British Journal of Mathematical and Statistical Psychology (2009), 62, 621–640

The British Psychological Society

www.bpsjournals.co.uk

A Box–Cox normal model for response times

R. H. Klein Entink*, W. J. van der Linden and J.-P. Fox

Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

The log-transform has been a convenient choice in response time modelling on test items. However, motivated by a dataset of the Medical College Admission Test where the lognormal model violated the normality assumption, the possibilities of the broader class of Box–Cox transformations for response time modelling are investigated. After an introduction and an outline of a broader framework for analysing responses and response times simultaneously, the performance of a Box–Cox normal model for describing response times is investigated using simulation studies and a real data example. A transformation-invariant implementation of the deviance information criterium (DIC) is developed that allows for comparing model fit between models with different transformation parameters. Showing an enhanced description of the shape of the response time distributions, its application in an educational measurement context is discussed at length.

1. Introduction

Recording response times (RTs) on test items is common practice nowadays. As a result, besides the response patterns, an additional source of information is available to test developers and testing agencies. For instance, RTs can be helpful in improving the design of a test or study the response behaviour of test takers. However, an appropriate statistical treatment of the RTs is required before making any inferences.

Response time experiments have been a major source of inferences about cognitive processes in experimental psychology (Luce, 1986). To illustrate the type of experiments and the kind of data that arise from them, we give the following three examples. Schmiedek, Oberauer, Wilhelm, Süss, and Wittmann (2007) performed experiments using simple speed tasks to study attention fluctuation and working memory. One of the experiments reported by these authors was a verbal classification task where participants had to classify single words into categories of animals or plants.

*Correspondence should be addressed to Rinke Klein Entink, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (e-mail: r.h.kleinentink@gw.utwente.nl).

DOI:10.1348/000711008X374126

Ratcliff and Rouder (1998) performed experiments to study stimulus discrimination, where participants had to classify the intensity of an array of pixels on a monitor as high or low. An example of a time pressure study of the well-known speed–accuracy tradeoff can be found in van der Lubbe, Jaśkowski, Wauschkuhn, and Verleger (2001). There, participants had to respond before the space between an inner circle and an outer circle was filled. Typically, experiments like these consist of many repetitions of the same simple task. The data that arise from such experiments are the RTs (usually of the order of milliseconds) and accuracy measures (correct/incorrect). For the joint analysis of RT and accuracy data traditional ANOVA methods have been used up till recently (van der Lubbe *et al.*, 2001), with inferences based on the mean RTs and the mean proportion correct scores. An approach that provides more detail on the analysis and relates RTs and accuracy explicitly is the diffusion model presented by Ratcliff (1978). For more recent references and approaches that are related to the diffusion model see, for instance, Ratcliff and Tuerlinckx (2002), Wagenmakers, van der Maas, and Grasman (2007), and Browne and Heathcote (2008).

In educational assessment, measurement had for a long time to be based on response accuracy only. This limitation was overcome with the introduction of computerized test adminstration, which made the accurate collection of RTs feasible. This means that an additional source of information on test items and test takers has become available. For instance, when students are not motivated for a test, this might lead to lower RTs as a result of guessing behaviour, something that cannot be easily seen from accuracy data alone. Therefore, there is a need to incorporate RTs into the analysis of test data and study responses and RTs simultaneously. However, there are some important differences in the data collection process compared to procedures in experimental psychology. First, in experimental psychology the RTs are linked directly to theoretical cognitive phenomena which are evaluated, for instance, using elementary two-choice tasks, whereas in educational measurement the tasks (items) are of a much higher cognitive complexity. As a result, the observed RTs are in the range of seconds up to some minutes. Where experiments measured in milliseconds need to take account of a lower bound on the RTs, this can safely be ignored in educational measurement due to the size of the measurements. Also, in educational assessment multiple items are administered that are answered only once, contrary to the within-subject replications found in experimental psychology. These differences lead to a somewhat different approach to the joint modelling of RTs and accuracy data on test items than that mentioned above.

In educational testing, item response theory (IRT) models have served as measurement models for a latent construct, ability, which is assumed to underlie the accuracy data. Very different from the diffusion model, RTs are not included in IRT models. Instead, it will be assumed that individual differences between test takers in their observed RTs result from differences in speed. That is, speed will be assumed to be the latent construct underlying the RTs and a separate measurement model is required for measuring it. At a higher (second) level the relationships between the two measurement models are modelled to account for possible dependencies between the RTs and the accuracy data. This leads to a framework of modelling that allows for the simultaneous analysis of RTs and accuracy data on test items. IRT models have been well developed, but models for RTs have had much less attention in the psychometric literature. In this paper, motivated by an empirical problem, we focus on models for RTs that are flexible in their distributional shapes and fit well into the framework for the simultaneous analysis of responses and RTs.

Typically, RTs are non-negative and, as a result, their distribution is positively skewed. Various types of distributions are able to describe such data and have been extensively studied, for instance, in the field of lifetime modelling. Examples are the Poisson, gamma, Weibull, inverse normal, exponential and lognormal distributions. For discussions on the use of these distributions for modelling RTs in psychometric applications, the reader is referred to Maris (1993), Roskam (1997), Rouder, Sun, Speckman, Lu, and Zhou (2003), Thissen (1983), van Breukelen (1995), Schnipke and Scrams (1997, 2002) and van der Linden (2006). In practice, it is difficult to determine which distribution would fit the RT data best. The lognormal model has been a convenient choice, with good results in terms of model fit (Thissen, 1983; Schnipke & Scrams, 1997; van der Linden, Scrams & Schnipke, 1999; van der Linden, 2006). Besides, it permits the use of the nice statistical properties of a normal model for the log-transformed RTs. A normal model easily allows for decomposition of the mean into item and person effects. Van der Linden (2006) introduced a lognormal model for describing RTs.

Nevertheless, the analysis of a computerized version of the Medical College Admission Test (MCAT) revealed that the log-transformed RTs do not always satisfy the normality assumption (Section 6.2). A Bayesian residual analysis indicated that the skewness of the RT distributions was not always captured well for the MCAT data. In such cases, it would be desirable to evaluate the fit of the model against other distributions. For instance, a gamma model might be more appropriate for describing the structure of the skewness (Maris, 1993). But fitting and evaluating different RT models can be laborious and is not desirable from a practical perspective. A more general approach for describing any RT distribution would be preferred.

The Box–Cox transformation has been widely used to model skewed distributions. For instance, it finds application in lifetime/failure time models in industry and in the empirical determination of functional relationships in the field of economics. Nonetheless, to the authors' knowledge, it has not found application in the psychometric literature of response time modelling. Therefore, the class of Box–Cox transformations is considered in this study. Using a whole class of transformations gives the researcher more freedom in analysing response time data. It allows one to choose an appropriate transformation in order to obtain normally distributed data. Box and Cox (1964) proposed a power transformation as a function of an unknown parameter ν, which contains the log-transform as a special case:

$$T^{(\nu)} = \begin{cases} \frac{T^{\nu}-1}{\nu} & (\nu \neq 0), \\ \log T_{ik} & (\nu = 0), \end{cases}$$

where T denotes the original time and $T^{(\nu)}$ denotes the Box–Cox transformed time. Note that the log transform for $\nu = 0$ is defined in order to obtain a family of transformations over a continuous range of ν.

