Applied Psychological Measurement
34(5) 327–347
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621609349800
http://apm.sagepub.com
SAGE

IRT Parameter Estimation With Response Times as Collateral Information

Wim J. van der Linden[1], Rinke H. Klein Entink[2], and Jean-Paul Fox[2]

Abstract

Hierarchical modeling of responses and response times on test items facilitates the use of response times as collateral information in the estimation of the response parameters. In addition to the regular information in the response data, two sources of collateral information are identified: (a) the joint information in the responses and the response times summarized in the estimates of the second-level parameters and (b) the information in the posterior distribution of the response parameters given the response times. The latter is shown to be a natural empirical prior distribution for the estimation of the response parameters. Unlike traditional hierarchical item response theory (IRT) modeling, where the gain in estimation accuracy is typically paid for by an increase in bias, use of this posterior predictive distribution improves both the accuracy and the bias of IRT parameter estimates. In an empirical study, the improvements are demonstrated for the estimation of the person and item parameters in a three-parameter response model.

Keywords

collateral information, empirical Bayes estimation, item response theory (IRT), hierarchical modeling, response time

Item response theory (IRT) models belong to a class of models that explain the data for each unit of observation (here: each combination of a test taker and an item) by different parameters. One of the main advantages of a hierarchical approach to this class of models is the borrowing of information on one parameter from the data collected for the units associated with the other parameters. This borrowing is possible through the assumption of a common distribution of the parameters at a second level in the statistical model for the data. The estimator of the parameter then typically compromises between the assumed distribution and the likelihood associated with the data. In doing so, it tends to strike a profitable balance between ignoring the data on the other parameters (= case of separate estimates) and the rather stringent assumption of all

[1] CTB/McGraw-Hill, Monterey, California, USA
[2] University of Twente, Enschede, The Netherlands

Corresponding Author:
Wim J. van der Linden, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940
Email: wimvanderlinden@ctb.com

parameters being equal (= case of pooled estimates). The profit typically occurs in the form of a higher efficiency of the inference at the cost of a less serious increase in bias.

One of the first examples in test theory demonstrating this principle is the classical true-score estimate based on Kelley's regression function,

$$E(T|X = x) = \rho_{XX'}x + (1 - \rho_{XX'})\mu_T, \tag{1}$$

where X is the observed score of the test taker, T the true score, μ_T the mean true score in the population of test takers, and $\rho_{XX'}$ the reliability of the test (Lord & Novick, 1968, section 3.7). An estimate of the test taker's true score, $\hat{\tau}$, is obtained by substituting estimates of μ_T and $\rho_{XX'}$ derived from the marginal distribution of the observed scores into Equation 1. The estimate is a compromise between $X = x$ as a direct estimate of τ and the estimate of the population parameter μ_T in the sense of a linear combination with weights $\rho_{XX'}$ and $1 - \rho_{XX'}$. But using a well-known variance partition in classical test theory, the estimate can be shown to be also equivalent to the precision-weighted average of x and $\hat{\mu}_T$ (Novick & Jackson, 1974, equation 9.5.11).

As discussed extensively in Novick and Jackson (1974, section 9.5), the Kelley estimate is an instance of the more formal problem of estimating multiple means simultaneously. Later examples of applications of the same principle of borrowing information are the estimation of multiple regressions in m groups from normal data in Novick, Jackson, Thayer, and Cole (1972) and the estimation of proportions in m groups from binomial data in Novick, Lewis, and Jackson (1973). An instructive empirical application of the estimation of multiple regressions is the often-cited study of the effects of coaching schools on SAT scores in Rubin (1981).

The nature of the ''distribution of the parameters'' has not yet been specified. In a frequentist approach, the units of observation are typically assumed to be sampled from a population and hence their parameters are taken to be random. The interpretation of Kelley's estimate in classical test theory belongs to this approach. From a Bayesian perspective, any density that approximates the empirical distribution of the parameters for the data set becomes a profitable common prior for the inference with respect to each of these parameters. The difference between a hierarchical model with a population distribution at the highest level and this empirical Bayes approach resides only in their motivation and interpretation; the more formal aspects of both approaches involve the same two-level structure. For an introduction to the empirical Bayes approach, see, for example, Carlin and Louis (2000, chap. 3).

To emphasize that the borrowed information is collected simultaneously with the direct information on the parameters, Novick and Jackson (1974, section 9.5) introduced the notion of *collateral information.* This term avoids the more temporal connotation in the Bayesian terminology of *prior information*, which suggests that the extra information should always be present before any data on the parameters is collected.

It should be noticed that the use of the term *information* differs from that elsewhere in scientific endeavors, where it is typically taken to imply that observations can be predicted from other variables. However, collateral information in the hierarchical sense does not require the presence of any predicting variables but is already available if the units of observation can be assumed to follow a common distribution. If the assumption holds, as soon as data are collected for the parameters of some of the units, information is received on all of them, for example, about their typical range of values.

In this article, first-level models are combined for the responses and the response times (RTs) by the test takers on the items with second-level models for the joint empirical distributions of their item and person parameters. As a result, it is not only possible to borrow information on the response parameter for one item from the responses collected for the other items but also from the RTs collected for the *same* item. The same borrowing is possible for the response parameter for each person. Because responses and RTs are always recorded simultaneously, the additional information in the RTs is literally collateral. Surprisingly, as shown later in this article, the fact that the collateral information is specific for the individual items and persons leads to improvement of *both the accuracy and the bias of the estimates*. In doing so, the information thus breaks the bias–accuracy tradeoff typical of more traditional hierarchical modeling.

The research in this article was motivated by the fact that now that computer-based testing has become more popular, and RTs in this mode of testing are automatically recorded, it would be imprudent to ignore such information. At the same time, there is a continuous need for item calibration from smaller samples and ability estimation from shorter tests. The latter is particularly helpful to maintain realistic time limits on batteries of multiple tests as they are used, for instance, in diagnostic testing for vocational guidance or instructional purposes. Before showing how joint hierarchical modeling of responses and RTs helps to exploit such information, a closer look is taken first at the role of collateral information in the more traditional problem of parameter estimation in a separate IRT model.

Collateral Information in IRT

The example considered is the estimation of ability parameter θ_j for a test taker j in a response model (e.g., the well-known three-parameter logistic model) of which the item parameters are already known. The case is met, for instance, when a test from a calibrated item pool is used to measure the abilities of several test takers.

Suppose the test takers are from a population with a normal distribution of ability $N(\mu_\theta, \sigma^2_\theta)$, of which the mean μ_θ and variance σ^2_θ have already been estimated with enough precision to treat them as known. Estimates of θ_j that capitalize on this information should be based on the posterior distribution

$$f(\theta_j|\mathbf{u}_j, \mu_\theta, \sigma^2_\theta) \propto f(\mathbf{u}_j|\theta_j)f(\theta_j|\mu_\theta, \sigma^2_\theta), \tag{2}$$

where $\mathbf{u}_j = (u_{1j,\ldots,}u_{nj})$ are the responses by test taker j on the n items in the test and $f(\mathbf{u}_j|\theta_j)$ is the probability of the observed response vector by the test taker under the response model.

The mean of this posterior distribution, which is often used as a point estimate of θ_j, is generally known to have a smaller mean square error than a classical estimate based on the probability of the observed data, $f(\mathbf{u}_j|\theta_j)$, only. The decrease is due to the information in the population density $f(\theta_j|\mu_\theta, \sigma^2_\theta)$ in the right-hand side of Equation 2, which shows, for instance, where the ability parameters in the population are concentrated and how much they are dispersed. The decrease is paid for by an increase in the bias of the ability estimate toward the mean of the population of test takers or the domain of items. This combination of effects is an example of the well-known bias–accuracy tradeoff in statistics. However, in hierarchical modeling, the tradeoff is exploited to work to our advantage; the improved accuracy is generally realized at a less serious increase in bias. The same principle can be shown to hold for the estimation of the item parameters.

Hierarchical Model

To profit fully from the information on the IRT parameters in the RTs, a model has to be adopted for the RTs and the common distribution of all person and item parameters has to be modeled as well. The result is a hierarchical framework with the IRT and RT models as first-level components and population and domain models for the IRT and RT parameters as second-level components. For a graphical illustration of the framework, see Figure 1.

