Statistics and Its Interface Volume 15 (2022) 91–104

A generalized semi-parametric model for jointly analyzing response times and accuracy in computerized testing

Fang Liu, Jiwei Zhang , Ningzhong Shi*, and Ming-Hui Chen*

The Cox proportional hazards model has been widely used for modeling response-time data in educational and psychological research. However, based on the Kaplan-Meier (KM) plots in an empirical example, we find that the proportionality of the hazard ratios does not seem to be an appropriate assumption, and there are considerable differences in survival rates among different items. To overcome such a problem, we consider a class of flexible nonproportional hazards models known as the generalized odds-rate hazards class of regression models. This class is general enough to include several commonly used models, including the proportional hazards model and the proportional odds model, as special cases. A fully Bayesian method is developed for parameter estimation and the deviance information criterion (DIC) and the logarithm of the pseudomarginal likelihood (LPML) are employed for model comparison. Simulation studies are conducted and a detailed analysis of the Programme for International Student Assessment (PISA) science data is carried out to further illustrate the proposed methodology.

Keywords and phrases: Cox model, DIC, GORH models, Item response theory (IRT), LPML, MCMC.

1. INTRODUCTION

In the computer-based testing, response-time data are often collected as a byproduct. The importance of response-time has attracted the attention of education psychologists. Response-time is useful since it can be an important source of information about the performance of subjects [20, 39]. The new research directions of inquiry have been opened up in order to analyze the response-time data, which include identifying test takers with aberrant response behavior [35, 38, 22, 46], selecting items adaptively in computerized testing [37, 45], controlling test administration time [42, 40], and assembling tests [41]. However, one of the premises to explore these applications of response-time is how an appropriate psychometric model can be built and what valuable psychological phenomena can be obtained to guide the practice.

There are a variety of distributions for the response-time proposed in the literature. The log-normal distribution, one of the most important parametric models, proposed by Furneaux [13] for the first time in the literature, is frequently used to model the response-time data in psychometrics [34, 39, 38, 12, 21]. The other parametrical models such as the Box-Cox normal model [18] and the shifted Weibull distribution [30] have been proposed. Although the parametric models have the advantage of conciseness [45], the distributional assumptions under these models may not hold for response-time data from real applications. Over the last two decades, the semi-parametric proportional hazards (PH) model [25, 26, 27, 19, 45] has played a dominant role in modeling response-time data. The flexibility of the model has been widely accepted by education psychologists compared to the distributional assumption being limited to a specific parametric model [45, 10, 48, 19, 17]. However, a drawback of the PH model is that for a given item, the hazard ratio for two examinees is constant over time. The hazard ratios of the two groups of examinees who have the same ability level are often different between the initial and later stages to respond the item with some limitations, for example, a limit maximum time. It is obviously not practical to assume the constant hazard ratio.

In this paper, we propose a class of flexible and item-specific nonproportional hazards models, namely, generalized odds-rate hazards models (GORH), which is an extension of the model presented in Range and Kuhn [25, 28]. The generalized odds-rate class of regression models is not new, which has been widely used to fit the survival data [3, 6, 24, 5, 32, 2]. GORH overcomes the limitations of the proportional hazards model and more flexible than the PH model. The class of GORH's is governed by a set of non-proportionality parameters, which are general enough to include the PH model and the proportional odds (PO) model as the special cases. GORH reduces to the model considered in Ranger and Kuhn [28] by setting the non-proportionality parameters to be the same across all items. In addition, the item-specific GORH model allows us to fit a different model for a different item. If γ_j approaches to 0, GORH reduces to the PH model, and if $\gamma_j = 1$, GORH reduces to the PO

*Co corresponding authors.

model for the jth item. Thus, the proposed item-specified GORH model provides a great flexibility to fit the response-time data from real applications. GORH can also be viewed as a mixture model of PH with a gamma frailty, the connection between GORH and the frailty model immediately leads us to develop a modeling strategy for the "baseline" hazard function, and to facilitate an easy implementation of an efficient computational algorithm to carry out Bayesian inference. Furthermore, when we construct the item-specific GORH with a piecewise constant baseline hazard function, the joint models of the response and response-time are not identifiable due to the trade off relationships between the speed parameter, time intensity parameter, time discrimination parameter and the baseline function. In this article, we carefully characterize this non-identifiability issue and then provide an easy solution to resolve this issue. We empirically show that the proportional hazards assumption does not hold for the Programme for International Student Assessment (PISA) science data. PISA is the Organization for Economic Cooperation and Development (OECD) Programme for International Student Assessment, which measures the 15-year-old subject's ability to use his/her reading, mathematics and science knowledge and skills to meet real-life challenges. More details can be found at http://www.oecd.org/pisa/. Thus, GORH is more desirable to fit the PISA data.

The rest of the article is organized as follows. Section 2 introduces commonly used PH and PO models and presents the detailed development and the properties of a new flexible and item-specific GORH model. Section 3 presents the IRT model, the hierarchical model framework, and the joint likelihood of the responses and response-time. Section 4 is devoted to the specification of priors, the construction and computation of Bayesian model comparison criteria. In Section 5, three simulation studies are conducted to examine the empirical performance of the Bayesian model selection criteria and the parameter recovery. An in-depth analysis of the PISA science data is carried out in Section 6. Section 7 concludes the article with a brief discussion.

2. IRT MODEL AND A GENERALIZED SEMI-PARAMETRIC RESPONSE-TIME MODEL

2.1 Item response theory model

Let y_{ij} denote the response of individual i answering item j. We assume a two parameter logistic model (2PLM) [43] for y_{ij} given by

(1)
$$p_{ij} \equiv P_j(\theta_i^*) = P(y_{ij} = 1 \,|\theta_i^*, a_j, b_j\,) = \frac{\exp\{a_j(\theta_i^* - b_j)\}}{1 + \exp\{a_j(\theta_i^* - b_j)\}}$$

for $i = 1, ..., N$ and $j = 1, ..., J$. In (1), the correct response probability is represented by p_{ij}, θ_i^* denotes the ability of individual i, and a_j and b_j indicate discrimination and difficulty parameters for item j, respectively. The discrimination parameters distinguish the subjects with different abilities in answering the items correctly while the difficulty parameters describe the extent of easiness for subjects to respond the items correctly.

2.2 A generalized semi-parametric response-time model

The response-time, as discussed in the literature, depend on the impact of two factors. On the one hand, the response-time may depend on individual's response speed, that is, the individual who answers quickly needs less time to complete the item. On the other hand, the response-time may depend on the difficulty level of the item, i.e., the more difficult item may require longer time to answer. Next, two most commonly used response time models in survival analysis, i.e., PH and PO, will be introduced. Moreover, our GORH and its properties will also be discussed.

2.2.1 Proportional hazards models

In the proportional hazards model, the hazard rate $\lambda(t)$ is used to characterize the response-time distribution, which denotes the probability that an event will occur in the next instant given that the event has not yet occurred. Let T be the random variable representing the response-time. Also let $f(t)$ be the probability density function of the response-time T and let $F(t)$ be the corresponding cumulative distribution function. Therefore, the survival function is defined as the probability that the response-time exceeds t, that is, $S(t) = P(T > t)$ for $t > 0$. The hazards rate can be defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \,|T \ge t\,)}{\Delta t} = -\frac{dS(t)/dt}{S(t)}.$$

Let τ_i^* be the latent speed parameter for individual i, and also let ς_j^* and ϕ_j represent the time intensity and the discriminating parameter of item j, respectively. Then, the hazard rate is given by

(2)
$$\lambda(t; \tau_i^*, \phi_j, \varsigma_j^*) = \lambda_0^*(t) \exp\left\{\phi_j(\tau_i^* - \varsigma_j^*)\right\},$$

where $\lambda_0^*(t)$ is the baseline hazard rate. The individual and item characteristics that affect the response-time into the exponent are introduced to make the parameter part more fully explaining the change of response-time. Given the individual speed parameter τ_i^* and the positive time discrimination parameter ϕ_j, the survival probability increases as the time intensity ς_j^* gets higher. Similarly, given the time intensity parameter ς_j^* and time discrimination parameter ϕ_j, the survival probability decreases as τ_i^* increases. In terms of the survival function $S(t; \tau_i^*, \phi_j, \varsigma_j^*)$, PH can be equivalently written as

(3)
$$\log[-\log\left\{S(t; \tau_i^*, \phi_j, \varsigma_j^*)\right\}] = \varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*),$$

where $\varphi_0(t) = \log\left\{\int_0^t \lambda_0^*(u)du\right\}$ and $\lambda_0^*(t)$ is defined in (2). Note that the linear predictor, $\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)$, is connected to the survival function via the log-log link.

