

## A SPEEDED ITEM RESPONSE MODEL WITH GRADUAL PROCESS CHANGE

YURI GOEGBEUR

K.U. LEUVEN AND UNIVERSITY OF SOUTHERN DENMARK

PAUL DE BOECK

K.U. LEUVEN

JAMES A. WOLLACK

UNIVERSITY OF WISCONSIN

ALLAN S. COHEN

UNIVERSITY OF GEORGIA

An item response theory model for dealing with test speededness is proposed. The model consists of two random processes, a problem solving process and a random guessing process, with the random guessing gradually taking over from the problem solving process. The involved change point and change rate are considered random parameters in order to model examinee differences in both respects. The proposed model is evaluated on simulated data and in a case study.

Key words: item response model, local item dependence, test speededness.

### 1. Introduction

Test speededness has been modelled using two alternative item response theory (IRT) approaches both of which assume a single point at which the examinee's response strategy switches to an alternative response strategy due to time limits being reached for the test (Bolt, Cohen & Wollack, 2002; Yamamoto, 1987; Yamamoto & Everson, 1997). In this paper we propose an alternative model in which the response strategies switch more gradually. We show that this alternative model is in fact a general case which subsumes both previous explanations of speededness and provides a more realistic view of test speededness. In addition, this alternative model provides an opportunity to consider modelling other psychological processes, particularly ones which may change gradually, such as learning or change in attitudes or preferences.

According to the latent trait approach, examinees are characterized by a possibly vector valued random variable  $\Theta$ , often called the ability. The ability is not directly observable, and one typically infers about it through a sequence  $Y' = (Y_1, \dots, Y_I)$  of scored items (correct/incorrect, coded  $Y_i = 1$  and  $Y_i = 0$ , respectively) usually referred to as the 'test', intended to measure  $\Theta$ . Observed values of  $Y$  and  $Y_i$  will be denoted by  $y$  and  $y_i$ , respectively. In the latent trait approach one postulates a model for the random vector  $Y$  conditional on a realization  $\theta$  of  $\Theta$ , i.e. one specifies the conditional distribution  $P(Y = y|\theta)$ . From this model one derives the univariate

The research reported in this paper was supported by IAP P5/24 and GOA/2005/04, both awarded to Paul De Boeck and Iven Van Mechelen, and by IAP P6/03, awarded to Iven Van Mechelen. Yuri Goegebeur's research was supported by a grant of the Danish Natural Science Research Council.

Requests for reprints should be sent to Yuri Goegebeur, Department of Psychology, K.U. Leuven, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: yuri.goegebeur@stat.sdu.dk

conditional probability functions  $P(Y_i = y_i|\boldsymbol{\theta})$ , with  $P_i(\boldsymbol{\theta}) := P(Y_i = 1|\boldsymbol{\theta}) = E(Y_i|\boldsymbol{\theta})$ , called the item characteristic curve (ICC), giving the probability that a randomly selected examinee with ability  $\boldsymbol{\theta}$  will answer item  $i$  correctly, as well as the joint marginal model for  $\mathbf{Y}$ ,

$$P(\mathbf{Y} = \mathbf{y}) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} P(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}) dG(\boldsymbol{\theta}),$$

where  $G$  denotes the joint distribution function of  $\boldsymbol{\Theta}$ . In classical IRT models one usually assumes: (i) *local item independence*, i.e. conditional on  $\boldsymbol{\Theta} = \boldsymbol{\theta}$  all responses in  $\mathbf{Y}$  are independent:

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}) &= \prod_{i=1}^I P(Y_i = y_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^I [P_i(\boldsymbol{\theta})]^{y_i} [1 - P_i(\boldsymbol{\theta})]^{1-y_i} \end{aligned}$$

for all possible vectors  $\mathbf{y}$  ( $2^I$  in total), leading to the ‘usual’ IRT equation

$$P(\mathbf{Y} = \mathbf{y}) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left\{ \prod_{i=1}^I [P_i(\boldsymbol{\theta})]^{y_i} [1 - P_i(\boldsymbol{\theta})]^{1-y_i} \right\} dG(\boldsymbol{\theta});$$

(ii)  $\boldsymbol{\Theta}$  to be one-dimensional; and (iii) a 1PL (Rasch, 1960), 2PL (Birnbaum, 1968) or 3PL (Birnbaum, 1968) model for  $P_i(\boldsymbol{\theta})$ .

In case the responses are driven by more factors or latent traits than those included in the model, say in  $\boldsymbol{\theta}$ , not all dependencies will be properly accounted for by the model under consideration, making the validity of the local item independence assumption questionable. It is well known that the Rasch model, and IRT models in general, are not robust with respect to violations of the local item independence assumption. The inclusion of items with local item dependence (LID) may result in contaminated estimates of test reliability, person and item parameters, standard errors and equating coefficients, see, for instance, Yen (1984), Thissen, Steinberg and Mooney (1989), Sireci, Thissen and Wainer (1991), Yen (1993), Wainer and Thissen (1996), Lee, Kolen, Frisbie and Ankenmann (2001) and Tuerlinckx and De Boeck (2001). Next to this, some research has been devoted to the development of tests or indices for the detection of violations of the conditional independence assumption, see van den Wollenberg (1982), Rosenbaum (1984, 1988), Yen (1984), Stout (1987, 1990), Stout et al. (1996), Chen and Thissen (1997), Douglas, Kim, Habing and Gao (1998) and Ip (2001). We refer to Bradlow, Wainer and Wang (1999) and Tuerlinckx and De Boeck (2004) for possible approaches to modelling local item dependencies.

Yen (1993) and Ferrara, Huynh and Michaels (1999) provide a detailed taxonomy of possible reasons for the existence of local item dependency. One of the most prevalent causes in educational testing is test speededness. Whenever tests are administered within fixed time limits there is the possibility that some examinees will have insufficient time to answer all questions. It is well known that the speed of performing a task is one of the more noticeable aspects in which individuals differ with respect to each other, besides the differences in the ability to perform a task correctly (Yamamoto & Everson, 1997). Speededness is typically an inadvertent source of LID in that the speed with which an examinee responds is not an important part of the construct of interest (Lord & Novick, 1968). Speededness manifests itself in that LID is usually present on items at the end of the test and examinees affected by speededness receive ability estimates that underestimate their true levels. In addition, speededness may cause certain items, particularly those administered late in the test, to have poorly estimated parameters (Douglas et al., 1998;

Oshima, 1994) making it difficult to maintain a scale over time (Wollack, Cohen & Wells, 2003). For multiple choice tests where the score is a function of the number of correctly answered items, a sensible strategy for an examinee running out of time is to quickly guess the answers to the remaining items. In this way the number of correct answers, and hence the score, can be increased by chance alone. From the above discussion it should be clear that for some examinees, the response profiles will, besides the ability to answer the item correctly, also reflect an influence due to the limited time test administration, i.e. responses, typically at the end of the test, will also be driven by test speededness effects. Traditional IRT models did not explicitly incorporate speededness in the construct of ability, which lead to contaminated estimates for the ability to perform a task correctly and for the item difficulty parameters of end-of-test items.

Item response theory models dealing with test speededness are relatively new. The hybrid model of Yamamoto and Everson (1997) uses multiple IRT models to describe the behaviour of examinees. A classical item response model is valid throughout most of the test but end-of-test items are answered randomly by some subset of examinees. The model identifies  $M$  possible latent classes, one for whom an item response model is valid for all items, and  $M - 1$  classes with an item response model describing answers to the first  $I - m$  items and random guessing on the last  $m$  items,  $m = 1, \dots, M - 1$ . Formally,

$$P_i^{(m)}(\theta_p^{(m)}) = \begin{cases} \frac{\exp(\alpha_i(\theta_p^{(m)} - \beta_i))}{1 + \exp(\alpha_i(\theta_p^{(m)} - \beta_i))}, & i \leq I - m, \\ c_i, & i > I - m, \end{cases}$$

with  $m = 0, \dots, M - 1$ . Clearly, speededness is unlikely to be so straightforward, as not all students will switch immediately to random guessing beyond some point.

Bolt et al. (2002) extend the mixture Rasch model proposed by Rost (1990) to distinguish latent classes of examinees according to the existence of speededness in their item response patterns. Ordinal constraints are imposed on the item difficulty parameters across classes so as to distinguish a class having no speededness effects from a class whose responses are affected by speededness. In particular, for items early in the test, the item difficulty parameters are constrained to be equal in the two classes; however, the item difficulty parameters of end-of-test items in the speeded class are constrained to be larger than the respective item difficulty parameters in the nonspeeded class. Let  $g$  denote a class indicator with  $g = 0, 1$  referring to the nonspeeded and speeded class, respectively, and let  $k$  denote the first item where the examinees experience the effects of test speededness. The mixture Rasch model can then be stated as

$$P_i^{(g)}(\theta_p^{(g)}) = \frac{\exp(\theta_p^{(g)} - \beta_i^{(g)})}{1 + \exp(\theta_p^{(g)} - \beta_i^{(g)})},$$

with

$$\begin{aligned} \beta_i^{(0)} &= \beta_i^{(1)} & \text{for } i < k, \\ \beta_i^{(0)} &< \beta_i^{(1)} & \text{for } i \geq k. \end{aligned}$$

The item difficulty estimates obtained in the nonspeeded class provide more suitable estimates of the Rasch difficulties of end-of-test items than the difficulties estimated using all examinees. Although this model has worked quite well at identifying test speededness (and has subsequently been extended to model speededness in the three-parameter model Bolt, Mroch & Kim, 2003), it does not allow for different examinees becoming speeded at different points in the test. Since such differences are plausible, in this paper, we propose a model that provides for this kind of transition as a random effect within examinees.

The remainder of this paper is organized as follows. In the next section we propose an item response model that accommodates the disadvantages of the hybrid model and the mixture Rasch model. The model can be seen as consisting of two random processes, a problem solving process (a classical IRT process) and a random guessing process, with the random guessing gradually taking over from the problem-solving process. In this paper we use the 3PL model for the problem solving component of the model. The involved change point and change rate are considered random parameters in order to model examinee differences in both respects. The model was first formulated by Wollack and Cohen (2004) as a model to simulate speededness data, but it will be treated here as a full-fledged model for test data which can also be estimated. In Section 3 we evaluate the performance of the model on the basis of a simulation study. The final section reports the results of applying the model to a mathematics placement test.

## 2. A Model for Speeded Test Data with Gradual Process Change

In this section we propose a new item response model for dealing with speeded test data. Under the model, responses to items early in the test are governed by a 3PL model. Beyond some point the success probability gradually decreases and eventually reduces to the success probability under random guessing. Both change point and change rate are examinee specific.