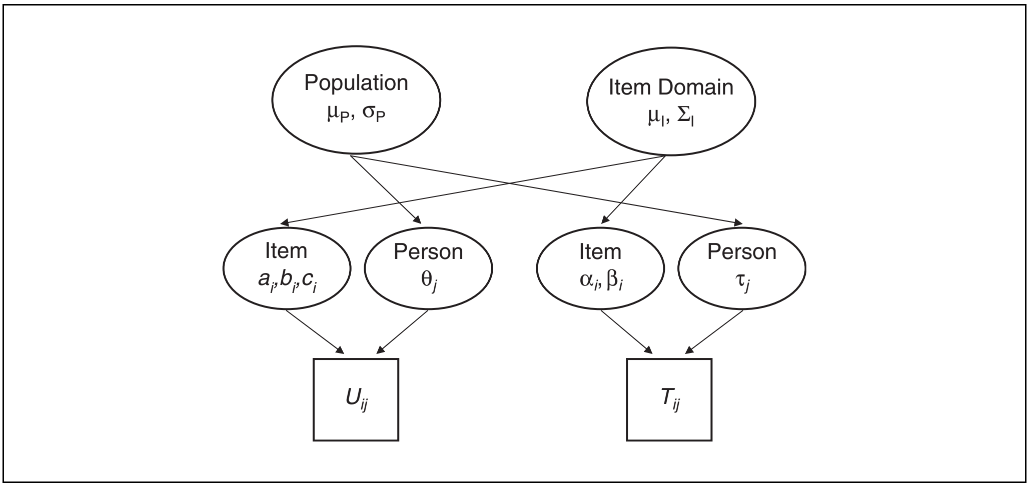

Figure 1 Graphical representation of hierarchical modeling framework for responses and response times

The RT model in Equation 4 below was proposed in van der Linden (2006) whereas the extension with the second-level models below was introduced in van der Linden (2007). These models are taken only as an example to demonstrate the benefits of using RTs as collateral information when estimating IRT parameters; for specific applications, such as tests with polytomous items or for misreadings, other models need to be substituted.

Response and RT Models

As the first-level model for the responses of test takers $j = 1, ..., N$ on items $i = 1, ..., n$, the three-parameter normal-ogive (3PNO) model is used, which gives the probability of a correct response on item i by person j as

$$P(U_{ij} = 1; \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\Phi(a_i(\theta_j - b_i)), \tag{3}$$

where $\Phi(\cdot)$ denotes the normal distribution function, θ_j is the ability parameter for test taker j, and a_i, b_i, and c_i are the discrimination, difficulty, and guessing parameters for item i, respectively.

Response–time distributions are often approximated well by lognormal distributions. Therefore, analogous to the IRT model in Equation 3, the RTs are modeled as a lognormal model with a speed parameter τ_j for test taker j and time intensity and discrimination parameters β_i and α_i for item i, respectively. Let T_{ij} denote the RT of test taker j on item i. The lognormal model posits that (van der Linden, 2006, 2007)

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))\right]^2\right\}. \tag{4}$$

Notice that except for the difference in sign, which is due to the negative relationship between time and speed, the interpretation of the two parameters for speed and time intensity in Equation 4 are analogous to those for the ability and item difficulty in Equation 3. However, unlike Equation 3, RT distributions have a natural zero and do not involve the estimation of any lower asymptote.

Population and Domain Models

The population model specifies the joint distribution of the person parameters θ and τ. It is assumed that the distribution is bivariate normal,

$$(\theta, \tau) \sim MVN(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{5}$$

where

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau) \tag{6}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{7}$$

Likewise, item parameters a_i, b_i, and c_i in the response model and α_i and β_i in the RT model are assumed to have a multivariate normal distribution,

$$(a, b, c, \alpha, \beta) \sim MVN(\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}), \tag{8}$$

where

$$\boldsymbol{\mu}_{\mathcal{I}} = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta) \tag{9}$$

and matrix $\boldsymbol{\Sigma}_{\mathcal{I}}$ has all variances and covariances between the item parameters as elements.

This hierarchical model is not yet fully identifiable. In addition to the usual lack of identification for a hierarchical IRT model, the parameters β_i and τ_j in the RT model are not identified; addition of a constant to β_i can be compensated by addition of the same constant to τ_j. Identifiability is obtained if one sets $\boldsymbol{\mu}_{\mathcal{P}} = \mathbf{0}$ and $\sigma_\theta^2 = 1$. The reason σ_τ^2 need not be fixed to a known constant is that all parameters in the RT model are on a time scale with a direct observable unit (e.g., second).

Alternative Component Models

As already noted, alternative response and RT models can be substituted for the three-parameter response and lognormal RT model in Equations 3 and 4. One set of candidates is the Poisson model for reading errors and gamma model for reading speed by Rasch (1960/1980). Both models have separate reading person and item parameters and therefore allow for the distinct types of population and item domain models in the preceding section (van der Linden, 2009). Other examples of response models are the well-known two-parameter logistic model and the Rasch model or models for polytomous responses. For alternative RT models based on a Weibull distribution, see Rouder, Sun, Speckman, Lu, and Zhou (2003) and Tatsuoka and Tatsuoka (1980), where Maris (1993) should be consulted for RT models based on gamma distributions. However, in addition to these distinct models for responses and RTs, hybrid models have been proposed that either model response distributions but have time parameters as well (e.g., Roskam, 1997; Verhelst, Verstralen, & Jansen, 1997; Wang & Hanson, 2005) or RT distributions but include response parameters (e.g., Thissen, 1983). These models do not lend themselves for substitution as first-level models in the framework above. For instance, substitution of a response model with RT parameters into (3) would double the presence of this type of parameter in the two first-level models and violate the inherent assumption of conditional independence between responses and RTs on which the framework is based (see below). A more complete review of alternative RT models is given in Schnipke and Scrams (2002); for more details on distinct and hybrid models for response and RT distributions, see van der Linden (2009).

Bayesian Estimation

In the empirical examples later in this article, the model parameters were estimated using Bayesian estimation with data augmentation and Gibbs sampling. The data augmentation was according to the method described in Albert (1992; see also Johnson & Albert, 1999), that is, with postulated continuous latent variables underlying the responses. Gibbs samplers are Monte Carlo Markov chain (MCMC) methods for sampling from the posterior distribution of the item parameters (e.g., Gelfand & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 2004, chap. 11). They do so by iteratively sampling from the conditional posterior distributions of one set of parameters given the previous draws from the distributions of all other parameters. For the current modeling framework, a conjugate normal-inverse-Wishart prior for the mean vectors and covariance matrices for the multivariate models in Equations 5 and 8 was specified (Gelman et al., 2004, section 3.6). This joint prior leads to sampling from conditional posterior distributions with known densities. For technical details of the estimation method, see Klein Entink, Fox, and van der Linden (2009) and van der Linden (2007). A package of procedures in the statistical language R that implements the method is described in Fox, Klein Entink, and van der Linden (2007).

Different Sources of Information

The same principle as in Equation 2 is demonstrated but this time for a test taker j with response vector $\mathbf{u}_j = (u_{1j}, \ldots, u_{nj})$ and RT vector $\mathbf{t}_j = (t_{1j}, \ldots, t_{nj})$. Again, without loss of generality, it is assumed that the second-level means, $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\mu}_{\mathcal{I}}$, and covariance matrices $\boldsymbol{\Sigma}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{I}}$, have already been estimated during item calibration. Consequently, θ_j and τ_j are the only unknown parameters.

The complication we are now faced with is an estimation problem under two separate models—a primary model for the responses and another for the RTs. To assess the improvement in the estimation of θ_j relative to Equation 2, it is attempted to factorize the posterior distribution of θ_j into a product with the primary model probability for the observed responses as a factor. A comparison between the remainder of this product and the prior distribution of θ_j in Equation 2 should allow assessing the improvement in the estimation of θ_j relative to Equation 2.

The posterior distribution of θ_j follows from the joint distribution of θ_j and τ_j given all known quantities

$$f(\theta_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) = \int f(\theta_j, \tau_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j. \tag{10}$$

For the integral, it holds that

$$\int f(\theta_j, \tau_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j \propto \int f(\mathbf{u}_j, \mathbf{t}_j|\theta_j, \tau_j)f(\theta_j, \tau_j|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j. \tag{11}$$

Hence, because of local independence between responses and RTs given (θ_j, τ_j),

$$f(\theta_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) \propto \int f(\mathbf{u}_j|\theta_j)f(\mathbf{t}_j|\tau_j)f(\theta_j, \tau_j|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j. \tag{12}$$

Observe that this is a different form of local independence than the usual independence of responses to items conditional on θ. For an empirical study confirming the plausibility of the assumption, see van der Linden and Glas (2010).