To illustrate the flexibility in shape of the Box–Cox density, consider response times T that follow a Box–Cox normal density with parameters (ν, λ, τ^2), $\nu \neq 0$, given by:

$$f(t) = t^{(\nu-1)} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\left(\frac{t^{(\nu)}-\lambda}{\tau}\right)^2\right).$$

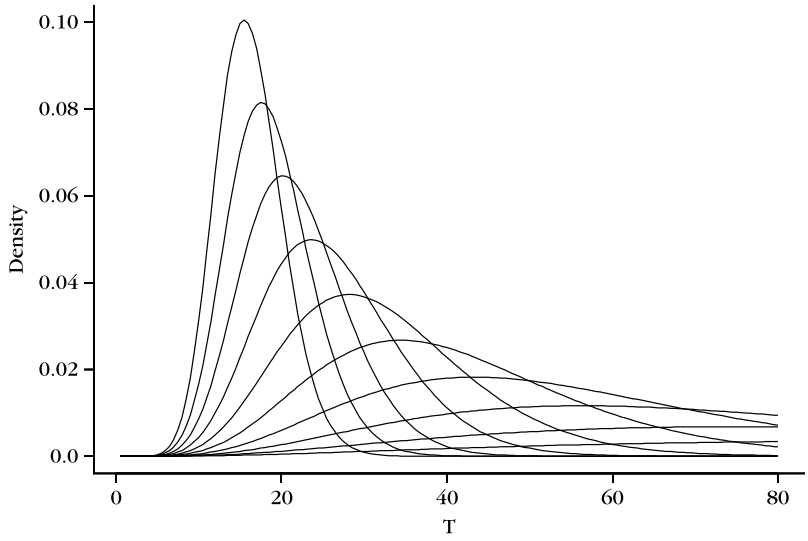

Figure 1. Box–Cox normal densities as a function of the transformation parameter $\nu \in (0, 1)$ and with fixed mean and variance $\lambda = 6$, $\tau^2 = 1$. The density with the leftmost and highest peak is $\nu = .05$, the flattest density is the one with $\nu = 1$.

An impression of the different shapes this density can take is given in Figure 1. For $(\lambda, \tau) = (6, 1)$, the density $f(t)$ is plotted for some values of $\nu \in (0, 1]$ over a range of $T \in [0, 80]$.

Besides flexibility, it is interesting to evaluate its benefits with respect to model fit as well as the interpretation of the parameters and its behaviour in a larger framework for the simultaneous analysis of responses and RTs. Below, the modelling framework is introduced first, followed by the method of estimation. Thereafter, the problem of how to obtain the moments of the Box–Cox distribution is discussed. These moments are helpful in characterizing, for instance, the skewness of the RT distributions. The presentation of a few tools for evaluating the fit of the model is then followed by empirical examples as well as simulation studies that address the research questions above. A discussion of the advantages and disadvantages of the Box–Cox approach for modelling RTs concludes this paper.

2. A framework for the simultaneous analysis of responses and response times

The interest of researchers is often focused on studying responses or response times alone. However, since both data sources contain information on the same items and test takers, it can be advantageous to study them simultaneously. For instance, the interest may be in the relationship between speed and accuracy of test takers or the testing of the common assumption that more difficult items also are more time intensive. Therefore, a framework that allows for modelling dependencies between responses and RTs is outlined here.

Measurement models at level 1 separate the variability in the observed responses and RTs into item and person effects. Just like ability, the speed of the test takers is assumed

to be the underlying construct for the RTs. Further, it is assumed that the speed and ability of the test takers are fixed during the test. This assumption leads to conditional independence of the responses and RTs of a test taker given the latent traits, which is a key feature of this model. At level 2, a correlation structure models the dependencies between the level 1 model parameters.

Not only can the Box–Cox normal model improve the description of the skewness of the data but, due to the transformation to normality, it also fits this hierarchical framework nicely. Contrary to a Weibull or gamma RT model, the Box–Cox model allows the use of easy to implement conjugate normal models for the item and person parameters at level 2, which enables a straightforward Gibbs sampling approach for estimation of the model parameters as well.

2.1. Response model

In IRT, it is assumed that the variability in observed response patterns on test items can be separated into item and person effects. Within an item, the variability between the responses of different test takers results from differences in their ability, denoted by θ. The higher one's ability, the higher the probability of giving a correct response. Within a test, there are differences between items regarding their difficulty. The probability that a test taker answers an item correctly depends on the difference between the difficulty b of the item and his or her ability. The way the item distinguishes between test takers of different ability is described by the discrimination parameter a.

Assuming that the probability that person $i = 1, \ldots, N$ answers item $k = 1, \ldots, K$ correctly ($Y_{ik} = 1$) follows the two-parameter normal ogive model,

$$P(Y_{ik} = 1|a_k, \theta_i, b_k) = \Phi(a_k\theta_i - b_k) \tag{1}$$

or, in its latent response formulation,

$$P(Y_{ik} = 1|a_k, \theta_i, b_k) = \int_0^\infty P(z_{ik}; a_k\theta_i - b_k)\mathrm{d}z,$$

where $Z_{ik} \geq 0$ when $Y_{ik} = 1$ and $Z_{ik} \leq 0$ otherwise. The model is given in its latent variable form for computational convenience, as introduced by Albert (1992).

2.2. Response-time model

Analogously, it is assumed that the variability in observed RT patterns on test items can be separated into item and person effects. For instance, it never happens that a group of test takers finish the test in the same time. Some persons are working faster than others. This leads to the assumption that, within an item, the variability in response times results from differences in speed of working of the test takers. Therefore, a personality trait for speed is introduced, denoted by ζ. That is, the speed parameter is assumed to be the underlying construct for the RTs, just as the ability parameter is for the responses. It is assumed that during a test, a person works at a fixed speed. In general, within a test, test takers do not spend equal time on the items. Some items require more time to solve (it is often assumed that this concerns the more difficult items). As an example, solving $2 + 5 = ?$ involves fewer steps than solving $2 + 5 + 7 = ?$ and, therefore, it can be expected that the latter is more time intensive. To represent these differences in time intensity of items, an item parameter λ is introduced. This parameter can be seen as the

time analogue of the difficulty parameter. The parameter ϕ reflects the way the item distinguishes between test takers of different speed levels.

The generalization to a Box–Cox normal model then leads to a linear model for the transformed RTs:

$$T^{(\nu)} = \begin{cases} \frac{T_{ik}^{\nu}-1}{\nu} \sim N(-\phi_k\zeta_i + \lambda_k, \tau_k^2) & (\nu \neq 0), \\ \log T_{ik} \sim N(-\phi_k\zeta_i + \lambda_k, \tau_k^2) & (\nu = 0). \end{cases} \tag{2}$$

For notational convenience, the superscript will be dropped and T_{ik} will denote the Box–Cox transformed time from now on.

2.3. Second-level models

At the second level of modelling, the person parameters are assumed to follow a multivariate normal distribution. Let $\boldsymbol{\xi}_i = (\theta_i, \zeta_i)$, then:

$$\boldsymbol{\xi}_i = \boldsymbol{\mu}_P + \mathbf{e}_P, \quad \mathbf{e}_P \sim N(\mathbf{0}, \boldsymbol{\Sigma}_P), \tag{3}$$

where $\boldsymbol{\mu}_P = (\mu_\theta, \mu_\zeta)$ and the covariance structure is specified by

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}.$$

Here, ρ denotes the covariance between the two person parameters. A positive estimate for ρ indicates a positive dependence between ability and speed, meaning that a person who works faster than average also tends to have an above-average ability.

Similarly, it can be assumed that the item parameters follow a multivariate normal distribution. Let $\boldsymbol{\Omega}_k = (a_k, b_k, \phi_k, \lambda_k)$; then

$$\boldsymbol{\Omega}_k = \boldsymbol{\mu}_I + \mathbf{e}_I, \quad \mathbf{e}_I \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I), \tag{4}$$

where $\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_\phi, \mu_\lambda)$ and the covariance structure is specified by

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{\phi a} & \sigma_{\phi b} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{\lambda a} & \sigma_{\lambda b} & \sigma_{\lambda\phi} & \sigma_\lambda^2 \end{bmatrix}.$$

This covariance structure allows for the investigation of dependencies between the item parameters. For instance, one can test the common assumption that more difficult items also take more time to solve. Doing so would amount to evaluating the null hypothesis $H_0 : \sigma_{b\lambda} = 0$ against the alternative $H_a : \sigma_{b\lambda} > 0$.

