PSYCHOMETRIKA—VOL. 83, NO. 1, 109–131
MARCH 2018
https://doi.org/10.1007/s11336-017-9590-9

A GENERALIZED SPEED–ACCURACY RESPONSE MODEL FOR DICHOTOMOUS ITEMS

PETER W. VAN RIJN

ETS GLOBAL

USAMA S. ALI

EDUCATIONAL TESTING SERVICE

SOUTH VALLEY UNIVERSITY

We propose a generalization of the speed–accuracy response model (SARM) introduced by Maris and van der Maas (Psychometrika 77:615–633, 2012). In these models, the scores that result from a scoring rule that incorporates both the speed and accuracy of item responses are modeled. Our generalization is similar to that of the one-parameter logistic (or Rasch) model to the two-parameter logistic (or Birnbaum) model in item response theory. An expectation–maximization (EM) algorithm for estimating model parameters and standard errors was developed. Furthermore, methods to assess model fit are provided in the form of generalized residuals for item score functions and saddlepoint approximations to the density of the sum score. The presented methods were evaluated in a small simulation study, the results of which indicated good parameter recovery and reasonable type I error rates for the residuals. Finally, the methods were applied to two real data sets. It was found that the two-parameter SARM showed improved fit compared to the one-parameter SARM in both data sets.

Key words: response times, speed–accuracy, scoring rules, item response theory, expectation–maximization, generalized residuals, saddlepoint approximations.

1. Introduction

The use of computer-based educational and psychological assessments has enabled recording of response times on test items. At least three approaches to modeling response times in this context can be distinguished (for overviews, see van der Linden, 2009; Lee & Chen, 2011; Goldhammer, 2015). In the statistical approach, the focus is on finding an appropriate distribution to model response time. Numerous studies are available that specify person-specific speed parameters to model the RT distribution. Various different distributions for response times have been considered, such as Weibull (Rouder, Sun, Speckman, Lu, & Zhou, 2003) and log-normal (van der Linden, 2007) distributions. In the psychological approach, the aim is to find a model that describes the underlying process between item presentation and item response. Examples of this approach are the diffusion model (Tuerlinckx & de Boeck, 2005) and the race model (Ranger, Kuhn, & Gaviria, 2015). In the third approach, the focus is on measurement, which means that speed is part of the construct being measured. One example in this approach is the speed–accuracy response model proposed by Maris and van der Maas (2012), which describes the scores that result from a scoring rule that incorporates both speed and accuracy of item responses. In this paper, we focus on this last approach and specifically on scoring rules that incorporate response time.

The approach of using scoring rules to derive psychometric models is not new. It is seen in item response theory (IRT) models in which case only the accuracy of the item response is used.

Correspondence should be made to Peter W. van Rijn, ETS Global, Amsterdam, The Netherlands. Email: pvanrijn@ets.org; pvanrijn@etsglobal.org

The most famous example is the Rasch model (Rasch, 1960), which follows if it is required that there be a minimal sufficient statistic for latent ability independent of item parameters (Andersen, 1973). The sufficient statistic in this case is the sum of correct item responses (or sum score). Similar models can be derived using more general scoring rules (Haberman, 2006). One typical example is formula scoring, which has been used for a long time in a number of high-stakes tests that consist of multiple-choice items and assigns a penalty for incorrect answers in order to discourage guessing (Thurstone, 1919).

There is a long history of modeling response latency in ability testing and experimental psychology (Spearman, 1927; Thurstone, 1937; Luce, 1986). A situation in which RTs can be useful in a psychometric model is when they provide additional information about ability. Roskam (1997, p. 187) noted that response speed can be as much an indicator of ability as response accuracy, but the relation between speed and accuracy depends on the particular domain being tested and the type of test being used. A distinction here can be made between speed tests and power tests. In speed tests, the items are easy and the time to complete the test is the variable of interest. In power tests, the items are of increasing difficulty and the number of correct responses is the variable of interest. In educational assessments, these two pure forms are not often used, and typically tests contain elements of both. We argue that RTs might be more useful for obtaining higher measurement precision (see also, van der Linden, 2008) when tests are intended to measure basic skills (e.g., arithmetic and spelling) rather than when they measure higher-order skills (e.g., writing an essay).