Factorizing $f(\theta_j, \tau_j|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$, the following is obtained

$$
\begin{aligned}
f(\theta_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) &\propto f(\mathbf{u}_j|\theta_j)\int f(\theta_j|\tau_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) f(\mathbf{t}_j|\tau_j) f(\tau_j|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j \\
&\propto f(\mathbf{u}_j|\theta_j)\int f(\theta_j|\tau_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) f(\tau_j|\mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})d\tau_j, \qquad (13)
\end{aligned}
$$

where the second step follows from the definition of the posterior distribution of τ_j as

$$
f(\tau_j|\mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) \propto f(\mathbf{t}_j|\tau_j) f(\tau_j|\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}). \qquad (14)
$$

For the integral in the second line of Equation 13, it holds that

$$
\int f(\theta_j|\tau_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) f(\tau_j|\mathbf{t}_j)d\tau_j \propto f(\theta_j|\mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}). \qquad (15)
$$

Notice that the right-hand side of Equation 15 is the conditional posterior density of θ_j given $\mathbf{t}_j$, that is, the probability of the test taker's ability θ_j given his or her speed τ_j integrated over the posterior distribution of τ_j given the response times $\mathbf{t}_j$.

Thus, it can be concluded that

$$
f(\theta_j|\mathbf{u}_j, \mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) \propto f(\mathbf{u}_j|\theta_j) f(\theta_j|\mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}). \qquad (16)
$$

The result has a simple form that is entirely analogous to Equation 2. It shows that when the RTs are used as collateral information, θ_j is estimated from the probability $f(\mathbf{u}_j|\theta_j)$ of the observed response vector $\mathbf{u}_j$ as when the RTs are ignored but with the original prior distribution of θ in Equation 2 replaced by the conditional posterior distribution of θ_j given the RTs $\mathbf{t}_j$ for the test taker.

More generally, the result also answers the earlier question of how to deal with the presence of two different models in the statistical inference for one kind of parameters in a hierarchical framework as in Equations 3 through 9. The solution is to keep the model with the primary parameters intact but absorb the second model in the conditional posterior density of the primary parameters given the information collected for the other parameters.

The result in Equation 16 helps identify *three* different sources of information on θ_j :

1. The information directly available in $\mathbf{u}_j$ in the first factor of Equation 16, that is, the regular model probability $f(\mathbf{u}_j|\theta_j)$ associated with the observed response vector.
2. The information summarized in the estimates of $\boldsymbol{\mu}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{P}}$ in the second factor. This information is derived from the vectors of responses and RTs in the entire sample of test takers. These estimates generalize the role of those of μ_θ and σ^2_θ in Equation 2.
3. The information in the shape of the conditional posterior distribution of the response parameters given the response times. Unlike the estimates of the population parameters in the preceding source of information, the information in this vector is unique for each individual test taker.

An analogous role for the RTs can be shown to hold for the estimation of the item parameters $\boldsymbol{\xi}_i = (a_i, b_i, c_i)$. For the sake of argument, let it be assumed that the person parameters are known. (Actually, in the Gibbs sampler, each time the conditional posterior distribution of $\boldsymbol{\xi}_i$ is sampled, all other parameters are treated as known.) Using the same derivation as before with $\boldsymbol{\mu}_{\mathcal{P}}$, $\boldsymbol{\Sigma}_{\mathcal{P}}$, θ_j, and $\mathbf{t}_j$ replaced by $\boldsymbol{\mu}_{\mathcal{I}}$, $\boldsymbol{\Sigma}_{\mathcal{I}}$, $\boldsymbol{\xi}_i$, and $\mathbf{t}_i$, respectively, the result is

$$f(\boldsymbol{\xi}_i|\mathbf{u}_i, \mathbf{t}_i, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) \propto f(\mathbf{u}_i|\boldsymbol{\xi}_i) f(\boldsymbol{\xi}_i|\mathbf{t}_i, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}), \tag{17}$$

where the distribution of the responses $\mathbf{u}_i$ given $\boldsymbol{\xi}_i$ is the one in Equation 3 for known ability parameters and the posterior distribution of $\boldsymbol{\xi}_i$ given $\mathbf{t}_i$, $\boldsymbol{\mu}_{\mathcal{I}}$, and $\boldsymbol{\Sigma}_{\mathcal{I}}$ is obtained analogous to Equation 14.

The treatment in Equations 10 through 17 was for the case of the second-level parameters having been estimated from an earlier sample with enough precision to treat them as known. But for the current argument, it does not matter if they were estimated along with the current θ_j and $\boldsymbol{\xi}_i$ parameters. The estimates of θ_j and $\boldsymbol{\xi}_i$ are constrained by the same type of conditional posterior distributions $f(\theta_j|\mathbf{t}_j, \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ in Equation 16 and $f(\boldsymbol{\xi}_i|\mathbf{t}_i, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}})$ in Equation 17 when the estimation procedure fits the estimates of $(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ and $(\boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}})$ simultaneously to the data set.

Bias–Accuracy Tradeoff

It is important to notice the consequences of the replacement of the common prior distribution of θ for all test takers in Equation 2 by the *individual* posterior distribution of θ_j in Equation 16. This posterior distribution not only has a smaller variance but, because of its conditioning on the RTs for each individual test taker, also tends to have a location closer to his or her true ability level. The impact of the former is a further increase of the posterior precision of θ_j ; the impact of the latter is a decrease in the bias of the posterior mean. The strength of both effects is dependent on the correlation between the person parameters in the response and RT models. Obviously, if the correlation is perfect, the same parameter is being estimated twice. The individual posterior distribution of θ_j would then tend to duplicate its response-based likelihood at the same location; that is, there would be double the information and no constraining by a common prior for θs at different locations.

As already argued, parameter estimation in traditional hierarchical IRT modeling is subject to the well-known bias–accuracy tradeoff in statistics. For this type of modeling, the addition of a population model for the distribution of the θ parameters leads both to an increase in the bias and a decrease in the lack of accuracy of their estimates. However, typically the net result—summarized in the mean square errors of the parameters—is positive.

On the other hand, the posterior density of θ_j in Equation 16 is conditional on the actual RTs by each test taker j and therefore serves as a prior with an individual location for each of them. Unlike the use of a population distribution as a common prior, whose location necessarily compromises between the true values of the individual θs, and hence produces a bias in their estimates, these individual priors avoid the necessity of such a compromise. In addition, they have smaller variances and are thus more informative.

Empirical Examples

Simulation studies were conducted to demonstrate the effect of the use of the collateral information in RTs on the estimation of the IRT model parameters. Simulated data were used because they allow comparing the estimates of the parameters against their true values. But before presenting these studies, a small example with empirical data is given to motivate interest in the effects of RTs.

Empirical Data

The data were from an item calibration study for a pool of items for a computerized version of a figural matrix test by Hornke and Habon (1986). The total data set was for 30,000 test takers and 456 dichotomously scored items, each with eight options. Each test taker answered only 12 items but the total data matrix was connected. In an earlier study, the model was fitted successfully and a correlation of $\rho_{\theta\tau} = -.61$ was found (for details, see Klein Entink, Kuhn, Hornke, & Fox, 2009).

The first two blocks of 667 test takers and 17 items from the data set were taken and the IRT parameters both without and with the RTs were estimated, and the same model as in the earlier study (2PNO model) was fitted. The Gibbs sampler was run for the full model with the earlier normal-inverse-Wishart priors with prior means $(\mu_{\theta_0}, \mu_{\tau_0}) = (0, 0)$ (for identification), prior means $(\mu_{a_0}, \mu_{b_0}, \mu_{\alpha_0}, \mu_{\beta}) = (1, 0, 2, 3)$, scale matrices $\boldsymbol{\Sigma}_{\mathcal{P}_0}$ and $\boldsymbol{\Sigma}_{\mathcal{I}_0}$ both equal to a diagonal matrix with elements 10, and two degrees of freedom. Observe that the choices of diagonal elements for the scale matrices and the degrees of freedom imply negligible prior information. For the run for the IRT model only, the same settings for the remaining prior parameters were used. Both times, the sampler was run for 10,000 iterations. The running time was approximately 15 minutes. Trace plots with the draws showed convergence after a burn-in of 500 iterations.

Figure 2 shows a plot of the two sets of θ estimates for the conditions without and with RTs (first panel) as well as a plot with the corresponding two sets of posterior standard deviations (*SD*s). The points in the first plot are around the identity line with a random variation due to estimation error typical of a test of 17 items. The second plot illustrates the effect of RTs in the form of a trend to lower *SD*s for the condition with the RTs. Figure 3 shows item parameter estimates also close to the identity line. Thus, the two types of estimation produce essentially the same parameter estimates but the use of RTs tends to improve their accuracy.