3. Bayesian estimation using an MCMC method

The model is estimated by means of a fully Bayesian approach using straightforward Markov chain Monte Carlo (MCMC) methods. In a Bayesian approach, inferences are made from the posterior distribution $p(\theta|x)$. Using Bayes' rule, the posterior is obtained

from the observed data x, a realization of $X \sim f(x|\theta)$, combined with available prior information, specified as $p(\theta)$. An introduction to Bayesian inference can be found, for instance, in Box and Tiao (1973).

Estimation of the posterior distributions of the model parameters requires evaluating integrals, which, analytically, for complex models, can be an impossible task. A solution to this problem is to use simulations to approximate the densities. Markov chain Monte Carlo methods, such as the Gibbs sampler (Geman & Geman, 1984) and the Metropolis–Hastings algorithm (Chib & Greenberg, 1998), are useful for drawing samples from the posterior distributions of the model parameters. Although computationally intensive, these methods remain straightforward when model complexity increases. Gelman, Carlin, Stern, and Rubin (2004) provide an introduction to MCMC methods; a more advanced text is Robert and Casella (1999).

Since our interest is in the Box–Cox normal RT model, the sampling steps for the transformation parameter and the RT model parameters are given explicitly below. Sampling of the other model parameters is outlined in Appendix A.

3.1. Identification

The RT model can be identified by setting $E(\zeta) = 0$, which fixes the mean. By specifying $\prod_{k=1}^{K} \phi_k = 1$, a tradeoff between σ_ζ^2 and ϕ_k is avoided. Identification of the hierarchical model can be obtained by fixing the location of the latent traits by $\boldsymbol{\mu}_P = \mathbf{0}$. Further, the scale of the ability trait can be fixed in two ways: by setting either $\sigma_\theta^2 = 1$ or $\prod_{k=1}^{K} a_k = 1$.

3.2. Sampling the Box–Cox parameter

A normal model for the transformed response times is assumed. The likelihood with respect to the original response times is given by

$$p(\mathbf{t}|\nu)\,\mathbf{J}(\nu, \mathbf{t}^*),$$

where $\mathbf{t}^*$ denotes the original response times and $\mathbf{J}(\nu, \mathbf{t}^*)$ the Jacobian of the transformation. For $\nu = 0$, the Jacobian equals t^{*-1}; when $\nu \neq 0$, it equals $t^{*(\nu-1)}$. Different priors for ν were studied by Box and Cox (1964), However, these were outcome dependent; that is, they were was dependent on the observations. Pericchi (1981) did propose non-informative priors for the transformation parameter that where not outcome dependent. However, these priors were derived in order to obtain analytic results on the value of ν.

A main problem is that there does not seem to exist a conjugate prior for ν (to the authors' knowledge), so a Gibbs sampling step for the parameter is not feasible. However, for a sampling-based approach, the choice of a family of priors is less critical. For that reason a Metropolis–Hastings (MH) step is proposed, the advantage being that any chosen prior for ν is easily implemented in the MH step. At iteration m, a new value ν^*, sampled from a proposal density $\phi(\nu^*|\nu)$, is accepted with probability

$$\min\left\{1, \frac{p(\nu^*|\mathbf{t})}{p(\nu^{m-1}|\mathbf{t})} \times \frac{\varphi(\nu^{m-1}|\nu^*)}{\varphi(\nu^*|\nu^{m-1})}\right\}, \tag{5}$$

otherwise $\nu^m = \nu^{m-1}$.

When the optimal transformation is the logartihm, the distribution should converge to $E(\nu) = 0$. However, although a posterior mean of approximately 0 can be obtained,

a value of $\nu^{(m)} = 0$ will practically never be sampled since it has probability 0. To accommodate the log-transform, consider a critical value C such that when $|\nu^{(m)}| \leq C$, then $\nu^{(m)} = 0$ with probability .5. Tuning of the value of C is required, whereby (based on our experience) a value of .05 can be considered a good starting point.

3.3. Sampling the item and person parameters

Below the conditional posterior distributions of the person and items parameters of the RT model are presented. Together with the sampling step for the transformation parameter, these steps constitute the MCMC algorithm for the RT model.

- The person speed parameters $\boldsymbol{\zeta}$ are the parameters of the linear regression of $-\mathbf{T}_i + \boldsymbol{\lambda}$ on $\boldsymbol{\phi}$. Assuming a normal prior $\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2)$, the resulting posterior is again normal with

$$\zeta_i|\mathbf{t}_i, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \nu \sim N\left(\frac{\sigma_\zeta^{-2}\mu_\zeta + \sum_{k=1}^K \tau_k^{-2}\phi_k(\lambda_k - t_{ik})}{\sigma_\zeta^{-2} + \sum_{k=1}^K \phi_k^2\tau_k^{-2}}, \left(\sigma_\zeta^{-2} + \sum_{k=1}^K \phi_k^2\tau_k^{-2}\right)^{-1}\right)$$

- The item parameters $(\boldsymbol{\phi}, \boldsymbol{\lambda})$ are the coefficients of the regression of $\mathbf{T}_k$ on $\mathbf{X} = (-\boldsymbol{\zeta}, \mathbf{1})$. Assuming a normal prior, $\phi_k, \lambda_k \sim N(\boldsymbol{\mu}_{\phi,\lambda}, \boldsymbol{\Sigma}_{\phi,\lambda})$, the posterior distribution is given by

$$\phi_k, \lambda_k|t_k, \tau_k^2, \boldsymbol{\zeta}, \nu \sim N\left(\frac{\boldsymbol{\Sigma}_{\phi,\lambda}^{-1}\mu_{\phi,\lambda} + \tau_k^{-2}\mathbf{X}^t\mathbf{t}_k}{\boldsymbol{\Sigma}_{\phi,\lambda}^{-1} + \mathbf{X}^t\mathbf{X}\tau_k^{-2}}, (\boldsymbol{\Sigma}_{\phi,\lambda}^{-1} + \mathbf{X}^t\mathbf{X}\tau_k^{-2})^{-1}\right)$$

- For the residual variance τ_k^2, a conjugate inverse gamma prior with parameters *Inv-Gamma* (g_1, g_2) is assumed. The posterior is then again an inverse gamma distribution with parameter $g_1 + N/2$ and scale parameter $g_2 + (\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))^t \times (\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))/2$.

4. Moments of the response-time distributions

We will use the first three moments about zero of the distributions to assess the differences between the lognormal and Box–Cox normal models. More specifically, it is expected that these models will differ in their third moment, which characterizes the skewness of the distribution. Therefore, only the estimation of the first three moments of the distributions is considered in this study.

How to obtain the moments of the lognormal distribution is well known. However, the moments of the Box–Cox normal distribution are not so straightforward to estimate, except for some specific transformations, such as $\nu = 2$ or $\nu = 0.5$. Freeman and Modarres (2006) studied the properties of the inverse Box–Cox transformation. Let $Y = (X^\nu - 1)/\nu$, $Z = (Y - \mu)/\sigma$ and $Y \sim N(\mu, \sigma^2)$. Then X is power-normal distributed, or $X \sim PN(\nu, \mu, \sigma^2)$. The authors derived the rth moment of X as

$$E(X^r) = (\nu\mu + 1)^{r/\nu} + \sum_{i=1}^{\infty} \frac{1}{i!}(\nu\mu + 1)^{r/\nu - i}\sigma^i E(Z^i)\prod_{j=0}^{i-1}(r - j\nu), \tag{6}$$

for $\nu \neq 0$. Moreover, they showed that these moments can be approximated by $E(X^r) \approx (\nu\mu + 1)^{r/\nu} + \sum_i(\sigma^i/(2^{i/2}(i/2)!))(\nu\mu + 1)^{r/\nu - i}E(Z^i)\prod_{j=0}^{i-1}(r - j\nu)$, where $i > 0$ and even. When $\nu = 0$ the moments of the lognormal distribution can be

approximated by $E(X^r) = \exp(r\mu + (r^2\sigma^2/2))$. These results will be used to approximate the moments of the distributions.