2.2.2 Proportional odds model

An alternative formulation of the survival function is

$$-\operatorname{logit}\left\{S(t;\tau_i^*,\phi_j,\varsigma_j^*)\right\} = \varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*), \tag{4}$$

which gives a widely used proportional odds regression model in survival analysis. The large sample properties of the estimates of the regression coefficients under this proportional odds regression model are examined in [6] and [49] in the simple two sample case.

2.2.3 Generalized odds-rate hazard model

We now derive a generalized form of the preceding response-time models. A natural generalization of (3) and (4) is

$$\operatorname{link}\left\{S(t;\tau_i^*,\phi_j,\varsigma_j^*)\right\} = \varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*),$$

where link$(\cdot)$ is a decreasing function, $\varphi_0(t)$ is a completely unspecified nondecreasing and differentiable function, which maps the positive half-line to the whole real line, so that $\lim_{t\to 0}\varphi_0(t) = -\infty$ and $\lim_{t\to\infty}\varphi_0(t) = \infty$. By taking

$$\operatorname{link}^{-1}\{w_{ij}(t)\} = \left[1 + \gamma_j \exp\{w_{ij}(t)\}\right]^{-\gamma_j^{-1}},$$

where $w_{ij}(t) = \varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)$ and $\gamma_j > 0$, we obtain the following survival function

$$S(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j) = \left[1 + \gamma_j \exp\left\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\right\}\right]^{-\gamma_j^{-1}} \tag{5}$$

and the corresponding probability density function

$$f(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j) = \frac{\exp\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\}\varphi_0'(t)}{\left[1 + \gamma_j \exp\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\}\right]^{1+\gamma_j^{-1}}}$$

for $t > 0$. Equation (5) defines a class of GORHs governed by an item-specific nonproportionality parameter γ_j. The PH and PO models can be obtained by taking $\gamma_j \to 0$ and $\gamma_j = 1$, respectively. In (5), $\varphi_0(t)$ controls the form of the baseline hazard type of function. The model in (5) belongs to a class of frailty transformation models considered in [50, 9]. The following propositions summarize the relationships between GORH and PH as well as between GORH and PO.

Proposition 1 *Suppose $S_{ij}(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j)$ is given by (5). Then, we have the following results:*

(1) *The hazard rate is given by*

$$\lambda\left(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j\right) = \frac{\varphi_0'(t)\exp\left\{\varphi_0(t)+\phi_j(\tau_i^*-\varsigma_j^*)\right\}}{\left[1+\gamma_j\exp\left\{\varphi_0(t)+\phi_j(\tau_i^*-\varsigma_j^*)\right\}\right]}$$

for $t > 0$, $\gamma_j > 0$, where $\varphi_0'(t) = \frac{d}{dt}\varphi_0(t) > 0$.

(2) *As $\gamma_j \to 0$,*

$$\lambda(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j) \to \varphi_0'(t)\exp\left\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\right\}$$

giving PH.

(3) *If $\gamma_j = 1$, then*

$$\frac{1 - S_{ij}(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j)}{S_{ij}(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j)} = \exp\left\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\right\}$$

giving PO.

Proposition 2 GORH can also be considered as a generalized proportional odds model. The class of models given by (5) implies

$$\frac{1 - S^{\gamma_j}(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j)}{S^{\gamma_j}(t;\tau_i^*,\phi_j,\varsigma_j^*,\gamma_j)} = \gamma_j \exp\left\{\varphi_0(t) + \phi_j(\tau_i^* - \varsigma_j^*)\right\}$$

for $t > 0$, $\gamma_j > 0$.

The proofs of Propositions 1 and 2 are straightforward. Let $\Lambda_0(t) = \exp\{\varphi_0(t)\}$, which can be viewed as the baseline cumulative hazard type of function. We rewrite $\Lambda_0(t) = \int_0^t \lambda_0^*(u)du$, where $\lambda_0^*(t)$ is a nonnegative function so that $\int_0^\infty \lambda_0^*(u)du = \infty$. The corresponding probability density function can be rewritten as

$$f_{ij}(t) = -S_{ij}'(t) = \frac{\exp\left\{\phi_j(\tau_i^* - \varsigma_j^*)\right\}\lambda_0^*(t)}{\left[1 + \gamma_j \exp\left\{\phi_j(\tau_i^* - \varsigma_j^*)\right\}\Lambda_0(t)\right]^{1+\gamma_j^{-1}}}. \tag{6}$$

We now discuss an identifiability issue between the baseline hazard type of function $\lambda_0^*(t)$ and the item time intensity parameters ς_j^*'s. To see this, we consider the following reparametrization:

$$\lambda_0(t) = \lambda_0^*(t)/c_0, \Lambda_0^*(t) = \int_0^t \lambda_0(u)du, \varsigma_j = \varsigma_j^* - \frac{\log(c_0)}{\phi_j}, \tag{7}$$

for any $c_0 > 0$. Then (6) is equivalent to

$$f_{ij}(t) = \frac{\exp\left\{\phi_j(\tau_i^* - \varsigma_j)\right\}\lambda_0(t)}{\left[1 + \gamma_j \exp\left\{\phi_j(\tau_i^* - \varsigma_j)\right\}\Lambda_0^*(t)\right]^{1+\gamma_j^{-1}}}. \tag{8}$$

Therefore, the baseline hazard type of function $\lambda_0^*(t)$ in (6) is not identifiable as the form of the density in (8) is identical to (6) after the reparametrization (7). In other words, $\lambda_0^*(t)$ is identifiable up to a constant multiplier $c_0 > 0$.

We consider a piecewise constant baseline hazard type of function for $\lambda_0(t)$. Specifically, we construct a finite partition of the time axis, $0 = s_0 < s_1 < s_2 < ... < s_V$, with $s_V > t_{ij}$ for all individuals and items. We then assume that $\lambda_0(t) = \lambda_v$ when $t \in (s_{v-1}, s_v]$ for $v = 1, ..., V$. To ensure model identifiability, we assume $\lambda_1 = 1$. This assumption implies that we take $c_0 = \lambda_1^*$ when $\lambda_0^*(t) = \lambda_v^*$, $t \in (s_{v-1}, s_v]$, for

$v = 1, ..., V$. We may also view λ_v as a relative baseline hazard rate over the baseline hazard rate defined on the first interval $(0, s_1]$ for $v = 2, \ldots, V$. We use this modeling strategy for the GORH models as well as the PH and PO models.

3. HIERARCHICAL MODEL FRAMEWORK AND LIKELIHOOD

3.1 Hierarchical model framework

It has been widely recognized that the response-time should be modeled with item responses associated with the item response theory model. Based on the existing response-time models [34, 44, 47, 11, 36, 39, 23], we employ the hierarchical modeling method to construct the higher level model, where the responses and response-times are modeled in the first level, respectively, and then assume the multivariate normal distribution to capture the relation between the participant's parameters.

The individual parameters $\boldsymbol{U_i} = (\theta_i^*, \tau_i^*)'$ are assumed to follow a bivariate normal distribution, that is,

$$\boldsymbol{U_i} \sim MVN(\boldsymbol{\mu_P}, \boldsymbol{\Sigma_P})$$

with mean vector and covariance matrix as follow, respectively,

$$\boldsymbol{\mu_P} = (\mu_\theta, \mu_\tau)', \boldsymbol{\Sigma_P} = \begin{pmatrix} \sigma_\theta^2 & \eta\sigma_\theta\sigma_\tau \\ \eta\sigma_\theta\sigma_\tau & \sigma_\tau^2 \end{pmatrix}.$$

To ensure identifiability, the following constraints are imposed. The locations and scales of θ_i^* and τ_i^* are fixed as $\mu_\theta = 0$, $\sigma_\theta^2 = 1$, $\mu_\tau = 0$, and $\sigma_\tau^2 = 1$, where the first two constraints are standard in the IRT parameter estimation and the latter two constraints fix the location and the scale of τ_i^*. In order to give a more clear interpretation and facilitate Bayesian computation, a reparameterization is adopted. To be specific, $(\theta_i^*, \tau_i^*)' = \boldsymbol{\Gamma}(\theta_i, \tau_i)'$ with $\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 0 \\ \sin\varphi & \cos\varphi \end{pmatrix}$. Then θ_i and τ_i are independent to each other and each follows a standard normal distribution $N(0, 1)$. Let $\eta = \sin\varphi$, which describes the correlation between the untransformed ability θ_i^* and speediness τ_i^*.

