t_{ij}) - \beta_j\tau_i) + y_{ij}\log P_j(\theta_i) + (1-y_{ij})\log(1-P_j(\theta_i))\Big], \end{aligned}$$

where f_ε is the density of the error term distribution and $P_j(\theta_i)$ is defined in equation (3.5). The DIC for the joint model is obtained via

$$DIC = \bar{D} + (\bar{D} - \hat{D}),$$

where $\bar{D} \approx \frac{1}{M}\sum_{m=1}^{M} D(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\tau}^{(m)}, \boldsymbol{\beta}^{(m)}, H^{(m)})$, and M is the number of iterations of the algorithm; and $\hat{D} = E(D) = D(\boldsymbol{t}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\beta}}, \hat{H})$.

As an additional measure of global fit, we also use Kullback-Leibler (KL) distance as a global fit index. KL distance measures the divergence between two probability density functions. In the current setting, for item j, we obtain a set of estimated error terms $\varepsilon_{ij} = \hat{H}_j(t_i) - \hat{\beta}_j\hat{\tau}_i, i = 1, ..., N$ for $i = 1, .., N$, from which we can estimate its density by kernel smoothing as

$$\hat{f}_j(x, h) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - \varepsilon_{ij}}{h}\right).$$

According to the model, the true errors ε_{ij}'s will follow some theoretical distribution $f_{0j}(\varepsilon_{ij})$. Therefore, we can calculate the KL distance between the empirical and theoretical distribution of ε_{ij} as

$$KL_j = \int \hat{f}_j(x) \log \frac{\hat{f}_j(x)}{f_{0j}(x)} dx.$$

This KL distance is averaged over all items to obtain an averaged KL distance that quantifies the overall fit of the model. Smaller KL distances indicate better fit. Considering the MCMC model estimation method, the averaged KL distance can be calculated at each point of the chain such that we can obtain the whole distribution of the averaged KL distance. At an item level, we can check the fit graphically by comparing the empirical and theoretical distributions of the error terms.

3.4 Simulation Study

Simulation studies were carried out to check the performance of the proposed MCMC estimation method as well as the recursive method for estimating the non-parametric transformation. As a starting point, we only consider the non-adaptive situation, in which each examinee takes the same set of items. Test length was set to be 20, and examinee sample size was set to be 200. We chose such small values to demonstrate that even with short test lengths and small sample sizes, the estimation can still be accurate. This situation is also seen in adaptive designs, in which each item is measured by a certain group of examinees rather than the whole sample, and thus the examinee sample size will not be very large. A total of $3 \times 4 = 12$ different

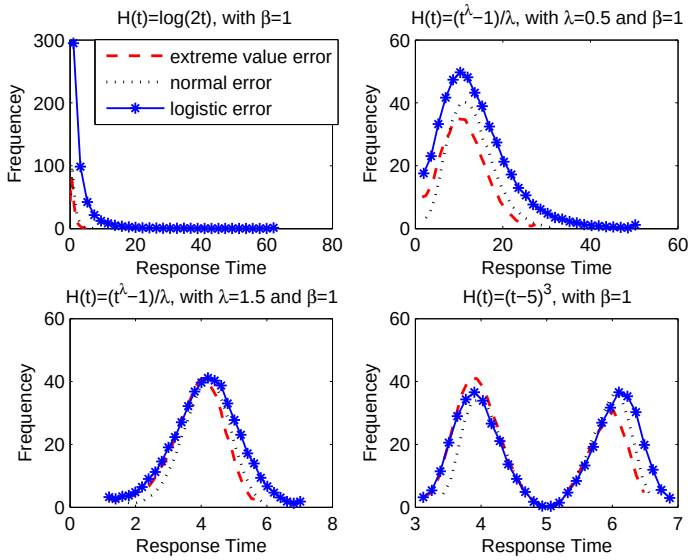


Figure 3.1: The possible RT distributions from different combinations of error term distribution and transformations

conditions were simulated. The first factor represents the distribution of the error term, it is either extreme value, normal, or logistic. The second factor indicates different monotone transformations. The first one was $H_j(t) = \log(\lambda_j t)$, with $\lambda \sim U(0.25, 1.5)$ as an item level parameter. The second one was a Box-Cox transformation, $H_j(t) = (t_j^{\lambda} - 1)/\lambda_j - 5$ with $\lambda \sim U(0.5, 1.5)$. We need to make sure $H(t)$ was on the real line, and thus we subtracted 5 from the transformation. However, other arbitrary value could be chosen. When $\lambda < 1$, it is a monotone concave function, and when $\lambda > 1$, it is a monotone convex function. The third transformation has a inflection point in the middle, i.e., $H_j(t) = \lambda_j(t-5)^3$ with $\lambda \sim U(0.5, 2)$. Again, 5 is chosen because it is the median of the response times, but other arbitrary values can also been chosen. The fourth one was a mixture of the above three transformations, with 7, 7,and 6 items belonging to log, Box-Cox, and inflection transformations. These transformations and corresponding parameters were chosen to produce realistic response time distributions. An illustration of the possible RT distributions are given in Fig 3.4, with each curve representing the shape of the histogram of the RT distributions. The curves were obtained by averaging over 100 replications.

As one might notice, all the transformations had supports on the positive real line, and the log transformation produced very skewed distributions. This type of RT distribution is seen when the items are very easy so that most examinees answer the items within a very short time, and only examinees with extremely low abilities (or low speed) tend to take a long time to finish. The Box-Cox transformation yielded either skewed or near symmetric RT distributions depending upon the value of the power transformation parame-

ter. The transformation with an inflection point yielded a bimodal distribution, and this distribution often indicates that examinees engage in two different strategies when answering the items. The examinees' latent traits (θ, τ) were generated from a bivariate normal distribution with mean $(0, 0)$ and covariance matrix $(1, 0.5; 0.5, 1)$.

Table 3.1: Bias of the estimates

		θ (IRT)	θ (IRT+RT)	β	τ	$\sigma_{\theta\tau}$	σ^2_θ	μ_θ
Normal Error	log	-0.031	-0.030	-0.029	0.051	0.098	0.331	-0.076
	Box-Cox	0.018	0.009	0.037	0.019	0.059	0.297	0.088
	Reflection	0.035	-0.030	0.056	0.059	0.091	0.298	0.081
	Mixed	0.029	0.021	0.081	0.063	0.044	0.341	0.006
Logistic Error	log	-0.017	-0.008	0.091	-0.018	0.043	0.271	0.007
	Box-Cox	-0.056	-0.045	-0.020	0.017	0.041	0.351	-0.006
	Reflection	0.059	0.037	-0.018	0.021	0.005	0.231	0.008
	Mixed	0.037	0.029	0.036	0.051	0.042	0.281	0.018
Extreme Value Error	log	0.005	-0.007	0.049	0.019	0.041	0.259	0.019
	Box-Cox	0.039	0.031	0.041	0.042	0.051	0.301	0.018
	Reflection	0.051	-0.029	0.039	-0.047	-0.019	0.244	0.015
	Mixed	0.003	0.005	0.047	-0.019	-0.029	0.237	0.109

Bias and mean squared error (MSE) were calculated to evaluate the closeness of the estimated parameters to their true values. For population parameter μ_θ, $\sigma_{\theta\tau}$ and σ^2_θ, only bias was calculated. Table 3.1 and 3.2 present the average bias and MSE of θ for both models under 12 simulations conditions. Both average bias and MSE were obtained over all examinees and all replications within a simulation condition. To show that with RT as collateral information, the estimation of θ will be more accurate, we present MSE calculated from both initial $\hat{\theta}^{(0)}$ (MLE of θ estimated from responses only) and final estimate of $\hat{\theta}$. The recovery of the non-parametric transformation was evaluated by the standardized version of the integrated absolute

Table 3.2: MSE for the parameters and absolute difference between non-parametric transformations

		θ (IRT)	θ (IRT+RT)	β	τ	$\delta(H)$
Normal Error	log	0.141	0.111	0.012	0.043	1.761
	Box-Cox	0.161	0.126	0.008	0.052	0.812
	Inflection	0.153	0.125	0.026	0.042	0.479
	Mixed	0.161	0.123	0.022	0.045	0.638
Logistic Error	log	0.169	0.138	0.078	0.147	1.623
	Box-Cox	0.164	0.109	0.027	0.153	0.809
	Inflection	0.161	0.131	0.029	0.159	0.244
	Mixed	0.157	0.106	0.046	0.118	0.471
Extreme Value Error	log	0.169	0.118	0.041	0.059	1.701
	Box-Cox	0.168	0.121	0.031	0.062	0.712
	Inflection	0.165	0.121	0.023	0.043	0.574
	Mixed	0.157	0.116	0.021	0.047	0.581

difference between the true $H_j(t)$ and its estimate $\hat{H}(t)$ for the j^{th} item

$$\delta(H)_j = \frac{\int |H_j(t) - \hat{H}_j(t)|dt}{\sqrt{\int |H_j(t)|^2 dt}}, \tag{3.24}$$

where the integration is taken over the possible RTs for item j. The denominator is added to remove the scale differences inherent in different transformations. The mean of $\delta(H)_j$ is also reported in Table 3.2.

As shown in Table 3.2, with normal errors and extreme value errors, the examinees' speed parameter τ were very accurately recovered with extremely small MSE. However, when the error term followed the logistic distribution, the MSE of τ was much larger because the approximation to the rank based likelihood (as in Equation 3.14) is less accurate. The bias results display similar patterns. All the bias values are acceptably small except for the bias of β in the logistic error model, and this is also due to the less-than-ideal approximation. The MSE of θ decreased when the response time information was considered, and this indicates that the response times provide useful collateral information to locate examinees' true abilities. The MSE of β was uniformly small with different error distributions. The various shapes of the monotone transformations did not affect the estimation results either. The parameter $\sigma_{\theta\tau}$ and μ_θ were recovered accurately with small bias across different conditions, whereas σ_θ^2 was often estimated with large positive bias. Considering that the ability variance will not affect our interpretations about the RT information as well as its relationships with responses, the results are still acceptable. The non-parametric transformation can also be accurately recovered by displaying small standardized integrated difference. Only the log-transformation showed slightly larger differences between the true and estimated transformations, and this is because the log transformation has a long and nearly flat tail that is relatively hard to capture, especially bearing in mind that only a few examinees will have extremely long RTs (see Figure). To further show that the unknown transformation can be accurately recovered, we present the true transformation versus estimated transformation for the normal error model in Figure 3.2.

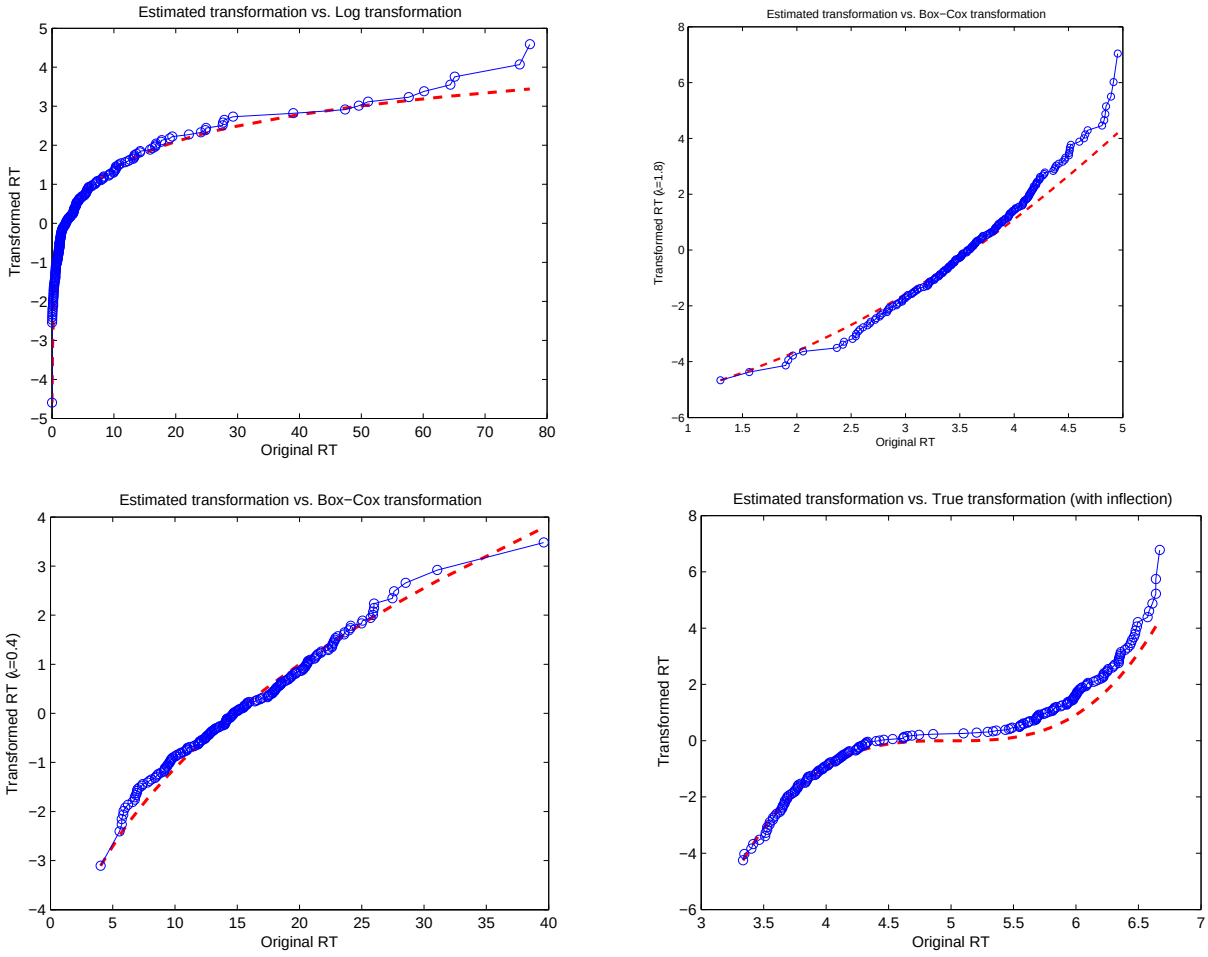


Figure 3.2: True vs. Estimated non-parametric transformation for normal error model

Chapter 4

Real Data Analysis

A dataset from a large scale high-stake computerized adaptive test is used to The dataset was comprised of $21,444$ examinees and 620 multiple choice items in total. It contained item responses, item response times, and item 3-PL parameters. The responses were scored as right or wrong. Response time was the time period from the onset of each item until examinees give the response to the item, and this is exactly the information we are interested in. The default test length was 37, but the number of items that each examinee answered ranged from 25 to 37. Because of the computerized adaptive version, each item was answered by different sets of examinees, and the number of examinees taking each item ranged from 6 to 489. We randomly sampled $3,000$ examinees from this population for analysis. However, we deleted 319 examinees because their RTs were not recorded; we further deleted 548 examinees because their total RTs were either too long (longer than 75 minutes) or because they failed to finish the whole test (i.e., test length was shorter than 37). Tests longer than 75 minutes occurred because some examinees took the test under non-standard accommodation settings. The resulting $2,061$ observations were used in the analysis. The original RTs were recorded in a millisecond scale, and for ease of calculation, we rescaled the RTs to the minute scale by dividing each RT record by $60,000$.