From these raw moments, the second central moment, which corresponds to the variance, and the third standardized moment, which is a measure of the skewness of the distribution, are obtained.

5. Evaluating model fit

Model fit will be evaluated using two methods: (i) Baysian residual analysis by evaluating the posterior probabilities under the model, and (ii) a deviance information criterion (DIC).

The transformed values t_{ij} are evaluated under their predictive density under the RT model. Subsequently, the probability $P(T_{ik} < t_{ik}|\mathbf{y}, \mathbf{t})$ can be approximated by

$$P(T_{ik} < t_{ik}|\mathbf{y}, \mathbf{t}) \approx \sum_{m=0}^{M} \Phi(t_{ik}|\zeta_i^{(m)}, \phi_k^{(m)}, \lambda_k^{(m)}, \nu_k^{(m)})/M$$

from the M iterations of the MCMC chain. Now, the probability integral transformation theorem (e.g. Casella & Berger, 2002) implies that under the true model these probabilities are distributed as $U(0, 1)$. This feature allows evaluation of the model fit. To do so, the calculated probabilities of the items are plotted against their expected values under the $U(0, 1)$ distribution. If the underlying distribution really is $U(0, 1)$, these plots should be approximately linear.

Graphical model checking can be very helpful in understanding in what way a fitted model departs from the data. However, graphical comparison of two competing models can be difficult when they are close. Also, the proposed graphical check does not penalize for model complexity. Therefore, the DIC (Spiegelhalter, Best, Carlin, & van der Linde, 2002), which does account for model complexity, should be estimated as well. Besides being a useful test statistic for model comparison, it has the advantage that it is easily obtained as a by-product of the MCMC chain.

The deviance $D(\mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta})$ is given by:

$$\begin{aligned} D(\mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta}) &= -2 \log p(\mathbf{t}|\boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta}) \\ &= N \sum_{k=1}^{K} \log(2\pi\tau_k^2) + \sum_{k=1}^{K} \sum_{i=1}^{N} (t_{ik} - (-\phi_k \zeta_i + \lambda_k))^2/\tau_k^2 \\ &\quad + \sum_{k=1}^{K} \sum_{i=1}^{N} \log J_{ik}, \end{aligned}$$

where J_{ik} denotes the Jacobian of the transformation, which is $(t_{ik}^*)^{-1}$ when $\nu = 0$ and $(t_{ik}^*)^{(\nu_k - 1)}$ when $\nu_k \neq 0$, with t_{ik}^* the original observation. The DIC is equal to the deviance plus a penalty term for model complexity, and is given by

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

with $\bar{D} \approx \frac{1}{M} \sum_{m=1}^{M} D(\mathbf{t}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\lambda}^{(m)}, \nu^{(m)}, \boldsymbol{\zeta}^{(m)}), m = 1, \ldots, M$, denoting the number of iterations of the algorithm, and $\hat{D} \approx E(D(\mathbf{t}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta})|\mathbf{t}^*)$. Spiegelhalter *et al.* (2002) report that when using the posterior median instead of the posterior mean to estimate

the DIC, the term for model complexity is invariant to transformations. However, the DIC is constructed from a likelihood-based term plus the correction for model complexity. Different transformations lead to different scales of the data and thus affect the likelihood. The Jacobian of the transformation is to guarantee that all DIC values correspond to one common scale (the original time scale). As a result, this DIC allows the comparison of model fit for different transformations.

6. Flexibility of the Box–Cox normal model

To illustrate the possibilities of the Box–Cox approach for modelling response times on test items, two examples are given here. In the first example, it will be shown that the Box–Cox normal model can approximate data resulting from Weibull, gamma and exponential models. The second example analyses an empirical data set and shows that model fit can be improved when the lognormal model is generalized to a Box–Cox normal model.

6.1. Approximation of Weibull, gamma and exponential data

The aim of this example is to show that if the true underlying distribution of the RTs is gamma, Weibull or exponential, the Box–Cox normal distribution can be a good approximation to the RTs.

For our example, we used the empirical mean and variance of the RTs of three items: The first was obtained from a Raven test taken by 300 German army recruits for which $(\text{mean}, \text{var}) = (64, 1766)$. The second was from a computerized version of the MCAT for which $(\text{mean}, \text{var}) = (190, 4904)$ seconds. Rouder *et al.* (2003) used a Weibull distribution to model reaction times and report a typical estimate for the shape parameter of 2. This value was used for the third item. The parameters for the gamma, Weibull and exponential distribution were chosen so as to correspond closely with the estimated means and variances of the selected items.

Subsequently, 10,000 data points were simulated under these models, and the Box–Cox normal model was fitted to the data. From the Box–Cox normal model parameter estimates obtained, the density function was plotted together with the density function of the true underlying distribution. The lognormal density was plotted on the same figures; see Figures 2–4. Furthermore, estimates of the DIC criterion as well as the moments of these distributions were obtained. Table 1 summarizes the results.

As can be seen from the figures, the Box–Cox normal model approximated the three chosen distributions quite well. Both the regions of highest density as well as the tails of the distributions are captured. Only for the exponential model did the lognormal and Box–Cox normal models have problems describing the density near 0. From Table 1 it can be seen that the means of the lognormal and Box–Cox densities were quite close, using (6). However, there were especially sharp differences in the skewness of the distributions. In all cases, the lognormal distribution was more skewed to the right than the Box–Cox normal distribution. For each distribution, one example is given in Figures 2–4. The lognormal model distribution was more peaked. According to the DIC criterion, the best descriptions of the data were obtained with the Box–Cox normal model. Of course, this does not prove that the Box–Cox model is well suited to approximating all possible gamma or Weibull models. However, the aim of this example was to show that, for a typical range of RTs, the Box–Cox model does provide a good approximation to such data.

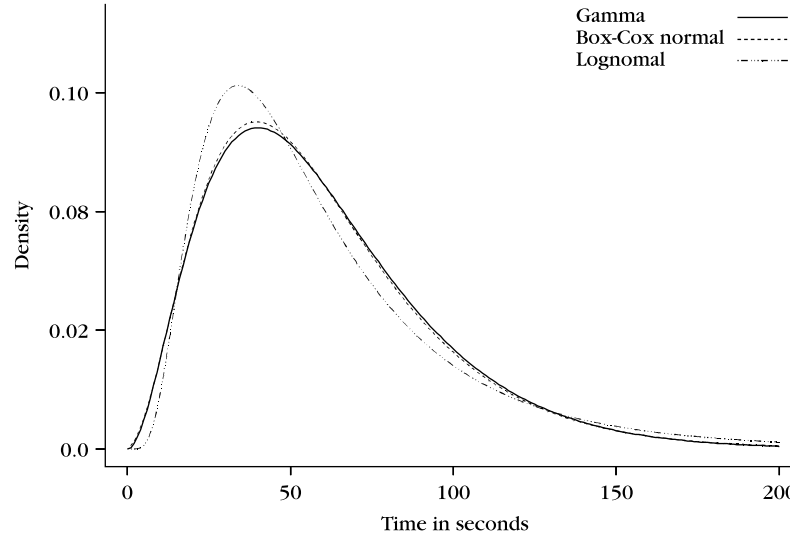

Figure 2. Density of the *Gamma*(3, 0.05) function, its lognormal and its Box–Cox normal approximation.