3.2 Joint likelihood of response and response-time

Let t_{ij} denote the response-time of individual i answering item j. Write $\boldsymbol{Y} = (y_{ij}; i = 1, \ldots, N, j = 1, \ldots, J)$ and $\boldsymbol{T} = (t_{ij}; i = 1, \ldots, N, j = 1, \ldots, J)$. For any t_{ij}, there exists v_{ij} such that t_{ij} belongs to the interval $\left(s_{v_{ij}-1}, s_{v_{ij}}\right]$. Hence, the baseline hazard type of rate at t_{ij} is $\lambda_0(t_{ij}) = \lambda_{v_{ij}}$. Then, the joint likelihood function can be written as

(9)

$$\begin{aligned} & L(\boldsymbol{Y}, \boldsymbol{T}|\boldsymbol{\Omega}) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{J}\Bigg(\lambda_{v_{ij}} \exp\{\phi_j(\theta_i \sin\varphi + \tau_i \cos\varphi - \varsigma_j)\} \\ &\times\Bigg[1 + \gamma_j\Bigg\{\lambda_{v_{ij}}(t_{ij} - s_{v_{ij}-1}) + \sum_{g=1}^{v_{ij}-1}\lambda_g(s_g - s_{g-1})\Bigg\} \\ &\exp\{\phi_j(\theta_i \sin\varphi + \tau_i \cos\varphi - \varsigma_j)\}\Bigg]^{-1-\gamma_j^{-1}\frac{\exp\{y_{ij}a_j(\theta_i - b_j)\}}{1+\exp\{a_j(\theta_i - b_j)\}}}\Bigg), \end{aligned}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\varsigma}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \varphi)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$, $\boldsymbol{a} = (a_1, \ldots, a_J)'$, $\boldsymbol{b} = (b_1, \ldots, b_J)'$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)'$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)'$, $\boldsymbol{\varsigma} = (\varsigma_1, \ldots, \varsigma_J)'$, and $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_V)'$ as $\lambda_1 = 1$.

4. BAYESIAN INFERENCE

In this section, we present the specification of the priors, the formulation and computation of the posterior distribution, and the development of the Bayesian model assessment measures for carrying out Bayesian inference under the proposed item-specific GORH model. Compared to the maximum likelihood based inference [26], the Baysian inference has several advantages. First, due to the recent advance in Bayesian computation, especially the development of Markov chain Monte Carlo (MCMC) sampling, it becomes computationally feasible to sample from the analytically intractable posterior distribution with the potentially high-dimensional latent ability and speed parameters. Second, MCMC sampling enables us to make exact inference without sorting to asymptotic calculations. In particular, under the maximum likelihood based approach, the variance estimates require asymptotic arguments and complicated derivations while the posterior standard deviations as well Bayesian credible intervals are the byproducts of MCMC sampling. Third, the Bayesian model assessment measures such as the deviance information criterion (DIC) [33] and the logarithm of the pseudomarginal likelihood (LPML) [14, 16] can be conveniently computed via MCMC sampling while the computation of the Akaike's information criterion (AIC) [1] is quite challenging in the presence of high-dimensional latent ability and speed parameters.

4.1 The prior and posterior distributions

For each of the discrimination parameter a_j and the time discrimination parameter ϕ_j, we assume a log-normal distribution, i.e., $\log a_j \sim N(0, 1)$ and $\log \phi_j \sim N(0, 1)$ for item j. The prior distribution for γ_j is assumed to be an inverse Gamma distribution $IG(\upsilon_1, \omega_1)$ with density $\pi(\gamma_j) \propto \gamma_j^{-(\upsilon_1+1)} \exp(-\omega_1/\gamma_j)$, where $\upsilon_1 = \omega_1 = 1$. A hierarchical normal prior $N(\mu_b, \sigma_b^2)$ is specified for the item difficulty parameter b_j, where $\mu_b \sim N(0, 10\sigma_b^2)$ and σ_b^2 follows

an inverse gamma distribution $IG(0.01, 0.01)$. A normal distribution $N(0, 100)$ is assumed for ς_j. In addition, the hyperparameters σ_θ^2 and σ_τ^2 are set to 1 to ensure identifiability. For the correlation term φ, we specify a uniform distribution, which is $U(-\pi/2, \pi/2)$. For the piecewise constant λ_v, we assume an inver-gamma distribution, i.e. $IG(.01, .01)$. These priors are denoted by $p(\theta_i)$, $p(\tau_i)$, $p(a_j)$, $p(b_j|\mu_b, \sigma_b^2)$, $p(\phi_j)$, $p(\varsigma_j)$, $p(\gamma_j)$, $p(\lambda_v)$, $p(\varphi)$, and $p\left(\mu_b, \sigma_b^2\right)$. Then, the posterior distribution can be written as

$$
\begin{aligned}
p(\boldsymbol{\Omega}\,|\boldsymbol{Y}, \boldsymbol{T}) \propto\, & L(\boldsymbol{Y}, \boldsymbol{T}\,|\boldsymbol{\Omega}) \prod_{i=1}^{N} p(\theta_i) p(\tau_i) \\
& \times \prod_{j=1}^{J} p(b_j|\mu_b, \sigma_b^2) p(\varsigma_j) p(a_j) p(\phi_j) p(\gamma_j) \prod_{v=2}^{V} p(\lambda_v) \\
& \times p(\varphi) p\left(\mu_b, \sigma_b^2\right),
\end{aligned} \tag{10}
$$

where the likelihood function $L(\boldsymbol{Y}, \boldsymbol{T}\,|\boldsymbol{\Omega})$ is defined in (9).

We develop the codes using an `R` package `NIMBLE` [7, 8] to sample from the posterior distribution given in (10). The key NIMBLE codes used in the simulation studies and the real data analysis are given in Section S.1 of the Supplementary Materials, http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0001/SII-2022-0015-0001-s002.pdf.

4.2 Bayesian model assessment

We consider DIC and LPML to compare different models. These two criteria are based on the log-likelihood functions evaluated at the posterior samples of model parameters. Therefore, DIC and LPML can be easily computed. Let $\left\{\boldsymbol{\Omega^{(1)}}, ..., \boldsymbol{\Omega^{(R)}}\right\}$, where $\boldsymbol{\Omega^{(r)}} = \{(\boldsymbol{\Omega_{ij}^{(r)}}, \varphi^{(r)}); i = 1, ..., N, j = 1, ..., J\}$, $\boldsymbol{\Omega_{ij}^{(r)}} = \left(\theta_i^{(r)}, a_j^{(r)}, b_j^{(r)}, \tau_i^{(r)}, \phi_j^{(r)}, \varsigma_j^{(r)}, \lambda_{vij}^{(r)}, \gamma_j^{(r)}\right)'$ for $r = 1, ..., R$, denote an MCMC sample from the posterior distribution in (10). Define

$$
\mathrm{Dev}_{ij}(\boldsymbol{\Omega}) = -2 \log f(y_{ij}, t_{ij}\,|\boldsymbol{\Omega}). \tag{11}
$$

The log-likelihood, $\log f(y_{ij}, t_{ij}\,|\boldsymbol{\Omega})$, is readily available from MCMC sampling outputs. The deviance information criterion is defined as

$$
\mathrm{DIC} = \mathrm{Dev}(\widehat{\boldsymbol{\Omega}}) + 2\mathrm{P_D},
$$

where $\widehat{\boldsymbol{\Omega}}$ denotes the posterior means of $\boldsymbol{\Omega}$, $\mathrm{Dev}(\boldsymbol{\Omega}) = \sum_{i=1}^{N}\sum_{j=1}^{J} \mathrm{Dev}_{ij}(\boldsymbol{\Omega})$ is the deviance function, and $\mathrm{P_D} = E\left[\mathrm{Dev}(\boldsymbol{\Omega})\,|\boldsymbol{Y}, \boldsymbol{T}\right] - \mathrm{Dev}(\widehat{\boldsymbol{\Omega}})$ is the effective number of model parameters. The model with a smaller DIC has a better fit to the data. For the ith subject and the jth item, we define $\mathrm{DIC_{ij}} = \mathrm{Dev_{ij}}(\widehat{\boldsymbol{\Omega}}) + 2\mathrm{P_{D_{ij}}}$, where $\mathrm{P_{D_{ij}}}$ is defined in a similar fashion as $\mathrm{P_D}$. Letting $U_{ij,\max} = \max_{1\le r\le R}\left\{-\log f(y_{ij}, t_{ij}\,|\boldsymbol{\Omega_{ij}^{(r)}})\right\}$, a Monte Carlo estimate of the conditional predictive ordinate [15, 4] is given by

$$
\begin{aligned}
\log\left(\widehat{\mathrm{CPO}_{ij}}\right) = & -U_{ij,\max} \\
& - \log\left[\frac{1}{R}\sum_{r=1}^{R} \exp\left\{-\log p(y_{ij}, t_{ij}\,|\boldsymbol{\Omega_{ij}^{(r)}}) - U_{ij,\max}\right\}\right].
\end{aligned} \tag{12}
$$

