PSYCHOMETRIKA—VOL. 72, NO. 3, 287–308
SEPTEMBER 2007
DOI: 10.1007/s11336-006-1478-z

A HIERARCHICAL FRAMEWORK FOR MODELING SPEED AND ACCURACY ON TEST ITEMS

WIM J. VAN DER LINDEN

UNIVERSITY OF TWENTE

Current modeling of response times on test items has been strongly influenced by the paradigm of experimental reaction-time research in psychology. For instance, some of the models have a parameter structure that was chosen to represent a speed-accuracy tradeoff, while others equate speed directly with response time. Also, several response-time models seem to be unclear as to the level of parametrization they represent. A hierarchical framework for modeling speed and accuracy on test items is presented as an alternative to these models. The framework allows a "plug-and-play approach" with alternative choices of models for the response and response-time distributions as well as the distributions of their parameters. Bayesian treatment of the framework with Markov chain Monte Carlo (MCMC) computation facilitates the approach. Use of the framework is illustrated for the choice of a normal-ogive response model, a lognormal model for the response times, and multivariate normal models for their parameters with Gibbs sampling from the joint posterior distribution.

Key words: hierarchical modeling, item-response theory, Gibbs sampler, Markov chain Monte Carlo estimation, speed-accuracy tradeoff, response times.

In addition to the responses on test items, the times needed to produce them are an important source of information on the test takers and the items. Their information may help us to improve such operational activities as item calibration, test design, item selection in adaptive testing, diagnosis of response behavior for possible aberrances, and the allowance of testing accommodations. These applications have become within our reach now that computerized testing with automatic recording of response times is replacing paper-and-pencil testing.

An important prerequisite for the use of response times is an appropriate statistical model for their distribution. Over the last two decades, different models for response times have been presented; a selection of them will be reviewed below. Several of these models appear to be influenced by the paradigm of experimental reaction-time research in psychology (see, e.g., Luce, 1986). Key features of the paradigm are: (1) the use of standardized tasks; (2) the equating of the subjects' speed on these task with their reaction times; and (3) experimental manipulation of the conditions under which the subjects operate. The paradigm is used, for instance, to tests hypotheses about underlying psychological processes or to decompose reaction times into the times needed for different processes or operations.

An important notion from reaction-time research with a special impact on the current modeling of response times on test items is that of a speed-accuracy tradeoff. The notion is based on the observation that when working on a task, a subject has the choice between working faster with lower accuracy or more slowly with higher accuracy. A typical form of this tradeoff is presented in Figure 1, where each of the hypothetical observations is for a different combination of speed and accuracy. Observe that speed has been presented as the independent variable but

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this paper are those of the author and do not necessarily reflect the policy and position of LSAC. The author is indebted to the American Institute of Certified Public Accountants for the data set in the empirical example and to Rinke H. Klein Entink for his computational assistance.

Requests for reprints should be sent to W. J. van der Linden, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: w.j.vanderlinden@utwente.nl

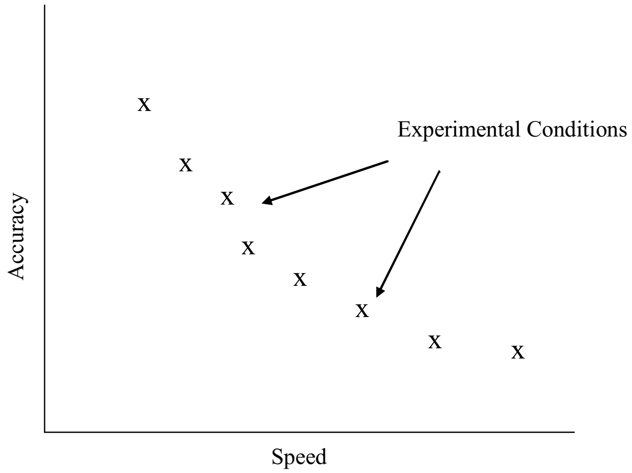


FIGURE 1.
Hypothetical example of a speed-accuracy tradeoff.

accuracy as the dependent variable; this choice is motivated by the fact that, when working on test items, a subject has control of his or her speed over a range of possible levels and has to accept the accuracy that is the result of the choice of speed (van der Linden, 2007a).

Generally, a speed-accuracy tradeoff can be described as a negative (nonlinear) correlation between the speed and accuracy levels at which a person is able to operate. These combinations can be observed, for instance, when a subject is instructed to repeat a task at different levels of speed. For the sequel of this paper, it is important to note that a speed-accuracy tradeoff is a *within-person* phenomenon. As will be shown later, some of the current models confound this level of observation with that of a single observation of a fixed person or a population of persons. This is incorrect; for example, it is perfectly possible for a population of persons to show a positive correlation between speed and accuracy while for each individual person the choice between speed and accuracy is constrained by the negative correlation in Figure 1.

In our review in the next section, we show how the parameter structure of some response-time models in the test-theory literature are the result of an attempt to incorporate a speed-accuracy tradeoff in the model. Other ideas from experimental reaction-time research with an impact on response-time modeling in test theory are the direct equating of time with the speed at which a person operates and the assumption of identically distributed times for a given person across tasks. We will argue that these ideas are inappropriate for response times on test items and then present a hierarchical framework for the analysis of speed and accuracy that is expected to better suit their specific nature. The framework also disentangles the levels of modeling that seem to be confounded in the current literature. Basically, it consists of an item-response theory (IRT) model, a model for the response-time distribution, and a higher-level structure to account for the dependences between the item and person parameters in these models. The framework is flexible in that we can substitute any IRT or response-time model that fits the format of the test items best. The same holds for the higher-level models for their parameters.

The "plug-and-play approach" allowed by this framework is greatly facilitated by a Bayesian treatment of their parameters with Gibbs sampling from their joint posterior distribution. The replacement of a model in the framework by a different plug-in leads only to the replacement of the corresponding steps in the Gibbs sampler. We will illustrate the treatment for the choice of a normal-ogive response model, a lognormal model for the response times, and multivariate normal models for their parameters. For an appropriate choice of prior distributions, all distributions in the sampler are known and its application becomes straightforward.

Current Models

Verhelst, Verstralen, and Jansen (1997) present a model for time-limit tests that is based on the assumption of a generalized extreme-value distribution of a latent response variable given the time spent on the item and a gamma distribution for the time. Capitalizing on the fact that the compound of these two distributions is a generalized logistic (Dubey, 1969), they arrive at the following model for the probability of a correct response on item i by person j,

$$p_i(\theta_j) = [1 + \exp(\theta_j - \ln \tau_j - b_i)]^{-\pi_i}, \tag{1}$$

where b_i is the difficulty parameter for item i, θ_j the ability parameters for person j, τ_j is interpreted as a speed parameter for person j, and π_i is an item-dependent shape parameter. For $\pi_i = 1$, the model reduces to a Rasch (1980) type model with $\xi_j = \theta_j - \ln \tau_j$ replacing the traditional ability parameter. Observe that $\theta_j - \ln \tau_j$ is person dependent only and governs the probability of a correct response; the accuracy at which a test taker operates is thus controlled by this composite parameter instead of the ability parameter θ_j in a regular IRT model. The authors of the model highlight the fact that it incorporates a speed-accuracy tradeoff. If a person decides to increase the speed τ_j, for fixed θ_j parameter ξ_j decreases and the effect is lower accuracy.

A similar model was derived by Roskam (1987; see also Roskam, 1997). His model is a Rasch model with its additive parameter structure extended with the logtime on the item as a regressor

$$p_i(\theta_j) = [1 + \exp(\theta_j + \ln t_{ij} - b_i)]^{-1}. \tag{2}$$

This model assumes a similar type of speed-accuracy tradeoff but now directly between the ability of the test taker and the actual time spent on a test item. Less time on the item indicates a higher speed and results in lower accuracy. Unlike the preceding model, (2) measures speed by the actual time spent on the item.

An entirely different type of model was introduced in Thissen (1983). This model assumes the following parameter structure for the logtime on an item:

$$\ln T_{ij} = \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij}, \tag{3}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma). \tag{4}$$

Parameters τ_j and β_i can be interpreted as the slowness of the test taker and the amount of time required by the item, respectively, whereas μ is a general level parameter, and a_i, θ_j, and b_i are the usual item-discrimination, ability, and item-difficulty parameters. The term $\rho(a_i\theta_j - b_i)$ represents a regression of the traditional parameter structure of a two-parameter (unidimensional) response model on the logtime with ρ as the regression parameter. The normally distributed random term ε_{ij} in (3) indicates that the model belongs to a lognormal family. Although they model the distribution response time instead of responses, for $\rho < 0$, (3)–(4) imply a similar tradeoff between speed (here: slowness) and accuracy as (1) and (2). But for $\rho > 0$ the relation reverses.

A model based on a Weibull distribution with a shift or location parameter was proposed in Rouder, Sun, Speckman, Lu, and Zhou (2003) and Tatsuoka and Tatsuoka (1980). The choice of a Weibull distribution is a classical one in industrial statistics, where it is used to model waiting times for a system failure as a function of the probabilities of a failure of its components. Rouder et al. posit a reaction-time distribution for person j on task i with density

$$f(t_{ij}) = \frac{\pi_j(t_{ij} - \psi_j)^{\pi_j - 1}}{\sigma_j^{\pi_j}} \exp\left\{ -\left(\frac{t_{ij} - \psi_j}{\sigma_j}\right)^{\pi_j} \right\}, \quad t_{ij} > \psi_j, \tag{5}$$

where ψ_j is a shift, σ_j a scale, and π_j a shape parameter. This choice of parametrization is motivated by the nature of the psychological processes typically studied in reaction-time experiments as well as its statistical tractability. For example, at reasonable values for the shape parameter, the left tail of the distribution in (5) falls off rapidly leading to better identifiability of all three parameters in reasonably sized samples. Tatsuoka and Tatsuoka drop the restriction on ψ_j and treat it as a location parameter, for which they substitute the average response time on the set of test items, $\bar{t}_j$.