6.2. Empirical example

For this example, the data of 405 test takers on 214 items from a computerized version of the MCAT were analysed. Six items were omitted from the data set because the algorithm showed convergence problems for them. For the remaining items, only a few observations were missing (less than 1%). These were assumed to be missing at random and were ignored in the estimation procedure. Preliminary analysis showed that the time discrimination parameter did not vary across items using the DIC criterion. The analyses reported below were therefore conducted under the restriction $\phi = 1$.

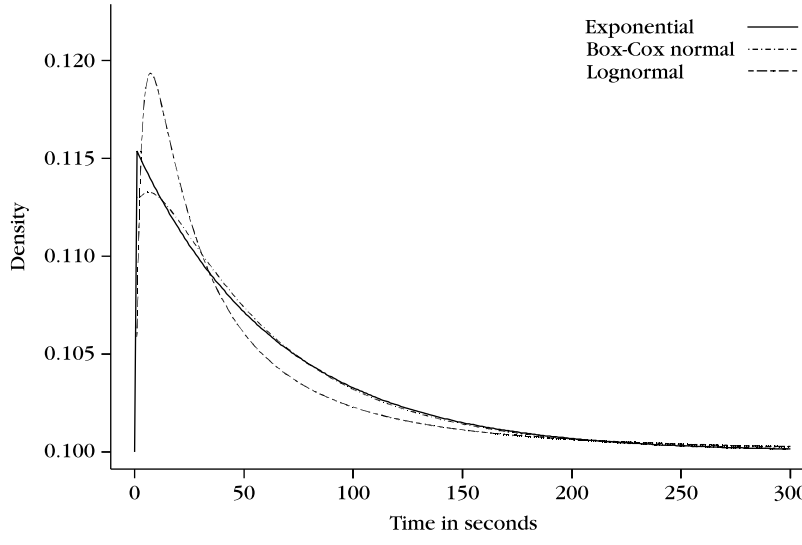

Figure 3. Density of the *Exponential*(1/64) function, its lognormal and its Box–Cox normal approximation.

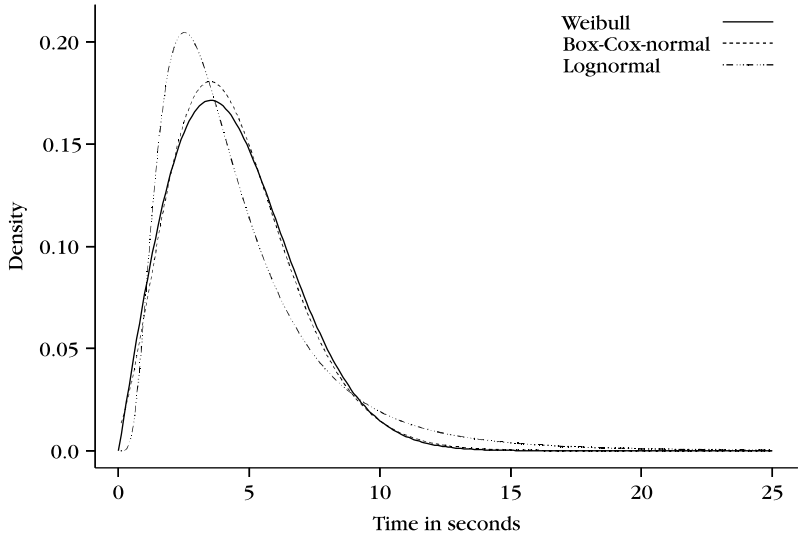

Figure 4. Density of the *Weibull*(2, 5) function, its lognormal, and its Box–Cox normal approximation.

Table 1. Parameter estimates, estimated moments (mean, variance and skewness) and DIC for the approximation of gamma, exponential and Weibull data

Distribution		Parameters			Moments			Model fit
Simulated	Approximated	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	*E(X)*	Var(*X*)	Skewness	DIC
Gamma(3, 0.05)	lognormal	3.92	0.64	0	61.0	1757	2.39	97231
	Box–Cox	7.89	2.04	0.31	59.8	1164	1.60	96591
Gamma(7.4, 0.04)	lognormal	5.1	0.39	0	176.9	5217	1.29	111411
	Box–Cox	12.1	1.78	0.30	175.6	4266	0.89	111221
Expon(1/64)	lognormal	3.59	1.27	0	81.2	26403	10.8	105128
	Box–Cox	6.27	3.06	0.27	59.4	3743	2.40	103373
Expon(1/200)	lognormal	4.22	1.28	0	154.2	97762	11.0	117721
	Box–Cox	8.34	3.58	0.26	126.9	17620	2.43	116022
Weibull(2, 5)	lognormal	1.32	0.63	0	4.59	10.44	2.46	45744
	Box–Cox	2.12	1.17	0.53	4.43	5.43	0.77	44066
Weibull(5, 6)	lognormal	1.65	0.31	0	5.48	3.11	0.99	38246
	Box–Cox	4.66	1.62	1.04	5.45	2.30	−0.03	36779

Two models were fitted to the data: model M_1 with the restriction $\nu = 0$ (lognormal model) and the more general Box–Cox model M_2 with $\nu \neq 0$. The prior for the person parameters was fixed at a mean of $\mu_\zeta = 0$ (for identification) and had a (low-informative) variance of $\sigma^2_\zeta = 10$. For the item parameters, (low-informative) priors $(\mu_\lambda, \sigma_\lambda) = (0, 10)$ were chosen. Since the values for ν are usually within the range of $[-1, 1]$, a slightly informative uniform $U(-4, 4)$ density was specified as prior. The models were estimated using 100,000 iterations of the MCMC algorithm, from which every 10th sample was stored. The reason for doing so is to reduce the autocorrelation between the draws of the transformation and item parameters. The draws of the transformation parameter affect the mean and variance of the distribution on the

transformed time scale and therefore influence λ_k and τ_k. It appeared sufficient to discard the first 1,000 stored samples and base the estimates of the model parameters and the model fit criteria on the remaining 9,000 samples. Rerunning the algorithm with different starting values confirmed convergence of the chains.

Table 2 gives the estimated DIC for each model. It can be seen that M_2 should be favoured over the more restricted model M_1. At the item level, the graphical model check suggested an improvement of model fit for the majority of the items for the Box–Cox model. The .95 highest posterior density region of the transformation parameter was estimated as (.19, .21). Since the DIC was calculated by summing the deviance terms over the items, it was straightforward to obtain the estimates of the DIC at the item level as well. From these results, it followed that model M_2 was selected by the DIC over M_1 for 174 of the 214 items.

Table 2. Estimated DIC values for the models of the MCAT analysis

Model		DIC
M_1	Lognormal	830132
M_2	Box–Cox	822847
M_3	Box–Cox (item-specific)	820963

The graphical posterior check suggested that the lognormal model assigned somewhat more weight to the middle region and somewhat less to the tails of the distributions. Plotting the estimated densities of the three models for some items confirmed this impression; the plots showed that the density of the Box–Cox model was less peaked near its highest density region and has somewhat wider tails than the lognormal model. On average, these difference resulted in an improved description of this data set.

6.2.1. Item-specific transformation parameters

The flexibility of the RT model may be improved further by making the transformation parameter item-specific. We explored this possibility mainly for theoretical reasons but observe that item-specific transformations also lead to item-specific time scales. As discussed below, we therefore expect the applicability to be low.

Using the DIC, the introduction of the item-specific transformation (model M_3) resulted in improved model fit for 150 items. Except for 17 items, the estimates of the DIC criterion suggested that the improvement in model M_3 relative to model M_1 was significant. The .95 highest posterior density regions of the transformation parameters $\boldsymbol{\nu}$ were consistent with these results as well (zero not being contained within these regions). Overall, the DIC for the complete data set decreased to 820,963 for M_3. Although an improvement, the decrease was smaller than for the transition from M_1 *to* M_2 (see Table 2).