Note that the maximum value adjustment used in $\log\left(\widehat{\mathrm{CPO}_{ij}}\right)$ plays an important role in numerical stabilization in computing $\exp\left\{-\log p(y_{ij}, t_{ij}\,|\Omega_{ij}^{(r)}) - U_{ij,\max}\right\}$ in (12). A summary statistic of the $\widehat{\mathrm{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and given by

$$
\mathrm{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J} \log\left(\widehat{\mathrm{CPO}_{ij}}\right).
$$

The model with a larger LPML has a better fit to the data.

5. SIMULATION STUDY

Two factors are considered in the simulation design. The first factor is the number of individuals, which is varied in two levels (N =500, 1000), and the second factor is the number of items, which is varied in two levels (J = 20, 40). The 2PLM and GORH with different γ_j's for different items and a constant baseline type of hazard function (i.e., $\lambda_1 = 1$ and $V = 1$) are used, respectively, to generate the response and response-time data. Four different levels (0.25, 0.75, 1.25, 1.75) are selected for γ_j, and each of these values accounts for 25% of the items. Three response time models, PH, GORH with same γ for all items (GORHS), and GORH with different γ_j's, are used to fit the simulated data.

The true values for the discrimination parameters a_j's is generated from a uniform distribution $U(0.5, 1.5)$ and the time discrimination parameters ϕ_j's is generated from a uniform distribution $U(0.7, 1.3)$ for item j. The latent ability parameter θ_i and the speed parameter τ_i are generated from $N(0, 1)$ for individual i, respectively. Further, let the correlation related parameter φ be 0.5. In addition, the true values of the difficulty parameter b_j and the time intensity parameter ς_j are also generated from $N(0, 0.5)$. To implement the MCMC sampling algorithm, chains of length 10000 with a burn-in period of 2000 are chosen. There are 500 replications for each simulation condition.

5.1 Simulation 1

This simulation aims to evaluate whether PH for the response-time is appropriate to fit the simulation data under the condition of four different sample sizes. The results of Bayesian model assessment for the DIC and LPML differences between the fitted model and the true model are shown in Table 1, where Q_1 and Q_3 denote the respective 25% and 75% quantiles of the DIC and LPML differences

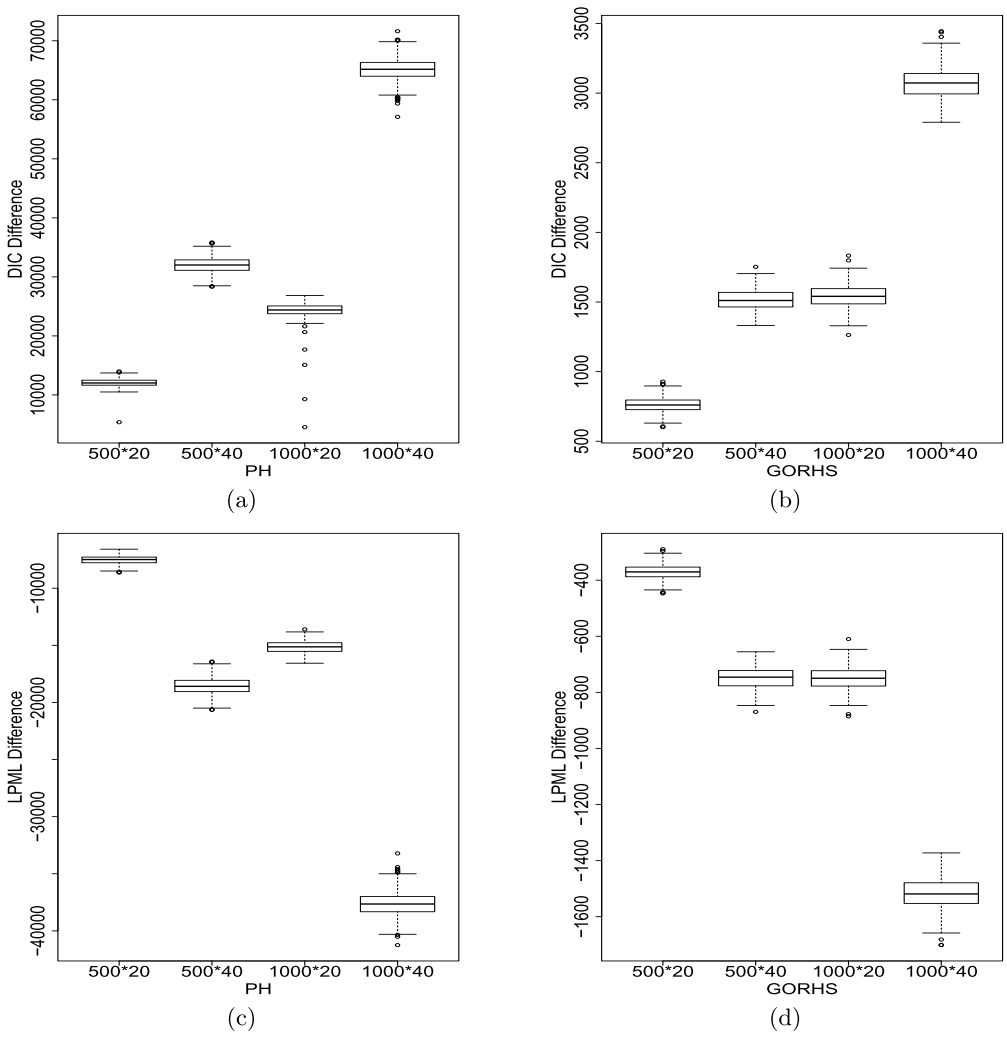

Figure 1. The boxplots of the DIC differences between PH and GORH (a), the DIC differences between GORHS and GORH (b), the LPML differences between PH and GORH (c), and the LPML differences between GORHS and GORH (d) in Simulation 1.

across simulation replicates. The corresponding boxplots are showed in Figure 1. Both Table 1 and Figure 1 show the PH model is the worst fitting model and the GORH model is the best fitting model. The above results indicate that traditional PH is not appropriate in fitting such simulated data compared with GORH and GORHS. Moreover, GORH can tailor each item to a different distribution rather than a rigid assumption that all items have the same non-proportional parameter, yielding a better fit than GORHS. In application, it may be necessary to introduce a new response-time model so that the model is more flexible by adjusting the non-proportional parameter. Our GORH just fills the gap.

As a part of this simulation study, for the simulated data with 500 subjects and 20 items from the GORH with different γ_j's, we also fit the GORH models with different piecewise constant baseline hazard type of functions. Specifically, we consider $V = 1$, $V = 10$, and $V = 20$ for the piecewise constant hazard type of function. Our goal is to examine the empirical performance of DIC and LPML in selecting the true model with $V = 1$. Among the 500 replicates, the counts for selecting the GORH models with V =1, 10, and 20 pieces are 434, 59 and 7, respectively, according to the DIC; and 479, 19 and 2, respectively, according to the LPML. These results indicate the good empirical performance of the DIC and LPML in selecting the true baseline hazard type of function under the GORH model. We note that the LPML slightly outperforms the DIC since the correct rate for selecting the true model by the LPML is 95.8%, which is higher than 86.8% by the DIC.