Following this reasoning, we aim to explore psychometric models that use RTs in straightforward scoring rules proposed by Maris and van der Maas (2012). They focused on a scoring rule that results in positive scores for fast correct responses and negative scores for fast incorrect responses. They derived a psychometric model in which the scoring rule is assumed to yield a sufficient statistic for ability, similar to the Rasch model. We present some extensions of their approach in terms of the scoring rule, the estimation method, and the assessment of model fit. To our knowledge, there is little research on the fit of speed–accuracy response models. Our main extension with respect to modeling is that, compared to the model of Maris and van der Maas (2012), we introduce an additional item parameter that describes the discriminating power of an item. Its equivalent in IRT is the discrimination parameter in the two-parameter logistic model (2PLM; Birnbaum, 1968). Our first motivation is analogous to that of Birnbaum (1968, p. 402) in evaluating the question: Do the item scores resulting from a scoring rule that includes both accuracy and speed differ in discriminating power?

A feature of the modeling approach by Maris and van der Maas (2012) is that it is assumed that items are administered under specific item-level time limits. This approach was recently advocated by Goldhammer (2015), who argued that this gives better control over individual differences in trade-offs between speed and accuracy by the test administrator. Also, it was argued that test takers should be well aware of the item-level time limits. We believe that this does not necessarily have to be case and aim to explore how the modeling approach works if there are no specific item-level time limits (e.g., if a limit is chosen based on the observed distribution of response time). Although this may sound somewhat unfair, we note that in the 2PLM (and other commonly used IRT models) test takers typically do not know in advance which items have the highest discrimination, that is, the exact scoring rule is not known. In this model, all things being equal, correct answers to items with higher discrimination lead to higher ability estimates than correct answers to items with lower discrimination. This issue becomes even more complex in case of multidimensional IRT models (see, e.g., Hooker, Finkelman, & Schwartzman, 2009; van Rijn & Rijmen, 2015). Furthermore, if item-level time limits are used during test administration, they may be either too tight or too loose. A feature of the model of Maris and van der Maas (2012) is that the time limit determines the item discrimination, but this can only be determined experimentally by using different time limits. However, it is not likely that the same time limit is equally useful for all items (unless the items are very homogeneous). Our second motivation is that we think that the addition of a discrimination parameter to the model can make the modeling approach more versatile without having to use different time limits in actual test administrations.

The paper is organized as follows. We will first present a somewhat more general scoring rule than used by Maris and van der Maas (2012). In the following section, we derive our model and some of its properties by making use of theory for exponential families. Then, an estimation method is presented that uses the expectation–maximization (EM) algorithm. We describe three types of standard errors and develop two approaches to assess model fit. The first approach is based on residual analysis suggested by Haberman and Sinharay (2013) in the context of unidimensional item response theory (IRT) models and can be used to evaluate the fit of individual items. In the second approach, we derive a saddlepoint approximation to the density of the sum score, which can be compared to the observed frequency distribution in order to evaluate overall model fit. Finally, the presented model, estimation method, and model fit approach are evaluated by means of a simulation study and illustrated by two applications to real data.

2. Scoring Rules

We make use of the following general scoring rule (Maris & van der Maas, 2012, Equation 55). For person i and item j, we have a score variable S_{ij} given by

$$S_{ij} = \left[w_{j0}(1 - X_{ij}) + w_{j1}X_{ij}\right](d_j - T_{ij}), \tag{1}$$

where w_{j0} and w_{j1} are known weights for incorrect and correct responses, respectively, X_{ij} is the dichotomously scored item response variable with zero for an incorrect response and one for a correct response, d_j is a known time limit, and T_{ij} is the response time variable. Several scoring rules that have been studied elsewhere can be obtained by specification of the weights. For example, the scoring rule used by van der Maas and Wagenmakers (2005), referred to as the correct item summed residual time (CISRT), is obtained by setting $w_{j0} = 0$ and $w_{j1} = 1$. Alternatively, Maris and van der Maas (2012) use a penalty for fast incorrect responses in their scoring rule referred to as the signed residual time (SRT) scoring rule. This rule is obtained by setting $w_{j0} = -1$ and $w_{j1} = 1$. However, with the above representation, asymmetric scoring rules can also be employed, for instance, by setting $w_{j0} = -1/3$ and $w_{j1} = 1$. This is akin to formula scoring for dichotomous selected-response items in that guessing behaviors are penalized (see, e.g., Lord, 1975), but not as strongly as in the SRT.

Maris and van der Maas (2012) argue that one could train students to comply with the scoring rule and thus to the underlying psychometric model. This is in contrast to the more traditional psychometric approach in which the model is inferred from student responses (Maris & van der Maas, 2012, p. 18). A possible benefit of the proposed scoring rule is that perfect (or zero) accuracy is less problematic, because the response time can aid in differentiating. This can be especially useful for shorter and easier tests.