Using  $p$  as an examinee index,  $p = 1, \dots, P$ , and  $i$  as an item index,  $i = 1, \dots, I$ , the model can be stated as

$$Y_{pi} | \theta_p, \eta_p, \lambda_p \sim \text{Bern}(\pi_{pi})$$

with

$$\pi_{pi} = c_i + (1 - c_i) P_i(\theta_p) \min \left\{ 1, \left[ 1 - \left( \frac{i}{I} - \eta_p \right) \right]^{\lambda_p} \right\}, \quad (1)$$

where  $c_i$  is a random guessing parameter,  $\eta_p$  ( $\eta_p \in [0, 1]$ ) represents the speededness point and  $\lambda_p$  ( $\lambda_p \geq 0$ ) the speededness rate of examinee  $p$ , and  $P_i(\theta_p)$  given by the 2PL model, i.e.

$$P_i(\theta_p) = \frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}.$$

The speededness point parameter  $\eta_p$  identifies the point in the test, expressed as a fraction of the number of items, where examinee  $p$  first experiences an effect due to speeding. For items with  $i \leq \eta_p I$  there is no effect of speeding. Once the examinee passes his/her speededness point,  $i/I - \eta_p$  is positive, resulting in a decrease of  $\min\{1, [1 - (i/I - \eta_p)]^{\lambda_p}\}$ . The rate of decrease of  $\min\{1, [1 - (i/I - \eta_p)]^{\lambda_p}\}$  is controlled by the parameter  $\lambda_p$ , with larger  $\lambda_p$  values resulting in a faster decrease. In Figure 1 we illustrate the role of  $\eta$  and  $\lambda$  by plotting the decay function  $\min\{1, [1 - (x - \eta)]^\lambda\}$  for some values of  $\eta$  and  $\lambda$ .

The rationale for the proposed model is as follows. Denote  $P_i(\eta_p, \lambda_p) = \min\{1, [1 - (i/I - \eta_p)]^{\lambda_p}\}$ . When examinee  $p$  encounters item  $i$ , he/she answers according to either a 3PL process or a random guessing process, with probabilities  $P_i(\eta_p, \lambda_p)$  and  $1 - P_i(\eta_p, \lambda_p)$ , respectively. Under random guessing the answer is correct with probability  $c_i$ . Under the problem solving process the examinee knows the answer with probability  $P_i(\theta_p)$ ; if ignorant, the examinee guesses at random. In Figure 2 we visualize the model with a decision tree. Clearly,

$$P(Y_{pi} = 1 | \theta_p, \eta_p, \lambda_p) = P_i(\eta_p, \lambda_p) P_i(\theta_p) + P_i(\eta_p, \lambda_p) [1 - P_i(\theta_p)] c_i + [1 - P_i(\eta_p, \lambda_p)] c_i,$$

which simplifies to (1).

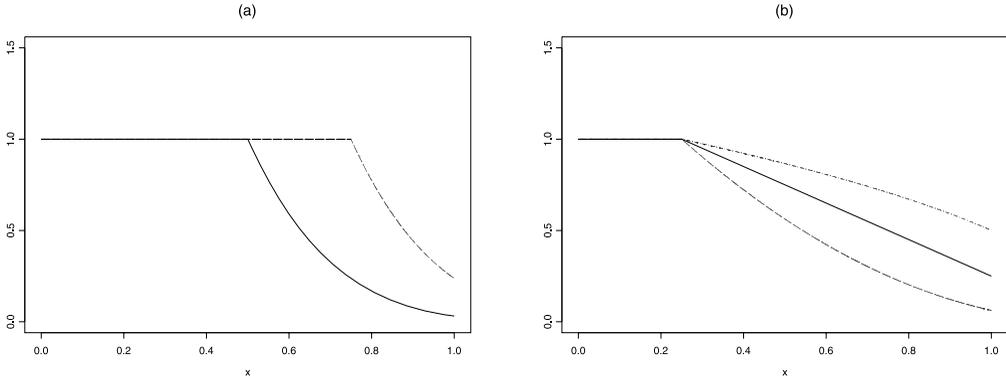


FIGURE 1.

(a)  $\min\{1, [1 - (x - \eta)]^\lambda\}$  for  $\lambda = 5$ ,  $\eta = 0.5$  (solid line) and  $\eta = 0.75$  (broken line); (b)  $\min\{1, [1 - (x - \eta)]^\lambda\}$  for  $\eta = 0.25$ ,  $\lambda = 1$  (solid line),  $\lambda = 2$  (broken line) and  $\lambda = 0.5$  (broken-dotted line).

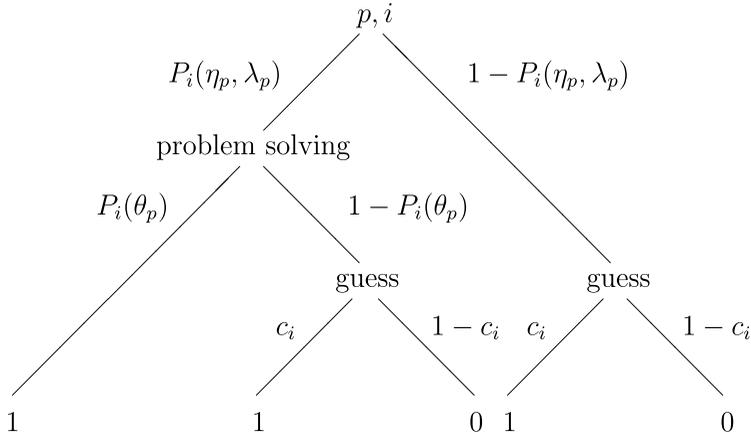


FIGURE 2.

Decision tree representation of speededness model.

Model (1) has some interesting limiting cases:

- if  $[1 - (i/I - \eta)]^\lambda = 0$  for  $i/I > \eta$  (this corresponds to the limiting case  $\lambda \rightarrow +\infty$ ), then (1) reduces to one of the speeded classes in the hybrid model, and speededness is modelled as random guessing
- in case  $\lambda = 0$  or  $\eta = 1$ , the proposed model reduces to the 3PL model
- in case  $\eta = 0$  and  $\lambda > 0$ , the examinee guesses at random at least to some degree from the first item up to the final item
- as with the 3PL model,  $c_i$  is the horizontal asymptote for  $\theta \rightarrow -\infty$ .

Note that the 3PL model is obtained as a limiting case for  $\lambda = 0$  (whatever the value of  $\eta$ ) or for  $\eta = 1$  (whatever the value of  $\lambda$ ) and hence is not uniquely identified within the proposed test speededness model. This of course may entail estimation difficulties, such as nonconvergence of the optimization algorithm or ill-conditioned observed information matrices, when model (1) is fitted to data that are not affected by test speededness. Estimation difficulties can also be expected if one is close to the identification limit, i.e. when test speededness effects come in very late. This is, of course, not completely unexpected since in this case the information available

for the estimation of the test speededness parameters is rather limited. We will illustrate this issue in Section 3.

As is usual in IRT, the parameter reflecting the examinee's ability to perform the task correctly is assumed to be normally distributed. The standard deviation for the ability distribution is fixed at the value of one to avoid the identification problem that otherwise would arise due to the presence of discrimination parameters. For the speededness parameters  $\eta_p$  and  $\lambda_p$  we make, without loss of generality, the following distributional assumptions:

$$\begin{aligned}\eta_p &\sim \text{Beta}(\alpha, \beta), \\ \lambda_p &\sim \log N(\mu_\lambda, \sigma_\lambda^2).\end{aligned}$$

The marginal distribution functions of  $\theta_p$ ,  $\eta_p$  and  $\lambda_p$  will be denoted by  $G_1$ ,  $G_2$  and  $G_3$ , respectively. Besides these marginal distribution functions we also have to specify the dependence structure. This will be done by a copula function  $C$ ; see Appendix 2 for more information on copulas. By using Sklar's theorem (Sklar, 1959) we have that the function  $G$ , defined as

$$G(\theta, \eta, \lambda) = C(G_1(\theta), G_2(\eta), G_3(\lambda)),$$

is a joint distribution function, with marginal distribution functions  $G_1$ ,  $G_2$  and  $G_3$ . Intuitively, a copula links marginal distribution functions together into a joint distribution function.

Concerning estimation, we restrict the discussion to the marginal maximum likelihood method. In the marginal maximum likelihood method the random effects are integrated out and the resulting likelihood is maximized with respect to the unknown parameters. Under (1), denoting the vector of unknown parameters by  $\xi$ , the marginal likelihood function is simply

$$L(\xi) = \prod_{p=1}^P \int_{\mathbb{R}} \int_0^1 \int_0^\infty \prod_{i=1}^I P(Y_{pi} = y_{pi} | \theta_p, \eta_p, \lambda_p) dG(\theta_p, \eta_p, \lambda_p). \quad (2)$$

The integrals involved in (2) can be numerically approximated by a quadrature method and the optimization can be performed using a standard Newton–Raphson algorithm. The SAS NLMIXED procedure fits nonlinear mixed models with multivariate normal random effect distributions. However, as long as  $G$  in (2) is characterized by a normal dependence structure (copula), NLMIXED can be used to fit model (1), whatever the functional form of the (continuous) marginal random effect distribution functions. Indeed, as shown in Proposition 1 (see Appendix 2), in case of a normal dependence function, appropriately chosen compositions of probability integral transforms and inverse probability integral transforms of the marginal distributions yield a multivariate normal distribution for the transformed random effects. Appendix 1 contains example SAS code. As an alternative to the SAS NLMIXED procedure, the authors developed a Fortran program to maximize (2) in the presence of a normal copula function. This program is built around the NAG library subroutines D01BBF and D01FBB for the numerical integration and E04UCF for the optimization (NAG, 1993). Under the test speededness model (1), and assuming a multivariate normal copula function, the vector of the unknown parameters is given by

$$\xi' = (\beta_1, \dots, \beta_I, \alpha_1, \dots, \alpha_I, c_1, \dots, c_I, \mu_\lambda, \sigma_\lambda^2, \alpha, \beta, \rho_{\theta\eta}, \rho_{\theta\lambda}, \rho_{\eta\lambda}),$$

where  $\rho_{XY}$  is used to denote the correlation between random variables  $X$  and  $Y$ . In some cases, besides  $\xi$ , the person specific effects  $\theta_p$ ,  $\eta_p$  and  $\lambda_p$  are also of special interest. Estimates of these parameters can be obtained from an empirical Bayes analysis of the postulated model. These empirical Bayes estimates then allow us to identify examinees affected by test speededness effects. More information on empirical Bayes estimation can be found in Appendix 3.