5.2 Simulation 2

This simulation study is conducted to evaluate the recovery performance of the posterior estimates under GORH

Table 1. The results of Bayesian model assessment in Simulation 1

Fitting Model	Sample Size	DIC Difference			LPML Difference		
		Median	Q_1	Q_3	Median	Q_1	Q_3
PH	500×20	12022.09	11646.16	12496.44	−7491.49	−7771.33	−7275.88
	500×40	32010.78	31112.01	32877.80	−18592.48	−19053.43	−18066.62
	1000×20	24391.94	23758.87	25079.43	−15135.10	−15526.83	−14775.41
	1000×40	65175.29	64000.73	66340.55	−37647.55	−38321.68	−36995.35
GORHS	500×20	760.34	727.14	796.06	−370.47	−387.70	−353.27
	500×40	1510.95	1464.81	1569.15	−745.82	−776.59	−722.04
	1000×20	1540.87	1486.77	1595.93	−749.67	−777.43	−722.89
	1000×40	3072.91	2995.07	3140.44	−1518.84	−1552.58	−1479.10

model. Five indexes are used to assess the accuracy of the parameter estimates. Let ϑ be the parameter of interest. We generate $M = 500$ data sets. Also, let $\widehat{\vartheta}^{(m)}$ and $\mathrm{SD}^{(m)}$ denote the posterior mean and the posterior standard deviation of ϑ obtained from the mth simulated data set for $m = 1, ..., M$. The bias and the mean squared error (MSE) are defined as $\mathrm{Bias} = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta)$ and $\mathrm{MSE} = \frac{1}{M}\sum_{m=1}^{M}(\widehat{\vartheta}^{(m)} - \vartheta)^2$. The simulation standard error (SE) is the square root of the sample variance of the posterior estimates over different simulated data sets, which is given by $\mathrm{SE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\vartheta}^{(m)} - \frac{1}{M}\sum_{\ell=1}^{M}\widehat{\vartheta}^{(\ell)}\right)^2}$, and the average of the posterior standard deviations is defined as $\mathrm{SD} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{SD}^{(m)}$. The coverage probability (CP) is computed as $\mathrm{CP} = \frac{1}{M}$ (the number of 95% HPD intervals containing ϑ in M simulated datasets). The true value, average absolute Bias, MSE, SE, SD and CP for the parameters of interest across items or subjects are reported in Table 2. The following conclusions are drawn. (i) The values of Bias and MSE for the item parameters decrease as the number of individuals increases from 500 to 1000 when $J = 40$. In most cases, the values of Bias and MSE for the item parameters are close to 0. Moreover, for the non-proportional parameters γ_j's that regulate the shape of the distribution, the values of Bias and MSE also decrease when the number of individuals increases. (ii) Based on the results of SD, we find that the values of SD for the item parameters decrease as the number of individuals increases from 500 to 1000, which indicates that the estimation becomes more accurate when the number of individuals increases. The results for the ability and speed parameters show the same trend with the increase of the number of items. Moreover, almost for all of the parameters, the values of SD and SE are similar, which indicates that the true variability of the posterior means across different replications is similar to the estimated variability of the posterior mean in each replication. Consequently, the values of CP are around 95% for almost all of the parameters under the four simulated conditions. All of the recovery results for all different settings are good. The detailed recovery results including the true values for the item parameters under 500×20 are given in Tables S.1 and S.2 of the Supplementary Materials.

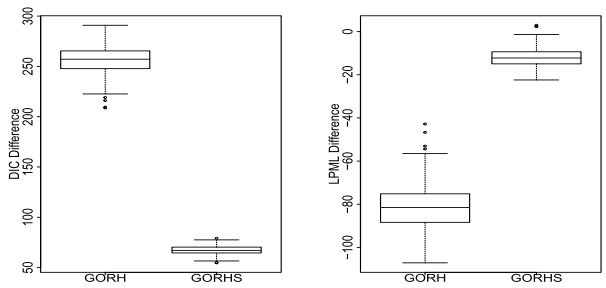

Figure 2. The boxplots in Simulation 3.

5.3 Simulation 3

This simulation focuses on the model performance of the PH, GORH and GORHS model when the data are generated from the PH model. We set $N = 500$ and $J = 20$, and fit the PH, GORH and GORHS models to examine the empirical performance of DIC and LPML and the posterior estimates under the PH model. Figure 2 shows the boxplots for the DIC and LPML differences between the fitted model and the PH model. The values of the median, Q_1 and Q_3 of the DIC differences are 257.13, 247.78, and 265.35, respectively, for comparing GORH to PH; and 67.08, 64.53, and 70.31, respectively, for comparing GORHS to PH. The values of the median, Q_1 and Q_3 of the LPML differences are −81.54, −88.37, and −75.20, respectively, for comparing GORH to PH; and −12.23, −14.93, and −9.35, respectively, for comparing GORHS to PH. These results show that the difference between the PH and GORH models is small, and there is almost no difference between the PH and GORHS models. Table 3 presents the recovery results of the parameters under the PH model, indicating good recovery results.

Table 2. The results of the posterior estimates under the GORH model in Simulation 2

Sample Size	Parameter	True	Bias	MSE	SD	SE	CP
500×20	a	.888	.014	.019	.139	.135	.951
	b	.456	.009	.017	.145	.128	.968
	ϕ	.881	.008	.007	.086	.082	.959
	ς	.401	.021	.017	.132	.122	.959
	γ	1.000	.012	.009	.092	.090	.946
	θ	.790	.183	.223	.477	.404	.950
	τ	.806	.199	.218	.467	.394	.948
	φ	.500	.013	.001	.051	.032	.992
500 × 40	a	.897	.014	.016	.128	.124	.951
	b	.390	.017	.015	.136	.120	.968
	ϕ	.901	.019	.007	.082	.077	.952
	ς	.446	.029	.016	.129	.118	.963
	γ	1.000	.011	.009	.091	.088	.948
	θ	.790	.102	.127	.362	.328	.951
	τ	.806	.118	.124	.355	.319	.949
	φ	.500	.020	.001	.044	.022	.996
1000 × 20	a	.888	.012	.009	.098	.095	.951
	b	.456	.012	.009	.102	.093	.964
	ϕ	.881	.011	.004	.061	.057	.956
	ς	.401	.027	.009	.092	.085	.950
	γ	1.000	.007	.005	.066	.064	.947
	θ	.794	.178	.220	.478	.404	.951
	τ	.798	.194	.215	.465	.395	.948
	φ	.500	.014	.000	.036	.024	.990
1000 × 40	a	.897	.012	.008	.090	.088	.949
	b	.390	.015	.008	.097	.088	.962
	ϕ	.901	.018	.004	.059	.056	.948
	ς	.446	.029	.008	.089	.082	.953
	γ	1.000	.006	.004	.064	.064	.947
	θ	.794	.097	.126	.362	.329	.951
	τ	.798	.113	.123	.353	.320	.949
	φ	.500	.016	.001	.032	.017	.996

Table 3. The results of the posterior estimates under the PH model in Simulation 3

Parameter	True	Bias	MSE	SD	SE	CP
a	.888	.015	.018	.138	.133	.955
b	.456	.029	.015	.131	.113	.965
ϕ	.881	.002	.002	.055	.047	.974
ς	.401	.033	.005	.076	.061	.972
θ	.790	.176	.212	.467	.394	.950
τ	.806	.143	.128	.357	.309	.946
φ	.500	.015	.001	.047	.028	.992

6. EMPIRICAL EXAMPLE

6.1 Data description

In this example, the 2015 computer-based PISA (Program for International Student Assessment) sciences data are used. Among the countries that have participated in the computer-based assessment of sciences, we choose the United States of America (USA). The original sample size is 658, and 110 students with Not Reached (original code 6) or Not Response (original code 9) are removed, where the Not Reached and Not Response (omitted) are treated as missing data. The sample size of the final data is 548 and both the responses and the response-time are available for all of these individuals. All 16 items are scored using a dichotomous scale. The descriptive statistics for this PISA data set are shown in Table 4. We find that the three items, DR442Q05C, DR442Q06C and CR442Q07S, have the lowest correct rates compared to the other items, and their values are 0.257, 0.232 and 0.285, respectively. Moreover, the three items with the highest correct rates are,

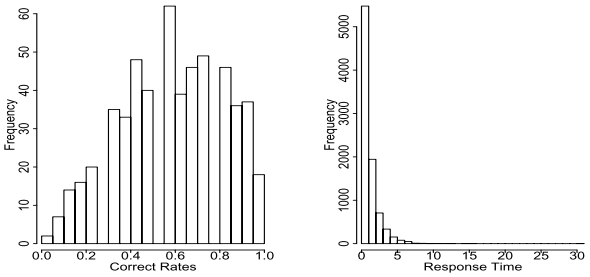

Figure 3. Frequency histograms of the correct rates and the response-times for 548 individuals.