Note that no assumptions have yet been made about the response time variable. For example, Maris and van der Maas (2012) treat response time as a continuous variable, but it can also be treated as a discrete variable (see, e.g., Ranger & Kuhn, 2012). In the most extreme case, T_{ij} is dichotomous and d_j is fixed at one. As we shall see, differential treatment of the response time variable leads to different models and estimation methods. Also, the metric of response time is arbitrary. For example, for continuous response times we can fix the geometric mean of item time limits d_j, $j = 1, \ldots, n$ to be 1. However, one needs to be aware that changing the metric of response time can also be accounted for by changing the weights.

3. Model Derivation

Two cases of statistical inference can be distinguished in specifying speed–accuracy response models: conditional and marginal inference. In conditional inference, models can be created in which the conditional distribution belongs to an exponential family. In marginal inference, models can be constructed which involve mixtures of exponential families or mixtures in which one or more components belong to an exponential family (Haberman, 2016). Our main focus will be on marginal inference. For a regular exponential family, we can write the probability of the observed data $\mathbf{z}$ given θ (or the likelihood of θ) as follows

$$p(\mathbf{Z} = \mathbf{z}|\theta) = \exp[\theta s(\mathbf{z}) - a(\mathbf{z}) - b(\theta)], \tag{2}$$

where $s(\mathbf{z})$ is a sufficient statistic for θ (the scoring rule in our case). Following Maris and van der Maas (2012), we assume that time T is continuous. We now derive the speed–accuracy response model (SARM).

For person i and item j, we can use the scoring rule in the following way to obtain the density of the score S_{ij} for item j

$$f(s_{ij}|\theta_i) = f(x_{ij}, t_{ij}|\theta_i) = \frac{\exp(s_{ij}\eta_{ij})}{C(\eta_{ij})}, \tag{3}$$

where $\eta_{ij} = \alpha_j\theta_i + \beta_j$, θ_i is a person parameter, α_j is an item slope parameter, β_j is an item intercept parameter, and $C(\eta_{ij})$ is a normalizing factor. We shall refer to this model as the two-parameter speed–accuracy response model (2P-SARM). If we drop the item index of the slope parameter α_j, the model of Maris and van der Maas (2012) is obtained, and we shall refer to this model as the one-parameter speed–accuracy response model (1P-SARM). The normalizing factor is different for different scoring rules, but for the general scoring rule in Eq. (1) it is

$$\begin{aligned} C(\eta_{ij}) &= \sum_{k=0}^{1} \int_0^{d_j} \exp(w_{jk}(d_j - t)\eta_{ij})\mathrm{d}t \\ &= \frac{\exp(w_{j0}d_j\eta_{ij}) - 1}{w_{j0}\eta_{ij}} + \frac{\exp(w_{j1}d_j\eta_{ij}) - 1}{w_{j1}\eta_{ij}}. \end{aligned} \tag{4}$$

Note that the normalizing factor is not defined if one of the weights is zero or if the linear predictor η_{ij} is zero. However, limits are easily found using l'Hôpital's rule. The normalizing factor plays an important role in further derivations.

The probability of a correct response or the item response function (IRF) is given by

$$\begin{aligned} E(X_{ij}|\theta_i) = p(X_{ij} = 1|\theta_i) &= \frac{1}{C(\eta_{ij})} \int_0^{d_j} \exp\left[w_{j1}(d_j - t)\eta_{ij}\right]\mathrm{d}t \\ &= \frac{\frac{\exp(w_{j1}d_j\eta_{ij})-1}{w_{j1}\eta_{ij}}}{\frac{\exp(w_{j0}d_j\eta_{ij})-1}{w_{j0}\eta_{ij}} + \frac{\exp(w_{j1}d_j\eta_{ij})-1}{w_{j1}\eta_{ij}}}. \end{aligned} \tag{5}$$

For the SRT scoring rule, the IRF simplifies to $p(X_{ij} = 1|\theta) = \exp(d_j\eta_{ij})/\left[1 + \exp(d_j\eta_{ij})\right]$, which is equivalent to the 2PLM. It can be seen that for the SRT scoring rule the item time limit

d_j can be absorbed in the item parameters ($d_j\eta_{ij} = d_j(a_j\theta_i + b_j) = d_j a_j\theta_i + d_j b_j = a'_j\theta_i + b'_j$). So, a feature of the model is that the metric of the item time limit is arbitrary when it comes to the accuracy of the responses. This situation is similar to assigning different weights to responses and fitting the 2PL model. The weights do not have any impact on model fit, because they are countered by the discrimination parameters. For example, if the weight of an item is doubled, then the estimated discrimination parameter will be halved.