### 3. Simulation Study

In this section we discuss the results of a small simulation study. Four data sets, each containing responses of 1000 examinees on 80 items were generated. The parameter values were taken from the mathematics test case study discussed in Section 4, this to ensure realistic settings for the simulation. This includes dependent examinee ability, speededness point and rate random effects. We examined the effect of manipulating the parameters  $\alpha$  and  $\beta$  of the test speededness point distribution. Sample 1 was generated under model (1) with the parameter values from the mathematics test data. Sample 2 was generated under model (1) with speededness point parameters  $\alpha = 2$  and  $\beta = 2$ , and Sample 3 was generated with speededness point parameters  $\alpha = 9$  and  $\beta = 2$ . Under these simulation conditions,  $E(\eta_p) = 0.31$  for Sample 1, 0.50 for Sample 2 and 0.82 for Sample 3. Finally, a fourth sample was generated from a 3PL model, i.e. no speededness was generated. All computations were performed with the Fortran/NAG implementation. Computation times varied between 50 and 100 hours per sample (on an Intel Pentium M, 2.13 GHz, 1 GB of RAM).

The effect of test speededness is illustrated in Figure 3(a), Figure 4(a), Figure 5(a) and Figure 6(a), where we plot the empirical proportions correct scores (solid lines) together with the theoretical ones (broken lines), given by

$$\begin{aligned} E(Y_{pi}) &= E[E(Y_{pi}|\theta_p, \eta_p, \lambda_p)] \\ &= E(\pi_{pi}) \\ &= c + (1 - c) \int_{\mathbb{R}} \int_0^1 \int_0^\infty P_i(\theta_p) P_i(\eta_p, \lambda_p) g(\theta_p, \eta_p, \lambda_p) d\theta_p d\eta_p d\lambda_p, \end{aligned} \quad (3)$$

in case of (1), and by

$$E(Y_{pi}) = c + (1 - c) \int_{\mathbb{R}} P_i(\theta_p) g_1(\theta_p) d\theta_p,$$

where  $g_1$  denotes the standard normal density function, in the case of the 3PL model. Clearly, test speededness decreases the probability of a correct answer for end-of-test items. Of course, the ultimate effect depends on the distribution of the speededness point and rate.

The fit of the speededness model can be evaluated by comparing the observed proportion correct scores with their model-based estimates, obtained by plugging the maximum likelihood estimates for the model parameters into (3). These estimated model-based proportions are drawn by broken-dotted lines in Figure 3(a), Figure 4(a), Figure 5(a) and Figure 6(a). Clearly, the estimated and empirical proportions correct scores are almost indistinguishable, indicating a very good fit of the model.

In Figure 3(b) and (c) we illustrate the effect of test speededness on the estimates for the item difficulties  $\beta_i^* = \alpha_i \beta_i$ . For items early in the test, the difficulty estimates obtained with the test speededness model (solid line) and the 3PL model (broken line) agree quite well. However, after a certain point the estimates obtained from fitting a 3PL model begin to diverge from those obtained under the speededness model. As is clear from the figures, ignoring test speededness causes upward biased estimates of the item difficulty estimates, a result that is consistent with the IRT literature. In Figure 3(d) and (e) we show the theoretical density functions (solid lines) of the speededness point and speededness rate, respectively, together with the fitted densities (broken lines). In Figure 4, Figure 5 and Figure 6 we present the corresponding estimation results for Sample 2, Sample 3 and Sample 4, respectively. As expected, in case the speededness effects come in later, the item difficulty estimates obtained under the 3PL model not only diverge later, but also to a lesser extent, from those obtained under (1). Moreover, the information available to estimate the speededness parameters is rather limited in case speededness effects come in late, leading to estimates showing high sampling variability, see, for instance, Figure 5(d).

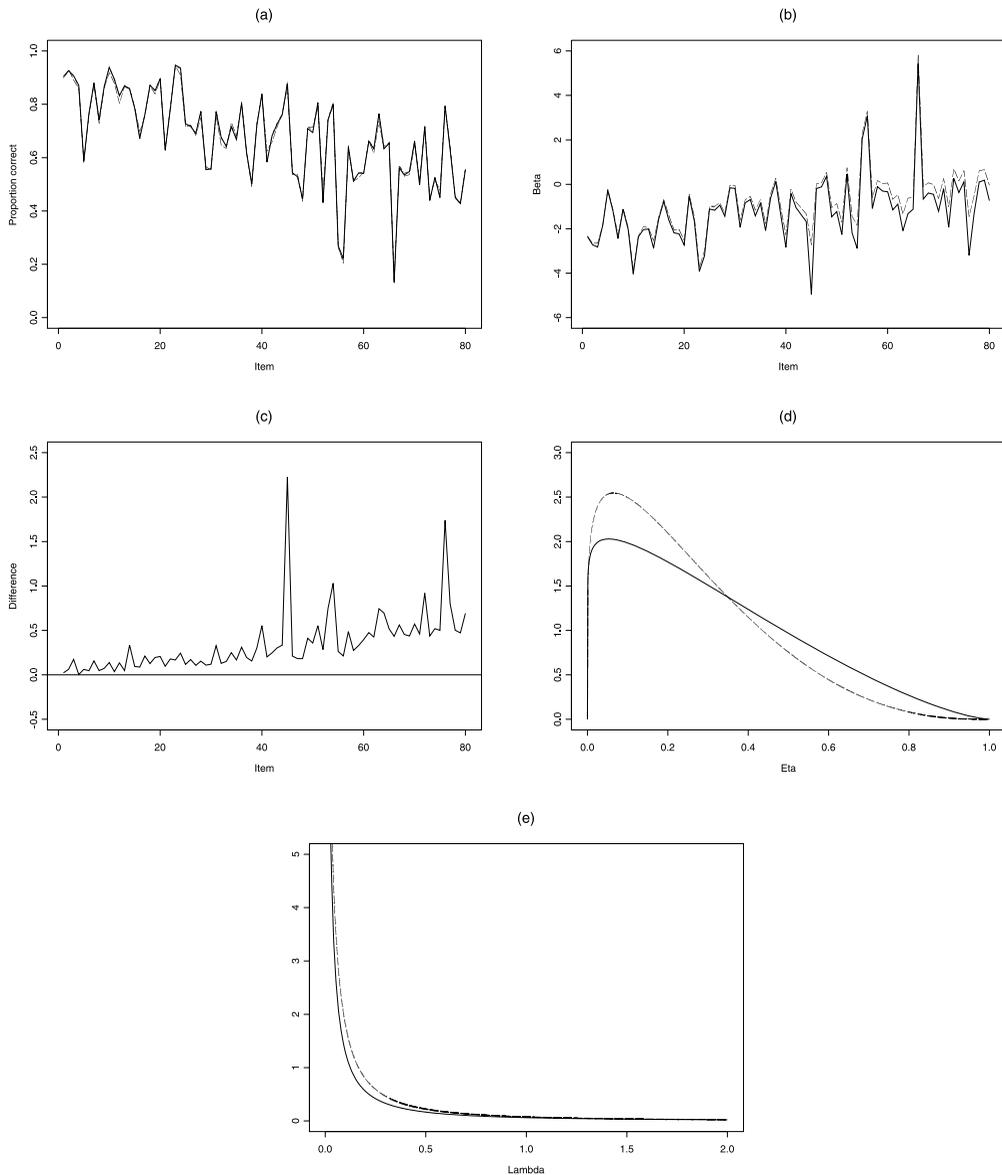


FIGURE 3.

Results for Sample 1: (a) proportion of correct scores versus item number: empirical (*solid line*), theoretical with true parameter values (*broken line*), theoretical with estimated parameter values (*broken-dotted line*); (b) estimated item difficulty parameters under (1) (*solid line*) and the 3PL model (*broken line*); (c) difference between item difficulty estimates; (d) density function of  $\eta$ : theoretical (*solid line*) and fitted (*broken line*); and (e) density function of  $\lambda$ : theoretical (*solid line*) and fitted (*broken line*).

In Table 1 we compare model (1) with the 3PL model in terms of  $-2 \log L$ , the Akaike information criterion (AIC) and the Schwarz Bayes information criterion (BIC). The 3PL model is nested in the test speededness model and hence its values for  $-2 \log L$  will always be larger than the ones for model (1). For all cases considered, AIC and BIC select the appropriate model, i.e. the test speededness model for Sample 1, Sample 2 and Sample 3 and the 3PL model for Sample 4.

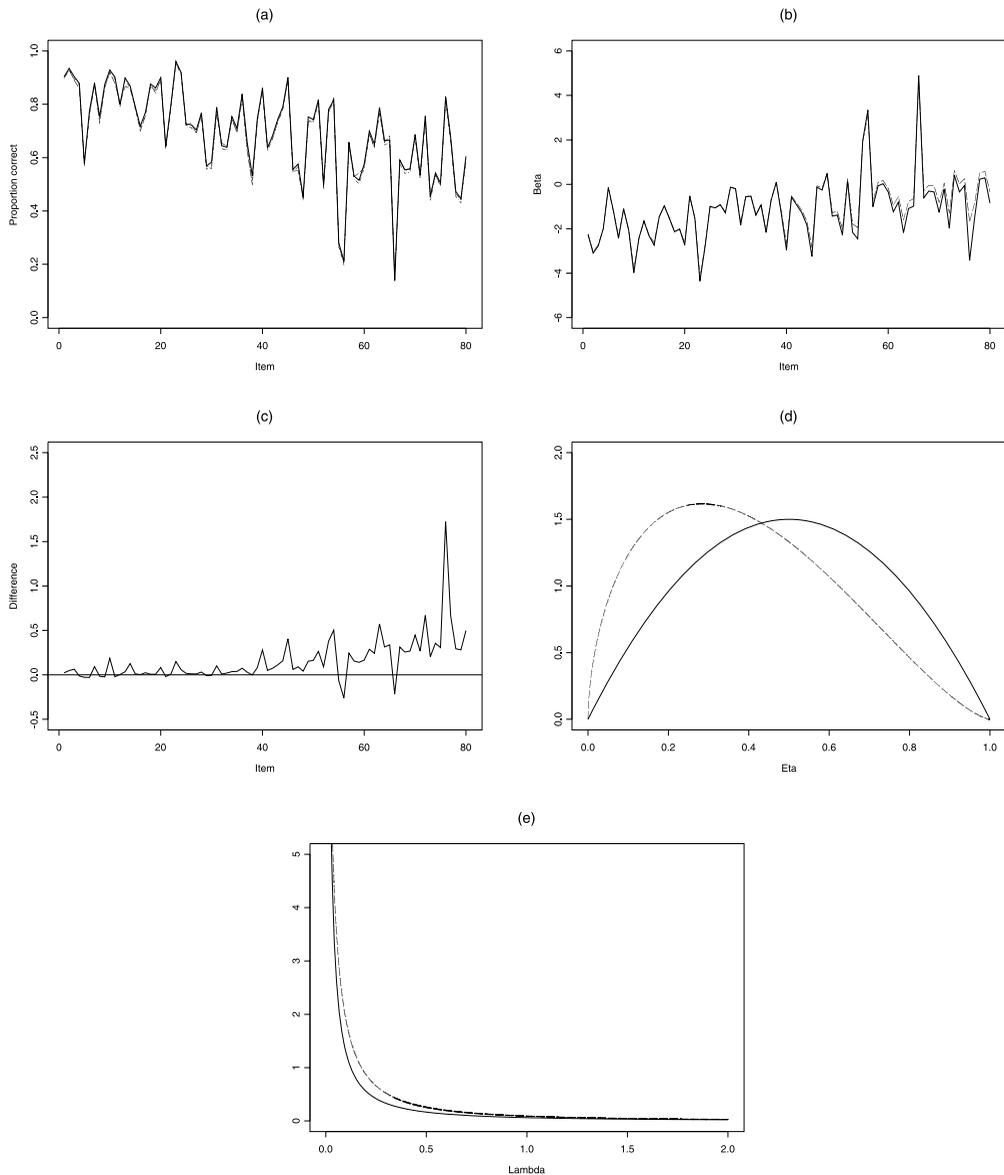


FIGURE 4.