Table 4. The descriptive statistics for PISA 2015 released computer-based sciences items

		Response-Time	
Item	Correct rate	Median	IQR
CR083Q01S	.542	.510	(.31,.95)
CR083Q02S	.836	.700	(.52,.98)
CR083Q03S	.752	.542	(.39,.81)
CR083Q04S	.666	.799	(.53,1.19)
DR442Q02C	.801	2.325	(1.54,3.70)
DR442Q03C	.765	1.594	(.93,2.44)
DR442Q05C	.257	1.379	(.94,1.96)
DR442Q06C	.231	2.498	(1.66,3.73)
CR442Q07S	.285	.358	(.25,.54)
CR245Q01S	.538	1.608	(1.00,2.08)
CR245Q02S	.600	.681	(.40,1.02)
CR101Q01S	.436	.528	(.27,1.22)
CR101Q02S	.876	.227	(.16,.38)
CR101Q03S	.577	.377	(.27,.52)
CR101Q04S	.801	.330	(.22,.68)
CR101Q05S	.487	.587	(.44,.83)

Note that response-time unit is minute.

respectively, CR101Q02S (0.876), CR083Q02S (0.836) and DR442Q02C (0.801). The three most time-consuming items are DR442Q06C, DR442Q02C and CR245Q01S. Their median response-times are respectively 2.498 minutes, 2.325 minutes and 1.608 minutes. In addition, the frequency histogram of the correct rates for 548 examinees and the corresponding frequency histogram of the response-times are shown in Figure 3.

6.2 Bayesian model comparison

Three models, PH, GORH and GORHS, are used to fit the data. Moreover, the three fitted models with various choices of the piece values (V) are considered, where $V = 5, 15, 25, 35, 45, 55, 65, 80, 100$. Following Ibrahim et al. [16], Rizopoulos [29], and Zhang et al. [51], we use the popular equally-spaced quantile partition (ESQP) to construct the partition of the time axis, $0 = s_0 < s_1 < s_2 < ... < s_V$, for the piecewise constant hazard type of function. In the estimation procedure, the same hyper-parameter values that are used as in Simulation are utilized to specify prior distributions for unknown parameters. In all of the Bayesian computations, we have run 80,000 MCMC samples with a burn-in of 30,000 iterations and thinned the MCMC samples for every 5 steps for each model to compute all posterior estimates.

According to DIC and LPML as shown in Table 5, we find that GORH with 65 pieces are the best fitting model compared to the other models. The values of DIC and LPML are 20210.5 and −10135.6, respectively. For the three fitted models with the same number of pieces, PH is always the worst, and the performance of GORH is always the best. Moreover, the fitting results do not show that the more pieces we use for the response-time model, the better fit will be achieved. The model fitting becomes worse after 65 pieces. This indicates that the more dimensional penalty kicks in as the number of pieces increases. The DIC and LPML measures achieve a balance between the goodness-of-fit and the model complexity. Next, we will analyze the PISA data based on GORH.

6.3 Analysis of item parameters

The estimated results for the item parameters are shown in Table 6. From Table 6, we see that the expected *a posteriori* (EAP) estimates of the fourteen item discrimination parameters are greater than one. This indicates that these items can well distinguish the differences between different abilities. In addition, the EAP estimates of the eleven difficulty parameters are less than zero, which indicates that the eleven items are slightly easier than the other items. The five most difficult items are items 8(DR442Q06C), 7(DR442Q05C), 9(CR442Q07S), 12(CR101Q01S) and 16 (CR101Q05S). The EAP estimates of the difficulty parameters for these five items are 1.029, 0.809, 0.758, 0.297 and 0.061, respectively. The corresponding correct rates for these five items shown in Table 4 are 0.231, 0.257, 0.285, 0.436 and 0.487, respectively. The most difficult five items have the lowest correct rates, which is consistent with our intuition. The eleven EAP estimates of the time intensity parameters are larger than two. According to the medians of the response-times in Table 4, the three items with the longest response-time are 8, 5 and 10, and the corresponding medians of the response-times are respectively 2.498, 2.325 and 1.608. The estimates of the time intensity parameters for these three items are 5.011, 5.631 and 6.283, respectively. This indicates that the factors of the items, such as the length of item, have a great influence on the response-time. Moreover, we find that the time intensity parameter for item 13 (CR101Q02S), ς_{13}, has the smallest estimated value of 0.012, while the difficulty parameter, b_{13}, also has the small estimated value of −1.668. This may be due to the fact that the item is so easy that the item can be answered correctly and quickly.

Based on the sixteen non-proportional parameters, we observe that the estimated values of all the parameters are

Table 5. DIC and LPML with different number of pieces (V) for PISA under different models

	DIC			LPML		
V	PH	GORHS	GORH	PH	GORHS	GORH
5	21678.4	21669.2	21478.3	−10995.9	−10910.6	−10739.8
15	21041.4	20813.7	20503.7	−10730.3	−10465.3	−10275.8
25	20989.0	20666.9	20362.9	-10700.7	−10387.5	−10208.4
35	20962.2	20591.8	20284.0	−10692.3	−10347.3	−10171.1
45	20962.3	20587.0	20280.1	−10691.3	−10348.0	−10170.0
65	20926.6	20524.6	20210.5	−10673.4	−10313.8	−10135.6
80	20962.2	20552.0	20230.9	−10694.2	−10329.2	−10146.1
100	20987.9	20570.0	20247.1	−10713.9	−10335.8	−10154.5

greater than zero, again confirming that the traditional PH is not appropriate. The estimates of the non-proportional parameters are different for different items, so it is obviously not appropriate to fit all items with the same non-proportional parameter. According to the results of the Monte Carlo standard error (MCSE), we see from Table 6 that the MCSEs for a, b, ϕ and γ are less than 0.011. This indicates that our algorithm is convergent. Moreover, the trace plots of the γ_j for some items and the corresponding autocorrelation plots indicate good convergence of the nonproportional parameters.

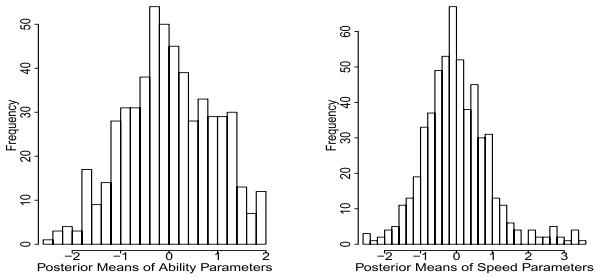

Figure 4. The histograms of the posterior estimates of ability and speed parameters.

6.4 Analysis of individual parameters

The histograms of the posterior estimates of ability and speed parameters are shown in Figure 4. Most of the estimated ability parameters are near zero. The number of the examinees with high ability (the estimates are between 1 and 2) is more than that of the examinees with low ability (the estimates are between −2 and −1). The histogram of the posterior means of the ability parameters is consistent with the frequency histogram of the correct rates (Figure 3), which once again verify that the results of estimation are accurate. Similarly, most of the estimated values of the speed parameters are also near zero. The number of the individuals with high speed is more than that of the individuals with slow speed, and some of them are larger than one. This may be due to the reason that the individuals adopt a rapid guessing strategy to answer the item, which requires a further in-depth analysis of the item and individual.

6.5 Further model assessment

Next, we analyze the data by using Kaplan-Meier (KM) plots. Based on the estimates of speed parameters, a subset of the individuals who answer the sixteen items are selected to form three groups, low speed, medium speed and high speed. Specifically, the low speed group consists of the individuals whose $\widehat{\tau}_i^*$'s are below 20% of all of the estimates of the speed parameters, the middle speed group consists of the individuals whose $\widehat{\tau}_i^*$'s are between 40% and 60% for all of the estimates of the speed parameters, and the high speed group consists of the individuals whose $\widehat{\tau}_i^*$'s are above the upper 20% for all of the estimates of the speed parameters. Then the KM plots of the response times stratified by the items for the lower, middle, and high speed groups are shown in Figures 5.

We analyze the impact of different non-proportional parameters on the survival curves for items 3, 4, 10, and 12 with $\gamma_3 = 0.516$, $\gamma_4 = 0.514$, $\gamma_{10} = 0.616$, and $\gamma_{12} = 2.072$, respectively. We see from Figure 5 that (i) for the lower speed group, the KM curves for items 10 and 12 are significantly different with a Wilcoxon test p-value of 0.000125; (ii) for the middle speed group, the KM curves for items 4 and 12 are significantly different with a Wilcoxon test p-value of 0.000390; (iii) for the high speed group, the KM curves for items 3 and 12 are significantly different with a Wilcoxon test p-value of 0.001494; and (iv) in all three plots, the two KM curves are crossed each other. If PH would fit the data well, the two KM curves between two items would never be crossed each other. These results indicate that PH is not appropriate for the PISA data, which is consistent with the findings based on DIC, LPML, and the posterior estimates of the non-proportional parameters.