The response time density is

$$f(t|\theta_i) = \frac{\sum_{k=0}^{1}\exp[w_{jk}(d_j - t)\eta_{ij}]}{C(\eta_{ij})}. \tag{6}$$

It is easily found that response times are uniformly distributed between 0 and d_j if η_{ij} is zero, even if the scoring rule is asymmetric. However, if the time limit increases without bound, the response time distribution is a negative exponential only if the scoring rule is symmetric (i.e., $w_1 = -w_0$). So, the 2P-SARM does not necessarily lead to different types of response time distributions compared to the model of Maris and van der Maas (2012), but it allows more flexibility in the shapes they can take from one item to the next. Now, the expected response time, which we shall refer to as the item time function (ITF), is

$$E(T_{ij}|\theta_i) = \frac{1}{C(\eta_{ij})}\sum_{k=0}^{1}\int_0^{d_j} t\exp[w_{jk}(d_j - t)\eta_{ij}]\mathrm{d}t \tag{7}$$

$$= \frac{1}{C(\eta_{ij})}\sum_{k=0}^{1}\frac{\exp(w_{jk}d_j\eta_{ij}) - w_{jk}d_j\eta_{ij} - 1}{w_{jk}^2\eta_{ij}^2} \tag{8}$$

Figure 1 shows ITFs for three items with time limit $d = 1$ and different item parameters. The maximum of the ITFs is always half the item time limit d_j and its location is $\theta = -\frac{\beta_j}{\alpha_j}$, in which we can recognize the item difficulty parameter of the 2PLM. The expected response time decreases when ability and difficulty do not match. In contrast to the IRF, we do see a different effect of d_j and α_j on the ITF.

The expected score, referred to as the item score function (ISF), for the SRT can be derived easily using exponential family theory, because $E(S_{ij}|\theta_i) = b'_j(\theta_i) = \frac{\partial \log C(\eta_{ij})}{\partial\eta_{ij}}$. It is also found by

$$E(S_{ij}|\theta_i) = \sum_{k=0}^{1}\frac{1}{C(\eta_{ij})}\int_0^{d_j} w_{jk}(d_j - t)\exp[w_{jk}(d_j - t)\eta_{ij}]\mathrm{d}t \tag{9}$$

$$= \frac{1}{C(\eta_{ij})}\sum_{k=0}^{1}\frac{\exp(w_{jk}d_j\eta_{ij})(w_{jk}d_j\eta_{ij} - 1) + 1}{w_{jk}\eta_{ij}^2} \tag{10}$$

$$= \frac{C'(\eta_{ij})}{C(\eta_{ij})}, \tag{11}$$

where

$$C'(\eta_{ij}) = \frac{\partial C(\eta_{ij})}{\partial\eta_{ij}} = \sum_{k=0}^{1}\frac{w_{jk}^2 d_j^2\exp(w_{jk}d_j\eta_{ij})(w_{jk}d_j\eta_{ij} + 1)}{2w_{jk}}. \tag{12}$$

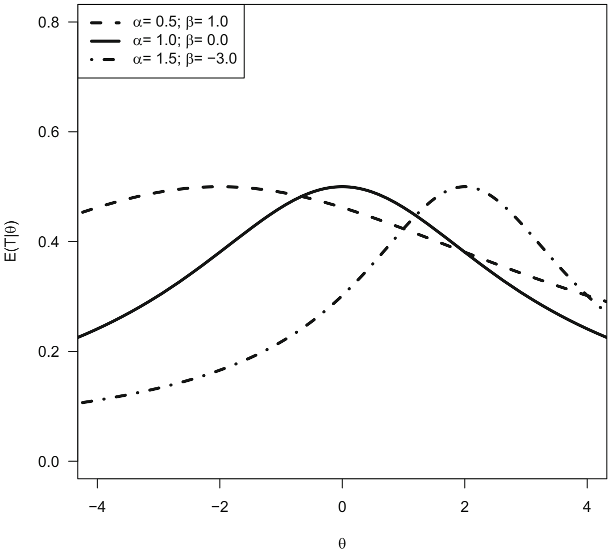

FIGURE 1.
Item time functions for three items.