Unlike the preceding three models, the two versions of the Weibull model are pure response-time models. They do not assume anything about an ability of the person or the features of the items. In fact, they do not even adopt any item parameters at all but treat the response times for a fixed person as identically distributed across items. This assumption seems reasonable for the experimental paradigm, for which Rouder et al. (2003) presented their model, but certainly does not hold for the case of response times on test items addressed in Tatsuoka and Tatsuoka (1980).

A response-time model that does account for differences between test items is that by Oosterloo (1975) and Scheiblechner (1979, 1985). They model response times as an exponential distribution with density

$$f(t_{ij}) = (\tau_j + \beta_i)\exp[-(\tau_j + \beta_i)t_{ij}], \tag{6}$$

with τ_j and β_i as person and item parameter. Since it holds for the exponential distribution that

$$E(T_{ij}) = \frac{1}{\tau_j + \beta_i}, \tag{7}$$

the parameters are interpreted by these authors as the speed of the person and the item, respectively.

The model in (6) is also derived from the waiting-time literature. It is known to represent the time for a Poisson process to produce its first success. Though behavior on some elementary tasks may be modeled as a Poisson process, we do not believe the model to be generally adequate for response times on test items. For instance, exponential distributions have their mode at $t_{ij} = 0$, which simply is not realistic for times on test items that typically run into tens of seconds or even minutes.

This review is not complete; for example, it does not include the Poisson model for reading speed by Rasch (1980), which has been studied extensively by Jansen (e.g., 1986, 1997a, 1997b; Jansen & Duijn, 1992), the additive and multiplicative gamma models by Maris (1993), the mixed and conditional logistic models for responses and response times in van Breukelen (2005), and the model for multivariate survival times with latent covariates by Douglas, Kosorok, and Chewning (1999). The models above have only been chosen to prepare our discussion in the next section. For a more complete review of response-time models for test items, see Schnipke and Scrams (2002).

Discussion

The first two models were motivated by the idea of a speed-accuracy tradeoff. The existence of such a tradeoff is supported by overwhelming evidence in reaction-time research. In testing, the tradeoff explains the typical behavior at the end of an unfortunately speeded test in the form of relatively large numbers of omitted responses and/or random guesses. But on a test with a reasonable time limit, unless the test taker changes his or her speed during the test, there is no necessity whatsoever to incorporate a tradeoff in a response-time model for a fixed person and a fixed set of test items. The only thing that counts is the actual level of speed at which the test taker has chosen to operate on the items. As Figure 1 illustrates, once the speed is fixed, accuracy is also fixed. For the typical hybrid type of test considered in this paper (see below), all we need is two *free* parameters to represent the test taker's speed and accuracy. Any attempt to constrain these parameters easily leads to a misspecification and, consequently, a less satisfactory empirical fit of the model.

Second, some of the models above can be viewed as the results of a confounding of the level of modeling. Three different levels should be distinguished:

(1) the within-person level, at which the value of the person parameters is allowed to change over time (e.g., due to a change of strategy or external conditions);
(2) the fixed-person level, at which the parameters remain constant; and
(3) the level of a population of fixed persons, for which we have a distribution of parameter values across persons.

The idea of incorporating a speed-accuracy tradeoff in the first two models seems to confound the within-person and fixed-person levels. The tradeoff can only become manifest as the result of a change of strategy or condition. But the two models are for a person with fixed levels of ability and speed. Another example of confounding occurs in Tatsuoka and Tatsuoka (1980), where the same Weibull model in (5) is used for the response times of a fixed person and a random person from a population. The same happens for a lognormal model in Schnipke and Scrams (1997, 1999).

In (2), speed is directly equated with the actual response time. Intuitively, speed on a test is a measure of the amount of labor accomplished in a time interval. Therefore, measuring speed as the time needed to answer a fixed set of items (or, conversely, the number of items completed in a fixed interval) is only appropriate if each item involves the same amount of labor—a condition that is approximated for the standardized tasks in experimental reaction-time research. But for test items, which may differ considerably in the amount of information processing and problem solving they involve, the only way to measure speed is with explicit time parameters that help us to disentangle the effects of the test taker's speed and the time consumingness of the items on the response time distributions. The models in (3)–(4) and (6)–(7) do contain such time parameters for the items. But they are absent in (1), (2), and (5).

The first two models above imply that a regular IRT model cannot be true unless it has a speed parameter for the test taker or uses the actual time spent on the item as a covariate. The third model has an analogous implication; it excludes the pure response-time models in (5) and (6)–(7) from being true because they have no response parameters. These implications do not need to have serious practical consequences. Models are never perfectly true, and a wrong model can approximate empirical data closely enough to yield valuable conclusions. But it is important to observe that, for the hierarchical framework below, no such direct implications exist. It has level-one models for the response and time distributions with separate sets of parameters. The only constraint on them is through second-level assumptions about the shape of their distributions in the population of test takers and the domain of items.

General Hierarchical Framework

The type of test modeled in this section is neither a pure speed nor a pure power test but the hybrid type of test typically administered in a computer-based testing program. Such tests do have items varying in difficulty; some of them will be difficult for a test taker and are likely to be answered incorrectly, whereas others typically result in correct answers. The items also differ enough in the amount of cognitive processing that they involve to yield different response-time distributions. Typically, the tests have a generous time limit; unless something special happens, the test takers are able to finish the test in time.

Key Assumptions

The first assumption is that of a test taker operating at a fixed level of speed. This assumption of stationarity excludes changes in behavior during the test due to learning, fatigue, strategy shifts, and the like. As already indicated, the assumption implies a fixed level of accuracy as well, which is a standard assumption underlying IRT modeling. This assumption of stationarity does not make the result less useful for test takers with small fluctuations in speed or even a minor trend; such violations can be detected by a residual analysis (van der Linden, Breithaupt, Chuah, & Zhang, 2007). Without a model based on the assumption of stationarity, possible changes and trends in the behavior of test takers are even likely to remain unnoticed.

Second, for a fixed test taker, both the response and the time on an item are assumed to be random variables. This assumption of randomness pervades test theory but, due to retention and/or learning, does not lend itself to direct experimental verification for test items with mental tasks. However, it is supported by empirical observations of variations in performance for persons repeating more physical tasks under identical conditions (e.g., Townsend & Ashby, 1983).

Third, we assume separate item and person parameters both for the distributions of the responses on the items and the time required to produce them. For a response model, it would be unusual to omit item parameters. As already indicated, we need item parameters in response-time models to account for the different amount of work (i.e., information processing and problem solving) they involve. We therefore follow the examples set in the models in (3)–(4) and (6)–(7), and adopt such parameters. An extremely practical consequence of this choice is that it allows us to compare the speed of test takers across tests with *different* items—a feature that is necessary, for instance, to control the level speededness of tests with an adaptive format (van der Linden et al., 2007; van der Linden, Scrams, & Schnipke, 1999).

The next assumption is that of conditional independence between the responses and the response times given the levels of ability and speed at which the test taker operates. This assumption may seem counterintuitive because both are nested within the same combination of test taker and test item. However, it follows from a heuristic argument analogous to that in IRT for the assumption of conditional independence between responses given θ (“ local independence”): For a fixed item, if a response model fits and the same holds for a response-time model, their person parameters capture all person effects on the response and response-time distributions. If these parameters are held constant, no potential sources of covariation are left and the response and the response time on an item become independent. It is important to distinguish the assumption from the traditional assumption of conditional independence between the responses of a fixed person across items. (This assumption is also made below; as well as an analogous assumption for the response times. But if the stationarity assumption is violated and a test taker changes his or her speed, e.g., when there is a threat of running out of time, these traditional assumptions will be violated but it is still possible for the responses and times on the individual items to remain independent.)

Finally, we model the relations between speed and accuracy for a population of test takers separately from the impact of these parameters on the responses and times of the individual test takers. The same will be done for the relations between the time and response parameters of the items. This approach allows us to capture such relations between the response and time parameters as the regression structure in (3) but at a higher level of modeling than the response-time model.

Levels of Modeling

The items are indexed by $i = 1, \dots, I$, and the test takers by $j = 1, \dots, J$. For test taker j, we have a response vector $\mathbf{U}_j = (U_{1j}, \dots, U_{Ij})$ and response-time vector $\mathbf{T}_j = (T_{1j}, \dots, T_{Ij})$ with realizations $\mathbf{u}_j = (u_{1j}, \dots, u_{Ij})$ and $\mathbf{t}_j = (t_{1j}, \dots, t_{Ij})$, respectively. On the first level, we specify both a response model and a response-time model for each combination of person and item. The relations between the parameters in these models are represented by two different second-level models. Together, these two levels constitute the empirical part of the framework. In the statistical treatment of the empirical model later in this paper, we add a third level with prior distributions for the second-level parameters or hyperparameters.

First-Level Models. We illustrate this level of modeling by choosing two specific models for the responses and times on the items; alternative choices are discussed below.