4.1 Marginal Distribution of Response Time

We first analyzed the marginal distribution of the response time for each item. The interest is to see (1) whether there exist a single parametric form that can explain the RT patterns of all the items; (2) whether the response time patterns change with the size of the item parameters, say, a- and b- parameters.

To investigate the marginal distribution of response time and its relationship with item parameters, we picked five items with varying level of b-parameters, and plotted their survival curve, cumulative hazard curve, and smoothed hazard rate curve in Figure 4.1. Similarly, another five items were chosen with varying level of a-parameters, and their response time patterns were displayed in Figure 4.2.

As Figure 4.1 shows, the cumulative hazard curves for all five items look similar, regardless of their item

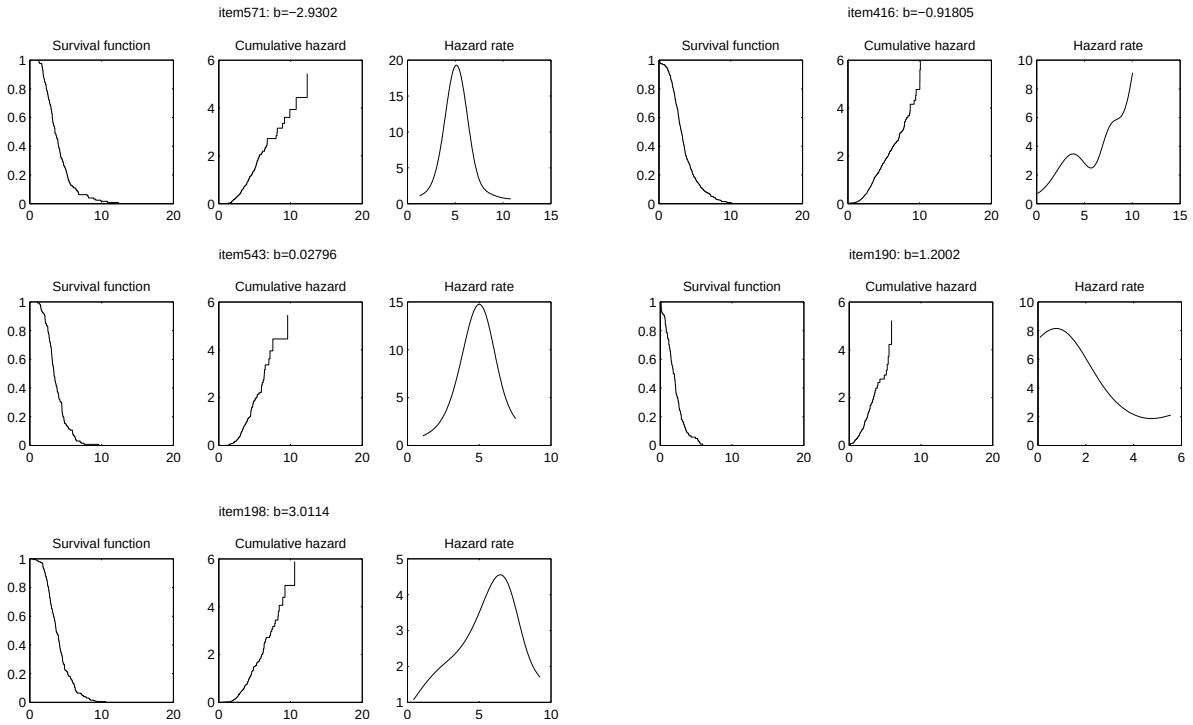


Figure 4.1: Response time distribution of items with increasing b-parameter

difficulty level. Although intuitively, harder items should take longer time to finish, and we would expect to see decreasing cumulative hazard by the increase of b-parameter. However, contrary to our expectation, the cumulative hazard curves in Figure 4.1 did not show a clear trend by b-parameter. The same argument holds for the a-parameter as well in Figure 4.2. This observation indicates that there might be no need to model the covariance structure between item time intensity and item 3PL parameters. The shapes of hazard curves, however, vary quite a bit and the variation is uncorrelated with the trend of the b-parameter or a-parameter. Due to these different shapes of the hazard rate, we need to employ a more flexible semi-parametric model. In addition, because of the adaptive feature, each item is answered by a group of examinees with limited ability levels, and this restriction of range might confound the relationship between item time intensity and difficulty. This indicates that analyzing the marginal distribution of RT alone might miss important information on the relationship between RT and underlying latent trait of interest, such as examinee's latent speed. This is the very reason that we need to use the regression models, the analysis of which will be presented in the next section.

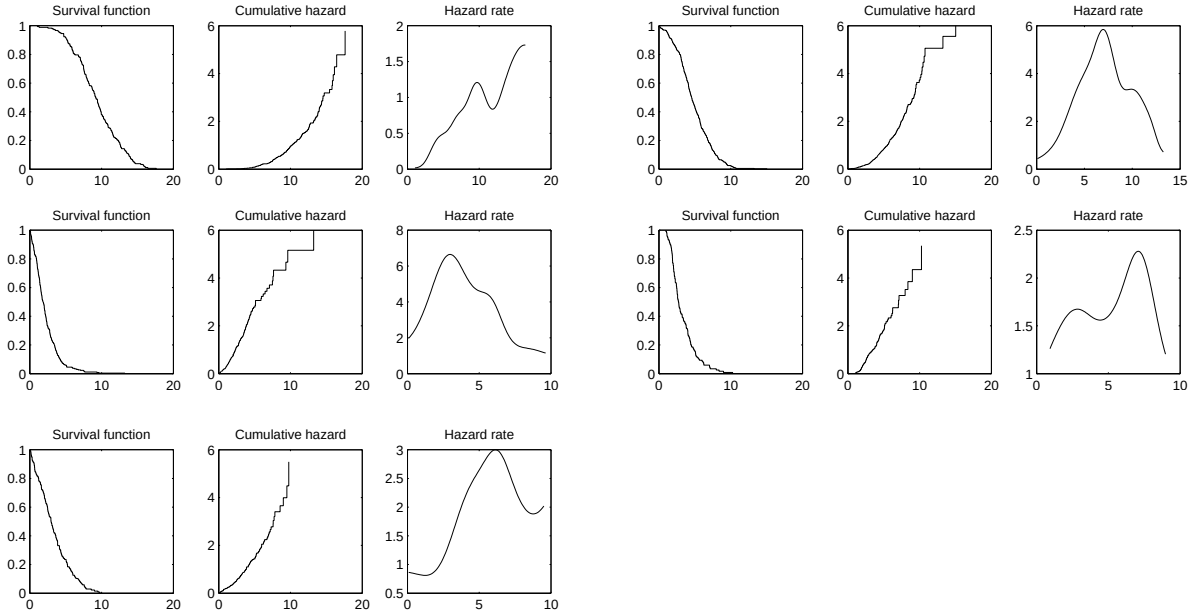


Figure 4.2: Response time distribution of items with increasing a-parameter

4.2 Model Selection

This section summarizes the fit and performance of various types of regression models, including linear transformation model with three error term distributions, and three simpler parametric models.

4.2.1 Linear Transformation Models with Different Error Distributions

Linear transformation model with three specific error term distributions were considered first, all the chains converged successfully. Summary statistics for the model parameters were given in Table 4.1.

Table 4.1: Summary statistics for estimated parameters under three different linear transformation models

		θ	τ	β	$\rho_{\theta\tau}$	σ^2_θ	μ_θ
Normal Error	Mean	0.639	-0.008	-0.005	0.548	1.766	0.646
	S.D.	1.134	0.713	0.572			
Logistic Error	Mean	0.635	0.008	0.591	0.417	1.878	0.625
	S.D.	1.149	0.788	1.964			
Extreme Value Error	Mean	0.645	0.003	-0.109	0.594	1.653	0.646
	S.D.	1.111	0.802	0.798			

The mean and standard deviation (S.D.) of examinees' ability estimates were very close across the models. The other parameters, however, differ quite a bit in terms of the mean and the S.D. The S.D. of β for the logistic error model was very large compared to the other two models because we found 9 out of 620 items had extremely large β values ($\beta > 6$). The correlation between examinees' latent ability and speed were

Table 4.2: Summary statistics for KL distances under three different linear transformation models

Model	DIC
Normal error model	-2.725×10^7
Logistic error model	-1.025×10^7
Extreme value error model	-2.745×10^7

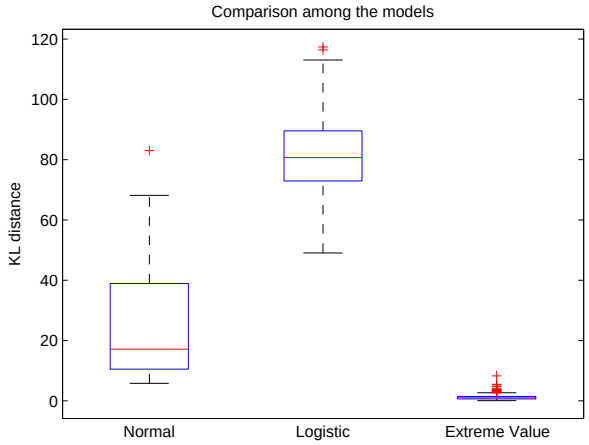


Figure 4.3: Boxplots of the KL distance measure for the three models

all larger than 0.4, and when the error terms follow extreme value distributions, the correlation is as high as 0.594. This result indicates that examinees with higher ability tend to answer the items faster, which is consistent with common sense. The DIC values for the three models were presented in table 4.2, and Figure 4.3 displayed the Boxplots of the KL distance measure for each model.

The KL distance was calculated on the last 1000 iterations, taking the first 3000 iterations as burn-in. The results showed that both DIC and KL distance measures favored the proportional hazard model, followed by the normal error model, whereas the logistic error model showed the poorest fit. Even so, if evaluating item level fit, we found that the PH model might not always be a best fit. For instance, Figure 4.4 presents one particular item that is best fitted with the logistic error model.

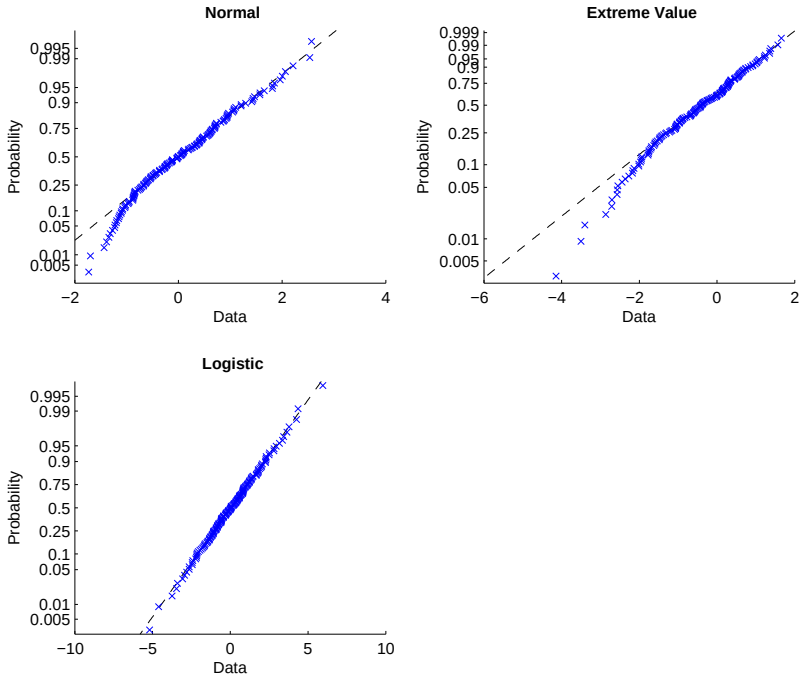


Figure 4.4: Item Fit Analysis for an item best fit by logistic model

4.2.2 Parametric vs. Semi-parametric Models

As Johnny von Neumann used to say "with four parameters I can fit an elephant and with five I can make him wiggle his trunk" (A meeting with Enrico Fermi, *Nature, 427*, 297, 2004), when we fit a model to a dataset, we need to avoid over-fitting. Therefore, in this section, we tried to explore whether the semi-parametric model is really necessary or a simpler parametric model is enough. Three parametric models considered in the study are: (1) the exponential model, with hazard function $h_{ij}(t|\tau_i) = \lambda_j \exp(\beta_j \tau_i)$; (2) the Weibull model, with hazard function $h_{ij}(t|\tau_i) = \gamma_j(\lambda_j t)^{\gamma_j - 1} \exp(\beta_j \tau_i)$; and (3) the lognormal model, with the response time density expressed as $f(t_{ij}) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\{-\frac{1}{2}[\alpha_i(\log t_{ij} - (\beta_i - \tau_j))]^2\}$ (van der Linden et al., 2009). These three models replaced the linear transformation model in the hierarchical framework. The MCMC algorithm was employed for model estimation. But instead of using partial likelihood, the traditional likelihood for response times (for the first two models for instance, the density for response time becomes $f(t) = h(t)\exp\{-\int_0^t h(s)ds\}$) was used. The parameters for the baseline hazard, such as λ and γ, were updated in separate chains in the MCMC algorithm. For the lognormal model, the complete algorithm introduced in van der Linden (2007) was used. In the lognormal model, because the item parameters α_j and β_j were interpreted as item time-intensity and time-discrimination parameter, van der Linden (2007) imposed a covariance structure on item parameters, and the covariance structure was estimated from the

real data as well.