In order to illustrate the effect of the Box–Cox transformation for this real-data example, three cases are given in Figures 5–7. These cases were chosen because they reflected a range of parameter values for ν_k. It can be seen that, for item 86, there was no noticeable difference between the two competing models even though the DIC criterion suggested a slight loss of model fit for the Box–Cox model. On the other hand, the Box–Cox model showed substantial improvement for item 4. For item 15, the result was between the two other items and pointed to a slight improvement in our description of the data.

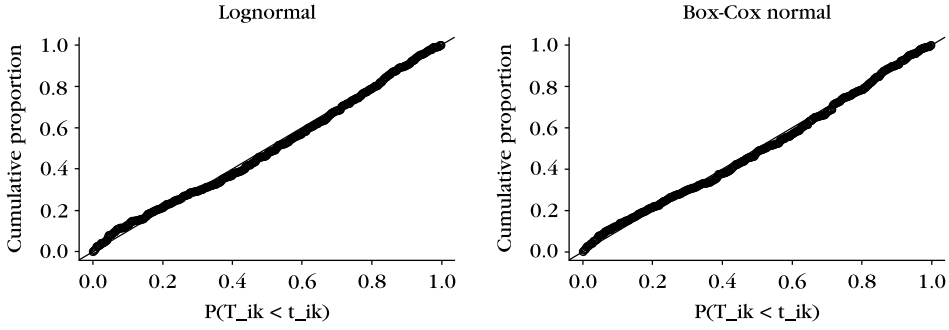

Figure 5. Cumulative probability plots of the posterior probabilities of item 86. Left: lognormal model, $DIC = 4751$, $\nu = 0$. Right: Box–Cox normal model, $DIC = 4761$, $EAP(\nu) = .05$, $.95$ $HPD(\nu) = [.00, .10]$.

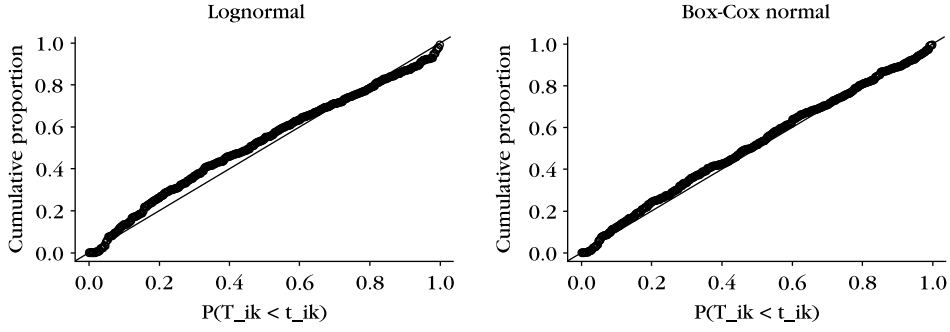

Figure 6. Cumulative probability plots of the posterior probabilities of item 15. Left: lognormal model, DIC=4189, $\nu = 0$. Right: Box–Cox normal model, $DIC = 4142$, $EAP(\nu) = .19$, $.95$ $HPD(\nu) = [.12, .24]$.

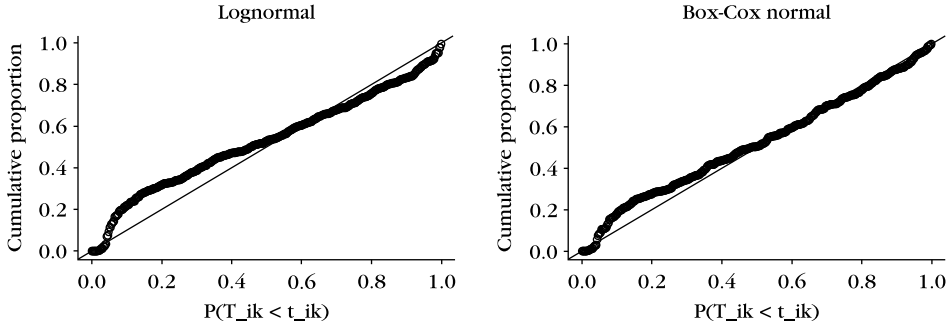

Figure 7. Cumulative probability plots of the posterior probabilities of item 4. Left: lognormal model, $DIC = 3918$, $\nu = 0$. Right: Box–Cox normal model, $DIC = 3810$, $EAP(\nu) = .26$, $.95$ $HPD(\nu) = [.21, .31]$.

7. Model interpretation and selection

It is interesting to determine the effects of the different transformations on the interpretation of the RT model parameters. These should help to guide our choice of model for different types of analyses.

Upon transformation, the RTs are assumed to follow a normal model. In educational testing, it is natural to assume variability across persons as well as items. Differences in the times required to solve the items are reflected by the time intensities λ_k of the items. That is, if item 1 is more time intensive than item 2, this will be reflected by $\lambda_1 > \lambda_2$. From (2), it can be seen that then the expected RTs on item 1 will be higher than those on item 2 as well: $E(T_{\lambda_1}) > E(T_{\lambda_2})$, which holds for every ζ. For the speed parameters, the relationship with the expected RTs is negative. That is, if person 1 has a speed of working ζ_1 greater than person 2 with speed ζ_2, then for every item $E(T_{\zeta_1}) < E(T_{\zeta_2})$. The discrimination parameter ϕ does not affect these relationships. It only controls the rate of decrease in expected RT on an item for one step of increase in speed of a test taker.

For $\nu_k = \nu$, $k = 1, \ldots, K$ (that is, one common transformation parameter for all items in the test), the interpretation above holds. All parameters are on the same transformed time scale, so it does not make any difference whether $\nu = 0$ or $\nu \neq 0$. Also, the sign of the relationships between $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ or $(\mathbf{a}, \boldsymbol{\phi})$, modelled at the second level, remains the same. Note, however, that the interpretations of time intensity and speed of the parameters do not hold for the original time scale. For instance, two items on the log-time scale both with $\lambda = 3$, but with $\tau_1^2 = 2$ and $\tau_2^2 = 4$, have a mean on the time scale of $\exp(3 + 2/2)$ and $\exp(3 + 4/2)$, respectively.

Things are different for item-specific transformations, that is, when $\nu_k \neq \nu$, $k = 1, \ldots, K$. These transformations result in item-specific scales. As a result, it is impossible to interpret differences between the item parameter estimates directly as differences between item characteristics. To do so, an extra scaling step would be required. Observe that for RTs on multiple tests, each with their own transformation, the same problem occurs and a scaling step would be required as well. Although the scale of the item varies under transformations, it can be seen that for two persons with $\zeta_1 > \zeta_2$ their expected RTs (on any item) are still ordered by $E(T_{\zeta_1}) < E(T_{\zeta_2})$, regardless of the transformation. So, by definition, the ranking of the speed parameters is invariant under these transformations. Thus, item-specific transformations do affect the scale of the population distribution, σ_ζ^2, as well as the covariance between ability and speed. But they do not lead to interpretative difficulties for the speed parameters or the dependency between ability and speed.

In practice, however, difficulties might arise in the case of missing data. Even when the missing data are ignorable, the analysis may still result in different scales for different test takers: as the scale is item-specific, a test taker who misses an item immediately works on a different speed scale.

In conclusion, the following practical guidelines can be given:

- The case of a common transformation parameter for all items in a test maintains the interpretation of the RT model parameters. It gives the researcher the freedom to fit different distributional shapes to the RT data and admits comparisons between the person and the item parameter estimates for the test.
- When the interest is in parameter estimates for multiple tests, the transformation parameter should be restricted to be common to all tests. Then all parameters are on the same scale, and no additional equating of scale is necessary to make comparisons.
- More general item-specific transformations are usually of main interest when the focus is on inferences with respect to the ranking of the person parameters. The item parameters are not directly comparable and would require rescaling to a common scale first. An example where the item-specific transformation might be of interest is the study of possible aberrant behaviour of test takers, for which van der Linden and

Guo (2008) presented an approach based on residual analysis. Then, the focus is on the individual person–item combinations and model fit is important to avoid misleading conclusions.