In order to find the variance, we need the second moment of the score, which is $E\left(S_{ij}^2|\theta_i\right) = C''(\eta_{ij})/C(\eta_{ij})$, where the second derivative is

$$C''(\eta_{ij}) = \sum_{k=0}^{1} \frac{w_{jk}^3 d_j^3 \exp(w_{jk} d_j \eta_{ij}) \left(w_{jk}^2 d_j^2 \eta_{ij}^2 + 4 w_{jk} d_j \eta_{ij} + 2\right)}{6 w_{jk}}. \tag{13}$$

Now, the variance is

$$\text{Var}(S_{ij}|\theta_i) = E\left(S_{ij}^2|\theta_i\right) - E(S_{ij}|\theta_i)^2 = \frac{C''(\eta_{ij})}{C(\eta_{ij})} - \frac{C'(\eta_{ij})^2}{C(\eta_{ij})^2}. \tag{14}$$

Figure 2 shows ISFs for three items with time limit $d = 1$ and different item parameters.

Under the assumption that the item scores are independent conditional on θ, we can write the probability of response vector **s** given θ as

$$f(\mathbf{S} = \mathbf{s}|\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{n} f(S_j = s_j|\theta, \alpha_j, \beta_j) \tag{15}$$

$$= \prod_{j=1}^{n} \frac{\exp\left[s_j(\alpha_j\theta + \beta_j)\right]}{C(\eta_{ij})} \tag{16}$$

$$= \exp(\mathbf{s}'\boldsymbol{\alpha}\theta)\exp(\mathbf{s}'\boldsymbol{\beta})p_C(\boldsymbol{\eta}). \tag{17}$$

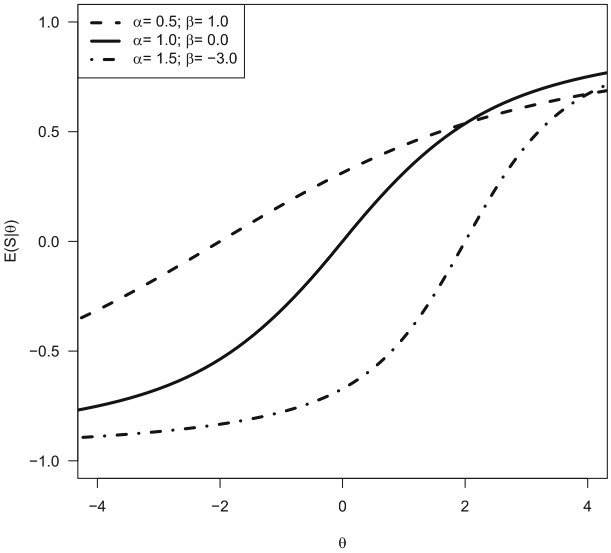

FIGURE 2.
Item score functions for three items.

The exponential form of Eq. (2) can now be recognized: $s(\mathbf{z}) = v = \sum_j \alpha_j s_j$ is the sufficient statistic, $a(\mathbf{z}) = \mathbf{s}'\boldsymbol{\beta}$, and $b(\theta) = \log(p_C(\boldsymbol{\eta})) = \log\left(\prod_{j=1}^{n} C(\eta_{ij})^{-1}\right)$. We shall use both the weighted sum score $v = \sum_j \alpha_j s_j$ and the unweighted sum score $u = \sum_j s_j$. The exponential form makes it straightforward to find other results, such as the cumulant-generating function of the sufficient statistic V (the weighted sum score; Butler, 2007, Eq. 5.3)

$$K_V(\tau) = b(\tau + \theta) - b(\theta), \tag{18}$$

which can be used to find the moments of the distribution of V. For example, the rth moment is

$$\mu_r = \left. \frac{\partial^r \exp[K_V(\tau)]}{\partial \tau^r} \right|_{\tau=0}. \tag{19}$$

In addition, the Fisher information of θ and the item information functions (IIFs) are

$$I(\theta) = b''(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} \alpha_j^2 \text{Var}(S_j|\theta). \tag{20}$$

Figure 3 shows IIFs for three items with time limit $d = 1$ and different item parameters. The item information for the speed–accuracy models can be larger than that for IRT models (see Maris & van der Maas, 2012, Fig. 6), but this also depends on the exact scoring rule and the item time limit d_j.