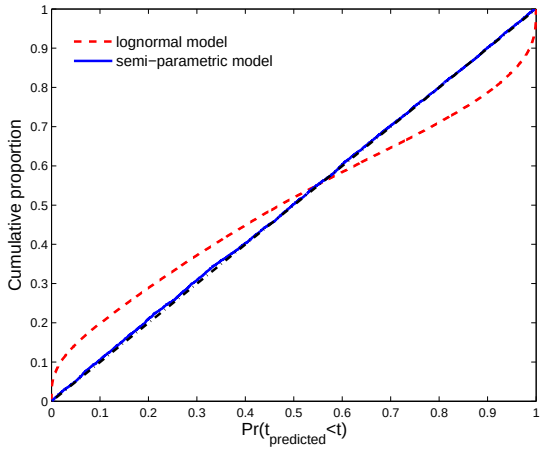
Results

Model fit was checked via the two approaches introduced in Chapter 2. We first checked the global fit of the semi-parametric model against van der Linden (2007)'s lognormal model, as shown in Figure 4.5(a). This figure shows the cumulative distribution of the predictive probabilities for the observed response times in Equation (2.17) for all person-item combinations in the data set. The data set was large enough (55,500 data points) to expect the empirical distribution to coincide with the identity line. The impression from Figure 4.5(a) is that the semi-parametric model fitted the data much better than the lognormal model. Therefore, the lognormal model was not considered in the following discussion. However, one potential useful result from the lognormal model is that the covariance matrix on item parameters had all off-diagonal elements within the range of [-0.1, 0.05]. This indicates that the item parameters were nearly independent, although this conclusion should be made with caution because of the model misfit.

Model fit was further checked by the residuals defined in Equation (2.19) on three proportional hazard models. We calculated $\varepsilon_{ij}, i = 1, ..., n_j$ for each item $j = 1, .., 620$ separately, got the kernel smooth (KS) density estimation of ε_{ij}'s and calculated the KL distance between the KS density and the extreme-value density. Figure 4.5(b) presents the Box plot of the KL distance for each item under three models. Each box represents the distribution of the KL distance for 620 items. It is apparent that the semi-parametric model generated the smallest KL distance, followed by the Weibull model. The exponential model yielded the largest KL distance because it is the most restrictive model. Notice that the semi-parametric model had far more parameters than the exponential or Weibull model, thus Figure 4.5(b) was not surprising, but in real applications, practitioners may decide between model adequacy and parsimony.

We further drew the distribution plot of ε_{ij} obtained from hierarchical Cox PH model against the extreme-value distribution. Similar to the Q-Q plot, points tightly along a line indicate a good fit. Figure 4.6 presents the fit plots for six items. The three items on the left are the ones with the best fit and the three items on the right are the ones with the worst fit. All these six items were answered by more than 150 examinees, and had reliable parameter estimates. As demonstrated by Figure 4.6, some of the items were fitted quite well, but some were not. This might be because the current model assumes that the hazards are proportional, and this assumption might not hold for some items. This suggests the need for a more general model that relaxes such an assumption, for example, the linear transformation model (Cuzick, 1988) with the error distributions completely unspecified.

(a) Global fit of the semi-parametric model vs. lognormal model



(b) Box-plot of the KL distance under three models

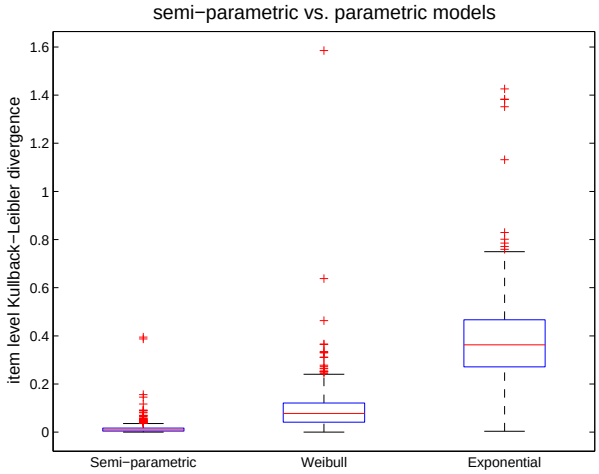


Figure 4.5: Diagnostic plots for four different models

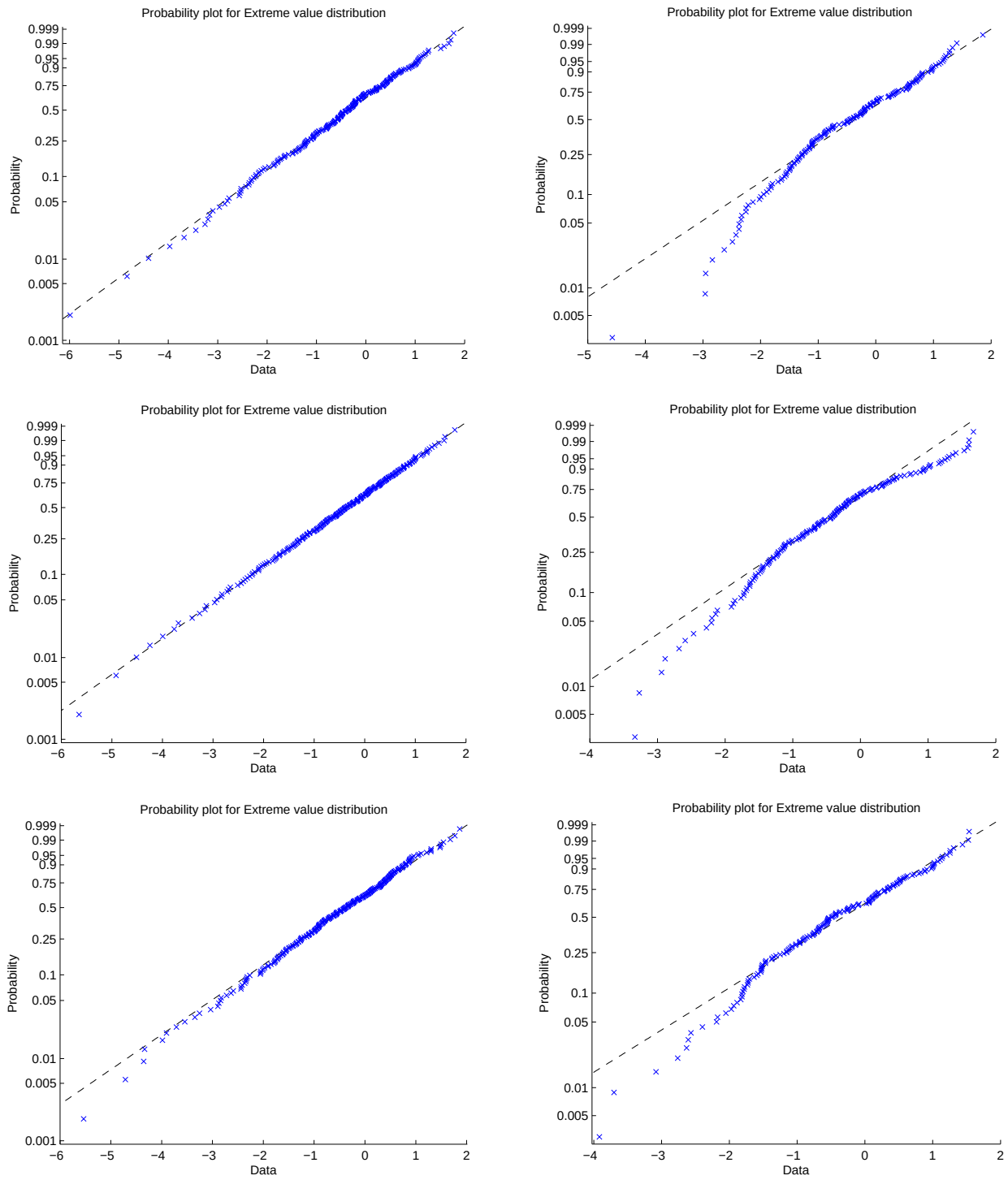


Figure 4.6: Bayesian latent residual diagnosis for six randomly selected items

4.3 Parameter Estimation

Because the model fit checking indicated that the hierarchical proportional hazard model fitted the data relatively best, in this subsection, we will only discuss the model calibration result from this model. Figures 4.7 and 4.8 display the traceplot and autocorrelation plot for the item/examinee parameters. As demonstrated by these figures, the chains reached equilibrium successfully, and the autocorrelation drops within the acceptable ranges after couple of lags.

To show how examinees' RT changes with the latent speed, we presented histograms[1] of $\hat{\beta}$, as shown in Figure 4.9.

The majority of items had βs falling in the interval [0,1]. One third of the items had negative $\hat{\beta}$'s, which indicates that for those items, more capable examinees tended to have longer RTs. There are at least two possible explanations for this phenomenon. First, these items are only exposed to a certain group of examinees with a small range of abilities or speed parameters rather than a representative group. For instance, if a relatively difficult item is given to a representative group, and if all examinees use a similar strategy to solve the item (i.e., none of them randomly guesses), then $\hat{\beta}$ will most likely be positive; however, if only high-ability examinees answer the item, within the restricted sample, the $\hat{\beta}$ estimate might be negative or near zero. To further verify this possibility, we explored two items that had negative $\hat{\beta}$, against two items with positive $\hat{\beta}$ in Figure 4.10. The items with negative $\hat{\beta}$ tended to have less skewed RT distributions, much higher cumulative hazards, and were given to examinees with high abilities with narrower ability ranges. Within such restricted groups, examinees with high speed in general might happen to answer those items slightly slower, possibly because they employed different solution strategies. In fact, we also computed the variance of the θ's for each item, and the items with negative $\hat{\beta}$ had smaller variances than those items with positive $\hat{\beta}$, further indicating that items with negative $\hat{\beta}$ were given to a more restricted sample. Second, the items with negative $\hat{\beta}$ typically were answered by a smaller number of examinees, and thus they had larger posterior variance for $\hat{\beta}$, roughly 0.15 whereas the posterior variance for the other itemsare only 0.02. Also, the model did not fit these items as well as the items with positive $\hat{\beta}$ because they tended to have larger KL distances for residuals (mean is around 0.023) than the other items (with mean around 0.013).

[1] The β's were obtained from Cox PH model parameterization, instead of from linear transformation model parameterization, so positive $\hat{\beta}$ indicates that examinees with higher latent speed tend to answer that item with shorter response time.

