LSAC RESEARCH REPORT SERIES

■ A Hierarchical Framework for Modeling Speed and Accuracy on Test Items

Wim J. van der Linden
University of Twente, Enschede, The Netherlands

■ Law School Admission Council
Research Report 05-02
October 2005

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

In the analysis of data for the Law School Admission Test (LSAT) and other similar standardized tests, a mathematical model called item response theory (IRT) is commonly used to estimate both the characteristics of the test questions (items) and the ability level of the test takers. Such analyses are based on the test takers' correct and incorrect responses to the test items. When a test is administered in a computerized mode, the capability of recording the amount of time a test taker has spent on each item provides us with additional information about the test-taking experience of individuals as well as the characteristics of items.

The practical goal of this paper is to use response times on test items as an additional source of information in estimating the abilities of the test takers when the test is delivered in a computerized mode. It is only possible to use this additional source when we have a mathematical model that (i) relates the speed at which a test taker works to his/her ability (accuracy) on the test and (ii) separates the test taker's speed on the test from the time intensities of the items.

A hierarchical framework of modeling is introduced that has two different levels, one for the individual test taker and one for the population of test takers. Each level includes two components: one to model speed and the other to model accuracy. At the level of the individual test taker, the framework models the test taker's responses to items (correct or incorrect) and the time he or she spent on each item. Both component models have separate parameters for the item and person effects. At the second level, the framework has a model for the population of test takers that explains how the speed and accuracy of the test takers tend to be related. In addition, it has an item-domain model that relates the time intensities of the items to such features as their difficulties.

A method is applied to estimate all unknown parameters from the responses and times on the test items. A description of how the second level model can be used to predict a test taker's accuracy from his speed or the difficulties of the items from their time intensities is also included.

Abstract

Current modeling of response times on test items has been influenced by the experimental paradigm of 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. Other response-time models seem to be unclear as to the level of parameterization they represent. A hierarchical framework of modeling is proposed to better represent the nature of speed and accuracy on test items as well as the different levels of dependency between them. The framework allows a "plug-and-play approach" with alternative choices of response and response-time models to deal with different types of test items as well as population and item-domain models to represent key relations between 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 the population and item domain with Gibbs sampling from the joint posterior distribution.

Introduction

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, adaptive item selection, 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. Some of these models appear to be influenced by the experimental paradigm of 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 tasks with their reaction times, and (3) experimental manipulation of the conditions under which the subjects operate. The paradigm is used, for instance, to test hypotheses about underlying psychological processes or to decompose reaction times into the times needed for certain separate 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 is an independent variable but accuracy a dependent variable; a subject has control of his or her speed over a range of possible levels but has to accept the accuracy that is the result of a choice of speed.

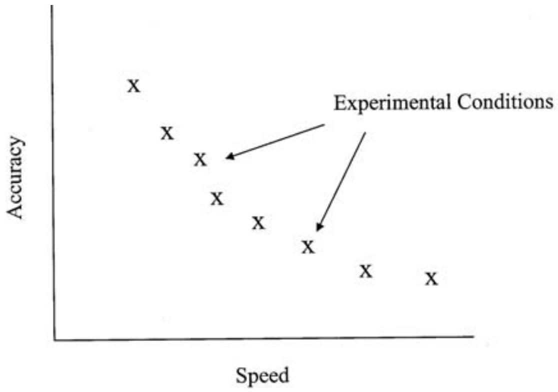

FIGURE 1. *Example of a speed-accuracy tradeoff*

Generally, a speed-accuracy tradeoff can be described as a negative (nonlinear) correlation between the speed and accuracy levels at which a person can 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 is 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, the framework consists of an item response theory (IRT) model and a model for the response-time distribution along with a higher-level structure to account for the dependences between their parameters. 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 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 an individual 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, Verstraalen, and Jansen (1997) present a model 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 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 the term $\theta_j - \ln \tau_j$ 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 as opposed to the ability parameter θ_j in a regular IRT model.

The authors highlight the fact that the model 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. Also, note that the model is only for the response distribution on the item. Since it nevertheless does have a time parameter estimated from the response times, it implies that, in spite of their history of success in operational testing, IRT models that ignore response times can never fit response distributions.