As an response model, the three-parameter normal-ogive (3PNO) model is adopted. That is, each response variable is assumed to be distributed as

$$U_{ij} \sim f(u_{ij}; \theta_j, a_i, b_i, c_i), \tag{8}$$

where $f(u_{ij}; \theta_j, a_i, b_i, c_i)$ denotes a Bernoulli probability function with success parameter

$$p_i(\theta_j) \equiv c_i + (1 - c_i)\Phi(a_i(\theta_j - b_i)), \tag{9}$$

where $\theta_j \in \Re$ is the ability parameter for the test taker j, $a_i \in \Re^+$, $b_i \in \Re$, and $c_i \in [0, 1]$ are the discrimination, difficulty, and guessing parameters for item i, respectively, and $\Phi(\cdot)$ denotes the normal distribution function.

For the response times, a lognormal model is chosen:

$$T_{ij} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i), \tag{10}$$

with

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\right\}, \tag{11}$$

where $\tau_j \in \Re$ is the speed parameter of test taker j and $\beta_i \in \Re$ and $\alpha_i \in \Re^+$ represent the time intensity and the discriminating power of item i, respectively. The lognormal family seems an appropriate choice because it has the positive support and a skew required for response-time distributions. The parametrization in (11) resembles that of the usual (unidimensional) models for dichotomous responses, such as in (9), except for a guessing parameter, which is not needed because time has a natural lower bound at $t = 0$. The model does not have the regression structure of the lognormal model in (3). In addition, because α_i is item dependent, it is more flexible than (3) in that it allows for differences between the variances of the logtimes on different items. The model showed excellent behavior in earlier studies; for reports on fit analyses as well as several other aspects of the model, see van der Linden (2006; in press), van der Linden and Guo (2006) and van der Linden et al. (2007b).

The vector with the parameters for person j is denoted as $\boldsymbol{\xi}_j = (\theta_j, \tau_j)$, the vector with the parameters for item i as $\boldsymbol{\psi}_i = (a_i, b_i, c_i, \alpha_i, \beta_i)$, and we use $\boldsymbol{\psi} = (\boldsymbol{\psi}_i)$. Because of the conditional independence of U_{ij} and T_{ij} given (θ, τ), the sampling distribution of $(\mathbf{U}_j, \mathbf{T}_j)$, $j = 1, \dots, J$, follows from (8) and (10) as

$$f(\mathbf{u}_j, \mathbf{t}_j; \boldsymbol{\xi}_j, \boldsymbol{\psi}) = \prod_{i=1}^{I} f(u_{ij}; \theta_j, a_i, b_i, c_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i). \tag{12}$$

Second-Level Models. One model describes the joint distribution of the person parameters in a population, $\mathcal{P}$, from which the test takers can be assumed to be sampled. We refer to this model as the *population model.*

The values of $\boldsymbol{\xi}_j$ are assumed to be randomly drawn from a multivariate normal distribution over $\mathcal{P}$; that is,

$$\boldsymbol{\xi}_j \sim f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{13}$$

where the density function is

$$f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) = \frac{\left|\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}\right|^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_{\mathcal{P}})^T \Sigma_{\mathcal{P}}^{-1}(\boldsymbol{\xi}_j - \boldsymbol{\mu}_{\mathcal{P}})\right] \tag{14}$$

with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau), \tag{15}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{16}$$

A second model captures the relations between the item parameters. It does so by specifying a joint distribution for the item parameters in the domain of items, $\mathcal{I}$, that the test represents. We refer to this model as the *item-domain model.* Analogous to (13)–(16), parameter vector $\boldsymbol{\psi}_i$ has a multivariate normal distribution

$$\boldsymbol{\psi}_i \sim f(\boldsymbol{\psi}_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) \tag{17}$$

with density function

$$f(\boldsymbol{\psi}_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) = \frac{\left|\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right|^{1/2}}{(2\pi)^{5/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\psi}_i - \boldsymbol{\mu}_{\mathcal{I}})^T \Sigma_{\mathcal{I}}^{-1}(\boldsymbol{\psi}_i - \boldsymbol{\mu}_{\mathcal{I}})\right], \tag{18}$$

mean vector

$$\boldsymbol{\mu}_{\mathcal{I}} = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta), \tag{19}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{I}} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{a\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}. \tag{20}$$

For the full model, the sampling distribution in (12) has to be extended to

$$f(\mathbf{u}, \mathbf{t}; \boldsymbol{\xi}, \boldsymbol{\psi}) = \prod_{j=1}^{J} \prod_{i=1}^{I} f(\mathbf{u}_j, \mathbf{t}_j; \boldsymbol{\xi}_j, \boldsymbol{\psi}_i) f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) f(\boldsymbol{\psi}_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}). \tag{21}$$

Identifiability

To establish identifiability, we suggest the constraints

$$\mu_\theta = 0, \quad \sigma_\theta^2 = 1, \quad \mu_\tau = 0. \tag{22}$$

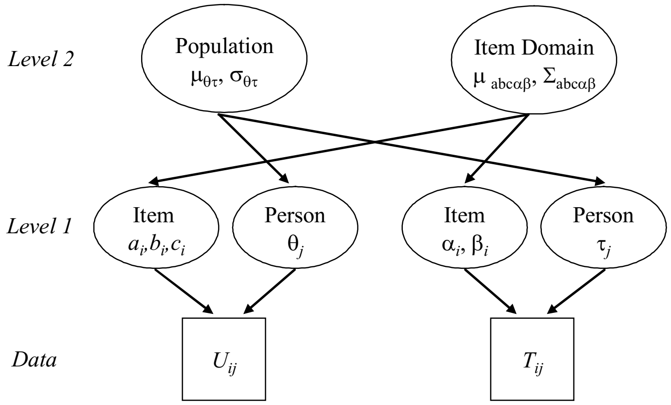


FIGURE 2.
A hierarchical framework for modeling speed and accuracy on test items.

The first two constraints are usual in IRT parameter estimation. The third constraint fixes the zero of τ and, hence, helps us to remove the tradeoff between β_i and τ_j from (11). Unlike θ, we do not need a constraint to fix the scale of τ or any other of the time parameters. Their scales are automatically fixed by the time unit in which $\ln t_{ij}$ is measured.

The last constraint allows us to equate μ_β to the average expected response time over the population and item domain and to interpret τ_j as a deviation from this average (van der Linden, 2006). Also, the full set of constraints allow us to keep all covariances between the item and person parameters, which typically are the second-level quantities of interest, as free parameters.

Alternative Models

A graphical representation of the hierarchical framework in the preceding section is given in Figure 2. The same framework can be specified with other plug-ins for the component models; the only condition is that the lower-level models have both person and item parameters.

As a response model, we can choose any current IRT model that would fit the items best. For instance, if the items are polytomous, a graded response model or (generalized) partial credit could be chosen. If the items appear to measure more than one ability dimension, a choice of a multidimensional response model becomes necessary. For a review of these and other options, see van der Linden and Hambleton (1997).

We do not necessarily expect the response format or dimensionality of the items to have an impact on their time distributions; the nature of the problems formulated is expected to be their main determinant. Besides, the two lower-level models fit independently; it is possible to replace the response model but keep the response-time model (or conversely).

Our choice of the lognormal model was mainly motivated by a "distribution-fitting approach." The lognormal density has the right support and skew for response-time distributions whereas the basic parameters in (11) give it enough flexibility to capture the main differences in time between persons and items. Nevertheless, if more becomes known about the processes underlying the responses, a different model may become attractive.

For example, if the items are simple and the problem solving has the features of a Poisson process, the exponential model in (6)–(7) could be fitted. This model already has the type of parametrization required by the framework. Rouder et al. (2003) explain why psychological processes with a sensory and problem-solving component may fit a Weibull distribution. To make their model appropriate for the framework, the following reparametrization might be helpful:

$$f(t_{ij}) = \pi\alpha_i^{\pi}(t_{ij} - (\beta_i - \tau_j))^{\pi-1}\exp\{-[\alpha_i(t_{ij} - (\beta_i - \tau_j))]^{\pi}\}, \quad t_{ij} > \beta_i - \tau_j, \tag{23}$$

with τ_j, α_i, and β_i parameters with the same interpretation as in (11) and π a general shape parameter. (If necessary, the shape parameter can be chosen to be item dependent.) Other choices with a more psychological motivation can be derived from the gamma models in Maris (1993).

The choice of a multivariate normal as a second-level model has the advantage of means and covariances as descriptive parameters of the population and item-domain distributions we are interested in. Besides, they give us closed-form expressions for the regression of the response and time parameters on one another—a feature that allows us to use response times as valuable collateral information in a response problem, and the other way around (van der Linden, 2007a; van der Linden, Klein Entink, & Fox, 2006; van der Linden & Guo, 2006). Because of these advantages, if their fit is unsatisfactory, rather than fitting members of different families of models, we recommend transforming some of their parameters. In fact, the transformations

$$a^* = \ln a, \tag{24}$$

$$c^* = \operatorname{logit} c, \tag{25}$$

can be used to improve the ranges of these and account for the skewness of typical empirical distributions of the guessing and discrimination parameters.