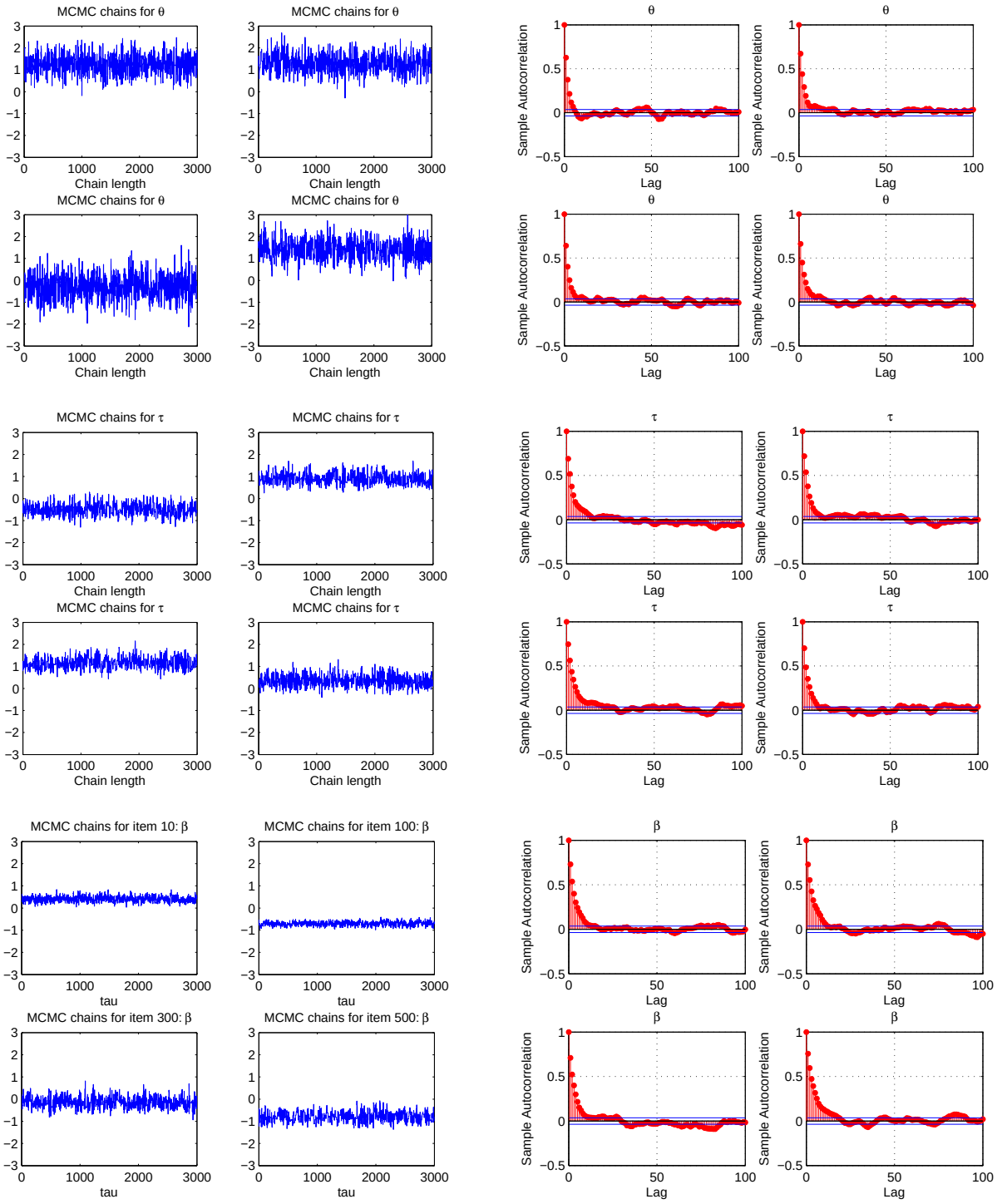


Figure 4.7: Traceplots and Autocorrelation plots for θ, τ, and β parameters

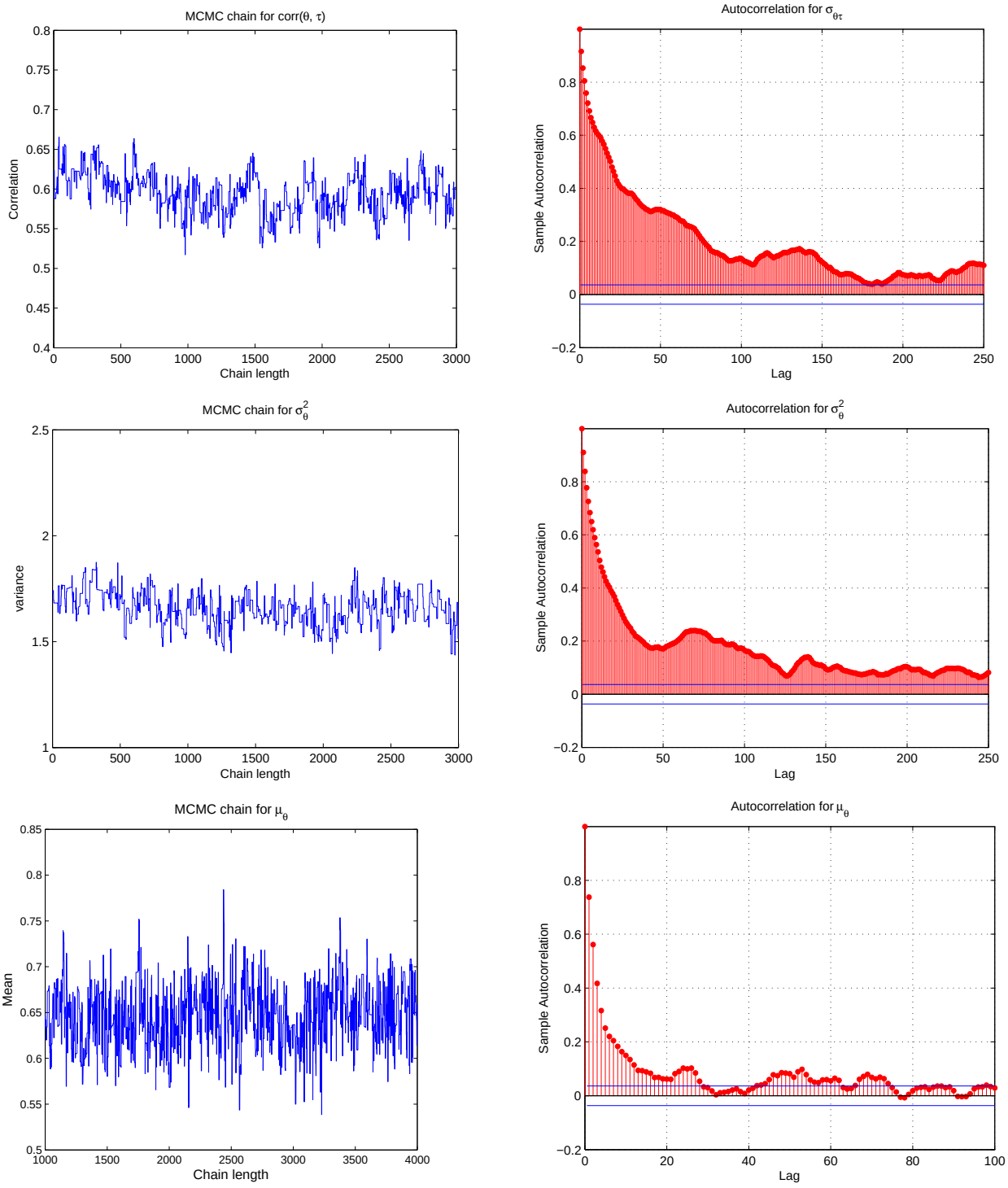


Figure 4.8: Traceplots and Autocorrelation plots for population ($\sigma_{\theta\tau}$, σ^2_θ, and μ_θ) parameters

4.3.1 Fitting Cumulative Baseline Hazard with B-splines and Parametric Functions

The baseline cumulative hazards calculated from the Brewslow estimator are also provided in Figure 4.10. When fitting the B-spline, The degree of the B-spline basis was set to be 3, and 3 inner knots were chosen to construct the basis. R functions `bs` in `"splines"`package were used to carry out the B-spline fitting, and function `lm` was used to regress the B-spline bases on the Breslow estimation results through linear models. The B-spline curve was plotted against the Breslow estimator for the four items, as presented in Figure 4.10. It shows that the B-spline curves fit well with the points estimated from the non-parametric Breslow estimator, and therefore, we can largely reduce the number of parameters needed to adequately recover the entire cumulative baseline hazard estimate.

In some cases, besides using B-splines, the shape of the cumulative hazard for certain items could be summarized in a simpler parametric form, such as Weibull function. Recall that the cumulative hazard function of Weibull distribution is $H(t) = \lambda t^{\alpha}$, thus two parameters, λ and α, need to be estimated in this regard. These two parameters were estimated through a generic function `nlinfit` (nonlinear least square fitting) in MATLAB. We presented the fitted curve as well as the Breslow estimated curve for two representative items below in Figure 4.11. As one can notice, the parametric curves actually fitted the non-parametric baseline cumulative hazard well, with most of the red open circles surrounding the blue curves

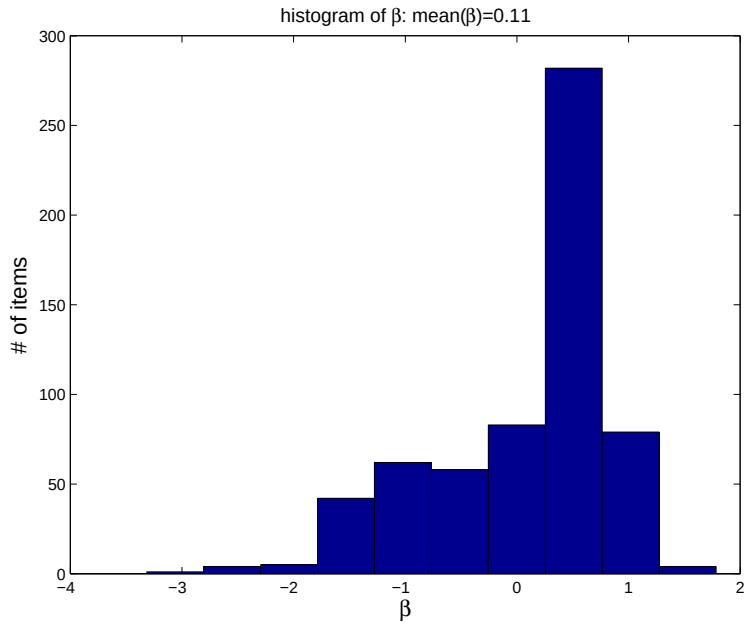
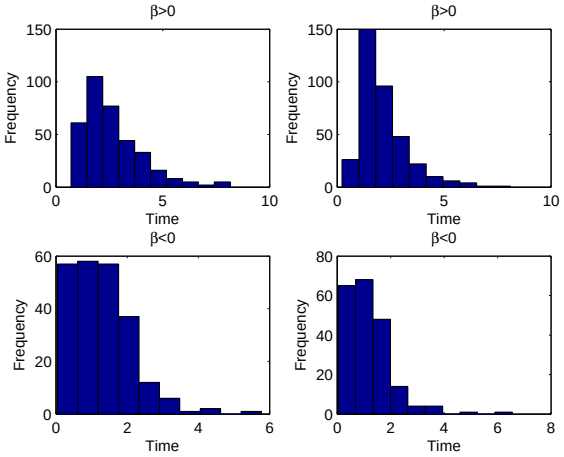
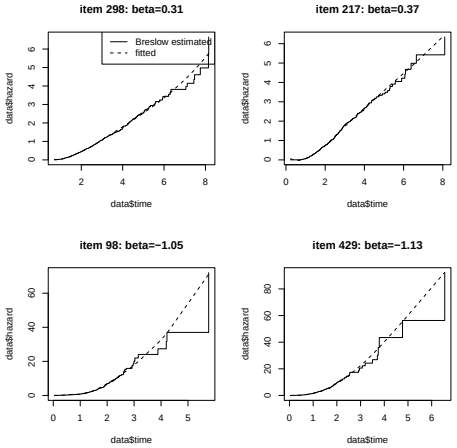


Figure 4.9: $\hat{\beta}$ distribution estimated from the hierarchical proportional hazard model

(a) *RT distribution for four items*



(b) *Estimation of the cumulative hazard*



(c) *θ distribution for the corresponding items*

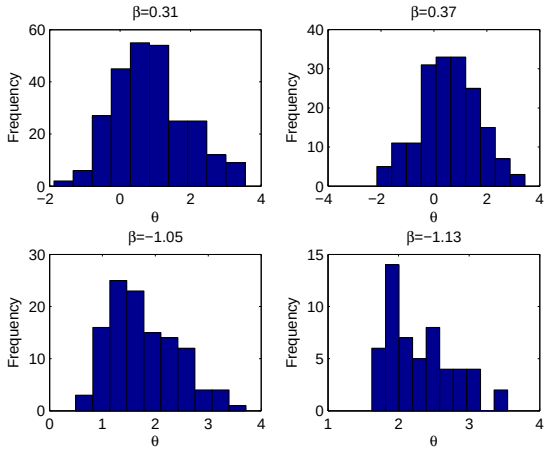


Figure 4.10: Illustration the RT histogram, cumulative baseline hazard, and examinees' ability distribution for four items

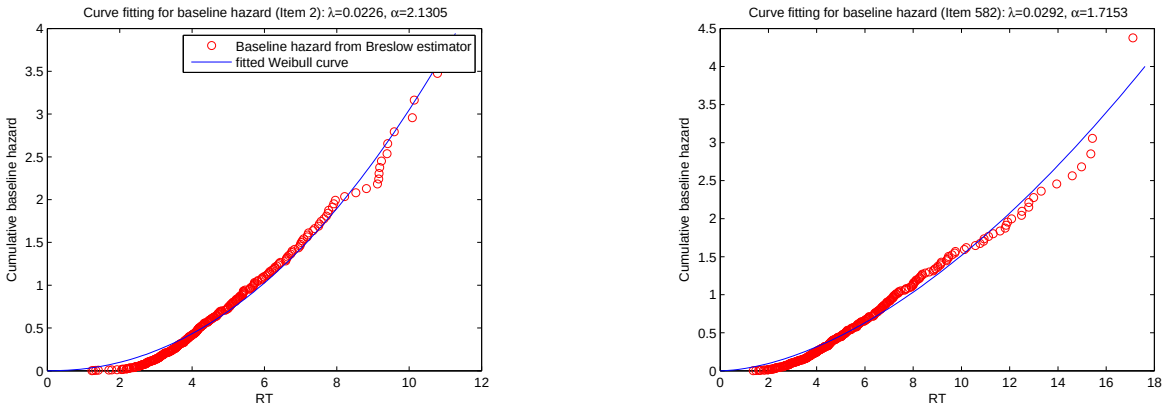


Figure 4.11: Curve fitting for baseline cumulative hazard: item 2(left) and item 582(right)

tightly.