Results for Sample 2: (a) proportion of correct scores versus item number: empirical (*solid line*), theoretical with true parameter values (*broken line*), theoretical with estimated parameter values (*broken-dotted line*); (b) estimated item difficulty parameters under (1) (*solid line*) and the 3PL model (*broken line*); (c) difference between item difficulty estimates; (d) density function of  $\eta$ : theoretical (*solid line*) and fitted (*broken line*); and (e) density function of  $\lambda$ : theoretical (*solid line*) and fitted (*broken line*).

#### 4. Application to Mathematics Placement Test

Data from Form 1 of the 2004 administration of a mathematics placement test at a large, selective Midwestern university were analyzed for test speededness using model (1). The data set contains response profiles of 3447 students. The mathematics placement test included 75 operational and 10 pilot items covering mathematics basics, college algebra and trigonometry

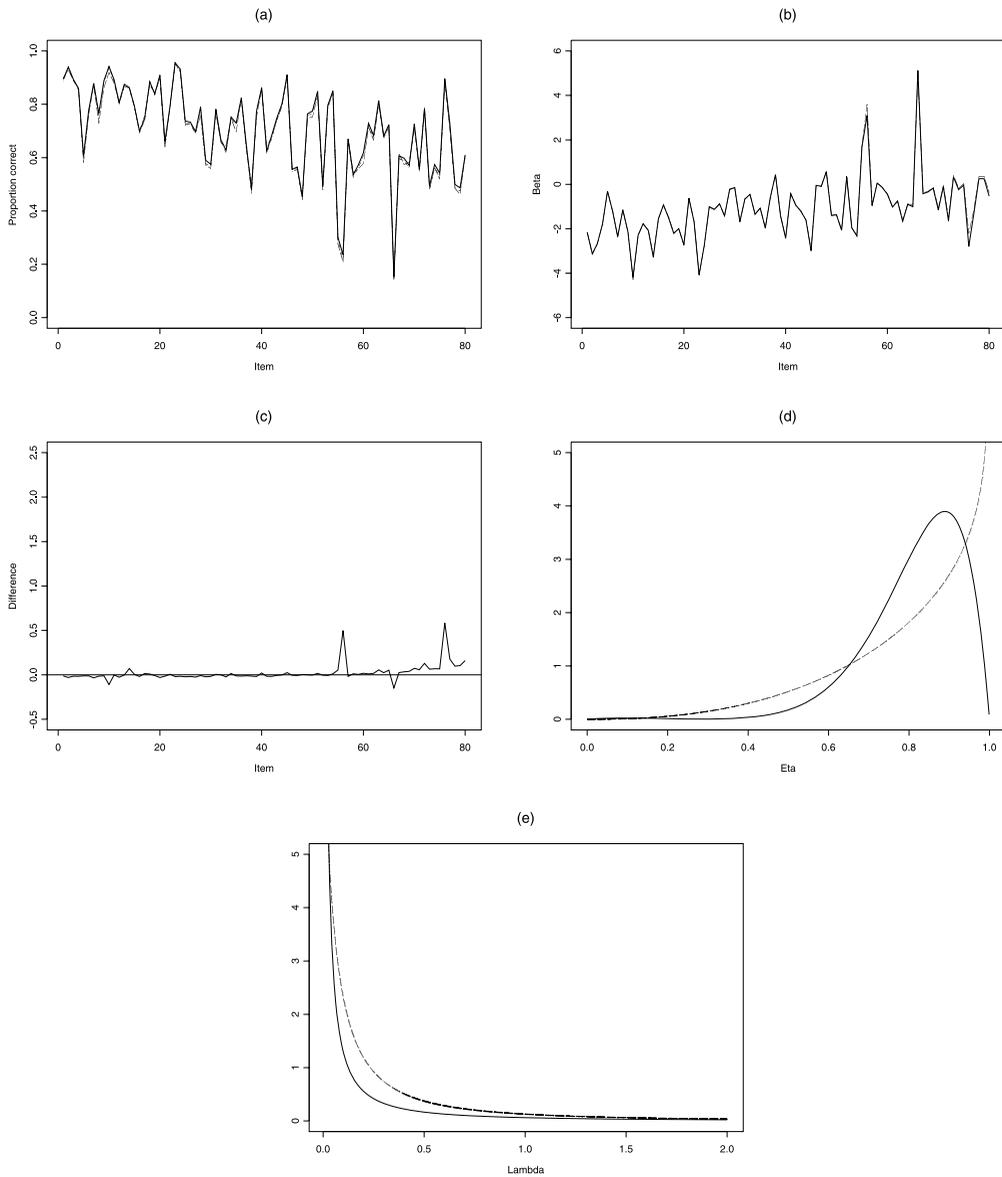


FIGURE 5.

Results for Sample 3: (a) proportion of correct scores versus item number: empirical (*solid line*), theoretical with true parameter values (*broken line*), theoretical with estimated parameter values (*broken-dotted line*); (b) estimated item difficulty parameters under (1) (*solid line*) and the 3PL model (*broken line*); (c) difference between item difficulty estimates; (d) density function of  $\eta$ : theoretical (*solid line*) and fitted (*broken line*); and (e) density function of  $\lambda$ : theoretical (*solid line*) and fitted (*broken line*).

and is designed to be completed in 90 minutes. All items had five alternatives. Because the item-total correlations for the last five pilot items (locations 45, 55, 65, 75 and 85) were poor, these items were dropped, resulting in an analysis of 80 items.

In Table 2 we compare the test speededness model and the 3PL model, both with a common guessing parameter  $c_i = c$ ,  $i = 1, \dots, I$ , in terms of  $-2 \log L$ , AIC and BIC. As is clear, all criteria indicate the test speededness model as the most appropriate one to describe these data.

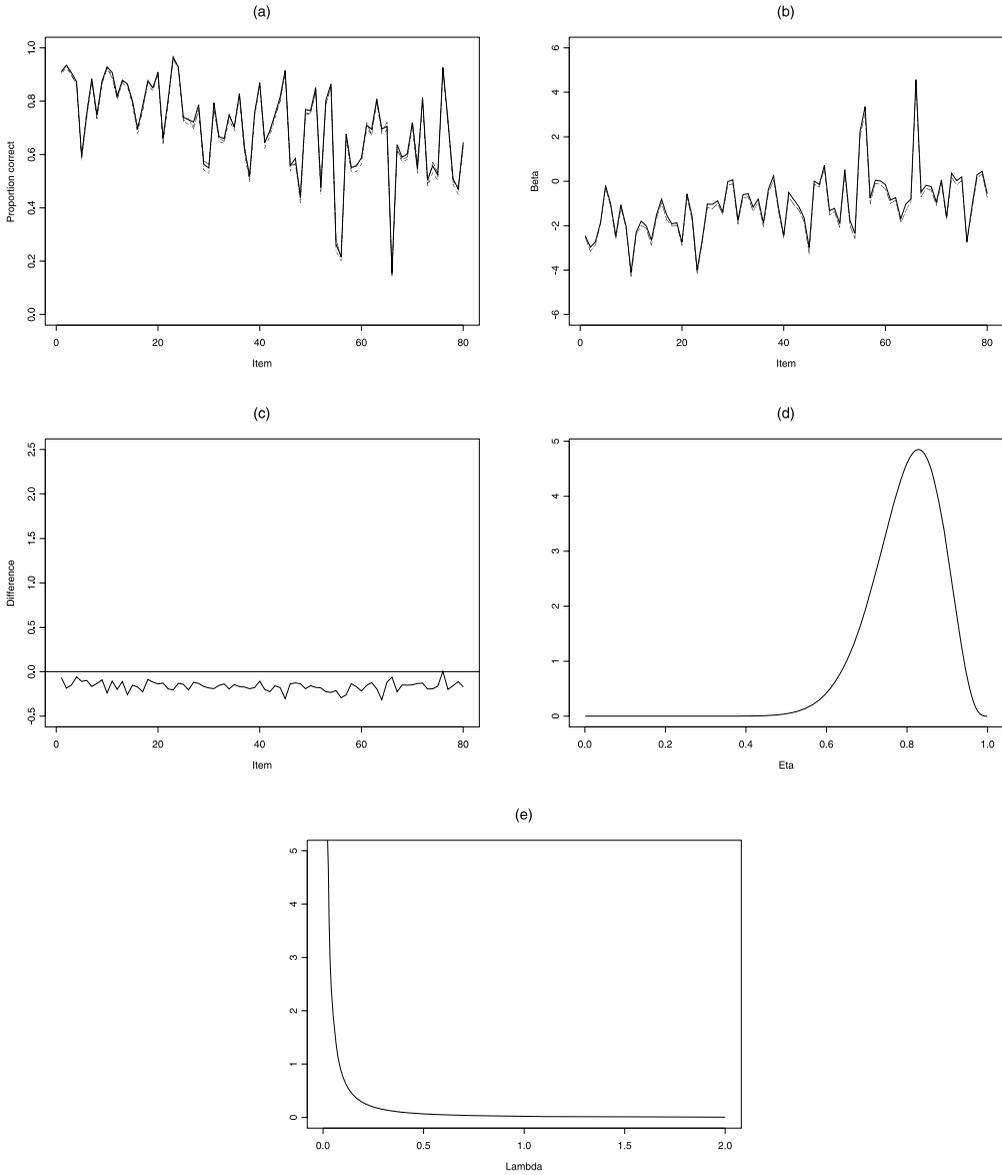


FIGURE 6.

Results for Sample 4: (a) proportion of correct scores versus item number: empirical (*solid line*), theoretical with true parameter values (*broken line*), theoretical with estimated parameter values (*broken-dotted line*); (b) estimated item difficulty parameters under (1) (*solid line*) and the 3PL model (*broken line*); (c) difference between item difficulty estimates; (d) fitted density function of  $\eta$ ; and (e) fitted density function of  $\lambda$ .

TABLE 1.

Goodness-of-fit of the test speededness model versus the 3PL model with a common guessing parameter.

	Sample 1		Sample 2		Sample 3		Sample 4	
	Speeded	3PL	Speeded	3PL	Speeded	3PL	Speeded	3PL
$-2 \log L$	75995	76924	74714	75317	73664	74054	74496	74508
AIC	76331	77246	75059	75639	73999	74376	74832	74830
BIC	77156	78036	75875	76429	74824	75166	75657	75620

TABLE 2.