For the choice of some first-level models, the total number of parameters may involve a complexity too great to deal with by the second-level models. If so, a simple strategy is just to ignore some of the less interesting relations between these first-level parameters at the second level. For example, we may neither be interested in the variance of c nor in its covariances with any of the other item parameters. Removal of them would reduce the number of free hyperparameters in (19)–(20) by five. The framework should then be treated statistically by choosing a prior distribution directly for the first-level parameter for which the hyperparameters have been omitted. If a low informative prior is chosen, the impact of the removal of these hyperparameters on the remaining portion of the second-level model is negligible. We will illustrate the procedure when choosing a prior distribution for guessing parameter c_i in (30) below.

Dependence between Observed Scores and Times

The assumption of conditional independence between responses and times in this hierarchical framework does not imply anything for the relations between the scores and times on test items that can be observed in samples of test takers. As shown in Figure 2, covariation between observed scores and times can have two different origins: (1) the entries in the covariance matrix $\Sigma_{\mathcal{P}}$ of the person parameters; and (2) the entries in the covariance matrix $\Sigma_{\mathcal{I}}$ of the item parameters.

Depending on these entries, almost any pattern of correlation between observed responses and time can be produced. For example, if ability and speed correlate positively but all correlations between the item parameters are negligible, we will observe a positive correlation between observed scores and times in a sample of persons. But if some of the item parameters are negatively related, the correlation may vanish or even become negative if it is calculated between number-correct scores and total time on sets of items. The differences between conditional independence and possible patterns of dependence between observed scores and times illustrate what is more generally known in statistics as Simpson's paradox.

The dependence between observed scores and times becomes particularly unpredictable if different persons take different sets of items. An interesting example arose in an earlier study of differential speededness in computerized adaptive testing (van der Linden, Scrams, & Schnipke, 1999; see also the report in van der Linden, 2005, Sect. 9.5). In this study, there was no correlation between θ and τ but we nevertheless found a substantial positive correlation between the ability of the test takers and the actual amount of time spent they spent on the test. The reason was a positive correlation between the difficulty and time intensity of the items ($\rho_{b\beta} = .65$). Because an adaptive item-selection algorithm tends to give more difficult items to the more able students, a positive correlation between the observed times and ability levels arose. Thus, in order to predict the dependencies between test scores and times, in addition to the two covariance matrices, we also need to account for the sampling design for the persons and items. An advantage of the hierarchical framework above is that it does so automatically as long as the missing item administrations are missing at random (MAR).

Because of the unpredictability of observed correlations between test scores and times in samples of test takers, it may be misleading to take these correlations as descriptive and relate them to the features of the items or the scores of the test takers (Swanson, Featherman, Case, Luecht, & Nungester, 1999; Swanson, Case, Ripkey, Clauser, & Holtman, 2001). Instead, we should use the descriptive correlation between the test takers' abilities and speed provided by the framework.

Priors Distributions

To illustrate the treatment of less interesting first-level parameters that was discussed in the section on Alternative Models, we leave c_i out of the item-domain model in (17)–(19) and specify priors directly for these parameters.

As priors for the population and item-domain models, we choose (independent) normal/inverse-Wishart prior distributions; that is,

$$\boldsymbol{\Sigma}_{\mathcal{P}} \sim \text{Inverse-Wishart}\left(\boldsymbol{\Sigma}_{\mathcal{P}0}^{-1}, \nu_{\mathcal{P}0}\right), \tag{26}$$

$$\boldsymbol{\mu}_{\mathcal{P}} \mid \boldsymbol{\Sigma}_{\mathcal{P}} \sim \text{MVN}\,(\boldsymbol{\mu}_{\mathcal{P}0}, \boldsymbol{\Sigma}_{\mathcal{P}}/\kappa_{\mathcal{P}0}), \tag{27}$$

$$\boldsymbol{\Sigma}_{\mathcal{I}} \sim \text{Inverse-Wishart}\left(\boldsymbol{\Sigma}_{\mathcal{I}0}^{-1}, \nu_{\mathcal{I}0}\right), \tag{28}$$

$$\boldsymbol{\mu}_{\mathcal{I}} \mid \boldsymbol{\Sigma}_{\mathcal{I}} \sim \text{MVN}\,(\boldsymbol{\mu}_{\mathcal{I}0}, \boldsymbol{\Sigma}_{\mathcal{I}}/\kappa_{I0}), \tag{29}$$

where $\nu_{\mathcal{P}0} \geq 2$ is a scalar degrees-of-freedom parameter, $\boldsymbol{\Sigma}_{\mathcal{P}0}$ is a 2×2 (positive definite symmetric) scale matrix for the prior on $\boldsymbol{\Sigma}_{\mathcal{P}}$, and $\boldsymbol{\mu}_{\mathcal{P}0}$ and $\kappa_{\mathcal{P}0}$ are the vector with the means of the posterior distribution and the strength of prior information about these means, respectively. The parameters for the prior distributions of $\boldsymbol{\Sigma}_{\mathcal{I}}$ and $\boldsymbol{\mu}_{\mathcal{I}}$ are defined analogously.

We assume a common prior distribution for the guessing parameters in the first-level model in (9):

$$c_i \sim \text{beta}(\gamma, \delta), \quad i = 1, \ldots, I. \tag{30}$$

Because of this separate treatment, we will use the notation $\boldsymbol{\psi}_i = (a_i, b_i, \alpha_i, \beta_i)$, and $\mathbf{c} = (c_i)$.

For this choice of prior distributions, the joint posterior distribution of the parameters factors as

$$f(\boldsymbol{\xi}, \boldsymbol{\psi}, \mathbf{c}, \mu_{\mathcal{P}}, \mu_{\mathcal{I}}, \Sigma_{\mathcal{P}}, \Sigma_{\mathcal{I}} \mid \mathbf{u}, \mathbf{t}) \propto \prod_{j=1}^{J}\prod_{i=1}^{I} f(u_{ij}; \theta_j, a_i, b_i, c_i) f(t_{ij}; \tau_j, \alpha_i, \beta_i)$$
$$\times f(\boldsymbol{\xi}_j; \boldsymbol{\mu}_{\mathcal{P}}, \Sigma_{\mathcal{P}}) f(\boldsymbol{\psi}_i, c_i; \boldsymbol{\mu}_{\mathcal{I}}, \Sigma_{\mathcal{I}}) f(\boldsymbol{\mu}_{\mathcal{P}}, \Sigma_{\mathcal{P}}) f(\boldsymbol{\mu}_{\mathcal{I}}, \Sigma_{\mathcal{I}}) f(\mathbf{c}). \tag{31}$$

Parameter Estimation

Given the specifications in (8)–(20) and (26)–(30), Bayesian estimation of the model parameters with the Gibbs sampler is attractive. The sampler iterates through draws from the full conditional distributions of one block of parameters given all remaining parameters. The conditional distributions of the blocks of parameters can be derived from (31).

For the version of the 3PNO model in (9) without guessing parameter c_i and independent priors for the other parameters, Albert (1992) introduced Gibbs sampling with data augmentation. An extension of the full model 3PNO model was suggested in Johnson and Albert (1999, Sect. 6.9). The suggestion was further developed in Béguin and Glas (2001) and Fox and Glas (2001). Bayesian estimation with Gibbs sampling of the lognormal model in (11) was used in van der Linden (2006). Gibbs sampling for a version for the 3PNO model for item families with normal distributions of the ability parameters and multivariate normal distributions of the item parameters is given in Glas and van der Linden (2006).

The proposed implementation of the Gibbs sampler uses several elements from these references. For a summary of its steps, see the Appendix.

Empirical Example

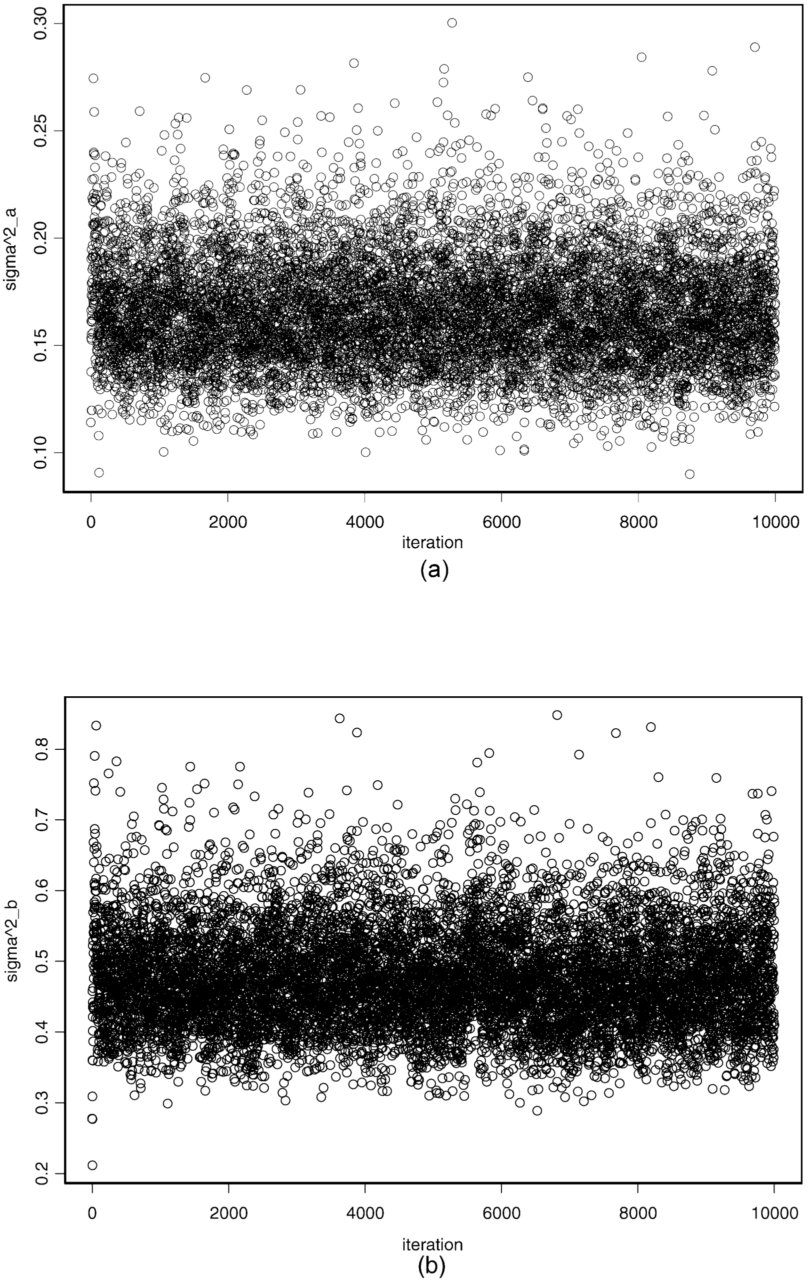
The hierarchical model was applied to a test from the computerized CPA Examination, which is administered by the American Institute of Certified Public Accountants (AICPA) as part of its certification program. The test had a multistage format with a first stage with one subtest of moderate difficulty and two subsequent stages with a subtest of moderate and high difficulty. We received a data set from the AICPA with the responses and response times for a sample of 1104 test takers on 96 items. The items were operational items that had been shown to have a good fit to the three-parameter logistic (3PL) model in (9). In an earlier study, we found an excellent fit of the same items to the lognormal model in (11) (van der Linden et al., 2007). Although we estimated all parameters in (8)–(20) simultaneously, this time our interest was in the estimation of the covariance matrices $\mathbf{\Sigma}_{\mathcal{P}}$ and $\mathbf{\Sigma}_{\mathcal{I}}$ for the person and item parameters. How to diagnose the fit of these second models will be the topic of a future study.