4.3.2 Recovering the Response Time Distribution: Survival Curves

A straightforward way to check whether the proposed semi-parametric PH model fits the data at item level is to plot the Kaplan-Meier survival curve obtained from the observations against the expected survival curve generated from the model. The rationale is Kaplan-Meier estimator is a non-parametric estimator directly calculated from the data, thus it genuinely reflects the true survival pattern of an item, whereas the Cox model imposes the proportional hazard assumption as well as a fixed exponential link function. If these two curves are close, that means Cox PH model is an appropriate choice. In addition, we also added the expected survival function curve calculated from the proportional hazard model when imposing a Weibull parametric function on the cumulative baseline hazard. Specifically, for item j, the expected survival curve is estimated by

$$S_j(t) = \exp[-\exp(\hat{\beta}_j\hat{\tau})H_{0j}(t)]. \tag{4.1}$$

Figure 4.12 displays the Kaplan-Meier curve, expected survival function from semi-parametric Cox model and expected survival function from Cox model with parametric cumulative baseline hazard, for the same two items presented in Figure 4.11. As one can see in Figure 4.12, both red circles and blue circles are centered around the K-M curve, especially when RT is short. In general, red circles are relatively more closer to the K-M curve. This is reasonable because when imposing the parametric function on the baseline hazard, there might introduce some additional misfit to the model. Notice that we adopted a two-stage method, that is, we first fitted semi-parametric Cox model to the entire dataset and then tried the parametric form on the cumulative baseline hazard. Apparently, if we directly fitted the Weibull regression model on the dataset,

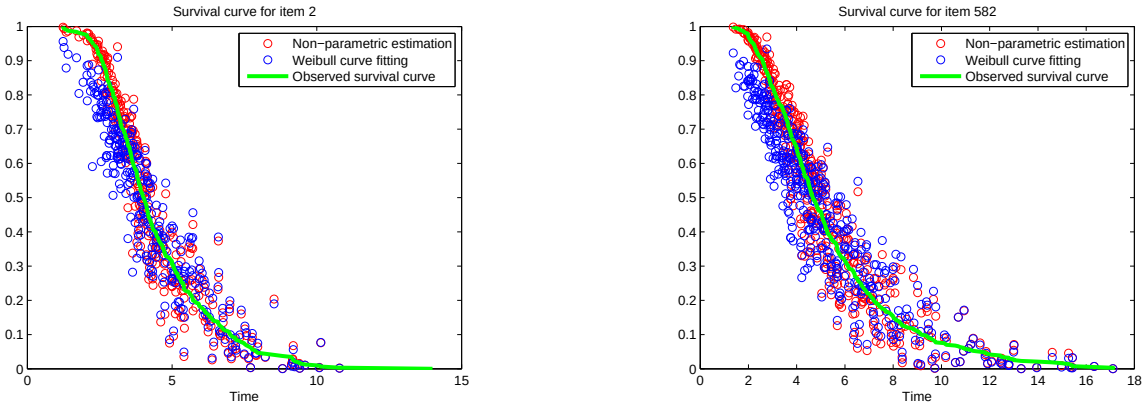


Figure 4.12: Observed vs. Expected survival curves

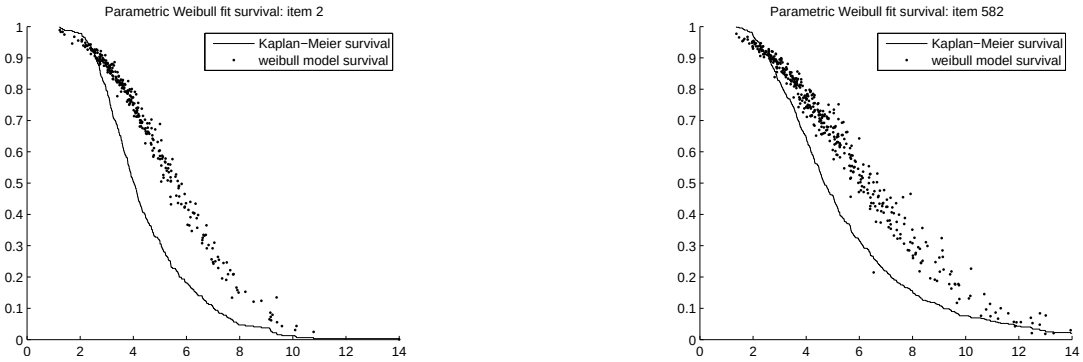


Figure 4.13: "True" and estimated survival curves from parametric Weibull model

the model data fit will be poor (as seen in Figure 4.5 for global fit), and the fit on individual items will also be deteriorated, as shown in Figure 4.13.

4.4 Further Model Diagnosis

Two key assumptions of the model are (1) the local independence of response time and response accuracy; and (2) stationarity assumptions. In this section, these two assumptions are checked via either hypothesis testing or descriptive statistics. The other two independence assumptions (see Equation (2.5) and (2.6)) are standard in IRT modeling, and we just simply assume they are satisfied.

4.4.1 Local Independence Assumption

An easy descriptive approach to check the local independence is to check the point-biserial correlation between responses and response times, with or without conditioning on the the estimated $\hat{\theta}$ and $\hat{\tau}$'s. If conditional correlation decreases significantly to near 0, one can conclude that the local independence assumption

might be satisfied. To calculate the conditional correlation, for each item, we first grouped examinees into K relatively homogeneous groups via *k-means clustering*, with the number of clusters, K, determined such that the number of examinees within each group was at least 5. We then calculated the point-biserial correlation within each group then obtained the mean for each item. By doing so, the weighted mean squared correlation, averaged over all items weighted by the sample size for each item, decreased from 0.13 to 0.07. In other words, the conditional correlation between item responses and RTs were nearly ± 0.27. The correlation was still not 0 because either the $\hat{\theta}$ and $\hat{\tau}$ estimation might contain measurement error; or some items were answered by as few as 6 test takers. Furthermore, within each "homogeneous" cluster, the θ and τ value could still vary substantially.

A more rigorous way to check the conditional independence assumption is via hypothesis testing. The conditional independence in Equation (2.7) can equivalently be expressed as

$$f(t_{ij}|y_{ij}, \tau_i) = f(t_{ij}|\tau_i),\ y_{ij} = 0, 1 \tag{4.2}$$

for all i and j. According to van der Linden and Glas (2010), this assumption is preferred over the alternative, $f(y_{ij}|t_{ij}, \theta_i) = f(y_{ij}|\theta_i)$ for all i and j. Two reasons were given. First, we only need to check whether the two conditional distributions of T_{ij} given $Y_{ij} = 0$ or 1 are equal rather than checking the equality of an entire family of distributions of Y_{ij} given the continuous measure of $T_{ij} = t_{ij}$. Second, the estimation of τ_i is typically more accurate than the estimation of θ_i (see Tables 2.1 and 3.2) because the continuous response time data is expected to contain more information than the binary response accuracy data.

If this assumption is violated, the response time model could be modified as

$$h_j(t_{ij}) = h_{0j}(t) \exp(\beta_j \tau_i + \lambda_j u_{ij}), \tag{4.3}$$

thus the assumption check reduces to check the significance of λ_j for item j. The null hypothesis would be

$$H_0 : \lambda_j = 0,$$

whereas the alternative hypothesis is $H_1 : \lambda_j \neq 0$. Similar to van der Linden and Glas (2010), we assumed the item parameters, including β_j and h_{0j}, were pre-calibrated. Thus, for a given item, the parameters that need to be estimated are $\boldsymbol{\tau} = (\tau_1, ..., \tau_N)$ and λ_j. The likelihood can be rewritten as

$$l(\boldsymbol{\tau}, \lambda_j) = \log \left\{ \prod_{i=1}^{N} \left[f(t_{ij}|\tau_i, \beta_j, \lambda_j) \prod_{l=1; l \neq j}^{J} f(t_{ij}|\tau_i, \beta_j) \right] \right\} \tag{4.4}$$

where

$$f(t_{ij}|\tau_i, \beta_j, \lambda_j) = H_j(t_{ij}) \exp^{\beta_j \tau_i + \lambda_j u_{ij}} \exp^{-\exp^{\beta_j \tau_i + \lambda_j u_{ij}} H_j(t_{ij})},$$

and

$$f(t_{il}|\tau_i, \beta_l) = H_j(t_{ij}) \exp^{\beta_j \tau_i} \exp^{-\exp^{\beta_j \tau_i} H_j(t_{ij})}.$$

Typically there are three ways of conducting this hypothesis test:

1. Likelihood ratio test: One has to fit both the null model and the alternative model in (4.3), and compare the differences between the two likelihoods.

2. Wald test statistic: One can construct the test statistics as $\chi^2 = \frac{(\hat{\lambda}_j)^2}{\text{var}(\hat{\lambda})}$, and this involves estimating the unknown parameter λ_j. It becomes slightly complicated when the latent covariates $\boldsymbol{\tau}$ need to be estimated as well.

3. Lagrange Multiplier (LM) test: This test is relatively easy to compute because only the null model needs to be estimated. Specifically, assume the null model has parameters $\boldsymbol{\eta}_1$ and the alternative model has additional parameters $\boldsymbol{\eta}_2$. Assume the hypothesis is $H_0 : \boldsymbol{\eta}_2 = 0$ against $H_1 : \boldsymbol{\eta}_2 \neq 0$. The LM test is defined as

$$LM(\boldsymbol{\eta}) = \boldsymbol{h}(\boldsymbol{\eta})' \boldsymbol{H}(\boldsymbol{\eta}, \boldsymbol{\eta})^{-1} \boldsymbol{h}(\boldsymbol{\eta}) \mid_{\boldsymbol{\eta}_1 = \hat{\boldsymbol{\eta}}, \boldsymbol{\eta}_2 = 0}, \tag{4.5}$$

   where $\boldsymbol{h}(\boldsymbol{\eta}) = \frac{\partial \ln L(\boldsymbol{\eta}; x)}{\partial \boldsymbol{\eta}}$, and $\boldsymbol{H}(\boldsymbol{\eta}, \boldsymbol{\eta})$ denotes an observed information matrix with elements $h(\eta_p, \eta_q) = -\frac{\partial^2}{\partial \eta_p \partial \eta_q} \ln L(\boldsymbol{\eta}; x)$. The maximum likelihood estimate of $\boldsymbol{\eta}_1$ is $\hat{\boldsymbol{\eta}}_1$, and x represents the data. One apparent advantage of LM test is the unknown parameter $\boldsymbol{\eta}_2$ does not need to be estimated, and thus the LM statistic is generally straightforward to calculate. The LM statistic follows a chisquare distribution with degree of freedom equal to the number of components in $\boldsymbol{\eta}_2$.

As suggested in van der Linden and Glas (2010), we will adopt the LM statistic to check the local independence assumption. For item j, the LM statistic is constructed as

$$LM(\lambda_j) = \frac{h(\lambda_j)^2}{h(\lambda_j, \lambda_j) - \boldsymbol{H}(\boldsymbol{\tau}, \lambda_j)' \boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\tau})^{-1} \boldsymbol{H}(\boldsymbol{\tau}, \lambda_j)} \mid_{\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}, \lambda_j = 0}, \tag{4.6}$$

where $\boldsymbol{H}(\boldsymbol{\tau}, \boldsymbol{\tau})$ is an $n_j \times n_j$ diagonal matrix with n_j denoting the number of examinees answering the j^{th}

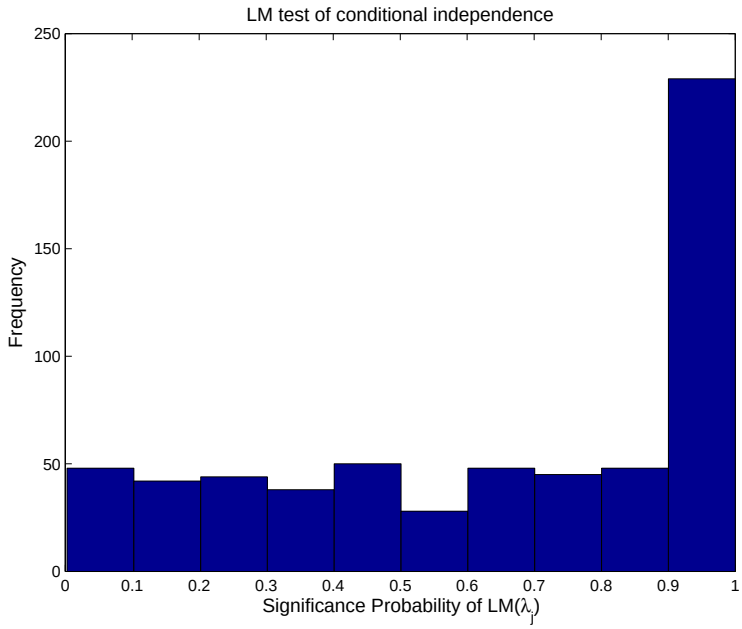


Figure 4.14: Lagrange Multiplier probability for conditional independence

item. By plugging in the likelihood function in Equation (4.4) into (4.6), we have

$$\begin{aligned}
-\boldsymbol{H}^{ii}(\boldsymbol{\tau},\boldsymbol{\tau}) &= -\beta_j^2 H_j(t_{ij})\exp\left(\beta_j\tau_i+\lambda_j u_{ij}\right) - \sum_{l=1;l\neq j}^{J}\beta_j^2 H_j(t_{ij})\exp\left(\beta_j\tau_i\right) \\
-h(\lambda_j) &= \sum_{i=1}^{n_j} u_{ij} - \sum_{i=1}^{n_j} H_j(t_{ij})\exp(\beta_j\tau_i+\lambda_j u_{ij})u_{ij} \\
-h(\lambda_j,\lambda_j) &= -\sum_{i=1}^{n_j} u_{ij}^2 H_j(t_{ij})\exp(\beta_j\tau_i+\lambda_j u_{ij}) \\
-\boldsymbol{H}^{i}(\boldsymbol{\tau},\lambda_j) &= -H_j(t_{ij})\exp(\beta_j\tau_i+\lambda_j u_{ij})\beta_j u_{ij}
\end{aligned}$$

In the calculation, replace τ_i with its corresponding MLE that were obtained by maximizing the likelihood function in (4.4). For the 620 items in the item bank, the LM(λ_j)'s are presented in Figure (4.14). Only 43 items have probabilities significant at 5% level. This again supported our conclusion that this local independence between θ and τ assumption was satisfied.