A similar model is derived in 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 speed-accuracy tradeoff directly between the ability of the test taker and the actual time spent on a test item; less time on an item indicates a higher speed and results in lower accuracy. The model also implies that none of the existing response models in IRT can fit a response distribution. In addition, 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 the lognormal family.

This model does not seem to be motivated by a speed-accuracy tradeoff. Instead, its main motivation seems to be the observation that more able persons tend to work faster; a higher ability θ_j implies a lower expected logtime on the item. As indicated in the introduction, this observation is not necessarily inconsistent with a speed-accuracy tradeoff; the observation of a positive correlation between speed and accuracy holds at the level of a population of persons, whereas a speed-accuracy tradeoff is a within-person relation. Also, note that the item-difficulty parameter b_i plays a role analogous to θ_j in (3); the expected logtime on a difficult item is higher than on an easy item.

Analogous to (1) and (2), the model in (3)–(4) implies that a response-time distribution on an item can never be modeled adequately if any of the features of the item or person usually parameterized in IRT are ignored.

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 seems to be inspired by its success 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

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

where ψ_j is a shift, σ_j a scale, and π_j a shape parameter. This choice of parameterization is motivated by the nature of the psychological processes typically studied in reaction-time experiments. 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 the 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. presented their model, but does certainly not hold for the case of response times on test items addressed in Tatsuoka and Tatsuoka.

A response-time model that does account for differences between test items is one 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 the person and item parameters. 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 & Duin, 1992); the additive and multiplicative gamma models by Maris (1993); 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. But, 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 are 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.

Also, the idea to incorporate a speed-accuracy tradeoff in a model can be viewed as a confounding of the level of modeling. Three different levels should be distinguished: (1) the within-person level, where each observation is nested in a series of parameter values changing over time; (2) the fixed-person level, where all parameters are constant; and (3) the level of a population of fixed persons, for which we have a distribution of parameter values. The first two models above seem to confound the within-person and fixed-person levels. The third model is for a fixed person but seems to be built on an observation that can only hold at a population level (positive correlation between speed and accuracy). 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, (average) speed is a measure of an amount of ground covered in a time interval. Therefore, measuring speed as the count of items completed in a time interval or the simple sum of response times is only appropriate if each item has the same time distribution, 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 disentangle a person's speed from the effects 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).

As already indicated, the first two models above imply that models can never fit a response distribution on an item unless they have a time parameter for the test taker. This implication is not consistent with the history of success of the use of regular response models in operational testing. The third model has an analogous implication for the time distribution. On the other hand, (5) and (6)-(7) are pure response-time models. In the hierarchical framework below, the models for the response and time distributions are fitted independently. They have no common parameters or other constraints that would allow us to predict the fit of one model from the fit of the other.

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 information processing and problem-solving 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 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 appropriate for test takers with small fluctuations in speed or even a minor trend; such violations can be detected by a residual analysis. Without a model based on the assumption of stationarity, possible changes and trends in the behavior of test takers might 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 used to produce them. For a response model, it would be unusual to omit item parameters. Analogously, we follow the examples set in the models in (3)–(4) and (6)–(7) and introduce item parameters in the response-time model. 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 condition that is frequently met in computer-based testing.

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 daring because both are produced by the same test taker on the same test items. However, the assumption follows from a heuristic argument analogous to that in IRT for the assumption of conditional independence between responses given θ (”local independence”): 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.

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 hierarchically higher level of modeling.

If this paper makes any contribution, it might be the replacement of the attempt to incorporate a speed-accuracy in response-time modeling with the assumption of conditional independence between response and response time given speed and accuracy, as mentioned above. As already indicated, this assumption is counterintuitive because both the response and the time are nested within the same person. It is important to distinguish the assumption from the traditional assumption of conditional independence between the responses given accuracy across items. (This assumption, as well as an analogous assumption for the response times, can also be found below.) If the stationarity assumption is violated and a test taker changes his or her speed because of lack of time, these traditional assumptions will be violated, but it is perfectly possible for the responses and times on the individual items to remain conditionally independent.

Empirical Levels of Modeling

The items are indexed by $i = 1,..., I$, and the test takers by $j = 1,..., J$. For test taker j, we have a response vector $\mathbf{U}_j = (U_{1j},..., U_{Ij})$ and response-time vector $\mathbf{T}_j = (T_{1j},..., T_{Ij})$ with realizations $\mathbf{u}_j = (u_{1j},..., u_{Ij})$ and $\mathbf{t}_j = (t_{1j},..., t_{Ij})$, 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 a response model, the 3-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

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

$\theta_j \in \Re$ is the ability parameter for 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 $\Phi(.)$ 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

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

where $\tau_j \in \Re$ is a speed parameter for 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 parameterization 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 an earlier study; for a report on the fit analysis as well as several other aspects of the model, see van der Linden (2006).