Simulated Data

Obviously, improvements in the accuracy of the parameter estimates are dependent on the second-level correlations in the hierarchical framework in Equations 3 through 9. In the simulation studies, the correlation between the speed and ability parameters, $\rho_{\theta\tau}$, was the focus. (The effects for the other parameters are analogous.) The effects of the use of the RTs on the IRT parameter estimates were therefore evaluated for a range of alternative sizes of $\rho_{\theta\tau}$. In a review of empirical studies with estimates of $\rho_{\theta\tau}$, they were found to have values in the range $[-.65, .30]$ for tests as disparate as the Armed Services Vocational Aptitude Battery (ASVAB), Certified Public Accountant Exam, Graduate Management Admission Test, and a test of quantitative and scientific proficiencies for colleges students (van der Linden, 2009). For the correlation between the item difficulty and time intensity parameter, $\rho_{b\beta}$, the range was $[-.33, .65]$.

To assess the improvements of the estimation of the item and person parameters in the response model separately, two studies with a different setup had to be conducted. One study addressed the gain in statistical accuracy of the ability estimates due to the collateral information in RTs. The other addressed the same issue for the estimation of the item parameters. For both studies, the setup was conservative in that only the correlation between the ability and speed parameters was varied. The gains reported therefore do not include any additional impact of the RTs through the joint distribution of the item parameters in Equations 8 and 9.

The accuracy of the estimates of θ depends on the length of the simulated test and its item parameters. The item parameters had a typical range of values (see below). The test length was set equal to $n = 30$ items, which was neither too long nor overly short relative to real-world tests. On the other hand, when estimating the θs both from the responses and the RTs, the effective number of observations on the parameter is longer. Roughly speaking, the effective test length should then be between 30 ($\rho_{\theta\tau} = 0$) and 60 items ($\rho_{\theta\tau} = 1$). Besides, notice that for a fixed correlation, the *relative* amount of information on θ in the responses and RTs remains identical if the test length increases; each additional item entails one extra response and one

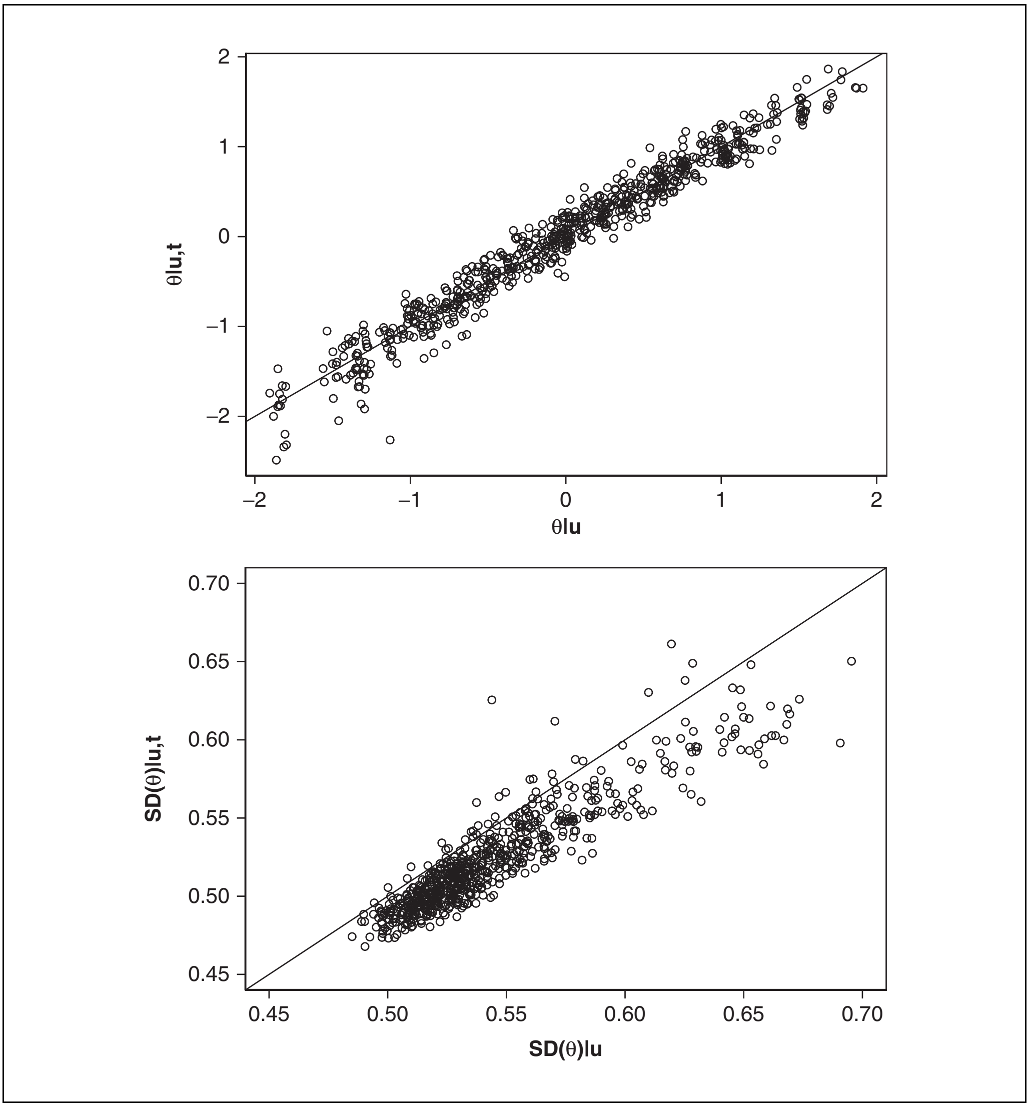

Figure 2 Estimates of θ and their posterior standard deviations (*SD*s) for a figural matrix test without and with response times

RT. But of course, eventually estimation error vanishes, and although their relative importance does not change, so will the absolute size of the contribution of the RTs.

Likewise, for the item calibration study, it was only needed to focus on the size of the sample of test takers and its distribution of ability parameters. The standard normal was chosen as the simulated ability distribution. The sample size was $N = 300$. This is a small size for a real-world item calibration study but, analogous to the effective length in the preceding argument, the use of RTs increases the effective sample size to a higher level. The number of items does not have any impact on the accuracy of the estimates of the parameters of an individual item. Two different versions were needed of the second study, each with several conditions and multiple replications. However, the total running time for the project had to be kept manageable (the average time for one run was