4.4.2 Stationarity Assumption

The stationarity assumption claims that examinees' speed and ability are constant during the test. While constant ability is standard in item response modeling, the constant speed assumption needs to be checked.

For examinee i, we calculated the residual response time as

$$r_{ij} = \tilde{t}_{ij} - t_{ij} = \int S_{ij}dt - t_{ij} = \int \exp[-\exp(\hat{\beta}_j\hat{\tau}_i)\hat{H}_{0j}(t)]dt - t_{ij},$$

for $j = 1, 2, ..., 37$. We then conducted the Wald-Wolfowitz RUNS test on residual RTs for each examinee separately. The null hypothesis is that the residuals on different items were independent, i.e., there was no item position effect and the examinee's speed could be viewed as a constant. Out of 2036 test takers, only 86 were rejected, which implies that the stationarity assumption might hold. In addition, as in van der Linden et al. (2007), we plotted the residual response times against item position for four randomly chosen examinees in Figure 4.15. If the stationarity assumption holds, the residual should fluctuate around 0 along

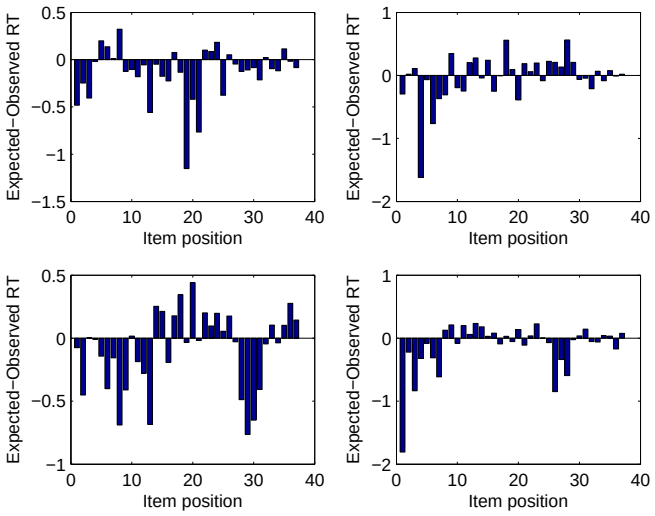


Figure 4.15: Mean residual times on the items as a function of their position in the test: four selected examinees

the tests. But as one can see, for some examinees, the residuals were uniformly negative at the beginning and positive toward the end. That means, they worked somewhat slower than expected at the beginning of the test and compensated toward the end. Our statistical analysis of this conflicted somewhat with our graphical analysis, but we do believe a slight position effect exists.

In addition, due to the adaptive feature of the dataset, the same item may appear in different positions for different examinees. We were interested to see whether there existed a position effect of the items, that is, whether the residuals of item response time changed with the item position in the test. Figure 4.16 displays the residual plots for four randomly selected items. The results showed that the residuals displayed similar patterns regardless of the item position in the tests.

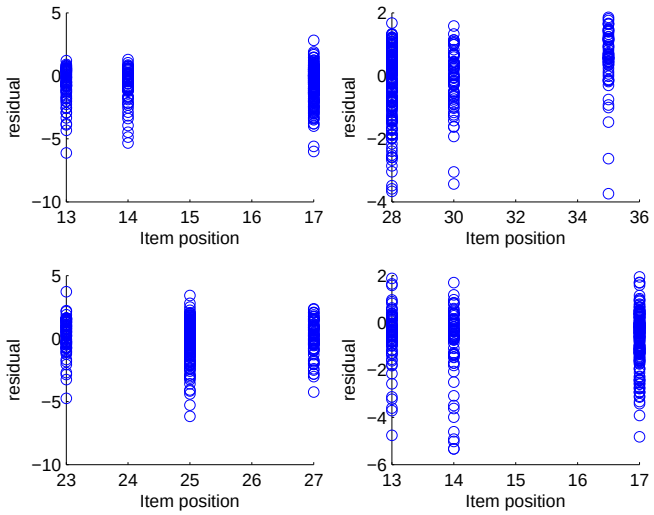


Figure 4.16: Mean residual times on the items as a function of their position in the test: four selected items

Chapter 5

Alternative Methods of Semi-Parametric Model Estimation

Besides proposing two new semi-parametric models for analyzing response time and response accuracy simultaneously, one major contribution of the dissertation is developing two-stage methods for model calibration. It employs a "divide-and-conquer" strategy. In the first stage, only the information regarding the rank of the response times is used to estimate the parametric part of the model; in the second stage, the non-parametric part is approximated conditioning on the "known" parameters. One limitation of the method is that it requires a known approximation to the marginal rank-based likelihood (see Equation (3.14)). The approximation, however, might be either hard to derive for certain error term distributions, or may depart significantly from its true distribution. In this chapter, we brainstorm for other possible estimation techniques that inherit from advanced statistical computing, yet have not been applied to the latent linear transformation modeling.

5.1 Modeling Non-Parametric Transformation Through Incomplete Beta Function

This method unifies the estimation of both the non-parametric transformation and parameters within a fully Bayesian MCMC algorithm. Different from the MCMC algorithm introduced in Section 3.2.3, we now need to update the unknown monotone transformation in a separate Markov chain as well. The incomplete beta function, defined as

$$\mathrm{IB}(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt,$$

can be used as a basis to approximate an arbitrary monotone function defined on $(0, +\infty)$. The specific idea is presented below.

Suppose that

1. h is a strictly increasing differentiable transformation from $(-\infty, +\infty)$ onto $[0, 1]$; one possible choice is $h(x) = \frac{\exp(x)}{1+\exp(x)}$

2. J is a strictly increasing differentiable distribution function from $(0,\infty)$ onto $[0,1]$, $J(t) = h[g(t)]$. Also suppose,$g(t)$ is a function defined from $[0,\infty)$ onto $(-\infty,+\infty)$. With the above choice of $h(x)$, we have $g(t) = \log\left(\frac{J(t)}{1-J(t)}\right)$.

To avoid confusion, let us rewrite the linear transformation model of interest as

$$g_j(t_{ij}) = \beta_j \tau_i + \varepsilon_{ij}.$$

For item j, let g_0 be a base function for g_j and let $J_0(t) = h[g_0(t)]$ be the distribution function associated with g_0. A way to find an appropriate function J is to search in the family of discrete mixtures of beta densities (Diaconis & Ylvisaker, 1985), which provides a dense class for continuous densities on $[0,1]$. A member of this class is given by

$$J(t) = \sum_{k=1}^{K} w_k \mathrm{IB}(J_0(t)|c_k, d_K), \tag{5.1}$$

where K denotes the number of mixands, $w_k \leq 0$ and $\sum w_k = 1$. By representing J as in Equation 5.1, $g_j(t) = h^{-1}[J(t)] = \log\left(\frac{J(t)}{1-J(t)}\right)$ is directly available. Usually the number of mixands is chosen to be 4 (Walker & Wakefield, 1996), and $c_k = \lambda k$ and $d_k = \lambda(r+1-k)$ (Mallick, 1994). The initial form of J_0 can be obtained via cubic smoothing spline. A function called csaps in MATLAB was used.

With the above argument, for a given item, sampling g_j is equivalent to just sampling the weights $w = (w_1, ..., w_k)$ subject to the constraint that $\sum w_i = 1$. This is done by first transforming $(w_2, ..., w_k)$ via logit transformation, and then use a multivariate normal proposal centered on the logit transformation of the current weights. The new w_1 is fixed once the proposal on $(w_2, ..., w_k)$ are obtained. If $\sum_{i=2}^{k} w_i > 1$, we just simulate a new set of values from multivariate normal proposal again until the value of the resulting w_1 is reasonable. This proposal is accepted or rejected according to some probability obtained in the usual way for a Metropolis step.

One apparent advantage is that once the transformation g_j is known, we can construct the likelihood of β_j and τ_i in a usual way as follows,

$$L(\beta_j|\boldsymbol{\tau}, \boldsymbol{t}_j, g_j) = \prod_{i=1}^{N} f_\varepsilon(g_j(t_{ij} + \tau_i\beta_j)),$$

instead of resorting to partial likelihood or approximation of the marginal rank based likelihood. To show that this algorithm generates reasonable parameter estimation, we did a small scale simulation study. The linear transformation model introduced in Chapter 3 was considered, the error term was assumed to follow the logistic distribution yielding a proportional odds model. We were interested in this model because the

marginalized rank likelihood in (3.10) has a relatively poor approximation via (3.15), and thus it is interesting to see whether this alternative method could improve the model estimation. Examinee sample size was set to 250, and test length was chosen to 20. Responses and response times were simulated in the same way as in Chapter 3. As a preliminary check, we only tried two transformations, log transformation and Box-Cox transformation. The results are presented in Tables 5.1 and 5.2.

Table 5.1: Bias of the estimations

		θ (IRT)	θ (IRT+RT)	β	τ	$\sigma_{\theta\tau}$	σ_θ^2	μ_θ
Rank based likelihood	log	-0.017	-0.008	0.091	-0.018	0.043	0.271	0.007
	Box-Cox	-0.056	-0.045	-0.020	0.017	0.041	0.351	-0.006
Beta function based method	log	0.019	-0.013	0.052	0.023	0.055	0.249	0.021
	Box-Cox	0.041	0.037	0.043	0.051	0.049	0.298	0.021

Table 5.2: MSE of the estimations

		θ (IRT)	θ (IRT+RT)	β	τ	$\delta(H)$
Rank based likelihood	log	0.169	0.138	0.078	0.147	1.623
	Box-Cox	0.164	0.109	0.027	0.153	0.809
Beta function based method	log	0.171	0.128	0.069	0.142	3.413
	Box-Cox	0.168	0.117	0.043	0.172	1.987

Different from our expectation, the beta function based method did not generate more accurate results than the method introduced in Chapter 3, the MSE is even slightly larger. The non-parametric transformation, $H(\cdot)$, was not estimated accurately, with $\delta(H)$ nearly twice as large as the results from previous method. This is due to the imperfections induced by using incomplete beta function to approximate the unknown transformation.

5.2 Likelihood-Free Algorithm

In the linear transformation model, due to the unknown non-parametric transformation $H(\cdot)$, the likelihood function for response time $L(\boldsymbol{t}|\boldsymbol{\beta}, \boldsymbol{\tau})$ is analytically unavailable. A class of algorithms and methods has been developed to perform Bayesian inference in this setting, and they have been known as likelihood-free computation or approximate Bayesian computation (Beaumont, Zhang, & Balding, 2002; Beaumont, Cornuet, Marin, & Robert, 2009). As the names indicate, these methods circumvent the explicit evaluation of the likelihood by a simulation based approximation.

The underlying idea of likelihood-free methods may be simply encapsulated as follows. Let β represent the unknown parameter, and let y represent the data. For a candidate parameter β', a data set is generated from the model (i.e. the likelihood function) $x \sim \pi(x|\beta')$. If the simulated and observed datasets are similar

to some extent, so that $x \approx y$, then β' is a good candidate and it should be retained and forms as a part of the samples from the posterior distribution $\pi(\beta|y)$. This likelihood free idea can also be viewed from a data augmentation perspective, that is, it augments the target posterior from $\pi(\beta|y) \propto \pi(y|\beta)\pi(\beta)$ to

$$\pi_{\text{LF}}(\beta, x|y) \propto \pi(y|x, \beta)\pi(x|\beta)\pi(\beta), \tag{5.2}$$

where the simulated dataset x from $\pi(x|\beta)$ is viewed as auxiliary parameter, on the same space as $y \in \mathcal{Y}$. $\pi(y|x, \beta)$ no longer needs to be a likelihood function, but instead, it is a function that weights the posterior $\pi(\beta|x)$ with high values in regions where x and y are similar. Ultimate interest is typically in the marginal posterior

$$\pi_{\text{LF}} \propto \pi(\beta) \int_{\mathcal{Y}} \pi(y|x, \beta)\pi(x|\beta)dx,$$

integrating out the auxiliary dataset x.

5.2.1 Likelihood-free MCMC Samplers

A Metropolis-Hastings sampler maybe constructed to target the augmented likelihood-free posterior $\pi_{\text{LF}}(\beta, x|y)$ without directly evaluating the intractable likelihood (Marjoram, Molitor, Plagnol, & Tavare, 2003). Specifically, assume at a current state (β, x), a new parameter β' is drawn from a proposal distribution $q(\beta, \beta')$, and conditionally on β' a proposed dataset x' is generated from the model $x' \sim \pi(x|\beta')$. The probability of accepting a move from (β, x) to (β', x') within the Metropolis-Hastings framework is $\min\{1, \alpha\}$, where

$$\alpha = \frac{\pi_{\text{LF}}(\beta', x'|y)q[(\beta', x'), (\beta, x))]}{\pi_{\text{LF}}(\beta, x|y)q[(\beta, x), (\beta', x')]} = \frac{\pi_\epsilon(y|x', \beta')\pi(\beta')q(\beta', \beta)}{\pi_\epsilon(y|x, \beta)\pi(\beta)q(\beta, \beta')}, \tag{5.3}$$

such that the intractable likelihoods do not need to be evaluated in the acceptance probability evaluation in (5.3).