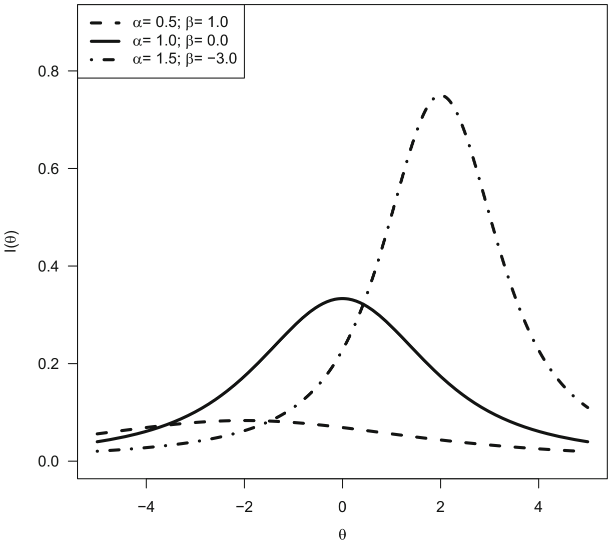

FIGURE 3.
Item information functions for three items.

Discrete time

If we let T_{ij} be discrete (e.g., seconds) or a discretization of time, the derivations are slightly different. Most notably, the normalizing factor $C(\eta_{ij})$ is then found to be

$$C(\eta_{ij}) = \sum_{k=0}^{1}\sum_{t=0}^{d_j} \exp\left[w_{jk}(d_j - t)\eta_{ij}\right]. \tag{21}$$

All previous derivations can be performed in a similar fashion with this normalizing factor, but this is not pursued here. We do note, however, that in this case conditional estimation of item parameters can be pursued if we let $\eta_{ij} = \theta_i + \beta_j$ (i.e., 1P-SARM), because the sum $\sum_{i=1}^{N} s_{ij}$ is a sufficient statistic for item parameter β_j.

Finally, we note that the treatment of time as either continuous or discrete does not have an impact on the probability of a correct response or IRF. Since, for discrete time, we can find the following IRF

$$\begin{aligned} p(X_{ij} = 1|\theta) &= \frac{1}{C(\eta_{ij})}\sum_{t=0}^{d_j} \exp\left[w_{j1}(d_j - t)\eta_{ij}\right] \\ &= \frac{\sum_{t=0}^{d_j} \exp\left[w_{j1}(d_j - t)\eta_{ij}\right]}{\sum_{t=0}^{d_j} \exp\left[w_{j0}(d_j - t)\eta_{ij}\right] + \sum_{t=0}^{d_j} \exp\left[w_{j1}(d_j - t)\eta_{ij}\right]}. \end{aligned} \tag{22}$$

It is now easily found that, for the SRT scoring rule, the above IRF again reduces to $\exp(d_j\eta_{ij})/\left[1 + \exp(d_j\eta_{ij})\right]$.

4. Estimation

4.1. Item Parameters

The focus in this paper is on developing maximum marginal likelihood estimation of the item parameters in order to estimate both the one- and two-parameter speed–accuracy response model. In order to determine the marginal likelihood, we assume a standard normal distribution for θ. The item parameter vector is defined as follows $\boldsymbol{\xi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$. The (complete data) likelihood is

$$L(\boldsymbol{\xi}; \mathbf{S}, \theta) = \prod_{i=1}^{N} f(\mathbf{s}_i|\theta) f(\theta), \tag{23}$$

where $f(\theta)$ is the standard normal density, while the marginal (or observed data) likelihood is

$$L(\boldsymbol{\xi}; \mathbf{S}) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} f(\mathbf{s}_i|\theta) f(\theta) \mathrm{d}\theta. \tag{24}$$

Furthermore, the posterior density of θ satisfies

$$f(\theta|\mathbf{s}_i) = \frac{f(\mathbf{s}_i|\theta) f(\theta)}{\int f(\mathbf{s}_i|\theta) f(\theta) \mathrm{d}\theta}. \tag{25}$$

Maris and van der Maas (2012) develop an EM-type algorithm in which they use quadratic minorization in order to simplify the estimation equations. Their algorithm is convenient because the parameter update has a closed form, but standard errors are not provided. We derive a regular EM algorithm and three sets of standard errors for the item parameter estimates. Let $\hat{\boldsymbol{\xi}}$ be the current estimate. To devise an EM algorithm, we need the so-called Q-function (Dempster, Laird, & Rubin, 1977), which is given by

$$Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}}) = \mathrm{E}_{\theta|\mathbf{S},\hat{\boldsymbol{\xi}}}(\log L(\boldsymbol{\xi}; \mathbf{S}, \theta)) \tag{26}$$

$$= \sum_{i=1}^{N} \int \sum_{j=1}^{n} \left[s_{ij}(\alpha_j \theta + \beta_j) - \log C(\eta_{ij}) - \log(\sqrt{2\pi}\sigma) - \frac{(\theta - \mu)^2}{2\sigma^2} \right] f(\theta|\mathbf{s}_i, \hat{\boldsymbol{\xi}}) \mathrm{d}\theta. \tag{27}$$