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 is denoted 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 $(\mathbf{U}_j; \mathbf{T}_j)$, $j = 1,..., J$ follows from (8) and (10) as

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

Second-Level Models

One model describes the joint distribution of the person parameters in a population P from which the test takers can be assumed to be sampled. We refer to this model as the *population model*.

Let $\xi_j = (\theta_j, \tau_j)$ be the vector with the person parameters in (8) and (10). We assume that ξ_j is randomly drawn from a multivariate normal distribution over P; that is,

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

where the density function is

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

with mean vector

$$\boldsymbol{\mu}_P = (\mu_\theta, \mu_\tau), \tag{15}$$

and covariance matrix

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

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 is intended to represent. We refer to this model as the *item-domain model*. Analogous to (13)–(16), for parameter vector ψ_i, we assume 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

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

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 completed as

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

Identifiability

To establish identifiability, we suggest the constraints

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

The first two constraints are usual in IRT parameter estimation. 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, these constraints allow us to keep all covariances between the item and person parameters, which typically are the 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.

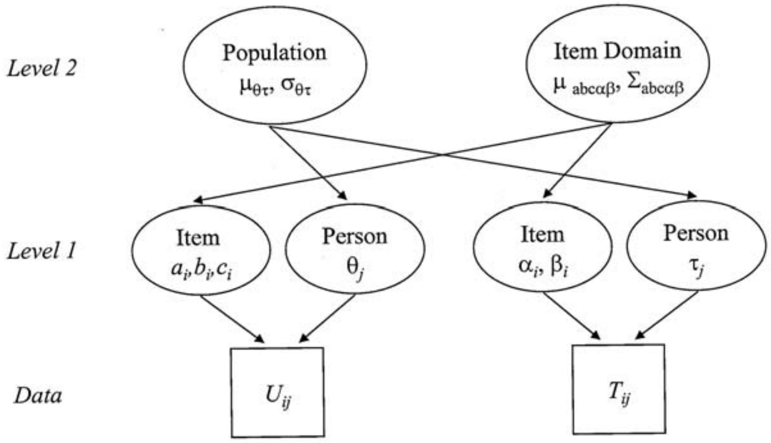

FIGURE 2. *A hierarchical framework for modeling speed and accuracy on test items*

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 the time distribution on the items; the nature of the problems formulated in them seems more important. Since the lower-level models fit independently, we need to focus on the possible impact of the format or dimensionality of the items on the time distributions only when fitting the response-time model.

The choice of the lognormal model in (8)–(9) was mainly motivated by a curve-fitting approach to response-time modeling. It has the right support and seems to have a good skew for response-time distributions whereas its basic parameters give it enough flexibility to capture the differences in time between persons and items. Nevertheless, if more is known about the processes underlying the responses, a different model may become possible.

For example, if the items are simple and the problem-solving process has the features of a Poisson process, the exponential model in (6)–(7) could be fitted. This model already has the type of parameterization 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 test items, the following reparameterization might be helpful:

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

with τ_j, α_i, and β_i parameters having the same interpretation as in (11) and π being 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. Also, they give us closed-form expressions for the regression of some of the parameters on others. Because of these advantages, if their fit is unsatisfactory, rather than fitting different models we recommend transforming some of their parameters. In fact, the transformations

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

$$c^* = \text{logit } c \tag{25}$$

$$\alpha^* = \ln \alpha \tag{26}$$

can be used to improve the range 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 to ignore some of the less interesting first-level parameters. For example, removal of one item parameter reduces the number of free hyperparameters in (19)–(20) by 5. Statistically, the framework should then be treated by choosing prior distributions directly for the omitted first-level parameter. If a low informative prior is chosen, the impact of the removal of a first-level parameter on the remaining portion of the second-level model is negligible. We will illustrate the procedure for choosing priors for guessing parameter c_i.