In computation, to improve the accuracy of Monte-Carlo approximation, one often calculates the acceptance probability as

$$\alpha \approx \frac{\frac{1}{S}\sum_S \pi_\epsilon(y|x'^s, \beta')\pi(\beta')q(\beta', \beta)}{\frac{1}{S}\sum_S \pi_\epsilon(y|x^s, \beta)\pi(\beta)q(\beta, \beta')}, \tag{5.4}$$

where $x'^1, ..., x'^S \sim \pi(x|\beta')$. The Monte-Carlo approximation becomes more accurate when S increases. A key component in the likelihood free method is the selection of π_ϵ. Two typical forms are constructed, and each allows some form of approximation to $\pi_{\text{LF}}(\beta|y)$. The first form is

$$\pi_\epsilon(y|x, \beta) = \frac{1}{\epsilon}K\big(\frac{|x - y|}{\epsilon}\big), \tag{5.5}$$

where K takes the form of standard smoothing kernel density centered at the point $x = y$. In this manner, $\pi_\epsilon(y|x, \beta)$ weights the intractable likelihood with high values in regions where the auxiliary and observed datasets are similar (i.e., $x \approx y$), and with low values in regions where they are different (Beaumont et al., 2002). The second form permits the comparison of the datasets, x and y, to occur through a low-dimensional vector of summary statistics $T(\cdot)$, and the function takes the following form

$$\pi_\epsilon(y|x, \beta) = \frac{1}{\epsilon} K\big(\frac{|T(x) - T(y)|}{\epsilon}\big). \tag{5.6}$$

It will provide regions of high values when $T(x)$ and $T(y)$ are close, and low values otherwise. When the summary statistics is sufficient for the unknown parameters β, then comparing the summary statistics of two datasets will be equivalent to comparing the datasets themselves.

To implement the likelihood-free idea in the linear transformation model estimation, when updating the chain of β_j, the regression parameter, we have the acceptance probability

$$\alpha(\beta_j, \beta_j') \approx \frac{\frac{1}{S}\sum_S \pi_\epsilon(y|x'^s, \beta_j')\pi(\beta_j')}{\frac{1}{S}\sum_S \pi_\epsilon(y|x^s, \beta_j)\pi(\beta_j)}, \tag{5.7}$$

where $q(\cdot, \cdot)$ is canceled out if a normal proposal function is used. $x'^1, ..., x'^S$ are simulated from F_ε, with mean shifted by $\beta^{(r-1)}\tau^{(r-1)}$, where $\tau^{(r-1)}$'s are the estimation from the $(r-1)^{\text{th}}$ iteration. However, the simulated data x and observed response time t are not comparable, due to the unknown transformation $H(\cdot)$. So we could transform x via an intermediate estimated $\hat{H}(\cdot)$ function before calculating π_ϵ in (5.6) or (5.5), and $\hat{H}(\cdot)$ can be obtained from the estimating equation method introduced in Chapter 3.

Chapter 6

Discussion and Future Work

Response times on test items are easily collected in modern computerized testing. Analyzing response time provides useful collateral information to further understand examinees' behaviors and item/test characteristics. A dozen non-linear latent trait models have been proposed in the past to model RTs exclusively or with responses simultaneously. Many of the models were based on the "distribution-fitting approach", such as the lognormal model that can capture the skewness of the RT distribution quite well. Although skewed distributions are widely seen in achievement testing, but will not hold for each item. For instance, if some test takers are engaging in two different response strategies, say, rapid guessing or actively answering the item, their response time distributions will be bimodal. Accordingly, we proposed semi-parametric models that are able to represent different kinds of RT distributions. In particular, the Cox PH frailty model introduced in Chapter 2 is a generalization of the exponential model, Box-Cox normal model (Klein Entink et al., 2009a), Weibull model (Rounder et al., 2003), and many other parametric models. The linear transformation model introduced in Chapter 3 is an even more general model that subsumes the lognormal model (van der Linden, 2006) and the Cox PH model as special cases. This new model contains the whole collection of possible functional forms between RT and latent covariates through various link functions. This generalized approach will save practitioners from a labor-intensive search for an adequte parametric model, and more importantly, it will provide an *individualized* fitting to each item.

6.1 Semi-parametric Modeling Approach

Since the 1972 publication of Cox's seminal article on statistical models for lifetime data, survival methods, especially those for continuous time data, have enjoyed increasing popularity in a variety of disciplines ranging from medicine and industrial testing to economics and sociology. Item response time analysis, a specific research topic in educational measurement, will also benefit from the advances in survival methods. In fact, the semi-parametric modeling approach in survival analysis opens another avenue for RT modeling. In Chapter 2, we proposed a new model, which can be viewed as an extension of the Cox PH model. In

the new model, examinees' latent speeds τ serve as covariates, and the regression parameter β controls the effect of τ on RTs. This model hinges on the assumption that examinees' latent speed determine their RTs directly. This new model assigns a separate speed parameter τ to account for the individual differences in speed, while allowing τ to be correlated with θ at the population level. The hierarchical framework (van der Linden, 2007) distinguishes the speed accuracy tradeoff within a person from the speed accuracy correlation across persons. Simulation studies show that the new model can be estimated accurately via an MCMC algorithm. One apparent advantage of the proposed model comes from its semi-parametric nature. The non-parametric baseline hazard is flexible enough to accommodate different shapes of RT distributions in real data. Once the non-parametric baseline hazard is recovered by the Breslow estimator, we can further fit it either with a parametric form or with a curve generated by B-spline bases, depending upon the specific shapes of the baseline hazard.

The estimation method proposed in Chapter 2 uses the partial likelihood that is motivated as resulting from integrating out the baseline cumulative hazard function with respect to a gamma process prior. Although Clayton (1991) also adopts a gamma process prior, he includes the cumulative baseline hazard as a "parameter" to be updated within each Markov chain. Sharef et al. (2010) advocated using B-splines on H_0 and update it in MCMC as well. An apparent advantage of their approaches is that inference can be made on the baseline hazard. However, with a somewhat complicated posterior distribution encountered here, it seems more beneficial to use a divide-and-conquer approach. That is, treat the non-parametric baseline hazard as a nuisance parameter and integrate it out first, and once the parameters are accurately calibrated, estimate the non-parametric hazard secondly.

Chapter 3 generalized the Cox model to the linear transformation model that creates further flexibility. One challenge lies in the model estimation. Current estimation methods (Kalbfleisch, 1978; Pettitt, 1982; Chen et al., 2002) are developed assuming the covariates are observed, yet RT modeling often involves latent covariates, such as θ or τ. Therefore new estimation techniques need to be built up for this particular limitation. Notice that the ranks of the observations remain unchanged by monotone increasing transformations, and this fact justifies the use of the "marginal likelihood of ranks" to make inference on unknown parameters (Kalbfleisch and Prentice, 1973; Pettitt, 1983). In the same chapter, we proposed to use the marginal likelihood of rank in MCMC for the estimation of β and τ, and then use the recursive algorithm (Chen et al., 2002) for estimating H. This two-stage estimation method is able to recover the true model parameters and the unknown transformation very well. Ranger and Kuhn (2012) also adopted the form of the linear transformation model, but rather than introducing an arbitrary monotone transformation, they

introduced a parametric link function as

$$\log\left[\frac{(1-P(y_{ij}=1|\tau_i))^{-c_j}}{c_j}\right]=\alpha_j+\beta_j\tau_i, c_j>0,$$

where the parameter c_j determines the shape of the link function. When $c_j=1$, one obtains a logit link, and when $c_j\to 0$, it becomes a complementary log-log link. Another difference of their model is that it only represents discrete time, with $y_{ij}=1$ when t_{ij} is less than a threshold. In this sense, our model is more powerful and utilizes RT information in a more straightforward fashion. We also provide model checking methods to help evaluating both global and item level fit.

The real data example shows that the proposed semi-parametric model tends to fit the data better than the more restricted lognormal model, or other parametric models. One less intuitive phenomenon is that for roughly one third of the items in the item bank, the $\hat{\beta}$ parameters are negative, indicating that examinees with higher speed in general tended to answer those items slower. This is because in adaptive testing, each item is assigned to a restricted sample, and within the sample, the relationship between actual response time and latent speed might reverse. Further studies should confirm the applicability of the new model for other types of test data (such as non-adaptive achievement tests). Another future direction is to further break down the latent speed parameter τ into different information processing components, because different examinees might employ different strategies when solving an item. Response caution also plays an important role in examinees' processing speed (van der Mass, Molenaar, Maris, Kievit, & Borsboom, 2011).

One limitation of the current estimation method (for the linear transformation model) is that we need to know the parametric form of the error term distribution beforehand. With different error distributions, the approximation to the rank-based likelihood changes substantially. However, based on the real data example given in the paper, a fixed error distribution is sometimes too restrictive, and it is often possible that some items are better fit with normal error model whereas others are more consistent with logistic error model. If that happens, we need to employ an even more flexible model with error distributions unspecified. This generalization will certainly introduce extra difficulty in estimation, Mallick and Walker (2003) provided a fully Bayesian estimation method for such generalized model with observed covariates, and they employed the Polya tree distribution as a prior for the unknown distribution F_ε. Their method might be a promising starting point for extending the current model with additional flexibility.

6.2 Application of Response Time

The objective of RT research is to improve the estimation accuracy of the examinees' abilities. This can be partially accomplished by ensuring that the test is of high quality. The test quality includes test fairness, test efficiency and so on. In particular, concerning test fairness, RTs allow us to formulate constraints on item selection (in adaptive tests) or test assembly (in linear form tests) that guarantee the multiple forms of a test to be equally speeded. In our modeling approach here, $\int_0^\infty S_j(t|\beta,\tau,h_0)dt$ is the expected time to answer the j^{th} item, and upon knowing this, the constraint related to RTs can be easily incorporated in item selection through the weighted deviation model (Stocking and Swanson, 1993) or the constraint weighted index (Cheng and Chang, 2009). Concerning efficiency, observe that a highly informative item can be quite time consuming, so it has less practical value compared to an equally or somewhat less informative items that require less time to complete. Therefore, instead of maximizing the raw item information, we can maximize the item information per time unit (Fan et al., 2012).

6.2.1 Redesign of Item Selection Algorithm in CAT

Traditional methods for item selection in computerized adaptive testing only focus on item information without taking into consideration the time required to answer an item. As a result, some examinees may receive a set of items that take a very long time to finish, and information is not accrued as efficiently as possible. With the well defined psychometric models on response time, one can take this information into the item selection.

Specifically, because the variance of $\hat{\theta}^{mle}$ is inversely related to the Fisher information, it motivates the procedure of selecting items to maximize Fisher information at the current ability estimate. For instance, index the bank of all possible items by $j = 1, 2, ..., J$, and suppose that m items have been administered, $Y_{j_1}, Y_{j_2}, ..., Y_{j_m}$. Let $S_m = \{j_1, j_2, ..., j_m\}$ denotes the indices for these items, and let $R_m = \{1, 2, ..., N\} \cap \bar{S}_m$ denotes the remaining items. Let $\hat{\theta}_m^{mle}$ denote the current ability estimate. The Maximum Information Criterion (MIC) selects the item which has highest information at $\hat{\theta}_m^{mle}$ as,

$$j_{m+1} = \max_l \{I_l(\hat{\theta}_m^{mle}) : l \in R_m\}. \tag{6.1}$$

However, MIC does not take into consideration the time required to answer an item. Such information is useful in that often a highly informative item can be quite time consuming. As a result, instead of maximizing raw item information $I_l(\hat{\theta}_m^{mle})$ in (6.1), Fan, Wang, Chang, and Douglas (2012) proposed a new criterion, Maximum Information per Time Unit (MICT). Rather than selecting the item with highest item

information at the current ability estimate, we can choose the next item based on

$$j_{m+1} = \max_{l} \left\{ \frac{I_l(\hat{\theta}_m^{mle})}{E[T_l|\hat{\tau}_m^{mle}]} : l \in R_m \right\}, \tag{6.2}$$

where T_l is the time required for the l item and $\hat{\tau}_m^{mle}$ is the maximum likelihood estimator of current the speed parameter τ. Under the lognormal model in van der Linden (2007), the expected time to answer the lth item is obtained by taking expectation of t_{nj} with respect to the density in Equation (1.9), treating α_j, β_j, and τ_n as known parameters for the density. Specifically,

$$E[T_l|\hat{\tau}_m^{mle}] = \int_{-\infty}^{\infty} \frac{\alpha_l}{t_l\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left[\alpha_l(\ln t_l - (\beta_l - \hat{\tau}_m^{mle}))\right]^2 \right\} dt_l$$

and it can be further simplified as

$$E[T_l|\hat{\tau}_m^{mle}] = \exp\left(\beta_l - \hat{\tau}_m^{mle} + \frac{1}{2\alpha_l^2}\right),\ l \in R_m.$$

The MLE of $\hat{\tau}_n^{mle}$ has a closed form expression as

$$\hat{\tau}_n^{mle} = \frac{\sum_{j\in R_m} \alpha_j^2(\beta_j - \log t_{nj})}{\sum_{j\in R_m} \alpha_j^2}. \tag{6.3}$$

Thus, if the lognormal model is employed for response time, it yields simple closed form solutions to the terms needed to implement the MICT. The semi-parametric model proposed in the current study can be used in a similar fashion, but more involved computation might be needed. Using this new item selection criterion, items with high information will tend to be chosen, but are less likely to be chosen if they require a great amount of time. By continually updating the speed parameter and the ability parameter, items may be chosen for an examinee that can assist in quickly accruing information about the examinee's ability. By doing so, an exam with a fixed amount of information required can be completed more quickly, possibly affording the chance to seek a higher level of information by adding more items. Also, an exam of fixed length can be completed more quickly.