The parameters were estimated using the proposed Gibbs sampler. The prior distributions for the population and item-domain model were those in (26)–(29) for $\mathbf{\Sigma}_{\mathcal{P}0}^{-1}$ and $\mathbf{\Sigma}_{\mathcal{I}0}^{-1}$ matrices with diagonal elements equal to 1 and 10 and off-diagonal elements equal to 0 and 1, respectively. Also, $\nu_{\mathcal{P}0} = 2, \kappa_{\mathcal{P}0} = 1, \nu_{\mathcal{I}0} = 4$, and $\kappa_{\mathcal{I}0} = 1$. From (22), $\boldsymbol{\mu}_{\mathcal{P}0} = (0, 0)$. In addition, $\boldsymbol{\mu}_{\mathcal{I}0}$ was set equal to $(1, 0, 1, 0)$. As noted by Patz and Junker (1999), an MCMC method may have difficulty dealing with the weak identifiability of the 3PL model. One of the reasons is a tradeoff between the a_i and c_i parameters. We therefore fixed the c_i parameter at .20, which was a value close to their Bilog estimates. Finally, the proposal density for the Metropolis–Hastings (MH) step was a normal centered at the previous draw with variance equal to .05.

As demonstrated by the traceplots in Figure 3, the sampler converged almost immediately. The speed of convergence may seem high for a hierarchical IRT model but has been observed in numerous runs with other data sets. Our explanation is the presence of the response time as a manifest variable in the component in (9). Unlike the fitting of a hierarchical IRT model, which typically involves dealing with a delicate tradeoff between all latent parameters, the response-time variable with its fixed physical units brings stability to the entire framework.

The elements of the covariance matrices were estimated from the output of the last 10,000 iterations. The correlations between the parameters calculated from the estimates are given in Table 1. They show that, for this data set, the more able test takers tended to work faster ($\rho_{\theta\tau} = .30$)

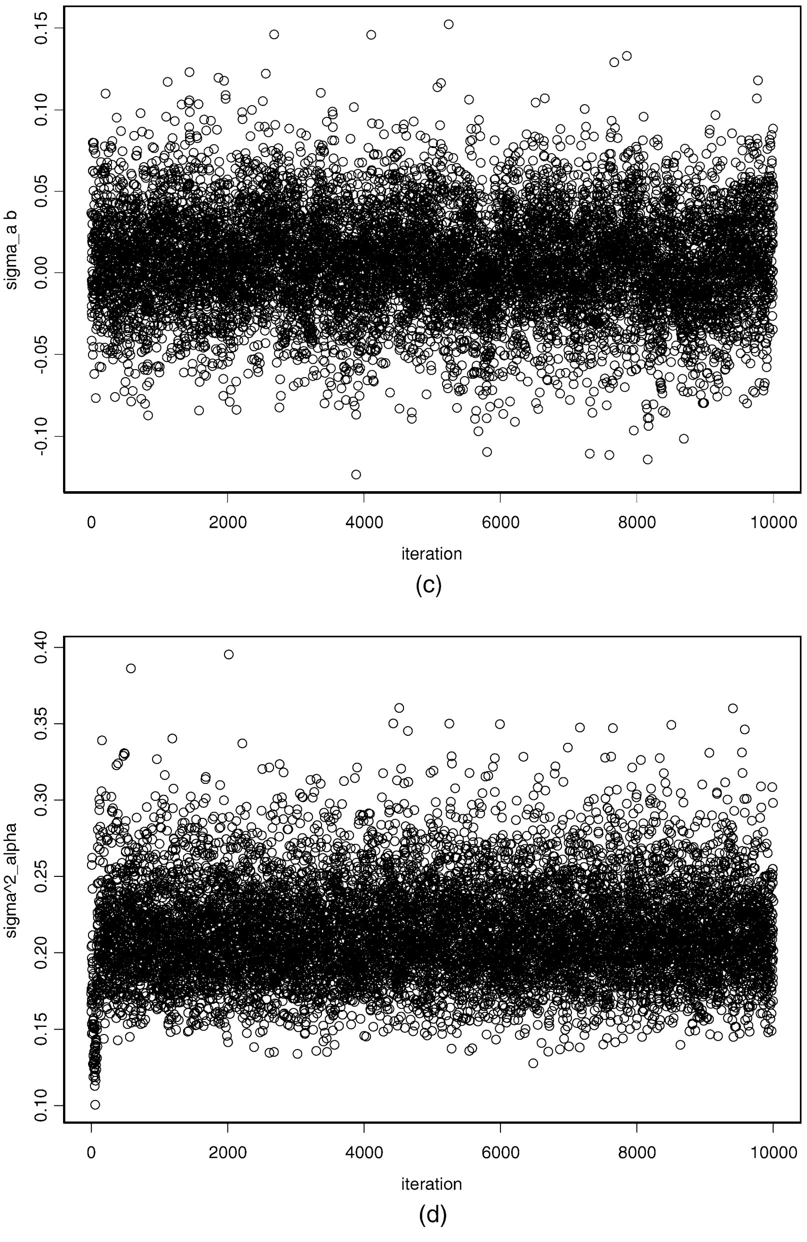
and the more difficult items tended to be more time intensive ($\rho_{b\beta} = .30$). Besides, there was some correlation between the difficulty and discrimination parameters in the response model ($\rho_{b\alpha} = .23$) and the time intensity and discrimination parameters in the response-time model ($\rho_{\alpha\beta} = .18$). All other correlations were negligible.



(a)

(b)

FIGURE 3.
Traceplots of the draws from the posterior distibutions of the covariance matrices of the item and person parameters.



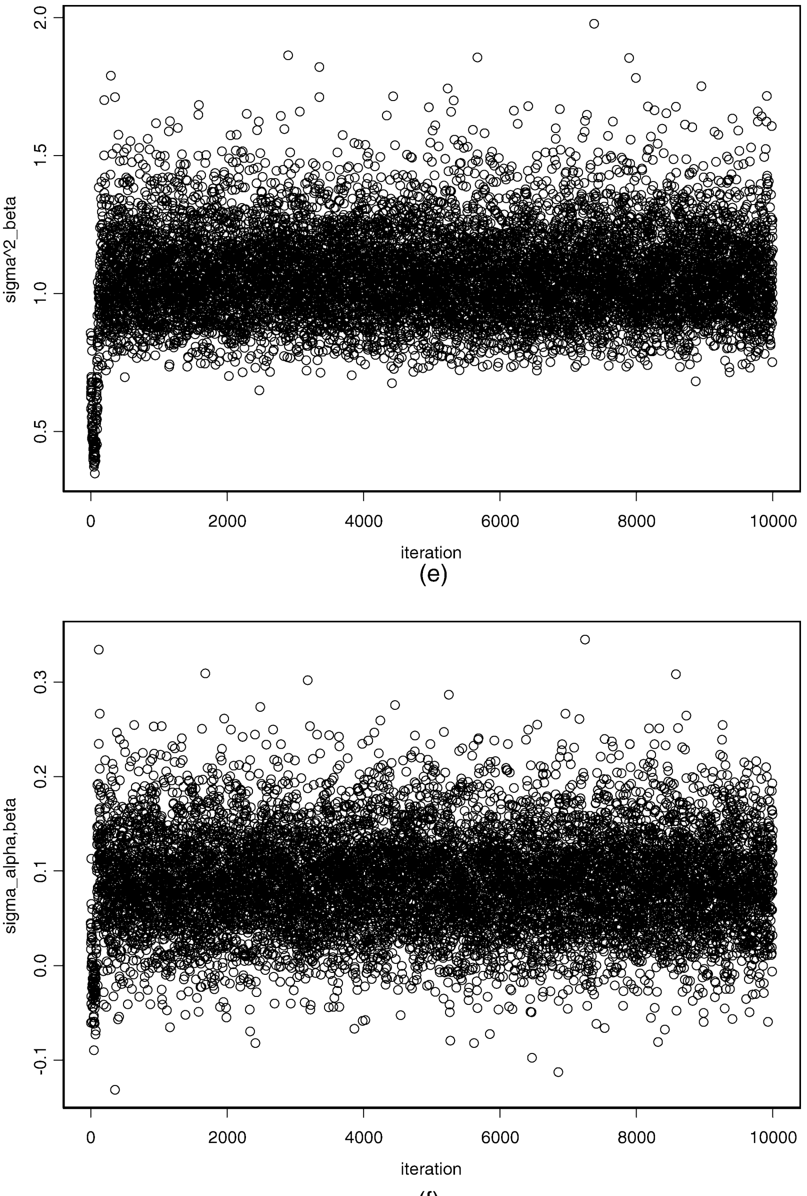
(c)