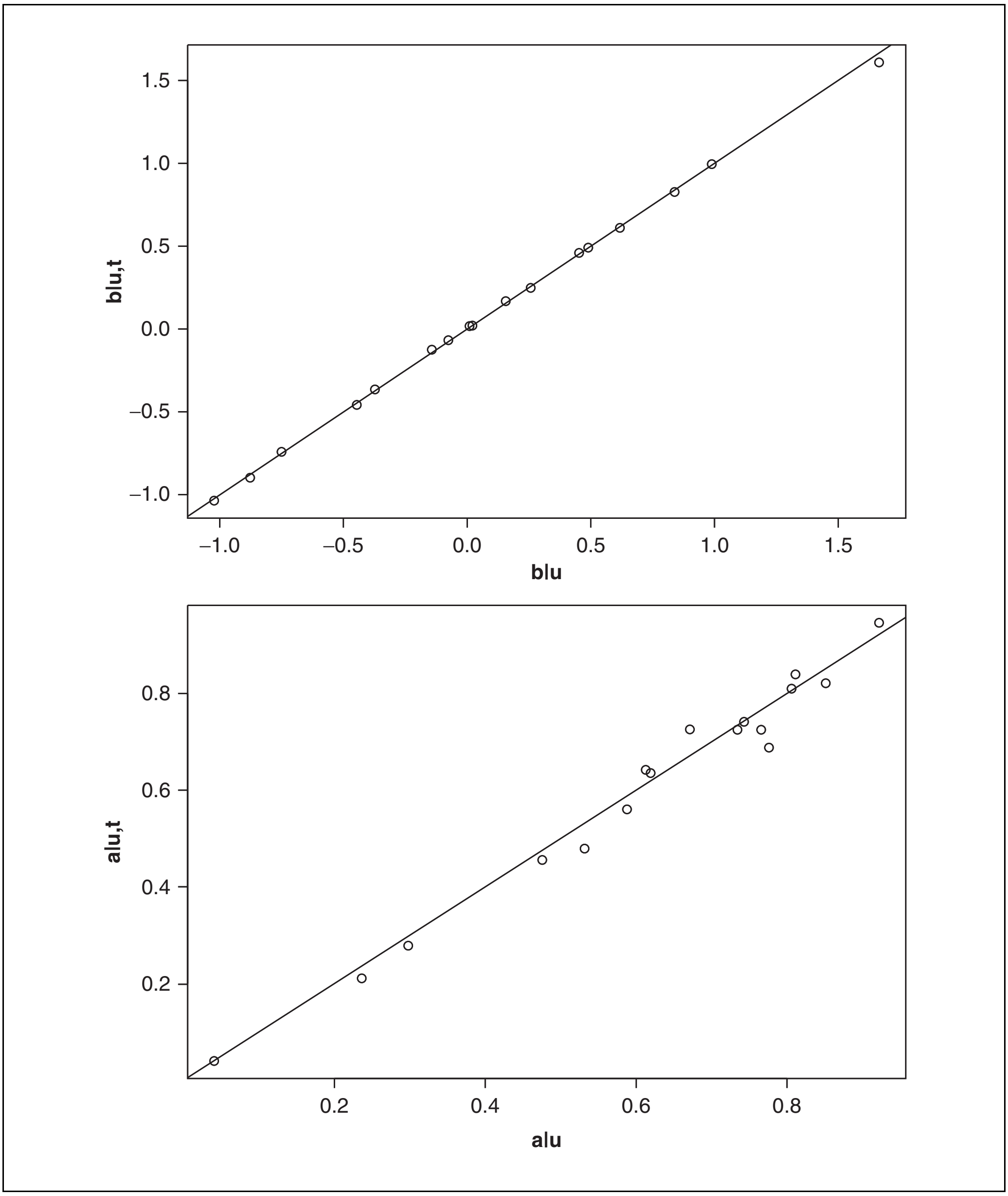

Figure 3 Estimates of the difficulty parameters b_i (upper panel) and discrimination parameters a_i (lower panel) for the items in a figural matrix test without and with response times

approximately 30 minutes). Given the above, small numbers of items were used in these studies, with evenly spread simulated parameter values to cover the relevant portions of their scales.

Study 1: Ability Estimation

Responses and RTs were simulated for $N = 1,000$ test takers on a 30-item test for five levels of correlation: $\rho_{\theta\tau} = 0$, .25, .50, .75, and 1. As the sign of the correlation has no impact on the amount of information about θ in the RTs (nothing changes if the scale of θ had been reversed), only positive correlations were used.

The parameters a_i were randomly drawn from $U(0.8, 1.2)$. To guarantee a test with uniform distribution of the item difficulties across θ, the difficulty parameters b_i were not sampled, but equally spaced values were used on $(-\frac{31}{15}, \frac{31}{15})$ with steps of $\frac{2}{15}$. The case of four-choice items was simulated and c_i was fixed at .25 for all items. The reason for this choice is comparability between the current and the next studies. Fixing the c_i parameter at a plausible value is common practice in the testing industry because it prevents the effects of possible tradeoffs between the a_i and c_i parameters—known to exist because of occasional weak identifiability of the 3PL model—during estimation. The next study assessed the impact of the collateral information in the RTs on the estimation of the item parameters, and the tradeoff would have contaminated the authors' assessment for the more interesting a_i parameters. The other two item parameters for the RT model were randomly drawn from $\alpha \sim U(1, 2)$ and $\beta \sim N(5, 1)$. This choice was motivated by earlier applications of the lognormal RT model with estimates of these parameters in the same ranges; see, for instance, the distribution of the estimates for a large data set from the AS-VAB in van der Linden (2006, figure 4, panel 1). Typical RTs for the simulated test takers would be 2 to 3 minutes per item.

As the correlation between θ and τ had to be manipulated, it was chosen to fix the θs for the test takers and then one τ for each θ was sampled from the conditional distribution with the desired correlation. The hierarchical framework assumes a normal bivariate distribution for (θ,τ) in Equation 5. The earlier identifiability constraints imply a standard normal marginal distribution of θ. The distribution was realized by fixing the values of the θs for the $N = 1,000$ test takers at the .001th, .002th, . . . quantiles of $N(0, 1)$. More formally, the values are defined as $\theta_p = \Phi^{-1}(p)$ for $p = .001, .002, \ldots, 1$, where Φ is the standard normal distribution function. This choice guaranteed a fine grid of quantiles θ_p values that covered the whole θ range and made it possible to assess the mean standard error (MSE) and bias of the estimators of θ with uniform precision across this grid. The speed parameters were randomly drawn from the conditional normal distributions of $\tau|\theta$ with the appropriate correlation $\rho_{\theta\tau}$, where the marginal distribution of τ was assumed to be $N(0, 1)$ as well. (One of the earlier identifiability constraints sets the mean of τ also equal to zero.) Given all these parameter values, vectors of responses and RTs were generated for each test taker. To estimate the bias and mean square error of the θ estimates, the entire setup was replicated 10 times for each of the five sizes of the correlation between θ and τ.

The ability parameters of the test takers were estimated for two different cases: First, all item parameters were assumed to be known and the ability parameters were the only parameters estimated (case of measurement using previously calibrated items). Second, both the item and person parameters were treated as unknown and estimated simultaneously (case of ability estimation in a calibration study). Because the results showed negligible differences between the MSEs and bias of the estimates of the θs for the two cases, the present study focuses on the first case.

The parameters were estimated using the Gibbs sampler referred to earlier. A noninformative version of the earlier normal-inverse-Wishart prior for the population model was used with mean vector equal to zero (for identification), a diagonal scale matrix with 10 elements, and 2 degrees of freedom. The Gibbs sampler was run for 10,000 iterations. A burn-in of 500 iterations was sufficient to reach convergence; autocorrelation between the draws appeared to be smaller than .10 after each 10th iteration.

The use of the RTs as collateral information was evaluated for the EAP estimates of θ (= mean of their posterior distributions), which for the Gibbs sampler were easily obtained as