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 observed scores and times on items or tests that can be observed in samples of test takers. As shown in Figure 2, observed scores and times have two different sources of covariation: (1) the entries in the covariance matrix Σ_{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 scores 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. The change from conditional independence to different patterns of dependence illustrate what is more generally known in statistics as Simpson's paradox.

The dependency 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 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.

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).

Parameter Estimation

For the specifications in (8)–(20), a full Bayesian treatment of the model with the Gibbs sampler is attractive.

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 to the full 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 of 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 (tentatively accepted). The treatment below uses several elements from these references.

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

Prior Distributions

For the population and item-domain models, we choose (independent) normal/Inverse-Wishart prior distributions; that is,

$$\boldsymbol{\Sigma}_{\mathrm{P}} \sim \text{Inverse-Wishart}(\boldsymbol{\Sigma}_{\mathrm{P}0}^{-1}, \nu_{\mathrm{P}0}), \tag{27}$$

$$\boldsymbol{\mu}_{\mathrm{P}} \mid \boldsymbol{\Sigma}_{\mathrm{P}} \sim \text{MVN}(\boldsymbol{\mu}_{\mathrm{P}0}, \boldsymbol{\Sigma}_{\mathrm{P}}/\kappa_{\mathrm{P}0}), \tag{28}$$

$$\boldsymbol{\Sigma}_{\mathcal{I}} \sim \text{Inverse-Wishart}(\boldsymbol{\Sigma}_{\mathcal{I}0}^{-1}, \nu_{\mathcal{I}0}), \tag{29}$$

$$\boldsymbol{\mu}_{\mathcal{I}} \mid \boldsymbol{\Sigma}_{\mathcal{I}} \sim \text{MVN}(\boldsymbol{\mu}_{\mathcal{I}0}, \boldsymbol{\Sigma}_{\mathcal{I}}/\kappa_{\mathcal{I}0}), \tag{30}$$

where $\nu_{\mathrm{P}0} \geq 2$ is a scalar degrees-of-freedom parameter, $\boldsymbol{\Sigma}_{\mathrm{P}0}$ is a 2 × 2 (positive definite symmetric) scale matrix for the prior on $\boldsymbol{\Sigma}_{\mathrm{P}}$, and $\boldsymbol{\mu}_{\mathrm{P}0}$ and $\kappa_{\mathrm{P}0}$ 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):

$$c_i \sim \mathrm{beta}(\gamma, \delta),\ i = 1,\ldots, I. \tag{31}$$

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

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

$$\times f(\xi_j; \boldsymbol{\mu}_{\mathrm{P}}, \boldsymbol{\Sigma}_{\mathrm{P}}) f(\boldsymbol{\psi}_i, c_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) f(\boldsymbol{\mu}_{\mathrm{P}}, \Sigma_{\mathrm{P}}) f(\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) f(\mathbf{c}). \tag{32}$$

The Gibbs sampler iterates through draws from the full conditional distributions of one block of parameters given all remaining parameters. In the next section, we specify these draws.

Gibbs Sampler

For convenience, the model in (9) is reformulated to have parameter structure $a_i\theta_j - b_i$. Let Z_{ij} be a latent variable underlying the response of test taker j on item i, with

$$Z_{ij} \sim \phi(z_{ij}; a_i\theta_j - b_i) \tag{33}$$

and $\phi(.)$ being 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 j does not know the answer. It is postulated that

$$\Pr\{Z_{ij} < 0\} \text{ if } \Pr\{W_{ij} = 0\};$$

$$\Pr\{Z_{ij} \geq 0\} \text{ if } \Pr\{W_{ij} = 1\}. \tag{34}$$

Step 1

The values z_{ij}, $i = 1,\ldots, I$, $j = 1,\ldots, J$ are drawn from their posterior distributions given $\mathbf{w} = (w_{ij})$, $\boldsymbol{\theta} = (\theta_j)$, and $\boldsymbol{\psi}$. From (33)–(34),

$$z_{ij} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\psi} \sim \{\phi(z_{ij}; a_i\theta_j - b_i)\} / \{\Phi(a_i\theta_j - b_i)^{1-w_{ij}} [1 - \Phi(a_i\theta_j - b_i)]^{w_{ij}}\}, \tag{35}$$

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}, $i = 1,\ldots, I$, $j = 1,\ldots, J$ are drawn from their posterior distributions given $\mathbf{u}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\mathbf{c}$.