(d)

FIGURE 3.
Continued

Discussion

The presence of both a response model and a response-time model gives the hierarchical framework large applicability in educational and psychological testing. In particular, the second-level link between their item and person parameters allows us to borrow information from the response times to improve testing routines that are traditionally based on the responses only, and conversely. Several studies that exploit this principle have already been completed or are in progress.

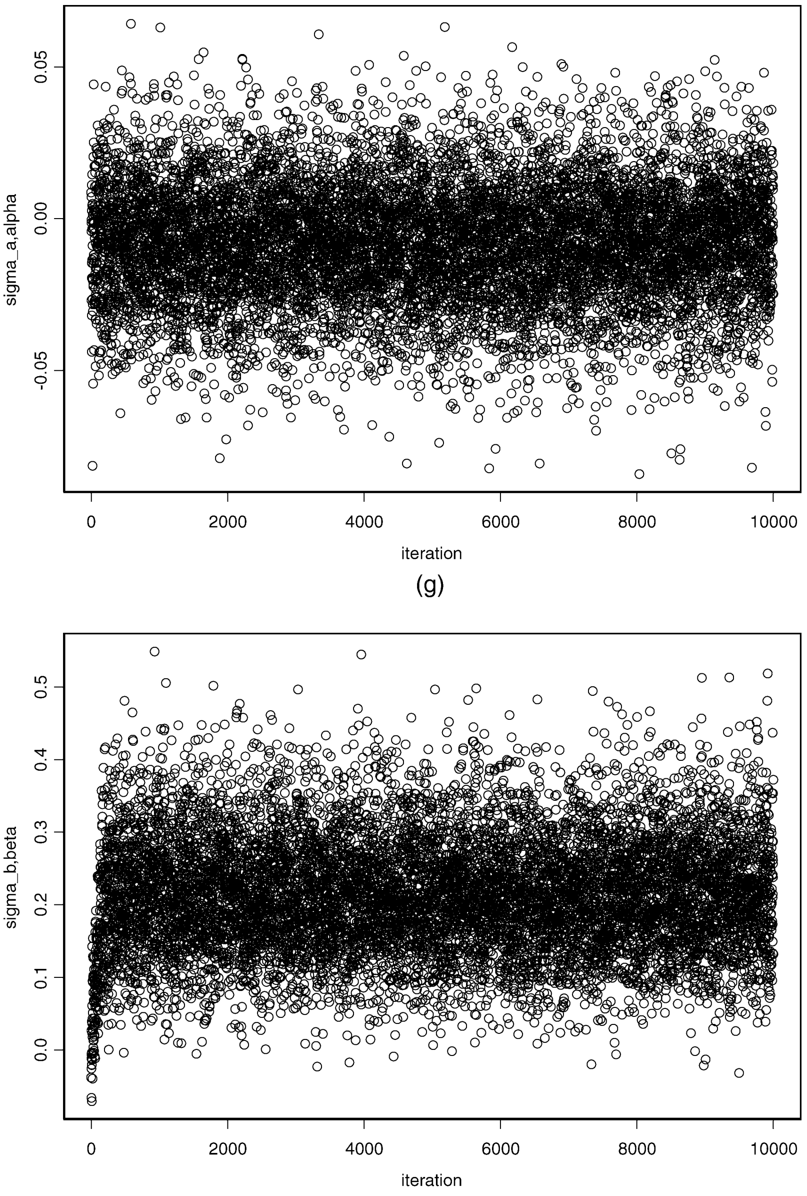


(e)

(f)

FIGURE 3.
Continued

For example, IRT parameter estimation is typically done separately from the estimation of any other item or person parameters. In one study, we showed that the simultaneous estimation of IRT parameters along with the other parameters in the framework in (8)–(20) may lead to a substantial increase in the accuracy of the estimated parameters (van der Linden et al., 2006). The best way to explain the increase is that, whereas traditional IRT parameter estimation is typically based on common priors for all item and person parameters, this estimation involves the simultaneous fitting of individual empirical priors for each parameter (= distribution of the parameter given the response times and the hyperparameters) while estimating them.



(g)

(h)

FIGURE 3.
Continued

The same principle can be used to improve the design of adaptive tests by selecting the items using the test taker's response times on the previous items in addition to the responses. The test then begins with a standard prior distribution for θ but after each next item we are able to retrofit the prior using a new response time. This retrofitting leads to a quick improvement of its location and spread, which in some cases may involve a reduction of the test length by some 50% (van der Linden, 2007b).

Response-time modeling is also necessary to deal with issues of speededness in testing. For example, it allows us to formulate constraints on item selection that guarantee multiple forms of

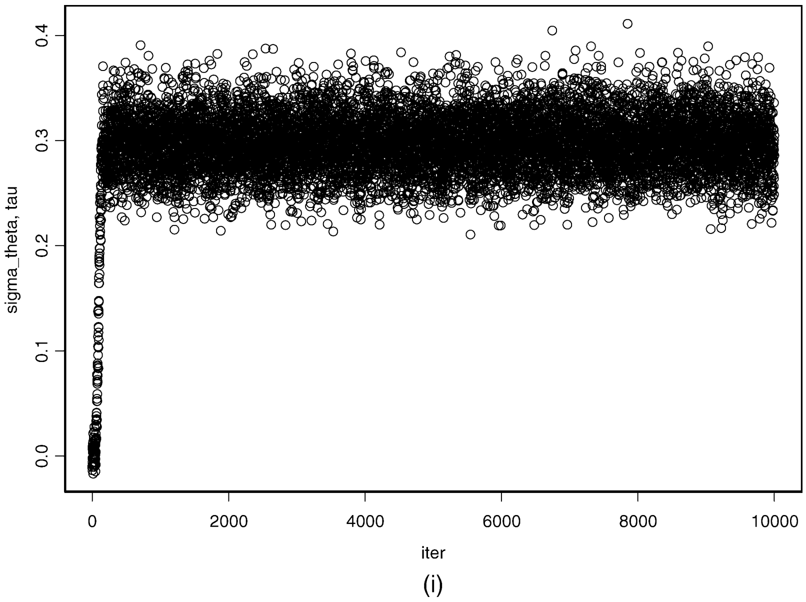


FIGURE 3.
Continued

TABLE 1.
Posterior means and 95% posterior highest posterior density intervals of correlations between model parameters.

Parameter	Posterior mean	95% HPD interval
$\rho_{\theta\tau}$	.30	(.24, .35)
ρ_{ab}	.03	(−.19, .23)
$\rho_{a\alpha}$	−.04	(−.24, .15)
$\rho_{a\beta}$	−.11	(−.32, .09)
$\rho_{b\alpha}$	.23	(−.03, .41)
$\rho_{b\beta}$	.30	(.01, .47)
$\rho_{\alpha\beta}$	.18	(−4.03, .37)

a test to be equally speeded. A special problem of speededness arises in adaptive testing, where different test takers get different items. Because items do vary in their time intensity (in empirical studies we have found them to differ easily by a factor of 5–8 across the item pool), these tests may suffer from differential speededness. Response-time modeling helps us to diagnose differential speededness (van der Linden et al., 2007) and to adjust an adaptive testing algorithm for it (van der Linden, 2005, Chap. 9; van der Linden et al., 1999).

As a final example, the detection of aberrant behavior on tests is mentioned. It is more advantageous to base the analysis of such behavior on response times than on the responses themselves, mainly because response times are continuous instead of binary, the procedure does not loose its power if the test and ability level of the test taker match (i.e., the probability of a correct response goes to a point close to .50), and it is hard for test takers to fake realistic response times on a typical test (van der Linden & Guo, 2006).

The hierarchical framework can also be used for analyzing reaction time data in psychological experiments. Unlike the traditional experimental paradigm, use of the model frees us from the necessity to use the same standardized tasks in one experiment. It also allows us to equate results from different experiments, even if they are obtained under different conditions of speed. In principle, when different conditions of speed exist in different experiments but all tasks have been calibrated, it is possible to replace the population model by a model for the within-person distribution of θ and τ and analyze the multiple data sets for a person from the experiments. The estimate of $\sigma_{\theta\tau}$ then allows us to study the tradeoff between speed and accuracy in these experiments.