Mathematics placement test data: Goodness-of-fit of the test speededness model versus the 3PL model with a common guessing parameter.

	Speeded	3PL
$-2 \log L$	252828	254996
AIC	253164	255318
BIC	254196	256308

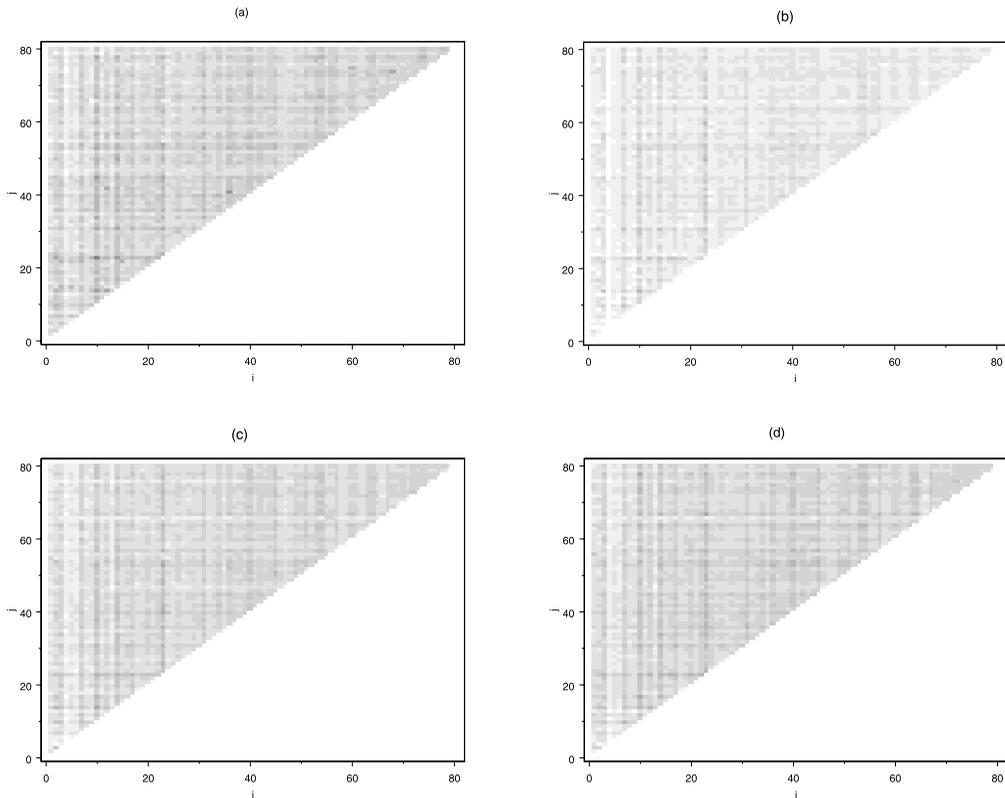


FIGURE 7.

Item pair log-odds ratio plots for: (a) the mathematics placement test data; (b) data simulated from the fitted 3PL model with common guessing parameter; (c) and (d) data simulated from the fitted test speededness model.

Further evidence in favour of the test speededness model can be obtained by comparing plots of empirical item pair log-odds ratios observed in the mathematics test data set and the respective plots obtained on data simulated from the models under investigation. In Figure 7 we show the item pair log-odds ratios of the mathematics placement test data (Figure 7(a)), together with the ones obtained on a data set simulated from the fitted 3PL model (Figure 7(b)) and for two data sets simulated from the test speededness model (Figure 7(c) and (d)). A darker value in the gray-scale matrix refers to a higher value for the empirical log-odds ratio, while a lighter value is chosen for a lower one. If a model explains the dependency structure of the data well, the observed gray-scale matrix should not differ systematically from simulated ones under the model. As is clear from Figure 7, the 3PL model does not account for all dependencies present in the data. On the other hand, the model for test speededness presented in this paper produces gray scale matrices that are almost indistinguishable from the one observed on the mathematics

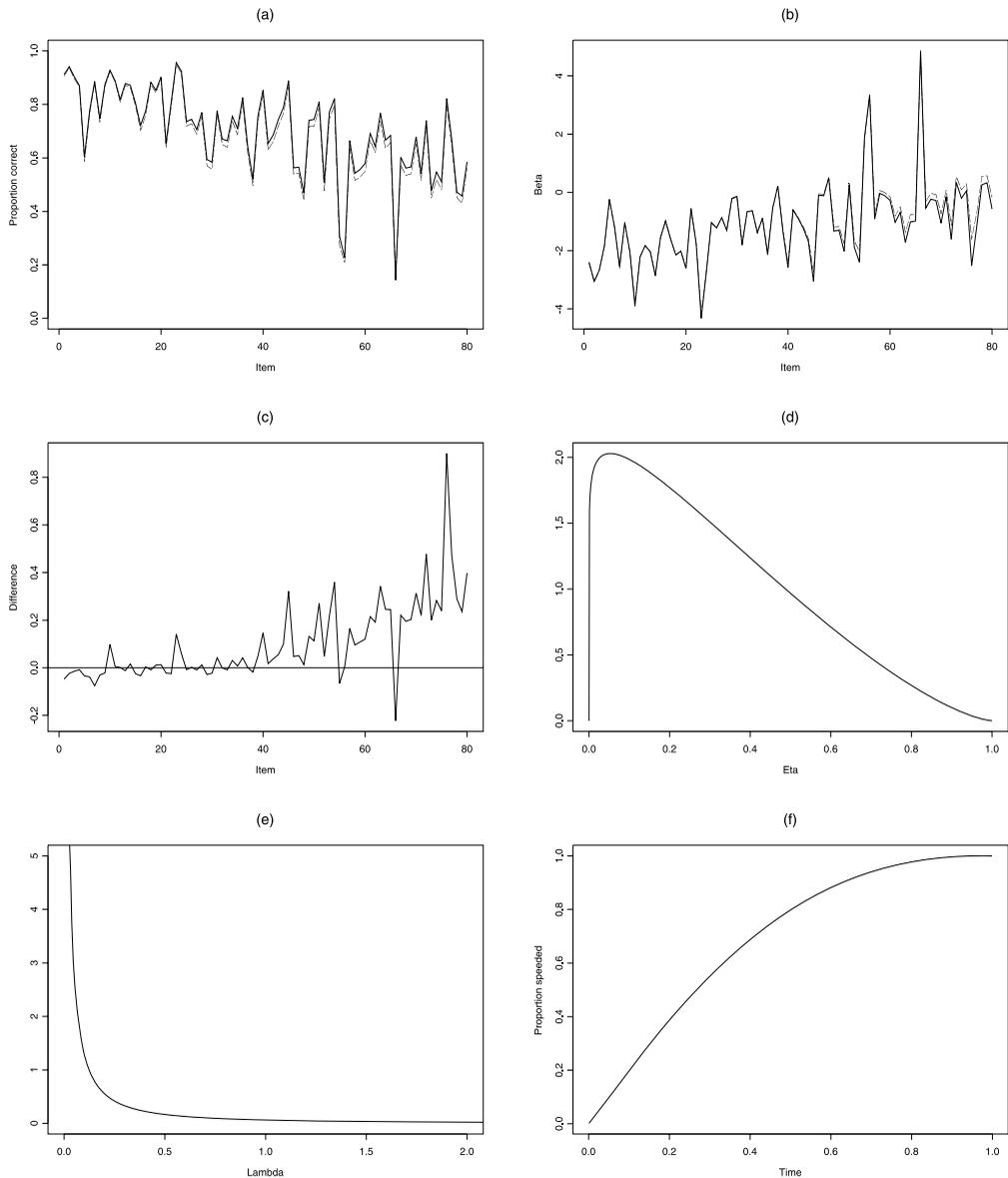


FIGURE 8.

Mathematics placement test data: (a) proportion of correct scores versus item number: empirical (*solid line*), theoretical with estimated parameter values (*broken line*); (b) estimated item difficulty parameters under (1) (*solid line*) and the 3PL model with common guessing parameter (*broken line*); (c) difference between item difficulty estimates; (d) fitted density function of  $\eta$ ; (e) fitted density function of  $\lambda$ ; and (f) proportion of the examinees experiencing test speededness effects as a function of time.

placement test data. Further note that the dependencies tend to become stronger as a function of the item positions. The darkening of the plots starts already quite early, say from items 20–30 on, an observation that is consistent with the estimates obtained for the parameters of the  $\eta_p$  density function. It is worthwhile mentioning that the log-odds plots discussed above indicated the need for the inclusion of discrimination parameters. In particular, model (1), when fitted without item discrimination parameters, yielded log-odds plots which, on average, reproduce the dependency

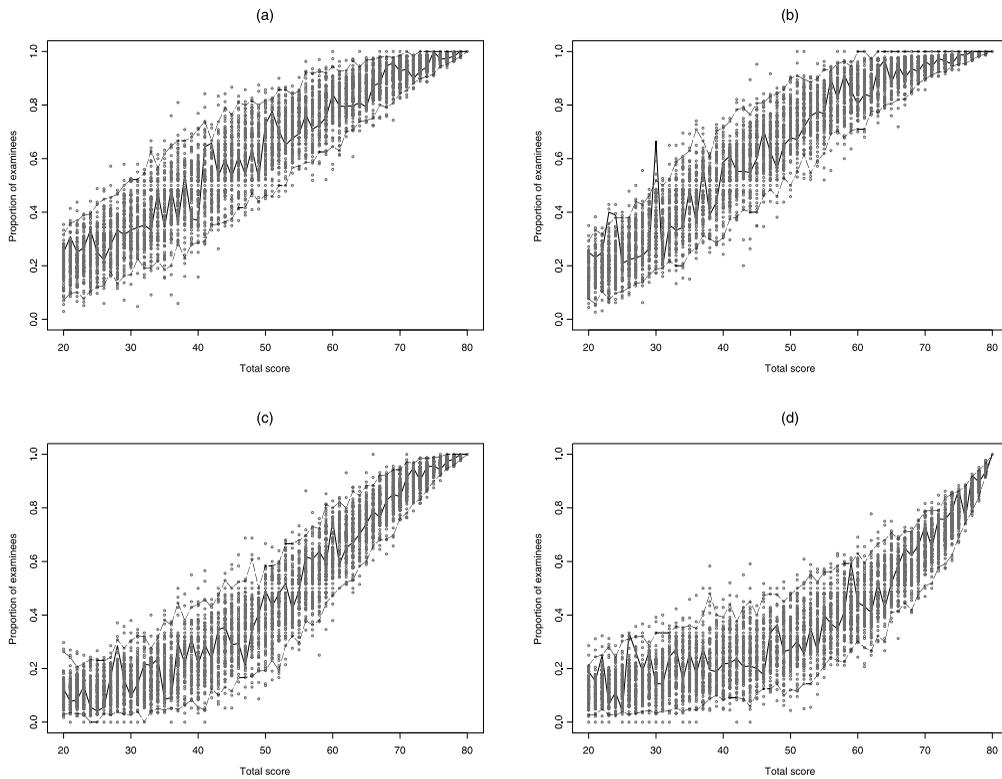


FIGURE 9.