The relevant derivatives for item parameters α_j and β_j for item j are easily found using $E(S_{ij}|\theta) = \frac{C'}{C}$ and the chain rule. They are given by

$$\frac{\partial Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}})}{\partial \alpha_j} = \sum_{i=1}^{N} \int \theta \left[s_{ij} - E(S_{ij}|\theta) \right] f(\theta|\mathbf{s}_i, \hat{\boldsymbol{\xi}}) \mathrm{d}\theta, \tag{28}$$

$$\frac{\partial Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}})}{\partial \beta_j} = \sum_{i=1}^{N} \int \left[s_{ij} - E(S_{ij}|\theta) \right] f(\theta|\mathbf{s}_i, \hat{\boldsymbol{\xi}}) \mathrm{d}\theta \tag{29}$$

$$\frac{\partial Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}})}{\partial \alpha_j \partial \alpha_j} = -\sum_{i=1}^{N} \int \theta^2 \mathrm{Var}(S_{ij}|\theta) f(\theta|\mathbf{s}_i, \hat{\boldsymbol{\xi}}) \mathrm{d}\theta, \tag{30}$$

$$\frac{\partial Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}})}{\partial\alpha_j\partial\beta_j} = -\sum_{i=1}^{N}\int \theta\mathrm{Var}(S_{ij}|\theta) f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta, \tag{31}$$

$$\frac{\partial Q(\boldsymbol{\xi}|\hat{\boldsymbol{\xi}})}{\partial\beta_j\partial\beta_j} = -\sum_{i=1}^{N}\int \mathrm{Var}(S_{ij}|\theta) f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta. \tag{32}$$

The maximization step is then performed with the Newton–Raphson procedure. Adaptive Gauss–Hermite quadrature to approximate the integrals is recommended (Naylor & Smith, 1982; Fahrmeir & Tutz, 2001).

Standard errors of the estimated item parameters can be obtained with the observed information matrix, which is (cf. Yuan, Cheng, & Patton, 2014, Eq. 11)

$$\begin{aligned}\mathbf{I}(\hat{\boldsymbol{\xi}}) = &-\sum_{i=1}^{N}\int \frac{\ell_i(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}'} f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta - \sum_{i=1}^{N}\int \frac{\ell_i(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}}\left[\frac{\ell_i(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}}\right]' f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta \\ &+\sum_{i=1}^{N}\left[\int \frac{\ell_i(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}} f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta\right]\left[\int \left(\frac{\ell_i(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}}\right)' f(\theta|\mathbf{s}_i,\hat{\boldsymbol{\xi}})\mathrm{d}\theta\right],\end{aligned} \tag{33}$$

where the derivatives are evaluated at $\hat{\boldsymbol{\xi}}$ and the first term is the Hessian matrix from the M-step with elements defined by Eqs. (30)–(32). The third term is sometimes referred to as the expected information matrix. If we denote this latter matrix by $\mathbf{J}(\hat{\boldsymbol{\xi}})$, we can obtain three sets of standard errors. The first set of SEs is obtained by inversion of the observed information matrix and the second set by inversion of the expected information matrix (Louis, 1982), and the third set is a sandwich estimator:

$$\mathrm{SE}(\hat{\boldsymbol{\xi}})^{(1)} = \sqrt{\mathrm{diag}\left[\mathbf{I}(\hat{\boldsymbol{\xi}})^{-1}\right]}, \tag{34}$$

$$\mathrm{SE}(\hat{\boldsymbol{\xi}})^{(2)} = \sqrt{\mathrm{diag}\left[\mathbf{J}(\hat{\boldsymbol{\xi}})^{-1}\right]}, \tag{35}$$

$$\mathrm{SE}(\hat{\boldsymbol{\xi}})^{(3)} = \sqrt{\mathrm{diag}\left[\mathbf{I}(\hat{\boldsymbol{\xi}})^{-1}\mathbf{J}(\hat{\boldsymbol{\xi}})\mathbf{I}(\hat{\boldsymbol{\xi}})^{-1}\right]}. \tag{36}$$

It is well known that if the model holds, the first two terms in Eq. (33) will be asymptotically equivalent. In this case, the three sets of SEs will converge. Equation (33) can also be used to either speed up the EM algorithm or devise a full Newton–Raphson algorithm. Finally, we note that supplemented EM (Meng & Rubin, 1991) can also be used, but this requires additional iterations.