Rewriting (8)–(9) as

$$p_i(U_{ij} = 1) \equiv \Phi(a_i\theta_j - b_i) + c_i[1 - \Phi(a_i\theta_j - b_i)] \tag{36}$$

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

$$f(w_{ij}; \mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{c}) = \begin{cases} 1 - w_{ij}, & \text{if } u_{ij} = 0, \\ K\Phi(a_i\theta_j - b_i)^{w_{ij}} [c_i(1 - \Phi(a_i\theta_j - b_i))]^{1-w_{ij}}, & \text{if } u_{ij} = 1, \end{cases} \tag{37}$$

with K as a normalizing constant equal to the right-hand side of (36).

Step 3

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

From (33), $z_{ij} + \beta_i = a_i \theta_j + \varepsilon_{ij}$ 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:

$$\theta_j \mid \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\mu}_{\mathrm{P}}, \boldsymbol{\Sigma}_{\mathrm{P}} \sim N\left(\frac{\sigma^{-2}_{\theta|\tau_j}\mu_{\theta|\tau_j} + \sum_{i=1}^{I} a_i (z_{ij} + b_i)}{\sigma^{-2}_{\theta|\tau_j} + \sum_{i=1}^{I} a_i}, \left(\sigma^{-2}_{\theta|\tau_j} + \sum_{i=1}^{I} a_i \right)^{-1} \right), \tag{38}$$

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

$$\mu_{\theta|\tau_j} = \mu_\theta + (\sigma_{\theta\tau}/\sigma^2_\tau)(\tau_j - \mu_\tau) \tag{39}$$

and

$$\sigma^2_{\theta|\tau_j} = \sigma^2_\tau - \sigma^2_{\theta\tau}/\sigma^2_\theta. \tag{40}$$

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 $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})$, $\boldsymbol{\theta}, \boldsymbol{\alpha} = (\alpha_i)$, $\boldsymbol{\beta} = (\beta_i)$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$.

(a_i, b_i) is a random parameter in the regression of $\mathbf{z}_j$ on $\mathbf{X} = (\boldsymbol{\theta}, -1)$, with 1 being 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:

$$a_i, b_i \mid \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \sim N\left(\frac{\boldsymbol{\mu}_{a,b|\alpha_i,\beta_i} \boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T \mathbf{z}_i}{(\boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T \mathbf{X})^{-1}}, \left(\boldsymbol{\Sigma}^{-1}_{a,b|\alpha_i,\beta_i} + \mathbf{X}^T \mathbf{X} \right)^{-1} \right), \tag{41}$$

where $\boldsymbol{\mu}_{a,b|\alpha_i,\beta_i}$ and covariance matrix $\boldsymbol{\Sigma}_{a,b|\alpha_i,\beta_i}$ follow directly from $\boldsymbol{\mu}_{\mathrm{P}}$ and $\boldsymbol{\Sigma}_{\mathrm{P}}$ 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 $x_i = \sum_{j=1}^{J} (u_{ij} | w_{ij} = 0)$. 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 (31),

$$c_i \mid \mathbf{u}_i, \mathbf{w}_i \sim \mathrm{beta}(\gamma + x_i, \delta + n_i - x_i). \tag{42}$$

Step 6

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

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:

$$\boldsymbol{\tau}_j \mid \mathbf{t}_j, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}_{\mathrm{P}}, \boldsymbol{\Sigma}_{\mathrm{P}} \sim N\left(\frac{\sigma_{\tau|\theta}^{-2}\mu_{\tau|\theta_j}^{2} + \sum_{i=1}^{I}\alpha_i^2(\beta_i - \ln t_{ij})}{\sigma_{\tau|\theta_j}^{-2} + \sum_{i=1}^{I} a_i^2}, (\sigma_{\tau|\theta_j}^{-2} + \sum_{i=1}^{I} a_i^2)^{-1} \right). \tag{43}$$