Mathematics placement test data: empirical ICCs of the real data (*solid line*) plotted against bootstrap results (*circles*) and pointwise 95% confidence intervals (*broken lines*): (a) item 16; (b) item 34; (c) item 60; and (d) item 79.

structure quite well, but failed to reproduce the ‘striped’ patterns as observed in the mathematics test data.

We now further evaluate the fit of the proposed test speededness model. In Figure 8(a) we plot the empirical (solid line) and estimated theoretical (broken line) proportions correct scores versus the item number. The proportions correct scores clearly tend to decrease when considered as a function of item number. This does not necessarily indicate test speededness as the items may simply be ordered according to item difficulty, with the more difficult items near the end of the test. Note, however, that the test speededness model produces an almost perfect fit to the data: in Figure 8(a) the estimated theoretical and empirical proportions correct scores are almost indistinguishable. This goodness-of-fit evaluation clearly only involves marginal probabilities and hence only gives a partial picture of the absolute model fit. To evaluate the absolute goodness-of-fit we have used a parametric bootstrap approach. In this we compare the empirical ICCs with those obtained from repeated sampling from the proposed test speededness model with parameters replaced by their maximum likelihood estimates. If the model really fits the data, the observed ICCs should be in line with the simulated ones. The bootstrap procedure was implemented with a uniform  $(-4, 4)$  distribution for the person ability parameters. This choice was made in order to obtain reliable estimates of the ICCs in the lower and upper ranges of ability. In Figure 9 we show for some items the empirical ICCs (solid lines) together with those obtained from 100 bootstrap iterations (dots), as well as pointwise 95% confidence intervals (broken lines). As is clear from this plot, except for a small number of scores, all empirical ICCs are contained in the confidence band based on the bootstrap samples, giving further evidence in favour of the

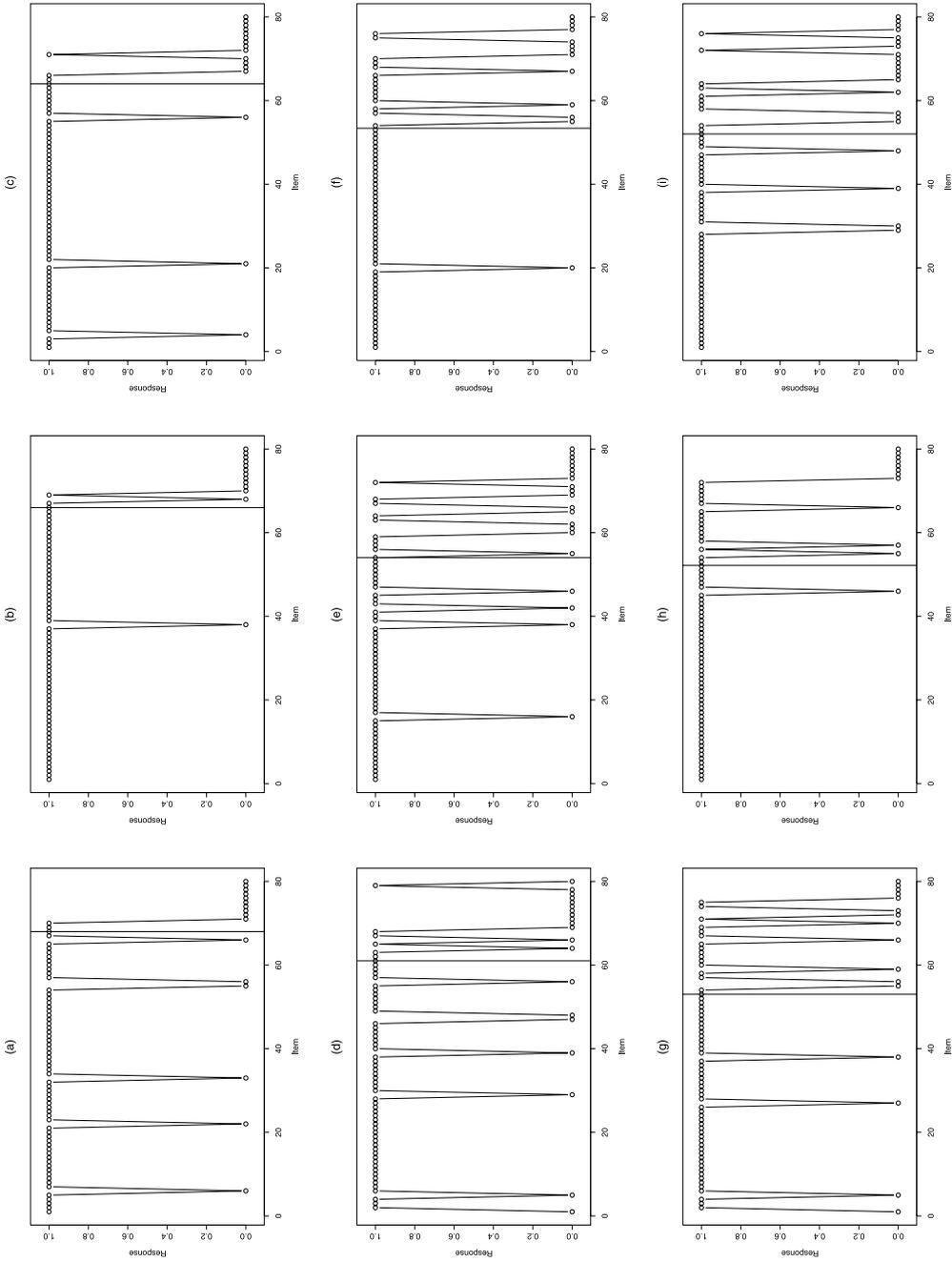


FIGURE 10.

Response profiles of the nine examinees with the largest estimated speededness point: **(a)** examinee 2780; **(b)** examinee 2493; **(c)** examinee 1606; **(d)** examinee 91; **(e)** examinee 2109; **(f)** examinee 1632; **(g)** examinee 850; **(h)** examinee 2075; and **(i)** examinee 2075.

TABLE 3.  
Empirical Bayes estimates for the random effects of the nine examinees with the largest speededness point.

Examinee	$\hat{\theta}_p$	$\hat{\lambda}_p$	$\hat{\eta}_p$
2493	1.0571	22.4043	0.8500
2780	1.8343	22.1902	0.8250
1606	1.4791	17.6195	0.8000
91	0.8175	8.6124	0.7625
2109	0.8360	4.8137	0.6750
1632	1.2097	1.8898	0.6673
354	0.8394	1.4091	0.6625
850	1.3136	1.6181	0.6518
2075	0.6190	3.7148	0.6500

TABLE 4.  
Empirical Bayes estimates of the random effects for the most able examinee (3418), an average examinee not affected by test speededness (53), an average examinee affected by test speededness (432) and the least able examinee (1920).

Examinee	$\hat{\theta}_p$	$\hat{\lambda}_p$	$\hat{\eta}_p$
3418	2.5011	$3.417 \times 10^{-6}$	0.3462
53	0.2107	$1.091 \times 10^{-6}$	0.0065
432	-0.2874	14.2562	0.6250
1920	-3.8906	$1.362 \times 10^{-6}$	$5.757 \times 10^{-9}$

model fit. The bootstrap goodness-of-fit results for the other items are similar to those given in Figure 9. To summarize the goodness-of-fit evaluation we can say that the speededness model fits the mathematics test data well: the model: (a) describes the dependency structure of the data; (b) fits the univariate marginal distributions; and (c) fits the conditional structure of the data.

Further estimation results are graphically represented in Figure 8. In Figure 8(b) and (c) we compare the estimates for the item difficulty parameters obtained under the test speededness model (1) with those obtained under the 3PL model with a common random guessing parameter. The  $\beta_i$  under (1) are in the range  $[-4.329; 4.855]$ . The estimated difficulties of end-of-test items are clearly larger under the 3PL model with common guessing parameter than under the test speededness model. Moreover, the difference between the two item difficulty estimates tends to increase in item number, see Figure 8(c). In Figure 8(d) and (e) we plot the fitted random effect density functions. For the speededness point parameter  $\eta_p$ , we obtained  $\hat{\alpha} = 1.080$  and  $\hat{\beta} = 2.441$ , yielding a Beta distribution with mean 0.307 and standard deviation 0.217. From a practical point of view this result can be interpreted as follows: assuming the postulated Beta distribution holds for the population of speededness points, one can estimate the proportion of the examinees that are speeded at some time point  $t^*$  in the test by the estimated distribution function of  $\eta$ , i.e. by  $\hat{G}_2(t^*)$ . We refer to Figure 8(f) for an illustration of this. Concerning the speededness rate  $\lambda_p$ , the estimates are  $\hat{\mu}_\lambda = -3.604$  and  $\hat{\sigma}_\lambda = 2.771$ , resulting in a log-normal distribution with mean 1.264 and standard deviation 58. Finally, the estimates for the correlation parameters obtained on the mathematics placement test agree with expectations:  $\theta$  and  $\eta$  show a positive correlation ( $\hat{\rho}_{\theta\eta} = 0.659$ ),  $\theta$  and  $\lambda$  a negative ( $\hat{\rho}_{\theta\lambda} = -0.178$ ), and  $\eta$  and  $\lambda$  a positive ( $\hat{\rho}_{\eta\lambda} = 0.319$ ). Moreover, all correlation parameters were found to be highly significantly different from zero on the basis of likelihood ratio tests.