6.2.2 Introduce Additional Covariates in the Model

We can introduce additional covariates in the model, such as examinees' demographic information, to better explain the variance in response time patterns. A survival model that is suitable for such a purpose is

$$h_j(t_{ij}|\tau_i, \boldsymbol{z}_i) = h_{0j}(t_{ij}) \exp(\beta_j \tau_i + \boldsymbol{\gamma}_j' \boldsymbol{z}), \tag{6.4}$$

or

$$H_j(t_{ij}|\tau_i, \boldsymbol{z}_i) = \beta_j \tau_i + \boldsymbol{\gamma}_j' \boldsymbol{z} + \varepsilon_{ij}, \tag{6.5}$$

where $\boldsymbol{Z}_i = (z_{i1}, ..., z_{ip})$ represents the observed covariates, such as gender, educational background, social economic status and such things. The current model can also be further extended to allow for the incorporation of explanatory variables to explain the variations in speed and accuracy between individuals who may be nested within groups. Such an effort is especially beneficial if the researchers want to pinpoint whether older people, or people with a certain disorder, tend to have decreased ability or slower information processing speed as opposed to younger people, or those without the disorder.

The model estimation methods introduced in Chapters 2 and 3 can be easily generalized to the new models. For instance, if model (6.4) is considered, it could be rewritten as $h_j(t_{ij}|\tau_i, \boldsymbol{z}_i) = h_{0j}(t_{ij}) \exp([\beta_j, \boldsymbol{\gamma}_j]'[\tau_i, \boldsymbol{z}])$, such that the partial likelihood for $[\beta_j, \boldsymbol{\gamma}_j]$ can be readily expressed as

$$L(\beta_j, \boldsymbol{\gamma}_j|\boldsymbol{\tau}, \boldsymbol{z}) = \prod_{i=1}^{N} \frac{\exp([\beta_j, \boldsymbol{\gamma}_j]'[\tau_i, \boldsymbol{z}_i])}{\sum_{p \geq i}^{N} \exp([\beta_j, \boldsymbol{\gamma}_j]'[\tau_p, \boldsymbol{z}_p])}.$$

In some cases, instead of imposing an item level coefficient $\boldsymbol{\gamma}_j$, one can also impose fixed or random slope for each observed covariate. Whether or not a certain covariate has a significant effect on the response time patterns can be checked via the Lagrange multiplier test introduced in Chapter 4.

6.2.3 Constructing RT Models for Cognitive Psychology

Current response time models are appropriate for tests such as achievement tests, attitude scales, or personality questionnaires. Cognitive and experimental psychologist, who often collect RTs as a major source of behavior data to make inference about the latent cognitive process, would require different modeling strategies. That is because in achievement testing, examinees normally have plenty of time to answer each item, and the current model assumes examinees operate at a constant ability and speed along the test. In cognitive experiments, however, subjects are often given the instruction such as "respond as quickly as you can", and in this regard, a speed-accuracy tradeoff should be included in the model. The model proposed

in Loeys et al.(2011) could be adopted, but again, instead of relying on log-transformation of response time, the semi-parametric models proposed in this dissertation can be better candidates.

6.3 Summary

Two classical models for continuous outcomes such as response times are the linear mixed model (Verbeke and Molenberghs, 2000) and the proportional hazard model (Cox, 1972). Though the lognormal model is chosen based on statistical convenience and goodness of fit rather than cognitive theory (Luce, 1986), the proportional hazard model is often popular in modeling response times in mathematical and experimental psychology due to its nice mathematical properties of hazard functions (Bloxom, 1985; Vorberg and Ulrich, 1987; Wenger and Gibson, 2004). However, no previous semi-parametric models with crossed random effects for persons and items have been proposed (Loeys et al., 2011). In this dissertation, instead of treating the random effects of persons and items in a non-hierarchial relationship as in a crossed random effect model (Raudenbush, 1993), we proposed a two-level semi-parametric model with random effects for both items and persons. In parallel with the explosion of hierarchical modeling in psychometrics, the field of diffusion models is also growing in cognitive theory and may offer alternative approaches for modeling response time and response accuracy simultaneously. The diffusion model offers the advantage of an interpretable process model as a measurement model, in contrast, the advantage of the psychometric models presented in this dissertation allow for relatively easy quantification of the correlation between person's latent speed and ability (Loeys et al., 2011), and between items' difficulty and time intensity. The semi-parametric models offer additional flexibility when analyzing real datasets, albeit at an additional level of model complexity. But, as the saying implies, "there is no such thing as a free lunch".

References

Andersen, PK, Borgan, Gill, R.D., & Keiding, N. (1992). *Statistical models based on counting processes.* Springer, New York.

Aslanidou, H., Dey, D., & Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods *The Canadian Journal of Statistics, 26* , 33–48.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barnard. G.A. (1962). "Some Logical Aspects of the Fiducial Argument," *Journal of the Royal Statistical Society, Series B (Methodological), 44*, 234–243.

Bassili, J. (1996). The how and why of resposne latency measurement in telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research*, (pp. 319–346). San Francisco: Jossey-Bass.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics, 162*, 2025–2035.

Beaumont, M. A., Cornuet, J. M., Marin, J. M., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika, 162*, 2025–2035.

Bennett, S. (1983). Analysis of survival data by the proportional odds model, *Statistics in Medicine, 2*, 273–277.

Bickel,P.J., & Ritov,Y. (1998). Local Asymptotic Normality of Ranks and Covariates in Transformation Models. *Unpublished Draft.*

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika, 50*, 383–397.

Box, G.E.P., & Cox, D.R. (1964). An Analysis of Transformations (with Discussion). *Journal of the Royal Statistical Society, 26*, 211–252.

Bradley, J.V. (1973). *Distribution Free Statistical Tests,* New Jersey: Prentice-Hall.

Breslow, NE. (1972). Discussion of the paper by D.R.Cox. *Journal of the Royal Statistical Society, series B, 34*, 216–217.

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

Cai, T., Hyndman, R., & Wand, M. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics, 11*, 784–798.

Casella, G. & Berger, L. (1990). *Statistical inference.* Pacific Grove, CA: Brooks/Cole.

Chang, H. (2004). *Computerized testing, E-rater, and generic algorithm: Psychometrics to support emerging technologies.* Invited Symposium, the 28th International Congress of Psychology, Beijing, China, August 8-13, 2004.

Chang, H. (in press). *Making computerized adaptive testing diagnostic tools for schools*. In R. W. Lissitz & H. Jiao (Ed.), Computers and their impact on state assessment: Recent history and predictions for the future. Information Age Publisher.

Cheng, S. C., Wei, L. J.,& Ying, Z. (1995). Analysis of Transformation Models With Censored Data. *Biometrika, 82*, 835–845.

Chen, K., Jin, Z., & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika, 89*, 659–668.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics, 48*, 61–72.

Clayton, D., & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A, 148*, 82–117.

Clayton, D., & Cuzick, J. (1986). The semi-parametric Pareto model for regression analysis of survival times. In *Papers on Semiparametric Models at the ISI Centenary Session* (R. D. Gill and M. N. Voors, eds.)19–30. Report MS-R8169, Center for Mathematics and Computer Science, Amsterdam.

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B (Methodological), 34*, 187–202.

Cox, D.R, & Oakes,D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Cox, D.R., & Snell, E.J. (1968). A General Definition of Residuals(with Discussion). *Journal of the Royal Statistical Society, Series B, 26*, 211–252.

Cuzick, J. (1988). Rank Regression. *The Annuals of Statistics, 16*, 1369–1389.

de Boor, C.A. (1978). *A Practical Guide to Splines*. Springer-Verlag.

Diaconis, P., & Ylvisaker, D. (1985). Quantifying prior opinion. In Bernardo, J. M. et al. (Eds.), *Bayesian Statistics*, North-Holland, Amsterdam, pp.133–156.

Doksum, K.A. (1987). An Extension of Partial Likelihood Methods for Proportional Hazard Models to General Transformation Models. *The Annuals of Statistics, 15*, 325–345.

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

Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association, 98*, 555–563.

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

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

Feigl, P., & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics, 21*, 826–838.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *Annals of Statistics, 10*, 101-113.

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

Gulliksen, H. (1950). *Theory of Mental Tests,* Hillsdale, NJ: Lawrence Erlbaum.

Gelman, A., Carlin, J.B., Stern, H., & Rubin, D.B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Gray, R.J. (1994). A Bayesin analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics, 50*, 244–253.

Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics, 53*, 230–242.

He, X., & Shi. P. (1998). Monotone B-Spline Smoothing. *Journal of the American Statistical Association, 93*, 643–650.

Henschel, V., Engel, J., Holzel, D., & Mansmann, U. (2009). A semiparametric Bayesian proportional hazard model for interval censored data with frailty effects. *BMC Methological Research Methodology,* 9. doi:10.1186/1471-2288-9-9.

Holden, R. & Kroner, D. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment, 4*, 170–173.

Jaeger, F. T. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Kahane, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.). *The Nature of Cognition* (pp. 323–384). Cambridge (MA): MIT

Kalbfleisch, J.D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of Royal Statistical Society (Series B), 40*, 214–221.

Kalbfleisch, J.D., & Prentice, R.L. (1973). Marginal Likelihoods Based on Cox's Regression and Life Model. *Biometrika, 60*, 267–278.

Kaplan, E.L., & Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *Journal of American Statistical Association, 53*, 457–481.

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

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.

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*, 54–57.

Lehmann, E.L. (1953). The power of rank tests. *Annals of Mathematical Statistics, 24*, 23–43.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling appraoch for reaction time and accuracy in psycholinguistic experiments. *Psychometrika, 76*, 487–503.

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

Mallick, B.K., & Walker, S. (2003). A Bayesian semiparametric transformation model incoporating frailties. *Journal of Statistical Planning and Inference, 112*, 159–174.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their applications as psychometric models for response times. *Psychometrika, 58*, 445–469.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

Mulder, J., &Fox, J.-P. (1985). Bayesian tests on components of the compound symmetry covariance matrix. *Statistical Computing.* DOI 10.1007/s11222-011-9295-3

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.

Pettitt, A.N. (1982). Inference for the Linear Model Using a Likelihood Based on Ranks. *Journal of the Royal Statistical Society, Series B (Methodological),* 44, 234–243.

Posner, M.J., & Boies, S.J. (1972). Components of attention. *Psychological Review, 78*, 391–408.

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

Ratcliff, R. (1988). Continuous versus discrete information processing: modeling accumulation of partial information. *Psychological Review, 95*, 238–255.

Ranger, J. & Ortner, T. (2011). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, DOI: 10.1111/j.2044-8317.2011.02032.x

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*, 321–349.

Roskam, E.E. (1997). Models for speed and time-limit tests. In w.j. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory*(pp. 187–208). New York: Springer.

Rounder, J.N., Sun, D., Speckman, P.L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions, *Psychometrika, 68*, 589–606.

Sargent, D.J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics, 54*, 1486–1497.

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

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

Schnipke, D.L. ,& Scrams,D.J. (2002). *Representing response time information in item banks.* (LSAC Computerized Testing Report No. 97-09). Newton, PA: Law School Admission Council.

Sharef, E., Strawderman, R.L., Ruppert, D., Cowen, M., & Halasyamani, L. (2010). Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic Journal of Statistics, 4*, 606–642.

Siem, F. (1996). The use of response latencies to enhance self-report personality measures. *Military Psychology, 8*, 15–27.

Singer, J.D. ,& Willett, J.B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18*, 155–195.

Sinharay, S., & Johnson, M. (2002). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models (ETS research report RR-0328).* Princeton, NJ: Educational Testing Service.

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

Smith, A.F.M., & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society, Series B, 55*, 3–23.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*, 583–593.

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

Therneau, T.M., Grambsch, P.M., & Fleming, T.R. (1990). Martingale based residuals for survival models. *Biometrika, 77*, 147–160.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J.Weiss(Eds), *New horizons in testing* (pp.179–203). New York: Academic Press.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annuals of Statistics, 22*, 1701–1762.

van Breukelen, G.J.P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359–391.

van der Linden, W.J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*, 21–29.

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., 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*, 365–384.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using response-time contraints to control for 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 respnse patterns in computerized adaptive testing. *Psychometrika, 68*, 251–265.

van der Mass, H.L.J., Molenaar, D., Maris, G., Kievit, R.A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118*, 339–356.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

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.

Vorberg, D., & Ulrich, R. (1987). Random search with unequal rates: serial and parallel generalizations of McGill's model. *Journal of Mathematical Psychology, 31*, 1–23.

Wang, C., Fan, Z., Chang, H., & Douglas, J. (under review). Semiparametric Modeling of Response Time in Computerized Testing. *Journal of Educational and Behavioral Statistics.*

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

Wenger, M. & Gibson, B. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology, 30*, 708–719.

Ying, Z., & Chang, H. (2005, April). *Modeling response latencies for computerized adaptive tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.