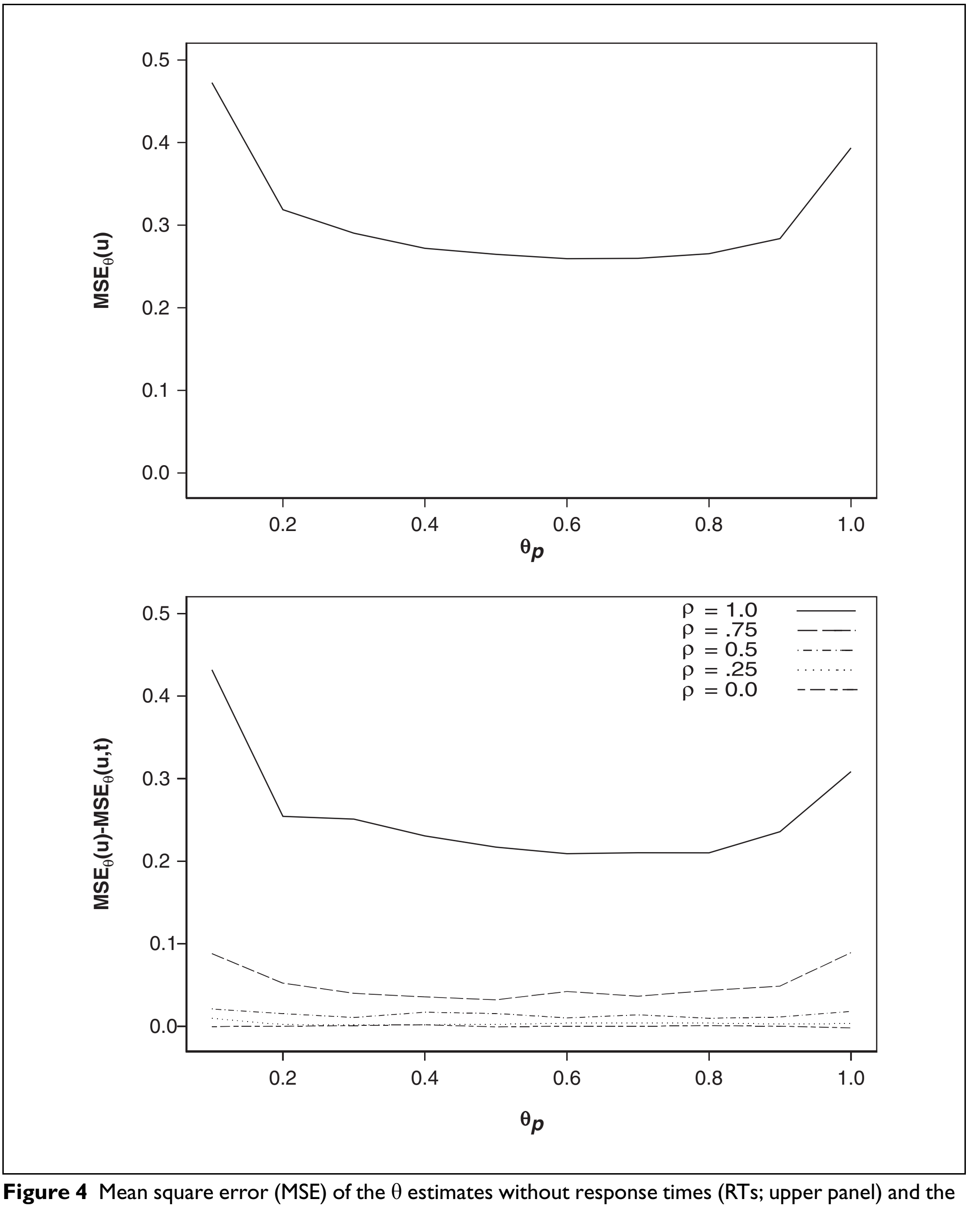

Figure 4 Mean square error (MSE) of the θ estimates without response times (RTs; upper panel) and the reduction in the MSE due to the use of RTs for different correlations $\rho_{\theta\tau}$ (lower panel)

Note: The results for the θ_ps are depicted at their p values; see the earlier text.

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\theta^{(m)}, \tag{18}$$

where $\theta^{(m)}$ was the draw from the posterior distribution of θ at iteration $m = 1, \ldots, M$ after burn-in. The bias and mean square error (MSE) were used as criteria, which were estimated as the average error, $\hat{\theta} - \theta$, and average squared error, $(\hat{\theta} - \theta)^2$, across replications for each of the simulated θ values.

Figure 4 shows both the MSE for the case of estimation without RTs (upper panel) and the reduction in MSE because of the use of the RTs , that is, $MSE_\theta(\mathbf{u}) - MSE_\theta(\mathbf{u}, \mathbf{t})$, over the grid of quantiles θ_p for the five conditions for $\rho_{\theta\tau}$ (lower panel). The baseline case in the upper panel has an MSE curve with a typical increase at the two ends of the scale. The MSE curves for the use of the RTs show the same general shape, which indicates that the reduction in MSE is greatest where it is needed most. The horizontal line at height zero for $\rho_{\theta\tau} = 0$ corresponds to the baseline case of fitting the IRT model without RTs. As the correlation increases, the reduction in the MSE becomes greater. Obviously, the more collateral information on the ability parameters in the RTs, the more accurate their estimates become. Notice that the θ_p values do not have the same metric as the original scale of θ but that the change has no impact on the vertical scale with the size of the MSEs. The transformation was required to average the MSEs over intervals with 10% of the test takers and create uniform precision for the plotted MSEs across the grid of θ_p values.

Figure 5 shows similar plots for the decrease in bias in the estimation of θ. For abilities below the population mean $\mu_\theta = 0$ (middle of the scale), the baseline in the upper panel shows a positive bias; above this point, the bias is negative. This pattern is typical of Bayesian estimation in IRT. Except for some sampling variation in the middle of the scale due to the number of replications, the pattern for the reduction of the bias is similar. It is reminded that the MSE of the estimators is equal to their variance plus the square of their bias. A comparison between the results in Figures 4 and 5 thus also gives an impression of the variance of the estimators.

On average, the improvement in accuracy and bias across grid of θ quantiles in these examples was some 5% for $\rho_{\theta\tau} = .5$ and 20% for $\rho_{\theta\tau} = .75$. The reduction was smaller than these averages for the abilities near the middle of the scale but larger toward the upper and lower ends of the scale, where the abilities are harder to measure.

Study 2: Item Calibration

The second study was to evaluate the use of the collateral information in the RTs in item calibration. Its setup had to be more complicated because (a) the accuracy of the item parameter estimation was not to be confounded with the estimation of the θs and (b) the effects of the RTs on the item parameters a_i and b_i were to be studied separately. For a Gibbs sampler, the first requirement can be realized by drawing the θs directly from the same fixed posterior distributions across all replications, which prevents variation in the impact of the posterior distributions on the estimation of the item parameters to show up as part of their MSE. The fixed posterior distributions of the θs were produced by running the Gibbs sampler first for response data for the same simulated test takers on an arbitrary test and recording the draws from the posteriors once the sampler had stabilized. These draws were then used in the main study. The second requirement was met by using a setup with a range of values for a_i parameters while holding the b_i parameters constant for all items and repeating the study with the roles of the parameters reversed.

More specifically, the setup was as follows. First, the posterior densities of the ability parameters of $N = 300$ test takers were obtained for a test of 10 items with parameters b_i equally spaced on $[-1.8, 1.8]$ and parameters a_i randomly drawn from $U(.8, 1.2)$. As the correlation between θ and τ had to be manipulated again, the same procedure with a fixed distribution of the θs matching $N(0, 1)$ and random sampling of one τ for each θ from their appropriate conditional distributions as in Study 1 was followed. This time, the distribution of the θs was realized by fixing the values for the 300 test takers at the .033th, .066th, . . . quantiles of $N(0, 1)$. Given all these parameters, vectors of responses and RTs were simulated for each of the test takers. The

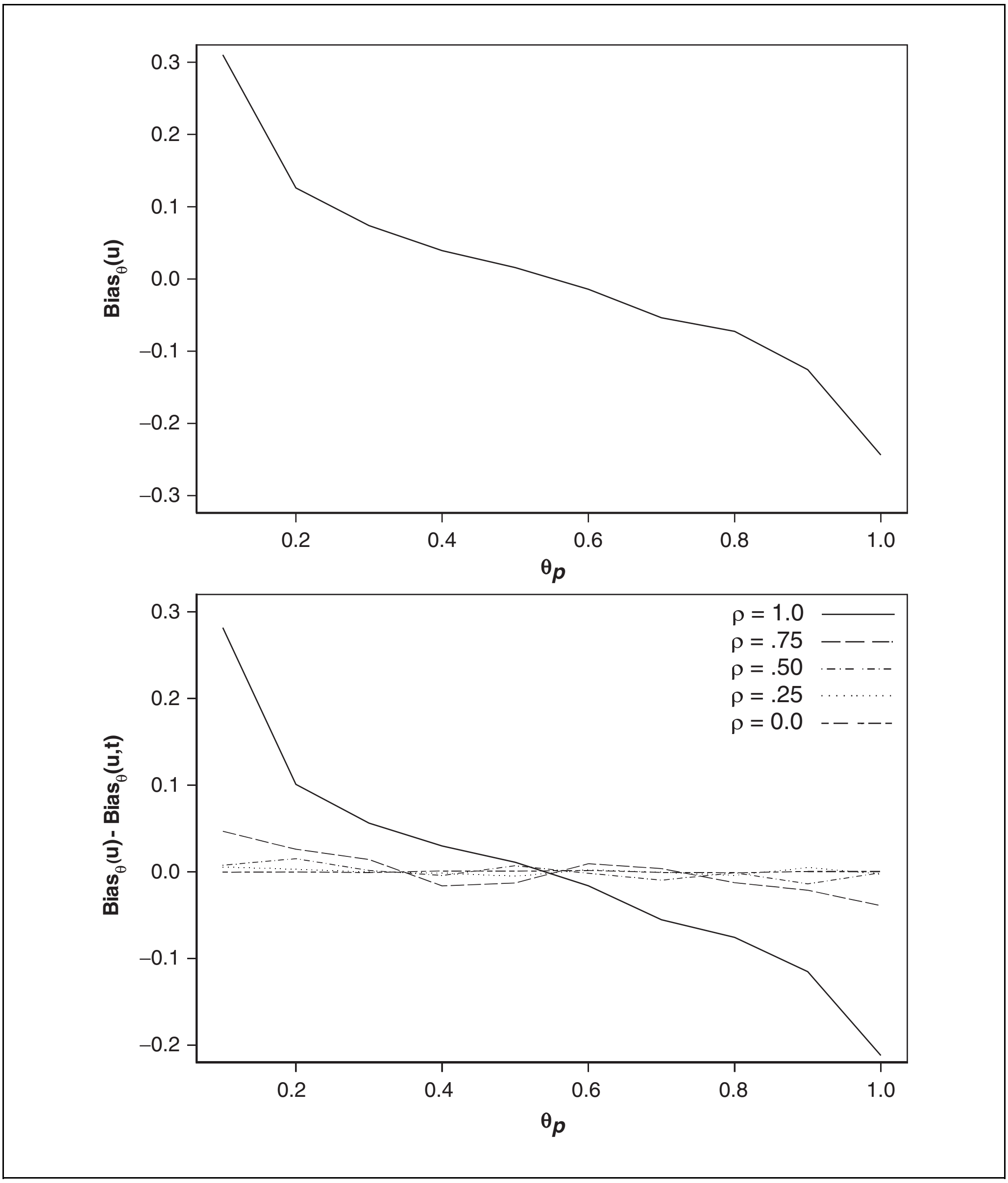

Figure 5 Bias of the θ estimates without response times (RTs; upper panel) and the reduction in the bias due to the use of RTs for different correlations $\rho_{\theta\tau}$ (lower panel)

Note: The results for the θ_ps are depicted at their p values; see the earlier text.