Appendix: Gibbs Sampler

To enable the implementation of the Gibbs samples, the model in (9) is reformulated to have parameter structure $a_i\theta_j - b_i$. The data augmentation involved the definition of a latent variable Z_{ij} underlying the response of test taker j on item i, with

$$Z_{ij} \sim \phi(z_{ij}; a_i\theta_j - b_i) \tag{32}$$

and $\phi(\cdot)$ the standard normal density. In addition, we assume indicator variables W_{ij} defined as $W_{ij} = 1$ if j knows the answer to item i and $W_{ij} = 0$ if (s)he does not know the answer. It thus holds that

$$\begin{aligned} Z_{ij} &< 0 \quad \text{if} \quad W_{ij} = 0, \\ Z_{ij} &\geq 0 \quad \text{if} \quad W_{ij} = 1. \end{aligned} \tag{33}$$

Step 1. The values z_{ij}, $i = 1, \dots, I$, $j = 1, \dots, J$, are drawn from their posterior distributions given $\mathbf{w} = (w_{ij})$, $\boldsymbol{\theta} = (\theta_j)$, and $\boldsymbol{\psi}$. From (32)–(33),

$$z_{ij} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\psi} \sim \{\phi(z_{ij}; a_i\theta_j - b_i)\}/\{\Phi(a_i\theta_j - b_i)^{1-w_{ij}}[1 - \Phi(a_i\theta_j - b_i)]^{w_{ij}}\}, \tag{34}$$

which is a normal density truncated at the left at $z_{ij} = 0$ when $w_{ij} = 0$ and at the right when $w_{ij} = 1$.

Step 2. The values w_{ij}, $i = 1, \dots, I$, $j = 1, \dots, J$, are drawn from their posterior distributions given $\mathbf{u}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\mathbf{c}$.

Rewriting (8)–(9) as

$$p_i(U_{ij} = 1) \equiv \Phi(a_i\theta_j - b_i) + c_i[1 - \Phi(a_i\theta_j - b_i)] \tag{35}$$

shows that $\Pr\{W_{ij} = 1 \mid U_{ij} = 1\} \propto \Phi(a_i\theta_j - b_i)$ and $\Pr\{W_{ij} = 1 \mid U_{ij} = 0\} = 0$. Therefore, the conditional posterior distribution of w_{ij} has density

$$f(w_{ij}; \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}) = \begin{cases} 1 - w_{ij}, & \text{if } u_{ij} = 0, \\ K\Phi(a_i\theta_j - b_i)^{w_{ij}}[c_i(1 - \Phi(a_i\theta_j - b_i))]^{1-w_{ij}}, & \text{if } u_{ij} = 1, \end{cases} \tag{36}$$

with K a normalizing constant equal to the right-hand side of (35).

Step 3. The person parameters θ_j, $j = 1, \dots, J$, are drawn from their posterior distributions given $z = (z_{ij})$, $\boldsymbol{\tau} = (\tau_j)$, $\boldsymbol{\mu}_{\mathcal{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$.

From (32), $z_{ij} + \beta_i = a_i\theta_j + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim N(0, 1)$. Therefore, θ_j is a parameter in the regression of $z_{ij} + \beta_i$ on a_i with a normal error term. Because θ_j is normally distributed with mean $\mu_{\theta|\tau_j}$ and variance $\sigma^2_{\theta|\tau_j}$, the conditional posterior distribution of θ_j is also normal:

$$\theta_j \mid \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P \sim N\left(\frac{\sigma^{-2}_{\theta|\tau_j}\mu_{\theta|\tau_j} + \sum_{i=1}^{I} a_i(z_{ij}+b_i)}{\sigma^{-2}_{\theta|\tau_j} + \sum_{i=1}^{I} a_i^2}, \left(\sigma^{-2}_{\theta|\tau_j} + \sum_{i=1}^{I} a_i^2\right)^{-1}\right), \tag{37}$$

where the conditional means and variances $\mu_{\theta|\tau_j}$ and $\sigma^2_{\theta|\tau_j}$ follow directly from $\boldsymbol{\mu}_{\mathcal{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$ in (15)–(16) as

$$\mu_{\theta|\tau_j} = \mu_\theta + \left(\sigma_{\theta\tau}/\sigma^2_\tau\right)(\tau_j - \mu_\tau) \tag{38}$$

and

$$\sigma^2_{\theta|\tau_j} = \sigma^2_\theta - \sigma^2_{\theta\tau}/\sigma^2_\tau. \tag{39}$$

The two expressions simplify because of (22).

Step 4. The item parameters $(a_i,\ b_i)$, $i = 1, \ldots, I$, are drawn from their posterior distributions given $z_i = (z_{i1}, \ldots, z_{iJ})$, $\boldsymbol{\theta}$, $\boldsymbol{\alpha} = (\alpha_i)$, $\boldsymbol{\beta} = (\beta_i)$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$

(a_i, b_i) is a random parameter in the regression of $\mathbf{z}_j$ on $\mathbf{X} = (\boldsymbol{\theta}, -\mathbf{1})$, with $\mathbf{1}$ a unit vector of length J. Because (a_i, b_i) has a bivariate normal conditional distribution given (α_i, β_i) with a mean $\boldsymbol{\mu}_{a,b|\alpha_i,\beta_i}$ and covariance matrix $\boldsymbol{\Sigma}_{a,b|\alpha_i,\beta_i}$, its posterior distribution is also bivariate normal:

$$a_i, b_i \mid \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \sim N\left(\frac{\boldsymbol{\mu}_{a,b|\alpha_i,\beta_i}\boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T\mathbf{z}_i}{\boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T\mathbf{X}}, \left(\boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T\mathbf{X}\right)^{-1}\right), \tag{40}$$

where the mean $\boldsymbol{\mu}_{a,b|\alpha_i,\beta_i}$ and covariance matrix $\boldsymbol{\Sigma}_{a,b|\alpha_i,\beta_i}$ follow directly from $\boldsymbol{\mu}_{\mathcal{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$ in (19)–(20).

Step 5. The guessing parameters c_i, $i = 1, \ldots, I$, are drawn from their posterior distributions given $\mathbf{u}_i$ and $\mathbf{w}_i = (w_i)$.

The number of test takers guessing on item i is $n_i = J - \sum_{j=1}^{J} w_{ij}$, whereas the number of correct guesses is $x_i = \sum_{j=1}^{J}(u_{ij} \mid w_{ij} = 0)$. Since c_i is the probability of a correct guess, it follows that x_i is binomially distributed with parameters n_i and c_i. From (30),

$$c_i \mid \mathbf{u}_i, \mathbf{w}_i \sim \text{beta}(\gamma + x_i, \delta + n_i - x_i). \tag{41}$$

Step 6. The person parameters τ_j, $j = 1, \ldots, J$, are drawn from their posterior distributions given $\mathbf{t}_j$, $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_{\mathcal{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$.

The density in (11) implies a normal distribution of $\ln t_{ij}$ with mean $\beta_i - \tau_j$ and variance α_i^{-2}. Hence, $\beta_i - \ln t_{ij}$ is normally distributed with mean τ_j and variance α_i^{-2}. Because τ_j is normally distributed with mean $\mu_{\tau|\theta_j}$ and variance $\sigma^2_{\tau|\theta_j}$, the posterior distribution of τ_j is also normal:

$$\tau_j \mid \mathbf{t}_j, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}} \sim N\left(\frac{\sigma^{-2}_{\tau|\theta_j}\mu_{\tau|\theta_j} + \sum_{i=1}^{I}\alpha_i^2(\beta_i - \ln t_{ij})}{\sigma^{-2}_{\tau|\theta_j} + \sum_{i=1}^{I}\alpha_i^2}, \left(\sigma^{-2}_{\tau|\theta_j} + \sum_{i=1}^{I}\alpha_i^2\right)^{-1}\right). \tag{42}$$

The conditional means $\mu_{\tau|\theta_j}$ and variances $\sigma^2_{\tau|\theta_j}$ follow directly from $\boldsymbol{\mu}_{\mathcal{P}}$, and $\boldsymbol{\Sigma}_{\mathcal{P}}$ in (15)–(16).

Step 7. The item parameters β_i, $i = 1, \ldots, I$, are drawn from their posterior distributions given $\mathbf{t}_j$, $\boldsymbol{\tau}$, $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\alpha}$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$.

Analogous to the preceding step, $\ln t_{ij} + \tau_j$ is normally distributed with mean β_i and variance α_i^{-2}. Because β_i is normally distributed with mean $\mu_{\beta|a_i,b_i,c_i,\alpha_i}$ and variance $\sigma^2_{\beta|a_i,b_i,c_i,\alpha_i}$, the posterior distribution of β_i is also normal:

$$\beta_i \mid \mathbf{t}_j, \boldsymbol{\tau}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha},\ \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \sim N\left(\frac{\sigma^{-2}_{\beta|a_i,b_i,c_i,\alpha_i}\mu_{\beta|a_i,b_i,c_i,\alpha_i} + \alpha_i^2 \sum_{j=1}^{J}(\ln t_{ij} + \tau_j)}{\sigma^{-2}_{\beta|a_i,b_i,c_i,\alpha_i} + J\alpha_i^2}, \left(\sigma^{-2}_{\beta|a_i,b_i,c_i,\alpha_i} + J\alpha_i^2\right)^{-1}\right). \tag{43}$$

The conditional means $\mu_{\beta|a_i,b_i,c_i,\alpha_i}$ and variances $\sigma^2_{\beta|a_i,b_i,c_i,\alpha_i}$ follow directly from $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$ in (19)–(20).