8. Discussion

Transformations to normality have obvious and much exploited advantages for the statistical modelling of non-normal data. For modelling response times in a psychometric application, the log-transform has proven to be useful. However, this study was motivated by a data set for which the lognormal model was not able to capture certain aspects of the data. Therefore, the class of Box–Cox transformations was considered, which allows for more flexibility in the description of the data. The examples illustrated how the Box–Cox transformation parameter affects the shape of RT distributions and, as a result, improves the description of the data.

In Section 2, the full modelling framework for responses and RTs on test items was developed to place the RT model in a broader context. A strong feature of the Box–Cox model is that its transformation of the data leaves the standard modelling framework intact.

In educational testing, it makes sense to decompose observed RTs into item effects (time intensity) and person effects (speed). Therefore, the parameters of the RT model presented in (2) have a clear interpretation in an educational context. Also, its conjugacy with the multivariate normal level-2 models for the person and item parameters allows for straightforward modelling of the dependencies between the parameters in the level-1 models (van der Linden, 2007; Fox, Klein Entink, & van der Linden, 2007).

Transforming the data instead of the model parameters provides the flexibility of using different distributional shapes for the RTs, while the MCMC algorithm is easily extended with an additional sampling step. On the other hand, for instance, the use of a more flexible three-parameter Weibull distribution instead of the Box–Cox transformation would require the replacement of the MCMC steps for the current normal RT model by much more complicated procedures since the conjugacy between the level-1 and level-2 models is then lost.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Education Statistics*, *17*, 251–269.

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

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Browne, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Thomson Learning.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*, 347–361.

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

Freeman, J., & Modarres, R. (2006). Inverse Box–Cox: The power-normal distribution. *Statistics and Probability Letters*, *76*, 764–772.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

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

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

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, *68*, 35–43.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review*, *9*, 438–481.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

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-Verlag.

Rouder, 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.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süss, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429.

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

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

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*, 583–639.

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

van Breukelen, G. J. P. (1995). Psychometric and information processing properties of selected response time models. *Psychometrika*, *60*, 95–113.

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

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J., & 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 constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Lubbe, R. H. J., Jaśkowski, P., Wauschkuhn, B., & Verleger, R. (2001). Influence of time pressure in a simple response task, a choice-by-location task, and the Simon task. *Journal of Psychophysiology*, *15*, 241–255.

Wagenmakers, E. J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin and Review*, 3–22.

Received 22 May 2008; revised version received 1 September 2008

Appendix A. Estimation of the hierarchical framework

This section briefly outlines the MCMC algorithm for the full hierarchical framework. A simulation study to illustrate the parameter recovery of this algorithm is given in Appendix B. The model can be identified by setting the means of the person parameters $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ to zero ($\boldsymbol{\mu}_P = \mathbf{0}$) and by specifying $\prod_{k=1}^{K} \phi_k = 1$ and $\prod_{k=1}^{K} a_k = 1$ which fixes the scale of the latent variables. Fox *et al.* (2007) provides a Gibbs sampler that uses identification for the ability scale by restricting $\sigma_\theta^2 = 1$, where the identifying restrictions are directly incorporated into the prior distributions.

A.1. Linear measurement models for augmented data

The vector of augmented data $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$ minus the vector of difficulty parameters, $\mathbf{b}^T$, and the similar vector of response times $\log \mathbf{T}_i = (\log T_{i1}, \ldots, \log T_{iK})$ minus the vector of time intensity parameters, $\boldsymbol{\lambda}^t$, are stacked in a vector $\mathbf{Z}_i^*$. Then, both measurement models can be presented as a linear regression structure,

$$\begin{aligned} \mathbf{Z}_i^* &= (\mathbf{a} \oplus -\boldsymbol{\phi})(\theta_i, \zeta_i)^t + \mathbf{e}_i \\ &= \mathbf{x}_I \boldsymbol{\xi}_i + \mathbf{e}_i \end{aligned} \tag{7}$$

where $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{C}_{2K})$ and $\mathbf{C}_{2K} = \mathbf{I}_K \oplus \mathbf{I}_K \boldsymbol{\tau}^2$.

Similarly, let $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{nk})^t$ and the vector of log response times, $\log \mathbf{T}_i = (\log T_{1k}, \ldots, \log T_{nk})^t$, to item k be stacked in a vector $\mathbf{Z}_k^*$. Define covariate matrices $\mathbf{H}_\theta$ and $\mathbf{H}_\zeta$ and $(\boldsymbol{\theta}, -\mathbf{1}_n)$ and as $(-\boldsymbol{\zeta}, \mathbf{1}_n)$, respectively. A regression structure for the item parameters can be presented as

$$\begin{aligned} \mathbf{Z}_k^* &= (\mathbf{H}_\theta \oplus \mathbf{H}_\zeta)(a_k, b_k, \phi_k, \lambda_k)^t + \mathbf{e}_k \\ &= \mathbf{x}_P \boldsymbol{\Omega}_k + \mathbf{e}_k \end{aligned} \tag{8}$$

where $\mathbf{e}_k \sim N(\mathbf{0}, \mathbf{C}_{2N})$ and $\mathbf{C}_{2N} = \mathbf{I}_N \oplus \mathbf{I}_N \tau_k^2$.

A.2. Hyperpriors

As a hyperprior for $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, a normal inverse Wishart distribution is chosen. That is,

$$\boldsymbol{\Sigma}_I \sim \textit{Inv-Wishart}_{\upsilon_I}(\mathbf{V}_I^{-1})$$

$$\boldsymbol{\mu}_I | \boldsymbol{\Sigma}_I \sim N(\boldsymbol{\mu}_{I0}, \boldsymbol{\Sigma}_I / \kappa),$$

where υ_1 and $\mathbf{V}_1$ are the degrees of freedom and scale matrix of the inverse Wishart distribution, $\boldsymbol{\mu}_{I0}$ is the prior mean and κ the number of prior measurements.

Similarly, as a hyperprior for $\boldsymbol{\Sigma}_\mathrm{P}$, an inverse-Wishart distribution is chosen. That is,

$$\boldsymbol{\Sigma}_P \sim \textit{Inv-Wishart}_{\upsilon_P}(\mathbf{V}_P^{-1})$$

where υ_P and $\mathbf{V}_P$ are the degrees of freedom and scale matrix of the inverse Wishart distribution. The mean $\boldsymbol{\mu}_P$ is fixed at $\mathbf{0}$ with probability 1 because of the identification restrictions.

A.3. MCMC algorithm

Estimation of all model parameters for the full hierarchical framework proceeds as follows:

Step 1. Sample augmented response data according to (1), given the values for the item and ability parameters.
Step 2. Sample values for the item parameter from $p(\boldsymbol{\Omega}_k|\mathbf{Z}_k^*, \boldsymbol{\xi}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ for $k = 1, \ldots, K$. A product of a normal likelihood and a normal prior again leads to a normal posterior distribution. So, from (8) and (4), it follows that

$$\boldsymbol{\Omega}_k \sim MVN(\boldsymbol{\mu}_{\Omega_k}, \boldsymbol{\Sigma}_{\Omega_k})$$

where $\boldsymbol{\Sigma}_{\Omega_k}^{-1} = \mathbf{x}_P^t \mathbf{C}_{2N}^{-1} \mathbf{x}_P + \boldsymbol{\Sigma}_I^{-1}$ and $\boldsymbol{\mu}_{\Omega_k} = \boldsymbol{\Sigma}_{\Omega_k}(\mathbf{x}_P^t \mathbf{C}_{2N}^{-1} \mathbf{Z}_k^* + \boldsymbol{\Sigma}_I^{-1} \boldsymbol{\mu}_I)$.
Step 3. Sample values for the ability speed parameters from a multivariate normal distribution. Analogous to Step 2, the full conditional posterior distribution is constructed from a multivariate normal likelihood, (7) and a multivariate normal prior distribution as

$$\boldsymbol{\xi}_i \sim MVN(\boldsymbol{\mu}_{\xi_i}, \boldsymbol{\Sigma}_{\xi_i})$$

where $\boldsymbol{\Sigma}_{\xi_i}^{-1} = \mathbf{x}_I^t \mathbf{C}_{2K}^{-1} \mathbf{x}_I + \boldsymbol{\Sigma}_P^{-1}$ and $\boldsymbol{\mu}_{\xi_i} = \boldsymbol{\Sigma}_{\xi_i}(\mathbf{x}_I^t \mathbf{C}_{2K}^{-1} \mathbf{Z}_i^* + \boldsymbol{\Sigma}_P^{-1} \boldsymbol{\mu}_P)$.
Step 4. For the residual variance τ_k^2, a conjugate inverse gamma prior with parameters *Inv-Gamma*(g_1, g_2) is assumed. The posterior is then again an inverse gamma distribution with parameter $g_1 + N/2$ and scale parameter $g_2 + (\mathbf{t}_k - (-\phi_k \boldsymbol{\zeta} + \lambda_k))^t (\mathbf{t}_k - (-\phi_k \boldsymbol{\zeta} + \lambda_k))/2$.
Step 5. Draw a new value for ν from a proposal density $\varphi(\nu^*|\nu)$ and accept the draw with the probability specified in (5).
Step 6. The hyperprior parameters are related to a multivariate normal model for the person parameters, $\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P$, or a multivariate model for the item parameters, $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$.

- The full conditional posterior distribution of $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ has a normal inverse Wishart distribution (e.g. Gelman *et al.*, 2004). It follows that

$$p(\boldsymbol{\mu}_I|\boldsymbol{\Sigma}_I, \boldsymbol{\mu}_0, \boldsymbol{\Omega}, \mathbf{V}_I) = N((\kappa \boldsymbol{\mu}_0 + K\bar{\boldsymbol{\Omega}})/(\kappa + K), \boldsymbol{\Sigma}_I/(K + \kappa))$$

  where $\bar{\boldsymbol{\Omega}} = \sum_k \boldsymbol{\Omega}_k/K$. Subsequently, the full conditional of $\boldsymbol{\Sigma}_I$ is an inverse Wishart with parameter $K + v_I$ and scale parameter $\mathbf{V}_I + \sum_k (\boldsymbol{\Omega}_k - \bar{\boldsymbol{\Omega}}) \times (\boldsymbol{\Omega}_k - \bar{\boldsymbol{\Omega}})^t + \frac{\kappa K}{\kappa + K}(\bar{\boldsymbol{\Omega}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{\Omega}} - \boldsymbol{\mu}_0)^t$.
- Similarly, the full conditional of $\boldsymbol{\Sigma}_P$ is an inverse Wishart with parameters $N + v_P$ and scale parameter $\mathbf{V}_P + \sum_n (\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}})^t + \frac{\kappa K}{\kappa + N}(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_{P0})(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_{P0})^t$.

Appendix B. Simulation

To illustrate the parameter recovery for the algorithm for the full hierarchical framework, a simulation study was performed. We simulated responses under the two-parameter logistic model and RTs under the RT model with $\nu = 0.3$ for 1,000 test takers answering 20 items. The ability and speed parameters were randomly drawn from $\theta_i \sim N(0, 1)$, $\zeta_i|\theta_i \sim N(0, 1)$ with $\rho = .5$ (see equation (3)). The item parameters were randomly drawn according to: $a_k \sim N(1, 0.1)$, $b_k \sim N(0, 1)$, $\lambda_k \sim N(10, 2)$ and the time

discrimination parameters were generated from $\phi_k \sim N(2, 0.3)$ and subsequently standardized to ensure that $\prod_{k=1}^{K}\phi_k = 1$.

The model was identified by setting $\boldsymbol{\mu}_P = \mathbf{0}, \sigma_\theta^2 = 1$ and $\prod_{k=1}^{K}\phi_k = 1$. The prior variance σ_ζ^2 was chosen to be non-informative and was set to 100, the prior covariance between ability and speed was chosen to be $\rho_0 = 0$. Priors for the item parameters were non-informative as well and were chosen to be $\boldsymbol{\mu}_{I0} = (1, 0, 1, 0)$ and a diagonal matrix with 10 on its diagonal for the prior covariance matrix. That is, we assumed prior independence between the response and RT model parameters.

The algorithm was run for 100,000 iterations of which every tenth was stored to account for autocorrelation induced by the Box–Cox transformation, since the transformation affects the mean and variance of the RT distribution. From the stored samples, the first 1,000 were discarded as burn-in. The simulated (true) values and the re-estimated values (expected a posteriori, EAP) plus standard deviations of the model parameters are given in Table B1. Graphical inspection of the re-estimated ability and speed parameters showed that their values were in good agreement with their true values. The EAP estimate of the transformation parameter was $E(\nu) = 0.309$ and its .95 highest posterior density region was estimated to be (0.293,0.323), which includes the true value of $\nu = 0.3$. It can be seen that for this example the parameter recovery of the algorithm was good, even with a moderate number of items.

Table B1. Simulated and re-estimated model parameters

a			b			α			β		
True	EAP	SD	True	EAP	SD	True	EAP	SD	True	EAP	SD
0.92	1.04	0.08	−1.66	−1.77	0.10	0.79	0.85	0.03	7.99	8.14	0.16
0.96	0.96	0.06	0.12	0.14	0.05	0.77	0.78	0.03	10.27	10.49	0.23
1.00	1.03	0.08	1.44	1.53	0.08	1.20	1.20	0.03	8.30	8.49	0.17
1.06	1.07	0.07	−0.96	−0.94	0.06	1.22	1.18	0.03	7.24	7.33	0.14
0.91	0.87	0.06	0.33	0.33	0.05	1.27	1.26	0.03	6.41	6.50	0.11
1.04	0.96	0.06	−0.49	−0.56	0.05	0.85	0.88	0.03	5.50	5.55	0.09
0.84	0.74	0.08	1.79	1.76	0.09	0.82	0.84	0.03	7.63	7.80	0.15
0.93	0.98	0.07	−0.88	−0.85	0.06	1.20	1.23	0.03	6.75	6.92	0.12
0.98	1.01	0.06	0.04	0.08	0.05	0.73	0.73	0.03	9.04	9.26	0.19
0.95	0.86	0.08	1.59	1.47	0.08	1.06	1.05	0.04	4.00	3.99	0.06
0.90	0.94	0.12	−2.60	−2.69	0.18	0.86	0.80	0.03	7.41	7.53	0.14
0.88	0.96	0.09	−1.69	−1.85	0.10	1.04	1.03	0.03	6.85	6.94	0.12
1.08	1.06	0.07	0.69	0.69	0.05	1.06	1.08	0.03	11.03	11.34	0.26
1.00	1.06	0.07	0.89	0.91	0.06	1.15	1.14	0.03	8.31	8.48	0.17
1.11	1.16	0.07	0.05	0.08	0.05	1.01	1.05	0.03	8.72	8.89	0.18
0.68	0.71	0.08	1.73	1.81	0.09	0.95	0.93	0.03	4.87	4.89	0.07
1.02	1.02	0.07	−0.99	−0.97	0.06	1.25	1.21	0.03	8.21	8.32	0.16
1.10	1.23	0.08	−1.26	−1.37	0.08	1.01	1.02	0.03	9.65	9.82	0.21
1.12	1.09	0.06	−0.03	−0.01	0.05	0.88	0.88	0.03	10.22	10.47	0.23
1.11	1.06	0.08	1.67	1.69	0.09	1.14	1.12	0.03	8.77	8.98	0.18