Gibbs sampler was then run with known item parameters and speed parameters to obtain 10,000 draws (after burn in) from the posterior distributions of θs, and all draws were saved for the main study. The procedure was repeated for $\rho_{\theta\tau} = 0$ (baseline), .5, .7, and .9. Observe that the posterior distributions of θs for each of these four conditions were for the same item parameters. Therefore, the comparison between these conditions was entirely fair as to the (arbitrary) choice of these parameters.

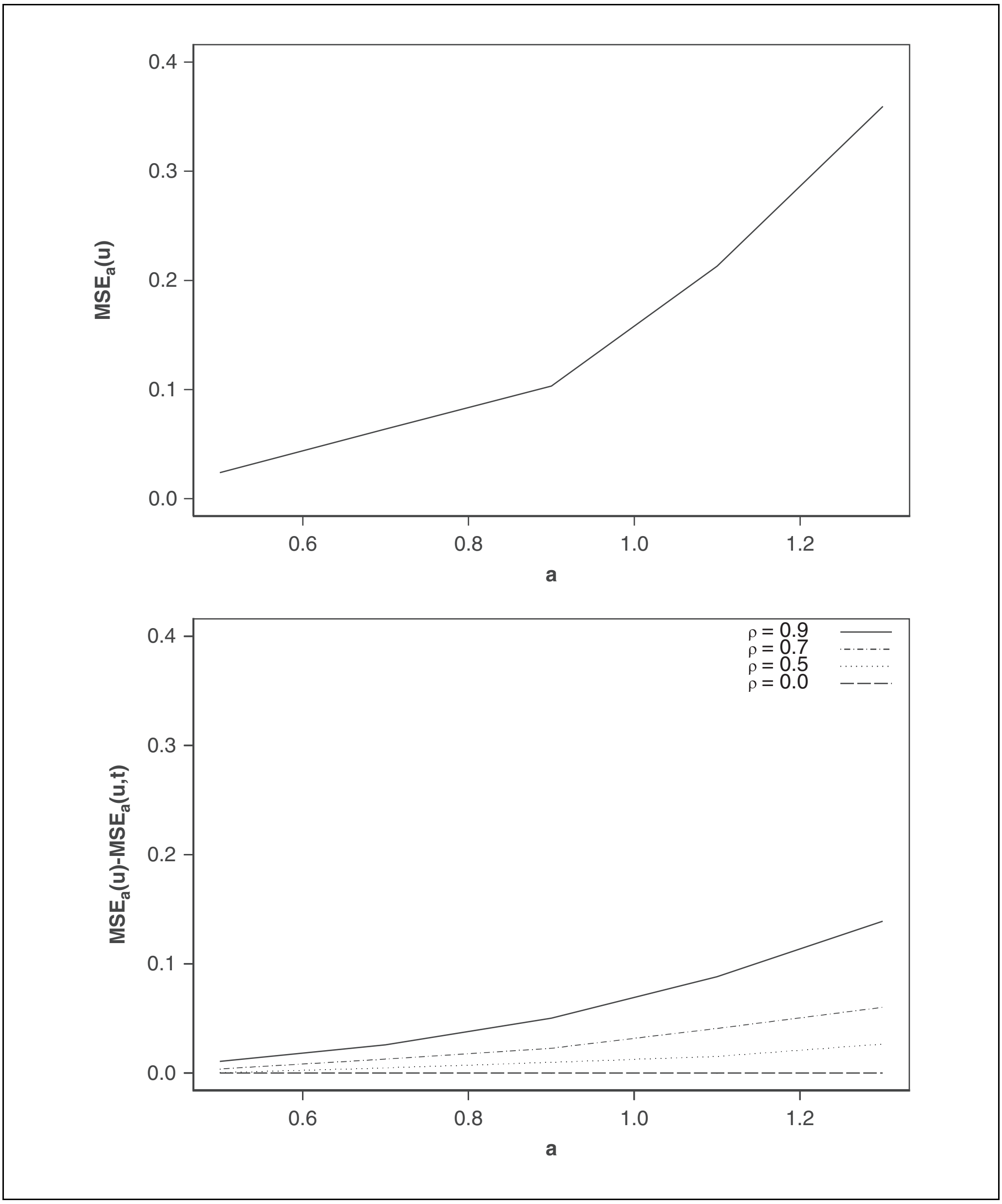

Figure 6 Mean square error (MSE) of the a_i estimates without response times (RTs; upper panel) and the reduction in the MSE due to the use of RTs for different correlations $\rho_{\theta\tau}$ (lower panel)
Note: The difficulty parameter is held constant at $b_i = 0$.

Second, the posterior densities of the ability parameters obtained in the first step were used to calibrate new items. From Study 1, it was known that the higher the correlation $\rho_{\theta\tau}$, the more accurate the ability estimates. Hence, it was also expected that higher correlations would lead to more accurate estimates of the item parameters. To check this expectation, for the same simulated test takers, response patterns $\mathbf{u}_2$ were generated for two different versions of a five-item