In addition, we also compare $\mathrm{DIC_{ij}}$ and $\mathrm{CPO_{ij}}$ under the competitive models for different speed groups. The observation with a smaller value of $\mathrm{DIC_{ij}}$ or a larger value of $\mathrm{CPO_{ij}}$ under one model supports that model over the other. The boxplots of $\mathrm{DIC_{ij}^{PH}} - \mathrm{DIC_{ij}^{GORH}}$ and $\mathrm{DIC_{ij}^{GORHS}} - \mathrm{DIC_{ij}^{GORH}}$ are shown in Figures 6 while the boxplots of $\log \frac{\mathrm{CPO_{ij}^{GORH}}}{\mathrm{CPO_{ij}^{PH}}}$

Table 6. The estimation results of item parameters for the PISA data

PARM	EAP	SD	MCSE	HPD	PARM	EAP	SD	MCSE	HPD
a_1	1.462	.170	.003	[1.131, 1.793]	b_1	-0.166	.084	.002	[−.325, .004]
a_2	1.198	.173	.004	[.873, 1.541]	b_2	−1.714	.201	.004	[−2.120, −1.356]
a_3	1.543	.193	.004	[1.170, 1.923]	b_3	−1.025	.113	.002	[−1.254, −.811]
a_4	1.484	.180	.003	[1.133, 1.836]	b_4	−.660	.096	.002	[−.845, −.472]
a_5	1.274	.170	.004	[.947, 1.615]	b_5	−1.421	.161	.004	[−1.743, −1.120]
a_6	1.788	.224	.005	[1.363, 2.232]	b_6	−1.010	.103	.002	[−1.204, −.806]
a_7	2.433	.318	.006	[1.870, 3.091]	b_7	.809	.081	.002	[.649, 0.967]
a_8	1.762	.221	.004	[1.340, 2.198]	b_8	1.029	.105	.002	[.826, 1.235]
a_9	1.990	.241	.004	[1.532, 2.471]	b_9	.758	.087	.002	[.591, .929]
a_{10}	.938	.128	.002	[.675, 1.175]	b_{10}	−.204	.112	.002	[−.433, .007]
a_{11}	1.467	.171	.003	[1.141, 1.810]	b_{11}	−.394	.086	.002	[−.562, −.224]
a_{12}	1.048	.134	.002	[.789, 1.312]	b_{12}	.297	.105	.002	[.092, .500]
a_{13}	1.696	.242	.005	[1.241, 2.174]	b_{13}	−1.668	.161	.004	[−1.974, −1.346]
a_{14}	1.360	.161	.003	[1.063, 1.690]	b_{14}	−.313	.090	.002	[−.483, −.132]
a_{15}	1.682	.219	.005	[1.253, 2.110]	b_{15}	−1.221	.122	.003	[−1.465, −.991]
a_{16}	.927	.125	.002	[.689, 1.182]	b_{16}	.061	.109	.002	[−.144, .287]
ϕ_1	.767	.092	.003	[.590, .953]	ς_1	2.208	.311	.013	[1.645, 2.841]
ϕ_2	.464	.072	.005	[.310, .592]	ς_2	5.934	1.000	.071	[4.243, 8.177]
ϕ_3	.562	.070	.004	[.428, .700]	ς_3	3.925	.556	.034	[2.888, 4.967]
ϕ_4	.819	.074	.004	[.673, .967]	ς_4	3.709	.369	.022	[3.038, 4.449]
ϕ_5	1.034	.112	.009	[.794, 1.248]	ς_5	5.631	.622	.051	[4.466, 6.923]
ϕ_6	1.292	.104	.007	[1.092, 1.495]	ς_6	3.642	.299	.021	[3.095, 4.260]
ϕ_7	1.079	.089	.007	[.917, 1.262]	ς_7	4.250	.350	.026	[3.585, 4.941]
ϕ_8	1.223	.123	.010	[.986, 1.463]	ς_8	5.011	.492	.039	[4.064, 5.954]
ϕ_9	.680	.074	.002	[.534, .823]	ς_9	1.767	.253	.012	[1.297, 2.287]
ϕ_{10}	.744	.104	.009	[.565, .955]	ς_{10}	6.283	.895	.076	[4.606, 7.978]
ϕ_{11}	.929	.080	.004	[.776, 1.080]	ς_{11}	2.737	.277	.016	[2.229, 3.265]
ϕ_{12}	1.079	.112	.003	[.858, 1.295]	ς_{12}	1.349	.184	.008	[.997, 1.709]
ϕ_{13}	.628	.083	.002	[.470, .794]	ς_{13}	.012	.194	.010	[−.399, .370]
ϕ_{14}	.776	.074	.002	[.632, .922]	ς_{14}	1.523	.206	.010	[1.138, 1.932]
ϕ_{15}	1.013	.094	.002	[.833, 1.200]	ς_{15}	.932	.142	.007	[.651, 1.208]
ϕ_{16}	.693	.075	.004	[.541, .836]	ς_{16}	3.301	.407	.024	[2.542, 4.113]
γ_1	1.281	.119	.005	[1.042, 1.506]	γ_9	.614	.073	.003	[.473, .756]
γ_2	.617	.074	.003	[.470, .762]	γ_{10}	.616	.114	.007	[.398, .845]
γ_3	.516	.069	.003	[.378, .650]	γ_{11}	.535	.076	.003	[.388, .680]
γ_4	.514	.072	.003	[.381, .661]	γ_{12}	2.072	.159	.007	[1.750, 2.371]
γ_5	1.174	.170	.011	[.844, 1.502]	γ_{13}	1.095	.091	.003	[.918, 1.277]
γ_6	.830	.120	.007	[.601, 1.069]	γ_{14}	.565	.067	.002	[.435, .694]
γ_7	.487	.088	.005	[.316, .662]	γ_{15}	1.166	.114	.005	[.949, 1.391]
γ_8	.926	.152	.009	[.636, 1.225]	γ_{16}	.643	.068	.003	[.520, .781]

Note: PARM denotes parameter, EAP is the expected *a posteriori* estimate, SD denotes the posterior standard deviation, MCSE denotes the Monte Carlo standard error, HPD denotes the 95% highest posterior density interval.

and $\log \frac{\text{CPO}_{ij}^{\text{GORH}}}{\text{CPO}_{ij}^{\text{GORHS}}}$ are shown in Figure 7 for item 4. We see from these boxplots that the GORH model fits the PISA data consistently better than both the PH model and the GORHS model. The similar boxplots for items 3 and 7 are shown in Figures S.1 to S.4 of the Supplementary Materials.

7. CONCLUSION AND DISCUSSION

In this paper, we propose a novel and flexible class of semi-parametric response-time models (GORH), which include the PH model and the PO model commonly used in survival analysis as special cases. A joint model is constructed by the new response-time model associated with the IRT model for fitting the binary item responses and the continuous response-times. In the analysis of the PISA data, we have empirically shown that GORH fits the data better than PH and GORHS according to the DIC and LPML criteria. Moreover, the posterior estimates of the nonproportional parameters γ_j's are not too close to 0, confirming that GORH is preferred over PH once again.

In this article, only DIC and LPML are considered. Other Bayesian model selection criteria such as marginal likeli-

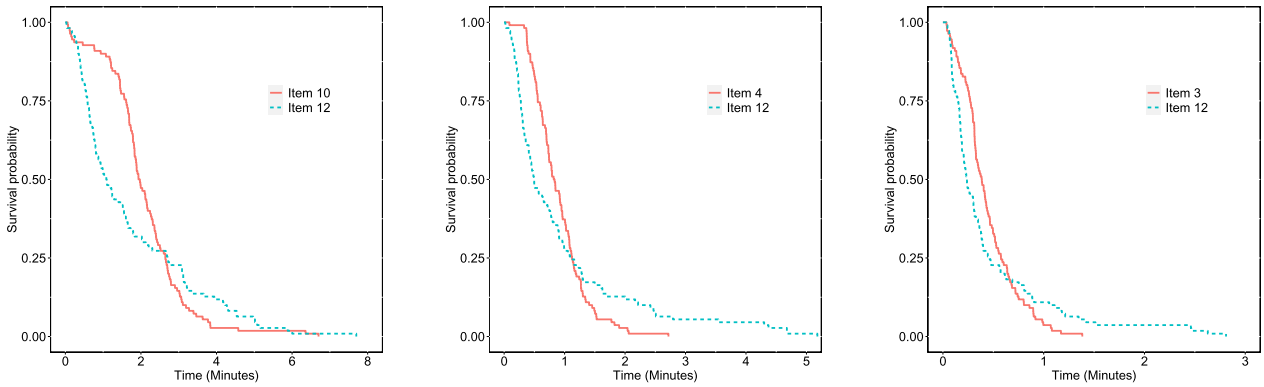

Figure 5. The survival curves of the response-times for the low, middle and high speed group.