Finally, we calculated the empirical Bayes estimates for the random effects  $\theta_p$ ,  $\eta_p$  and  $\lambda_p$ ,  $p = 1, \dots, P$ . These estimates allow us to identify outlying examinees, or examinees affected by speededness effects. In Table 3 we report the empirical Bayes estimates for the random effects of the nine examinees with the largest estimated test speededness point,  $\hat{\eta}_p$ . Figure 10 shows

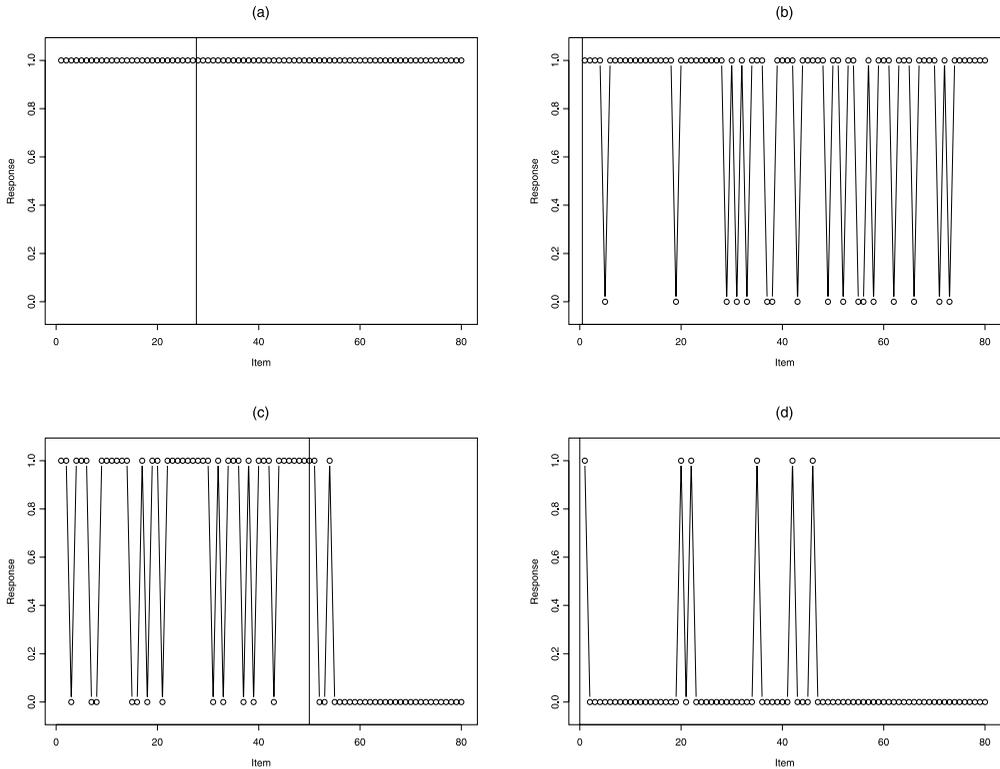


FIGURE 11.

Response profile of: (a) the examinee with the largest ability estimate (examinee 3418); (b) an examinee with an average ability estimate which is not affected by test speededness effects (examinee 53); (c) an examinee with an average ability estimate which is affected by test speededness effects (examinee 432); and (d) the examinee with the smallest ability estimate (examinee 1920).

the respective response profiles, with the vertical reference line representing the estimated test speededness point, i.e.  $80\hat{\eta}_p$ . Note, in particular, the role played by the test speededness point  $\eta_p$  and the speededness rate  $\lambda_p$ . Further, these profiles with a quite abrupt shift from correct to incorrect answers are all associated with examinees having a rather high ability, an observation that is consistent with Yamamoto and Everson (1997). In Table 4 we report the point estimates for the random effects of a very high, two average and a very low ability student; the corresponding response profiles are shown in Figure 11. Examinees with a very low ability, i.e. with almost all items incorrectly answered, do not of course exhibit test speededness effects (both  $\hat{\eta}_p$  and  $\hat{\lambda}_p$  very small). The same applies to students with a very high ability (almost all items correctly answered). Examinees with intermediate ability estimates may either show a speeded or a nonspeeded response profile.

## 5. Discussion and Conclusion

In this paper we proposed an IRT model dealing with test speededness. The model can be seen as consisting of two random processes, a classical IRT process and a random guessing process, with the random guessing process gradually taking over from the IRT process. Both change point and change rate are considered as random effects in order to model examinee differences in both respects. Moreover, the speededness random effects are allowed to be dependent

and besides may also depend on the ability of the examinee. The model improves on the hybrid model of Yamamoto and Everson (1997) in the sense that examinees do not switch immediately to random guessing once they become speeded. The model also extends the mixture IRT model approach (Bolt et al., 2002, 2003), by allowing examinees to become speeded at different points in the test. From the simulation study we may conclude that recovery of the parameter values of the test speededness model is rather good and that the model can be differentiated from the 3PL model by using information criteria such as AIC and BIC. Inference concerning the fixed effects of the proposed model can be drawn on the basis of likelihood ratio tests.

As mentioned in the simulation section, it can take a quite long time to estimate the model. These long computation times can to a large extent be traced back to the approximation of the three-dimensional integral when computing the marginal probability of a response profile, a computation that needs to be done for each of the examinees. These computations are essentially of the same nature for all examinees, and hence considerable gains can be achieved by running these computations in parallel. Another approach is to start with fitting a simple, unidimensional IRT model rather than the postulated model for test speededness, followed by an investigation of the need for more a complex model. This investigation can be done in a formal way on the basis of tests for unidimensionality, as proposed by, among others, Stout (1987, 1990), or using an informal graphical approach comparing, for instance, item pair log-odds plots for the data with similar plots for data sets generated from the fitted model. The latter approach is illustrated in Figure 7.

The model we presented is an instantiation of a more general category of models with gradual change in a series of repeated observations. For example, in a learning experiment one may start with guessing because one has no insight into how to solve the items, whereas later in the series a gradual shift may occur to a more appropriate strategy to actually solve the items, thanks to learning. This change process could be modelled in a way that is complementary to the speededness model, with a transition from guessing to solving instead of a transition from solving to guessing as for the case of speededness. From a general perspective, one may consider any transition between two strategies or two principles during a series of repeated observations, given that the two strategies or principles correspond to models that can also be estimated separately. This opens up a rather broad category of applications in psychology and in other disciplines.

The model considered assumes a dichotomous response (incorrect/correct) with test speededness gradually degrading responses towards incorrect answers. However, besides more frequent wrong answers, test speededness may also result in omitted answers. This omission may, next to test speededness, also depend on ability and hence the nonresponse is, using the terminology of Little and Rubin (1987), missing not at random (MNAR). An early attempt to model nonresponse in test data can be found in Lord (1983), where a trinomial response model (omit/incorrect/correct) is proposed with dropout being examinee specific. Extensions of this model including test speededness are worthwhile considering. We refer to Béguin (2003), Pimentel (2005) and Goegebeur, De Boeck, Molenberghs and del Pino (2006) for recent contributions. Next to these models, the selection and pattern-mixture models (see, e.g. Glynn, Laird & Rubin, 1986), two popular dropout models in the biomedical sciences, may also deserve attention in this respect. This is a topic of ongoing research.

#### Appendix 1. Example SAS Code

```
data geg;
infile 'c:\irm1\simul3nn.txt';
input y nr person x1-x80;
nr_n = nr/80;
run;
```

```

proc nlmixed data=geg method=gauss noad technique=newrap
maxiter=500 maxfu=5000 qpoints=5;
parms b1-b80=-1 a1-a80=1 c=.2 mlambda=0 slambda2=1 a=2 b=2 r12=0 r13=0
r23=0;

alpha =
a1*x1+a2*x2+a3*x3+a4*x4+a5*x5+a6*x6+a7*x7+a8*x8+a9*x9+a10*x10+
a11*x11+a12*x12+a13*x13+a14*x14+a15*x15+a16*x16+a17*x17+a18*x18+a19*x19
+a20*x20+
a21*x21+a22*x22+a23*x23+a24*x24+a25*x25+a26*x26+a27*x27+a28*x28+a29*x29
+a30*x30+
a31*x31+a32*x32+a33*x33+a34*x34+a35*x35+a36*x36+a37*x37+a38*x38+a39*x39
+a40*x40+
a41*x41+a42*x42+a43*x43+a44*x44+a45*x45+a46*x46+a47*x47+a48*x48+a49*x49
+a50*x50+
a51*x51+a52*x52+a53*x53+a54*x54+a55*x55+a56*x56+a57*x57+a58*x58+a59*x59
+a60*x60+
a61*x61+a62*x62+a63*x63+a64*x64+a65*x65+a66*x66+a67*x67+a68*x68+a69*x69
+a70*x70+
a71*x71+a72*x72+a73*x73+a74*x74+a75*x75+a76*x76+a77*x77+a78*x78+a79*x79
+a80*x80;

beta =
b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+b9*x9+b10*x10+
b11*x11+b12*x12+b13*x13+b14*x14+b15*x15+b16*x16+b17*x17+b18*x18+b19*x19
+b20*x20+
b21*x21+b22*x22+b23*x23+b24*x24+b25*x25+b26*x26+b27*x27+b28*x28+b29*x29
+b30*x30+
b31*x31+b32*x32+b33*x33+b34*x34+b35*x35+b36*x36+b37*x37+b38*x38+b39*x39
+b40*x40+
b41*x41+b42*x42+b43*x43+b44*x44+b45*x45+b46*x46+b47*x47+b48*x48+b49*x49
+b50*x50+
b51*x51+b52*x52+b53*x53+b54*x54+b55*x55+b56*x56+b57*x57+b58*x58+b59*x59
+b60*x60+
b61*x61+b62*x62+b63*x63+b64*x64+b65*x65+b66*x66+b67*x67+b68*x68+b69*x69
+b70*x70+
b71*x71+b72*x72+b73*x73+b74*x74+b75*x75+b76*x76+b77*x77+b78*x78+b79*x79
+b80*x80;

lambda=exp(mlambda+slambda2**.5*nlambda);
eta=betainv(probnorm(neta), a, b);
r=exp(alpha*theta-beta)/(1+exp(alpha*theta-beta));
s=(1-(nr_n-eta))*lambda;
if (s >= 1) then pr=c+(1-c)*r; else pr=c+(1-c)*r*s;
model y ~ binary(pr);
random theta nlambda neta ~ normal([0,0,0],[1,r12,1,r13,r23,1])
subject=person;
run;

```

## Appendix 2

We start by introducing the concept copula function as well as a fundamental theorem by Sklar (1959), stating that every joint distribution function can be decomposed into its marginal distribution functions and a copula function, i.e. a function describing the dependency structure.

**Definition 1.** An  $n$ -copula is a function  $C : [0, 1]^n \rightarrow [0, 1]$  with the following properties:

1. for every  $\mathbf{u} \in [0, 1]^n$  with at least one coordinate equal to 0,  $C(\mathbf{u}) = 0$
2. if all coordinates of  $\mathbf{u}$  are 1 except  $u_k$ , then  $C(\mathbf{u}) = u_k$
3. for all  $\mathbf{a}, \mathbf{b} \in [0, 1]^n$  with  $\mathbf{a} \leq \mathbf{b}$  the volume of the hyperrectangle with corners  $\mathbf{a}$  and  $\mathbf{b}$  is positive, i.e.

$$\sum_{i_1=1}^2 \cdots \sum_{i_n=1}^2 (-1)^{i_1+\cdots+i_n} C(u_{i_1}, \dots, u_{i_n}) \geq 0,$$

where  $u_{i_1} = a_i$  and  $u_{i_2} = b_i$ .

So essentially an  $n$ -copula is an  $n$ -dimensional distribution function on  $[0, 1]^n$  with standard uniform marginal distributions. The next theorem, due to Sklar, is central to the theory of copulas and forms the basis of the applications of that theory to statistics.