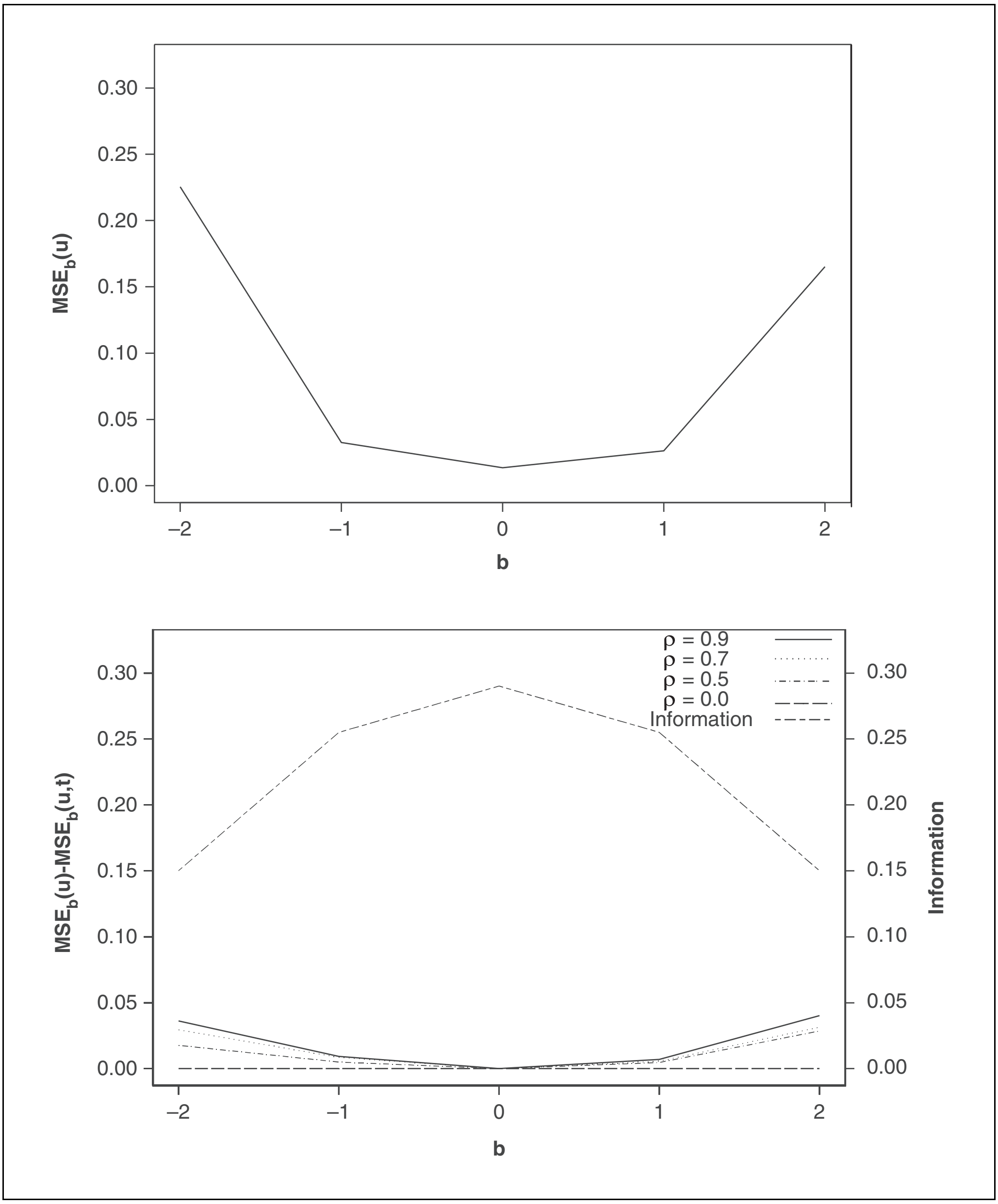

Figure 7 Mean square error (MSE) of the b_i estimates without response times (RTs; upper panel) and the reduction in the MSE due to the use of RTs for different correlations $\rho_{\theta\tau}$ (lower panel)
Note: The discrimination parameter is held constant at $a_i = 1$.

test: (a) $a_i = .5, .7, .9, 1.1,$ and 1.3 but $b_i = 0$ for all items and (b) $b_i = -2, -1, 0, 1,$ and 2 but $a_i = 1$ for all items. By first varying the a_i s while holding the b_i s constant and vice versa, it was possible to evaluate the effects of the correlation $\rho_{\theta\tau}$ on the two kinds of item parameters separately. For both cases, the item parameters were calibrated using the version of the Gibbs sampler with the draws obtained earlier from the posterior distributions of the ability parameters. As already indicated, the case of four-choice items was simulated with c_i fixed at .25 for all items.

As before, the MSE criterion was used to evaluate the EAP estimates of the a_i and b_i parameters, which were calculated analogously to Equation 18. For example, for the estimates of the a_i parameters, the focus was on the reduction in MSE due to the use of the RTs, that is, $MSE_a(\mathbf{u}) - MSE_a(\mathbf{u}, \mathbf{t})$. The evaluation of the estimates of the b_i parameters was entirely analogous.

The results are presented in Figures 6 and 7. The upper panels show the baseline cases of estimating the a_i and b_i parameters without RTs. These panels confirm the well-known facts that estimation error tends to be much higher for larger a_i parameters and at the lower and upper end of the scale for the b_i parameters. The lower panels show the reduction in MSE relative to the condition without the use of RTs as a function of $\rho_{\theta\tau}$. (Again, the line at height zero in the lower panel for $\rho_{\theta\tau} = 0$ corresponds with the baseline case without RTs). As these panels reveal, the greater absolute improvement in accuracy involved in the use of RTs is obtained for the larger discrimination parameters and more extreme difficulty parameters. Generally, in terms of percentages, the average improvements are in the same range as for Study 1—some 5% for $\rho_{\theta\tau} = .5$ and 20% for $\rho_{\theta\tau} = .75$.

The effect is nicely illustrated by the information curve in the lower panel, which represents the sum of the Fisher information about the difficulty parameter b_i across the grid of 300 θs in the simulation study (using the known a_i and c_i parameters). The lower the curve, the harder the difficulty parameter is to estimate. The curves with the reduction in MSE go up precisely where this curve goes down, exactly as should happen for a calibration study in practice.

Discussion

Since its inception, test theory has been hierarchical; the randomness of an observed score of an individual test taker has always been distinguished from that of his or her true score because of sampling from a population. In addition, for its statistical inference, test theory has been an early adopter of the Bayesian methodology. It therefore seems natural to broaden the traditional hierarchical (vertical) type of modeling of responses in IRT with the horizontal extension of RT modeling in this article.

Except for access to the software (for a download, see Fox et al., 2007), the practical implementation of the hierarchical model for the joint estimation of the response and RT parameters does not involve any special requirements. In computerized testing, the RTs on the items are automatically logged. Also, as the use of the MCMC method for the RT model does not involve any data augmentation, the increase in running time associated with joint estimation is only modest.

The average improvement in estimation in the examples with simulated data was some 5% to 20% for correlations $\rho_{\theta\tau}$ in the range from .5 to .75, with larger improvements toward the end of the θ and b_i scale and for the larger discrimination parameters a_i. It is important to emphasize that these improvements do not result from using some informative subjective priors for the person and item parameters. The only priors that need to be specified in the implementation of the model are for the (higher-level) means and covariance matrices of the population and item domain distributions, and these were chosen to be noninformative in the studies. Besides, the improvements in the estimation of the item parameters in Figures 6 and 7 were obtained only through the use of the collateral information in the RTs in the estimation of the abilities. The part of the hierarchical framework for the joint distribution of all item parameters in Equations 8 and 9 was not used, and additional beneficial effects due to their correlations were therefore not possible. Finally, it should be noted that all results were obtained for an arbitrary fixed test. In adaptive testing, the use of RTs in the estimation of the abilities leads to considerable improvement of the adaptation of the selection of the items to the test takers, and consequently to higher gains in the accuracy of ability estimation (van der Linden, 2008).

Improvement of parameter estimation has always been a concern of the testing industry; it makes test scores more informative and reduces the costs of item calibration. But there has also been a general reluctance to use other information than the test takers' performance on the test items, especially when the information is population dependent. This reluctance is understood but the following elements should be added to the discussion. First, RTs *are* part of the test takers' performance on the test items. Using them is not the same as, for example, the practice of regressing the test takers' abilities on background variables (socioeconomic status, type of school district, etc.) in large-scale assessments of educational achievements or any other type of information with only an indirect relation to the test performance. Therefore, less objection is expected against the use of RTs, particularly when estimating item parameters.

Second, the use of RTs does not change the construct or dimension measured by the test in any way. As demonstrated in Figure 2, the same parameter θ is estimated with and without the use of RTs as collateral information, only the accuracy changes.

Third, the modeling framework does require the specification of a second-level population distribution and may therefore seem to suggest some form of population-dependent test scoring. However, the role of the second-level distribution is different from that in traditional hierarchical estimation. For example, in Kelley's regression function and in Equation 2, the estimates are pulled toward the mean true score in the population of test takers, and different estimates are obtained for different populations. On the other hand, the θ estimate from Equation 16 is dependent only on the *conditional* distribution of θ given the test taker's speed τ. In particular, it does not depend on the population distribution of θ, and the same examinee working at the same speed can be expected to have the same estimator when being included in a different population.

Fourth, it is expected that the use of RTs as collateral information will not be an issue for ability estimation in low-stakes testing (e.g., diagnosis for remedial instruction in education). If it would be an issue in the more controversial area of high-stakes testing, the RTs could still be used jointly with the responses to optimize the test but a final score could be produced based on the responses only. An example is adaptive testing, where the items during the test can be selected using the θ estimates in this article but the final estimate could be inferred from the responses only. For this application, roughly the same reduction of test length has been found as for the MSEs of the θ estimates in the empirical example above (van der Linden, 2008).

Finally, one main conclusion from this article can be summarized by stating that in order to get better estimates of the test takers' abilities, the speed at which they have responded should be estimated as well. The reverse problem of estimating the test takers' speed was not discussed in this article. Because the modeling framework is symmetrical with respect to the two estimation problems, the reverse conclusion holds, too; to efficiently estimate how fast test takers respond to the items in the test, their speed should be estimated along with their ability.

At first sight, these conclusions seem counterintuitive. But they follow directly from the Bayesian principle of collateral information for the joint hierarchical modeling used in this research.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received the following financial support for the research and/or authorship of this article: funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this article are those of the authors and do not necessarily reflect the policy and position of LSAC.

References

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

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, FL: Chapman & Hall.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software, 20*(7), 1-14.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelman, A., Carlin, J. B., Stern, H., & Rubin, D. B. (2004). *Bayesian data analysis*. London: Chapman & Hall.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*, 369-380.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*, 21-48.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14*, 54-75.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in *m* groups: A cross validation study. *British Journal of Mathematical and Statistical Psychology, 25*, 33-50.

Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The estimation of proportions in *m* groups. *Psychometrika, 38*, 19-46.

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

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.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6*, 377-401.

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). Mahwah, NJ: Lawrence Erlbaum.

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.

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

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

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

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247-272.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*, 120-139.

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.

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