Step 8. The item parameters α_i, $i = 1, \ldots, I$, are drawn from their posterior distributions given $\mathbf{t}_j$, $\boldsymbol{\tau}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$.

From (17)–(18),

$$f(\alpha_i \mid \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) = \phi\left(\alpha_i; \mu_{\alpha|a_i,b_i,c_i,\beta_i}, \sigma^2_{\alpha|a_i,b_i,c_i,\beta_i}\right). \tag{44}$$

Hence, for the posterior distributions of α_i given $\mathbf{t}_j$, $\boldsymbol{\tau}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$,

$$f(\alpha_i \mid t_{ij}, \tau_j, \beta_i) \propto \prod_{j=1}^{J} f(t_{ij}; \tau_j, \alpha_i, \beta_i)\phi\left(\alpha_i; \mu_{a|a_i,b_i,c_i,\beta_i}, \sigma^2_{a|a_i,b_i,c_i,\beta_i}\right), \tag{45}$$

where the first factor is given in (10). Since the density has no closed form, we suggest an MH step: At iteration t, a value α^*_{it} is sampled from a proposal density $\varphi(\alpha_{it}, \alpha_{i(t-1)})$, which is accepted with probability

$$\min\left\{1, \frac{f(\alpha^*_{it} \mid t_{ij}, \tau_j, \beta_i)}{f\left(\alpha_{i(t-1)} \mid t_{ij}, \tau_j, \beta_i\right)} \times \frac{\varphi\left(\alpha_{i(t-1)}, \alpha^*_{it},\right)}{\varphi\left(\alpha^*_{it}, \alpha_{i(t-1)}\right)}\right\}; \tag{46}$$

otherwise the value at the preceding iteration is retained, that is, $\alpha_{it} = \alpha_{i(t-1)}$. The ratio of the posterior densities in (46) simplifies to

$$\left(\frac{\alpha^*_{it}}{\alpha_{i(t-1)}}\right)^J \exp\left\{-\frac{1}{2}\left[\left(\alpha^{*2}_{it} - \alpha^2_{i(t-1)}\right)\sum_{j-1}^{J}\left(\ln t_{ij} - (\beta_i - \tau_j)\right)^2 + \frac{\left(\alpha^*_{it} - \mu_{\alpha|a_i,b_i,c_i,\beta_i}\right)^2 - \left(\alpha_{i(t-1)} - \mu_{\alpha|a_i,b_i,c_i,\beta_i}\right)^2}{\sigma^2_{\alpha|a_i,b_i,c_i,\beta_i}}\right]\right\}. \tag{47}$$

Step 9. Population parameters $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{P}}$ are sampled from their posterior distribution given $\boldsymbol{\xi}$. Since the normal/inverse-Wishart prior is conjugate with the multivariate normal population model, the posterior distribution is also normal/inverse-Wishart family:

$$\boldsymbol{\Sigma}_{\mathcal{P}} \mid \boldsymbol{\xi} \sim \text{Inverse-Wishart}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{P}*}, \nu_{\mathcal{P}*}\right), \tag{48}$$

$$\boldsymbol{\mu}_{\mathcal{P}} \mid \boldsymbol{\xi}, \Sigma_{\mathcal{P}} \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{P}*}, \boldsymbol{\Sigma}_{\mathcal{P}}/\kappa_{\mathcal{P}*}), \tag{49}$$

where

$$\boldsymbol{\Sigma}_{\mathcal{P}*} = \boldsymbol{\Sigma}_{\mathcal{P}0} + \boldsymbol{S}_{\xi} + \tfrac{\kappa_{\mathcal{P}0}I}{\kappa_{\mathcal{P}0}+I}(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_{\mathcal{P}0})(\overline{\boldsymbol{\xi}} - \boldsymbol{\mu}_{\mathcal{P}0})^T, \tag{50}$$

$$\nu_{\mathcal{P}*} = \nu_{\mathcal{P}0} + I, \tag{51}$$

$$\kappa_{\mathcal{P}*} = \kappa_{\mathcal{P}0} + I, \tag{52}$$

$$\boldsymbol{\mu}_{\mathcal{P}*} = \tfrac{\kappa_{\mathcal{P}0}}{\kappa_{\mathcal{P}0}+I}\boldsymbol{\mu}_{\mathcal{P}0} + \tfrac{I}{\kappa_{\mathcal{P}0}+I}\overline{\boldsymbol{\xi}}, \tag{53}$$

and $\mathbf{S}_{\xi}$ is defined as

$$\mathbf{S}_{\xi} = \sum_{i=1}^{I}(\boldsymbol{\xi}-\overline{\boldsymbol{\xi}})(\boldsymbol{\xi}-\overline{\boldsymbol{\xi}})^{T}. \tag{54}$$

Step 10. The sampling of the item-domain parameters $\boldsymbol{\mu}_{\mathcal{I}}$ and $\boldsymbol{\Sigma}_{\mathcal{I}}$ from their posterior distributions given $\boldsymbol{\psi}$ is similar to (47)–(53) with $\boldsymbol{\mu}_{\mathcal{P}}$, $\boldsymbol{\Sigma}_{\mathcal{P}}$, $\boldsymbol{\xi}$, and J replaced by $\boldsymbol{\mu}_{\mathcal{I}}$, $\boldsymbol{\Sigma}_{\mathcal{I}}$, $\boldsymbol{\psi}$, and I.

Comment

In spite of the complexity of the framework, for the current choice of component models, Gibbs sampling is straightforward due to conjugacy between the model and prior distributions at each stage in the hierarchy. The only exception is Step 8 for the discrimination parameter in the response-time model. When carefully implemented, the proposed MH step for this parameter need not involve a substantial loss of efficiency of the sampler. An obvious strategy is to repeat Step 8 within each cycle until a new draw is accepted. For this and other issues related to the use of an MH step in a Gibbs sampler, see, for instance, Carlin and Louis (2000, Sect. 5.4).

References

Albert, J.H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17*, 261–269.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541–562.

Carlin, B.P., & Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman & Hall.

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

Dubey, S.D. (1969). A new derivation of the logistic distribution. *Naval Research Logistics Quarterly, 16*, 37–40.

Fox, J.P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271–288.

Glas, C.A.W., & van der Linden, W.J. (2006). Modeling variability in item parameters in item response models. *Psychometrika*. Submitted.

Jansen, M.G.H. (1986). A Bayesian version of Rasch's multiplicative Poisson model for the number of errors on achievement tests. *Journal of Educational Statistics, 11*, 51–65.

Jansen, M.G.H. (1997a). Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Statistics, 22*, 125–140.

Jansen, M.G.H. (1997b). Rasch's model for reading speed with manifest exploratory variables. *Psychometrika, 62*, 393–409.

Jansen, M.G.H., & Duijn, M.A.J. (1992). Extensions of Rasch's multiplicative Poisson model. *Psychometrika, 57*, 405–414.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal Data Modeling*. New York: Springer-Verlag.

Luce, R.D. (1986). *Response times: Their Roles in Inferring Elementary Mental Organization*. Oxford, UK: 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.

Oosterloo, S.J. (1975). *Modellen voor Reaktie-tijden* [*Models for Reaction Times*]. Unpublished master's thesis, Faculty of Psychology, University of Groningen, The Netherlands.

Patz, R.J., & Junker, B.W. (1999). Applications and extensions of MCMC in IRT: Mulitple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 34*, 342–366.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. (Original published in 1960)

Roskam, E.E. (1987). Toward a psychometric theory of intelligence. In E.E. Roskam & R. Suck (Eds.), *Progress in Mathematical Psychology* (pp. 151–171). Amsterdam: North-Holland.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 187–208). New York: Springer.

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

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

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

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

Schnipke, D.L., & Scrams, D.J. (1999). *Representing response time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.

Schnipke, D.L., & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C.N. Mills, M. Potenza, J.J. Fremer & W. Ward (Eds.), *Computer-Based Testing: Building the Foundation for Future Assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Swanson, D.B., Featherman, C.M., Case, S.M., Luecht, R.M., & Nungester, R. (1999, March). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Swanson, D.B., Case, S.E., Ripkey, D.R., Clauser, B.E., & Holtman, M.C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE, Step 1. *Academic Medicine, 76*, 114–116.

Tatsuoka, K.K., & Tatsuoka, M.M. (1980). A model for incorporating response-time data in scoring achievement tests. In D.J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236–256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

Townsend, J.T., & Ashby, F.G. (1983). *Stochastic Modeling of Elementary Psychological Processes*. Cambridge, UK: Cambridge University Press.

van Breukelen, G.J.P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359–376.

van der Linden, W.J. (2005). *Linear Models for Optimal Test Design*. New York: Springer-Verlag.

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. (2007a). *Conceptual Issues in Response-Time Modeling*. Submitted.

van der Linden, W.J. (2007b). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics, 32*.

van der Linden, W.J., & Guo, F. (2006). *Two Bayesian Procedures for Identifying Aberrant Response-Time Patterns in Adaptive Testing*. Manuscript submitted for publication.

van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

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

van der Linden, W.J., Klein Entink, R.H., & Fox, J.-P. (2006). *IRT Parameter Estimation with Response Times as Collateral Information*. Manuscript submitted for publication.

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

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G. (1997). A logistic model for time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 169–185). New York: Springer-Verlag.

Manuscript received 29 MAR 2006
Final version received 3 NOV 2006
Published Online Date: 8 AUG 2007