**Theorem 1** (Sklar, 1959). *Let  $\mathbf{X}' = (X_1, \dots, X_n)$  be a random vector with joint distribution function  $F_{\mathbf{X}}$  and marginal distribution functions  $F_i$ ,  $i = 1, \dots, n$ . Then there exists a copula  $C$  such that, for all  $\mathbf{x} \in \mathbb{R}^n$ ,*

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \dots, F_n(x_n)). \quad (4)$$

*If  $F_1, \dots, F_n$  are all continuous, then  $C$  is unique, otherwise  $C$  is uniquely determined on  $\text{Ran } F_1 \times \cdots \times \text{Ran } F_n$ . Conversely, given a copula  $C$  and marginal distribution functions  $F_1, \dots, F_n$ , the function  $F_{\mathbf{X}}$  as defined by (4) is a joint distribution function with margins  $F_1, \dots, F_n$ .*

As is clear, Sklar's theorem separates a joint distribution into a part that describes the dependence structure (the copula) and parts that describe the marginal behaviour (the marginal distributions). The second statement in the above theorem is very useful for modelling purposes as it gives a very convenient and flexible way to construct a joint distribution function. Indeed, all one has to do is to select models for the univariate marginal distributions and for the dependence structure. In our context, the latter property can be used to build a model for  $G$  in (2). Popular copula functions in this respect are, among others, the normal, Sarmanov (Lee, 1996), Clayton (Clayton, 1978), Frank (Frank, 1979) and Gumbel–Hougaard (Gumbel, 1960) copulas. For further details on copula functions we refer to Joe (1997) and Nelsen (1999).

The following proposition states that for a joint distribution with a normal copula function but arbitrary (continuous) marginal distributions, appropriately chosen compositions of probability and inverse probability integral transforms yield a multivariate normal distribution. The importance of the result stems from that fact that it provides the basis for using the SAS NLMIXED procedure, which allows only multivariate normal random effect distributions, in cases characterized by a normal copula function with otherwise arbitrary (continuous) distribution functions. The result is a special case of the general property that increasing transformations of the marginal distributions only affect the marginal distributions and not the dependence function.

**Proposition 1.** *Consider an  $n$ -dimensional random vector  $\mathbf{X}$  with joint distribution function  $G$  and continuous marginal distribution functions  $G_1, \dots, G_n$ . Assume that  $G$  is characterized by a normal dependence function (copula)  $C$ , i.e.*

$$G(x_1, \dots, x_n) = C(G_1(x_1), \dots, G_n(x_n))$$

with

$$C(u_1, \dots, u_n) = \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_n)} \frac{1}{(2\pi)^{n/2} |\mathbf{R}|^{1/2}} e^{-(1/2)\mathbf{z}'\mathbf{R}^{-1}\mathbf{z}} d\mathbf{z}, \quad (5)$$

in which  $\mathbf{R}$  denotes a (positive definite) correlation matrix and  $\Phi^{-1}$  is the inverse standard normal distribution function. Then the random variables

$$Y_i = \Phi^{-1}(G_i(X_i)), \quad i = 1, \dots, n,$$

are jointly distributed as multivariate normal.

*Proof:* Denote the joint distribution function of  $Y_1, \dots, Y_n$  by  $H$ . Then

$$\begin{aligned} H(y_1, \dots, y_n) &= P(Y_1 \leq y_1, \dots, Y_n \leq y_n) \\ &= P(\Phi^{-1}(G_1(X_1)) \leq y_1, \dots, \Phi^{-1}(G_n(X_n)) \leq y_n) \\ &= P(G_1(X_1) \leq \Phi(y_1), \dots, G_n(X_n) \leq \Phi(y_n)) \\ &= C(\Phi(y_1), \dots, \Phi(y_n)) \\ &= \int_{-\infty}^{\Phi(y_1)} \cdots \int_{-\infty}^{\Phi(y_n)} \frac{1}{(2\pi)^{n/2} |\mathbf{R}|^{1/2}} e^{-(1/2)\mathbf{z}'\mathbf{R}^{-1}\mathbf{z}} d\mathbf{z}, \end{aligned}$$

which is the distribution function of a multivariate normal distribution.  $\square$

### Appendix 3. Empirical Bayes Estimation of the Random Effects

Although the model estimation implies an estimate of the parameters of the marginal distribution of  $\mathbf{Y}$ , it is common practice in psychometrics to also calculate the estimations of the person parameters. These are in our case the ability parameters  $\theta_p$ , and the test speededness parameters  $\lambda_p$  and  $\eta_p$ ,  $p = 1, \dots, P$ . These random effects estimates give an idea about the between-subject variability, and hence provide information that is helpful for detecting special profiles, say outlying individuals, or groups of individuals evolving differently in time, in our context individuals affected by test speededness effects. To obtain estimates for the random effects, we need their conditional posterior distribution. For notational convenience we split  $\boldsymbol{\xi}$  into subvectors  $\boldsymbol{\xi}_1$  and  $\boldsymbol{\xi}_2$ , with  $\boldsymbol{\xi}'_1 = (\beta_1, \dots, \beta_I, \alpha_1, \dots, \alpha_I, c_1, \dots, c_I)$  and  $\boldsymbol{\xi}'_2 = (\mu_\lambda, \sigma_\lambda^2, \alpha, \beta, \rho_{\theta\eta}, \rho_{\theta\lambda}, \rho_{\eta\lambda})$ . Using Bayes' rule we have

$$p(\theta_p, \eta_p, \lambda_p | \mathbf{y}_p, \boldsymbol{\xi}) = \delta_p p(\mathbf{y}_p | \theta_p, \eta_p, \lambda_p, \boldsymbol{\xi}_1) p(\theta_p, \eta_p, \lambda_p | \boldsymbol{\xi}_2), \quad (6)$$

where  $\delta_p$  is the normalizing constant, i.e.  $\delta_p = 1/p(\mathbf{y}_p | \boldsymbol{\xi})$ , and with

$$p(\mathbf{y}_p | \theta_p, \eta_p, \lambda_p, \boldsymbol{\xi}_1) = \prod_{i=1}^I \pi_{pi}^{y_{pi}} (1 - \pi_{pi})^{1-y_{pi}}$$

and

$$p(\theta_p, \eta_p, \lambda_p | \boldsymbol{\xi}_2) = \frac{1}{(2\pi)^{3/2} |\mathbf{R}|^{1/2} \sigma_\lambda \lambda_p \phi(\Phi^{-1}(G_2(\eta_p)))} e^{-\boldsymbol{\gamma}'\mathbf{R}^{-1}\boldsymbol{\gamma}/2}. \quad (7)$$

In (7),  $\phi$  denotes the standard normal density function, and

$$\boldsymbol{y}' = \left( \theta_p, \Phi^{-1}(G_2(\eta_p)), \frac{\ln \lambda_p - \mu_\lambda}{\sigma_\lambda} \right).$$

Expression (7) is obtained by differentiating  $G(\theta_p, \eta_p, \lambda_p) = C(G_1(\theta_p), G_2(\eta_p), G_3(\lambda_p))$ , where  $C$  is given by (5), with respect to  $\theta_p$ ,  $\eta_p$  and  $\lambda_p$ , combined with the facts that  $G_1(\theta_p) = \Phi(\theta_p)$  and  $G_3(\lambda_p) = \Phi((\ln \lambda_p - \mu_\lambda)/\sigma_\lambda)$ . The mode of (6) is used as a point estimate for  $\theta_p$ ,  $\eta_p$  and  $\lambda_p$ . More specifically, the empirical Bayes estimate  $(\hat{\theta}_p, \hat{\eta}_p, \hat{\lambda}_p)$  is the value for  $(\theta_p, \eta_p, \lambda_p)$  that maximizes  $p(\theta_p, \eta_p, \lambda_p | \boldsymbol{y}_p, \boldsymbol{\xi})$ , in which the unknown parameters in  $\boldsymbol{\xi}$  have been replaced by their estimates obtained from the marginal maximum likelihood estimation.

#### References

- Béguin, A. (2003). A Bayesian estimation procedure for speeded tests. Paper presented at the 13th International Meeting of the Psychometric Society, Sardinia, Italy.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 394–479). Reading: Addison-Wesley.
- Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.
- Bolt, D.M., Mroch, A.A., & Kim, J.-S. (2003). An empirical investigation of the Hybrid IRT model for improving item parameter estimation in speeded tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.
- Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.
- Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*, 141–151.
- Douglas, J., Kim, H.R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*, 129–151.
- Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*, 119–140.
- Frank, M.J. (1979). On the simultaneous associativity of  $F(x, y)$  and  $x + y - F(x, y)$ . *Aequationes Mathematicae*, *19*, 194–226.
- Glynn, R.J., Laird, N.M., & Rubin, D.B. (1986). Selection modelling versus mixture modelling with nonignorable non-response. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 115–142). New York: Springer.
- Goegebeur, Y., De Boeck, P., Molenberghs, G., & del Pino, G. (2006). A local-influence-based diagnostic approach to a speeded item response theory model. *Applied Statistics*, *55*, 647–676.
- Gumbel, E.J. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, *9*, 171–173.
- Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*, 109–132.
- Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.
- Lee, M.-L.T. (1996). Properties and applications of Sarmanov's family of bivariate distributions. *Communications in Statistics: Theory and Methods*, *25*, 1207–1222.
- Lee, G., Kolen, M.J., Frisbie, D.A., & Ankenmann, R.D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, *25*, 357–372.
- Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477–482.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- NAG. (1993). *NAG Fortran library manual—Mark 19*. The Numerical Algorithms Group Limited.
- Nelsen, R.B. (1999). *An introduction to copulas*. New York: Springer.
- Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.
- Pimentel, J.L. (2005). *Item response theory modelling with nonignorable missing data*. PhD thesis, University of Twente, Enschede.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425–435.

- Rosenbaum, P. (1988). Item bundles. *Psychometrika*, *53*, 349–359.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.
- Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.
- Sklar, A. (1959). Fonctions de repartition a  $n$  dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.
- Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.
- Stout, W.F. (1990). A new item response theory modelling approach with application to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.
- Stout, W.F., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.
- Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, *26*, 247–260.
- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195.
- Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Exploratory item response models—a generalized linear and nonlinear approach* (pp. 289–316). New York: Springer.
- van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123–140.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice*, *15*, 22–29.
- Wollack, J.A., & Cohen, A.S. (2004). A model for simulating speeded test data. Paper presented at the annual meeting of the American Educational Research Association, San Diego.
- Wollack, J.A., Cohen, A.S., & Wells, C.S. (2003). The effects of test speededness on score scale stability. *Journal of Educational Measurement*, *40*, 307–330.
- Yamamoto, K. (1987). *A model that combines IRT and latent class models*. PhD thesis, University of Illinois, Champaign-Urbana.
- Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–99). New York: Waxmann.
- Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

*Manuscript received 1 SEP 2005*

*Final version received 1 JUL 2007*

*Published Online Date: 25 SEP 2007*