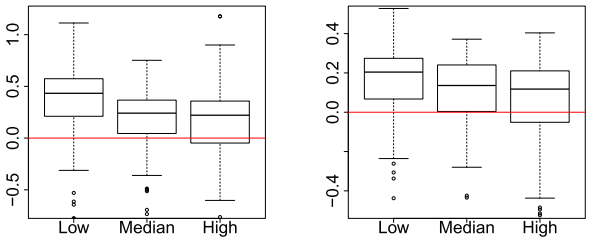

Figure 6. The boxplots of $\mathrm{DIC}_{\mathrm{i4}}^{\mathrm{PH}} - \mathrm{DIC}_{\mathrm{i4}}^{\mathrm{GORH}}$ *(left) and* $\mathrm{DIC}_{\mathrm{i4}}^{\mathrm{GORHS}} - \mathrm{DIC}_{\mathrm{i4}}^{\mathrm{GORH}}$ *(right) in different Speed Groups for Item 4.*

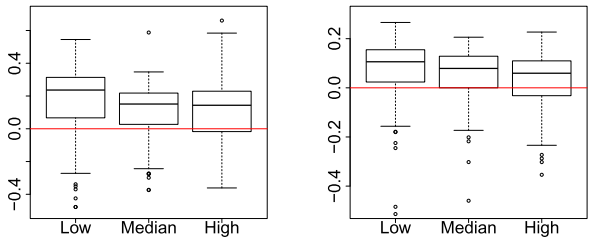

Figure 7. The boxplots of $\log \frac{\mathrm{CPO}_{\mathrm{i4}}^{\mathrm{GORH}}}{\mathrm{CPO}_{\mathrm{i4}}^{\mathrm{PH}}}$ *(left) and* $\log \frac{\mathrm{CPO}_{\mathrm{i4}}^{\mathrm{GORH}}}{\mathrm{CPO}_{\mathrm{i4}}^{\mathrm{GORHS}}}$ *(right) in different Speed Groups for Item 4.*

hoods may also be potentially useful for the joint framework of the IRT responses and the response-times. In addition, the joint models can be easily extended to fit the complex response-times and the item response data, for example, including the covariates information with multilevel structures, the polytomous test structure, and the detection of abnormal response behaviors and so on. These extensions are beyond the scope of this paper but they are currently under further investigation.

Received 20 April 2020

REFERENCES

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267–281). Budapest: Akademiai Kiado. MR0483125

[2] Banerjee, T., Chen, M.-H., Dey, D. K. and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis*, **13**, 241–260. MR2364529

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

[4] Chen, M.-H., Ibrahim, J. G. and Shao, Q.-M. (2000). Monte Carlo Methods in Bayesian Computation. New York: Springer. MR1742311

[5] Crowder, M. (1996). Some tests based on extreme values for a parametric survival model. *Journal of the Royal Statistical Society, Series B*, 417–424. MR1377841

[6] Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, **83**, 744–749. MR0963802

[7] de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F.,... Paganin, S. (2020). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. Retrieved from https://r-nimble.orghttps://github.com/nimble-dev/nimble.

[8] de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. and Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, **26**, 403–413. MR3640196

[9] De Castro, M., Chen, M.-H., Ibrahim, J. G. and Klein, J. P. (2014). Bayesian Transformation Models for Multivariate Survival Data. *Scand Stat Theory Appl*, **41**, 187–199. MR3181139

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

[11] Ferrando, P. J. and Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement*, **31**, 525–543. MR2408980

[12] Fox, J.-P. and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, **51**, 540–553.

[13] Furneaux, W. (1952). Some speed, error and difficulty relationships within a problem-solving situation. *Nature*, **170**, 37–38.

[14] Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160. MR0529531

[15] Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determinating using predictive distributions with implementation via sampling-based methods (with Discussion). *In Bayesian Statistics*, **4**, (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, & A.F.M. Smith). Oxford: Oxford University Press, pp. 147–167. MR1380275

[16] Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). Bayesian Survival Analysis. New York: Springer. MR1876598

[17] Kang, H.-A. (2017). Penalized partial likelihood inference of proportional hazards latent traits models. *British Journal of Mathematical and Statistical Psychology*, **70**, 187–208.

[18] Klein Entink, R. H., van der Linden, W. J. and Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical & Statistical Psychology*, **62**, 621–640. MR2750422

[19] Loeys, T., Legrand, C., Schettino, A. and Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, **67**, 304–327. MR3234177

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

[21] Man, K., Harring, J. R., Jiao, H. and Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, **43**, 639–654.

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

[23] Meng, X. B., Tao, J. and Shi, N. Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, **84**, 1–21. MR3169308

[24] Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society, Series B*, **33**, 169–175. MR0963714

[25] Ranger, J. and Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, **77**, 31–47. MR2868969

[26] Ranger, J. and Ortner, T. M. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, **65**, 334–349. MR2959080

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

[28] Ranger, J. and Kuhn, J. T. (2015). Modeling information accumulation in psychological tests using item response times. it Journal of Educational and Behavioral Statistics, **40**, 274–306.

[29] Rizopoulos D. (2010) "`JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data.`'' *Journal of Statistical Software*, **35**, 1–33.

[30] Rouder, J., Sun, D., Speckmanm P. L., Lu, J. and Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, **68**, 589–606. MR2272436

[31] Rouder, J., Tuerlinckx, F., Speckman, P., Lu, J. and Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, **15**, 1201–1208.

[32] Scharfstein, D., O., Tsiatis, A. A. and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis*, **4**, 355–391.

[33] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, (Statistical Methodology)*, **64**, 583–639. MR1979380

[34] Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.) *New horizons in testing: Latent trait theory and computerized adaptive testing*. pp. 179–203. New York: Academic Press.

[35] van der Linden, W. J. and van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, **68**, 251–265. MR2272379

[36] van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, **72**, 297–308. MR2361958

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

[38] van der Linden, W. J. and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, **73**, 365–384. MR2447320

[39] Van Der Linden, W. J. (2009). Conceptual issues in iesponse-time modeling. *Journal of Educational Measurement*, **46**, 247–272.

[40] van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, **33**, 24–41. MR2655405

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

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

[43] van der Linden, W. J. and Hambleton, R. K. (2013). Handbook of modern item response theory. Springer Science & Business Media.

[44] Verhelst, N. D., Verstralen, H. H. F. M. and 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. MR1856443

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

[46] Wang, C., Xu, G. and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, **83**, 223–254. MR3767020

[47] Wang, T. and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, **29**, 323–339. MR2190364

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

[49] Wu, C. O. (1995). Estimating the real parameter in a two-sample proportional odds model. *Annals of Statistics*, **23**, 376–395. MR1332572

[50] Zeng, D., Chen, Q. and Ibrahim, J. G. (2009). Gamma frailty transformation models for multivariate survival times. *Biometrika*, **96**, 277–291. MR2507143

[51] Zhang, D., Chen, M.-H., Ibrahim, J. G., Boye, M. E. and Shen, W. (2016) `JMFit`: A SAS Macro for Joint Models of Longitudinal and Survival Data. *Journal of Statistical Software*, **71**, 1–24. MR3274507

Fang Liu
School of Mathematics and Statistics, Northeast Normal University, Changchun, China
E-mail address: liuf853@nenu.edu.cn

Jiwei Zhang
School of Mathematics and Statistics, Yunnan University, Kunming, China
E-mail address: zhangjw713@nenu.edu.cn

Ningzhong Shi
School of Mathematics and Statistics, Northeast Normal University, Changchun, China
E-mail address: shinz@nenu.edu.cn

Ming-Hui Chen
Department of Statistics, University of Connecticut, Storrs, USA
E-mail address: ming-hui.chen@uconn.edu