The conditional means $\mu_{\tau|\theta_j}$ and variances $\sigma_{\tau|\theta_j}^2$ follow directly from $\boldsymbol{\mu}_{\mathrm{P}}$ and $\boldsymbol{\Sigma}_{\mathrm{P}}$ 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 $\mu_{\beta|a_i,b_i,\alpha_i}$ and variance $\sigma_{\beta|a_i,b_i,\alpha_i}^2$, the posterior distribution of β_i is also normal:

$$\beta_i \mid \mathbf{t}_j, \boldsymbol{\tau}, \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \sim N\left(\frac{\sigma_{\beta|a_i,b_i,\alpha_i}^{-2}\mu_{\beta|a_i,b_i,\alpha_i} + \alpha_i^2\sum_{j=1}^{J}(\ln t_{ij} + \tau_j)}{\sigma_{\beta|a_i,b_i,\alpha_i}^{-2} + J\alpha_i^2}, \left(\sigma_{\beta|a_i,b_i,\alpha_i}^{-2} + J\alpha_i^2\right)^{-1} \right). \tag{44}$$

The conditional means $\mu_{\beta|a_i,b_i,\alpha_i}$ and variances $\sigma_{\beta|a_i,b_i,\alpha_i}^2$ 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),

$$f(\alpha_i \mid \boldsymbol{\mu}_{\mathrm{P}}, \boldsymbol{\Sigma}_{\mathrm{P}}) = \phi(\alpha_i; \mu_{\alpha|a_i,b_i,\beta_i}, \sigma_{\alpha|a_i,b_i,\beta_i}^2). \tag{45}$$

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

$$f(\alpha_i \mid t_{ij}, \tau_j, \beta_i) \propto \prod_{j=1}^{J} f(t_{ij};\ \tau_j, \alpha_i, \beta_i)\phi(\alpha_i; \mu_{a|a_i,b_i,\beta_i}, \sigma_{\alpha|a_i,b_i,\beta_i}^2), \tag{46}$$

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

$$\max\left\{1, \frac{f(a_{it}^* \mid t_{ij}, \tau_j, \beta_i)}{f(\alpha_{i(t-1)} \mid t_{ij}, \tau_j, \beta_i)} \times \frac{\varphi(\alpha_{i(t-1)}, \alpha_{it}^*)}{\varphi(\alpha_{it}^*, \alpha_{i(t-1)})}\right\}; \tag{47}$$

otherwise, the value at the preceding iteration is retained; that is, $\alpha_{it} = \alpha_{i(t-1)}$. The ratio of the posterior densities in (47) simplifies to

$$\left(\frac{\alpha_{it}^*}{\alpha_{i(t-1)}}\right)^J \exp\left\{-\frac{1}{2}\left[(\alpha_{it}^{*2} - \alpha_{i(t-1)}^2)\sum_{j=1}^{J}(\ln t_{ij} - (\beta_i - \tau_j))^2 + \frac{(\alpha_{it}^* - \mu_{a|a_i,b_i,\beta_i})^2 - (\alpha_{i(t-1)} - \mu_{a|a_i,b_i,\beta_i})^2}{\sigma_{a|a_i,b_i,\beta_i}^2}\right]\right\}. \tag{48}$$

Step 9

Population parameters $\boldsymbol{\mu}_\mathrm{P}$ and $\boldsymbol{\Sigma}_\mathrm{P}$ 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 in the normal/Inverse-Wishart family:

$$\boldsymbol{\Sigma}_\mathrm{P} \mid \xi \sim \text{Inverse-Wishart}(\boldsymbol{\Sigma}_{\mathrm{P}*}^{-1}, \nu_{\mathrm{P}*}); \tag{49}$$

$$\boldsymbol{\mu}_\mathrm{P} \mid \xi, \boldsymbol{\Sigma}_\mathrm{P} \sim \text{MVN}(\boldsymbol{\mu}_{\mathrm{P}*}, \boldsymbol{\Sigma}_\mathrm{P} / \kappa_{\mathrm{P}*}); \tag{50}$$

where

$$\boldsymbol{\Sigma}_{\mathrm{P}*} = \boldsymbol{\Sigma}_{\mathrm{P}0} + \mathbf{S}_\xi + \frac{\kappa_{\mathrm{P}0} I}{\kappa_{\mathrm{P}0} + I}(\xi - \bar{\xi})(\xi - \bar{\xi})^T; \tag{51}$$

$$\nu_{\mathrm{P}*} = \nu_{\mathrm{P}0} + I; \tag{52}$$

$$\kappa_{\mathrm{P}*} = \kappa_{\mathrm{P}0} + I; \tag{53}$$

$$\boldsymbol{\mu}_{\mathrm{P}*} = \frac{\kappa_{\mathrm{P}0}}{\kappa_{\mathrm{P}0} + I}\boldsymbol{\mu}_{\mathrm{P}0} + \frac{I}{\kappa_{\mathrm{P}0} + 1}\bar{\xi}; \tag{54}$$

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

$$\mathbf{S}_\xi = \sum_{i=1}^{I}(\xi - \bar{\xi})(\xi - \bar{\xi})^T. \tag{55}$$

Step 10

The sampling of the item-domain parameters $\boldsymbol{\mu}_\mathcal{I}$ and $\boldsymbol{\Sigma}_\mathcal{I}$ from their posterior distributions given ψ is similar to (48)–(54) with $\boldsymbol{\mu}_\mathrm{P}$ $\boldsymbol{\Sigma}_\mathrm{P}$, ξ, and I replaced by $\boldsymbol{\mu}_\mathcal{I}$ $\boldsymbol{\Sigma}_\mathcal{I}$, $\boldsymbol{\psi}$, and N.

Discussion

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 for the discrimination parameter in the response-time model (Step 8). The proposed Metropolis-Hastings (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 Carlin and Louis (2000, sect. 5.4).

Goodness of Fit of the Lognormal Model

The two most important component models are the lower-level response and response-time models. Response models have been applied routinely in operational testing. An application of the lognormal model for the adaptive version of a test in the Armed Services Vocational Aptitude Battery (ASVAB) in van der Linden (in press) showed an excellent fit. Figure 3 gives an impression of the best-fitting and worst-fitting item in the set of 48 items calibrated in this study. Each plot shows the cumulative distribution of the probabilities of exceedance for the actual response times ("Bayesian *p*-values") for the 2,000 test takers in the sample. A perfect fit is obtained if the curve coincides with the identity line. The two plots seem to imply a satisfactory fit for the entire range of items in this study.

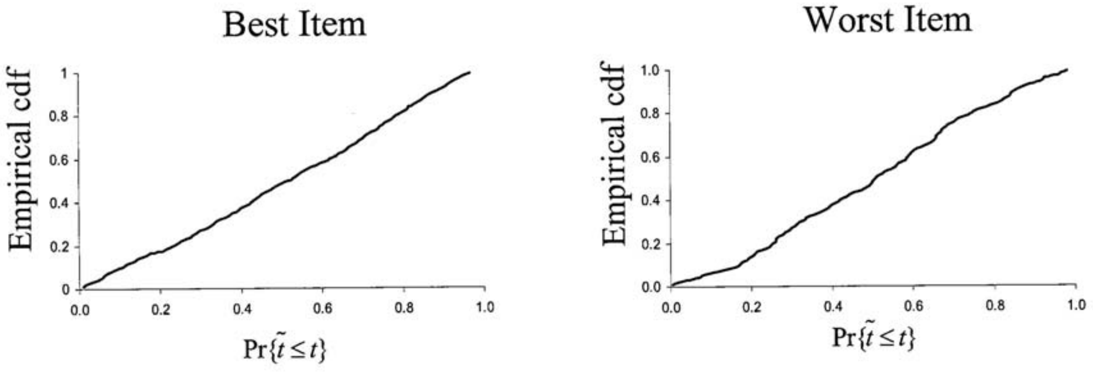

FIGURE 3. *Fit of the best- and worst-fitting item to the lognormal response-time model*

Concluding Comment

The presence of both a response model and a response-time model at the lower level gives the hierarchical framework large applicability in educational and psychological testing. In particular, the second-level link between the item and person parameters in the framework allows us to predict response-time parameters from the responses, and conversely.

The model can also be used for analyzing reaction time data in psychological experiments. It does not force us to use the same standardized tasks in one experiment. Further, it allows us to equate results from different experiments and to study the same type of task performed under different conditions of speed. In fact, if the same set of tasks is repeated under different conditions of speededness, a version of the framework with the population model replaced by a within-person model could be used to estimate a tradeoff between speed and accuracy.

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.

Beguin, 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. (tentatively accepted). Modeling item parameter variability in item response models. *Psychometrika*.

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.

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 reactie-tijden* [Models for reaction times]. Unpublished master's thesis, Faculty of Psychology, University of Groningen, The Netherlands.

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

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam and 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 Rep. 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: 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* (pp. 179–203). New York: Academic Press.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge, England: Cambridge University Press.

Verhelst, N. D., Verstraalen, 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.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

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., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

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