4.2. Person Parameters

After item parameters have been estimated, person parameters can be estimated. We discuss maximum likelihood (ML) and expected a posteriori (EAP) estimation of θ. Again using a standard exponential family result, the ML estimate $\hat{\theta}$ for weighted sum score w is the solution to

$$b'(\hat{\theta}) = w, \tag{37}$$

and its variance is $I(\hat{\theta})^{-1}$. Note that ML estimates of θ can even be obtained for zero and perfect accuracy as long as responses are given within the time limits (i.e., for perfect accuracy $t_j > 0$ and for zero accuracy $t_j < d_j$ for at least one item). The EAP estimate $\bar{\theta}$ for score pattern $\mathbf{s}$ is

$$E(\theta|\mathbf{s}) = \int \theta f(\theta|\mathbf{s})\mathrm{d}\theta, \tag{38}$$

with variance

$$\mathrm{Var}(\theta|\mathbf{s}) = \int \theta^2 f(\theta|\mathbf{s})\mathrm{d}\theta - \bar{\theta}^2. \tag{39}$$

A model-based reliability estimate for $\bar{\theta}$ can be found by defining the squared correlation between the EAP estimate and true ability as follows (Kim, 2012)

$$\rho^2_{\bar{\theta}\theta} = \frac{\mathrm{Var}[E(\theta|\mathbf{s})]}{E[\mathrm{Var}(\theta|\mathbf{s})] + \mathrm{Var}[E(\theta|\mathbf{s})]}. \tag{40}$$

5. Model Fit

We discuss two approaches to assessing model fit. In the first approach, we use residual analysis to evaluate item fit suggested by Haberman and Sinharay (2013) in the context of regular IRT models. In particular, the approach described by Haberman, Sinharay, and Chon (2013) to estimate residuals from estimated item response functions is used. We apply the approach for the 1P- and 2P-SARM, so that we can compute residuals for the item score function.

For the ISF, we can find the *observed ISF* following Haberman et al. (2013)

$$\tilde{S}_j(\theta) = \frac{N^{-1}\sum_{i=1}^{N} s_{ij} f(\theta|\mathbf{s}_i)}{n(\theta)}, \tag{41}$$

where

$$n(\theta) = N^{-1}\sum_{i=1}^{N} f(\theta|\mathbf{s}_i). \tag{42}$$

We refer to the model-based function as the *fitted ISF* (see Eq. 9), which we denote here as $\hat{S}_j(\theta)$. A generalized residual for the ISF is obtained by letting $\Delta_{S_j}(\theta) = \tilde{S}_j(\theta) - \hat{S}_j(\theta)$ and

$$z\left(\Delta_{S_j}(\theta)\right) = \frac{\Delta_{S_j}(\theta)}{\sigma\left(\Delta_{S_j}(\theta)\right)}, \tag{43}$$

where an estimate of the variance of the residual is (Haberman & Sinharay, 2013, Eq. 46)

$$s^2\left(\Delta_{S_j}(\theta)\right) = [Nn(\theta)]^{-2}\sum_{i=1}^{N}\left[f(\theta|\mathbf{s}_i)\left[s_{ij} - \hat{S}_j(\theta)\right] - \left[\mathbf{h}_j(\theta)\right]' \nabla\ell_i(\hat{\boldsymbol{\xi}})\right]^2, \tag{44}$$

where

$$\mathbf{h}_j(\theta) = \mathbf{J}(\hat{\boldsymbol{\xi}})^{-1}\sum_{i=1}^{N} f(\theta|\mathbf{s}_i)\left[s_{ij} - \hat{S}_j(\theta)\right]\nabla\ell_i(\hat{\boldsymbol{\xi}}). \tag{45}$$

The residual converges in distribution to a standard normal variable (Haberman & Sinharay 2013; Haberman et al., 2013).

In the second approach, we evaluate the fit of the model by inspecting observed and fitted distributions of the sum scores. These sum score distributions can be found for the item responses, item responses times, and item scores, but we focus on that for the item scores. The conditional probability of the unweighted sum score $u = \sum_j s_j$ is given by

$$f(u|\theta, \boldsymbol{\xi}) = \sum_{\mathbf{s}:u} f(\mathbf{S} = \mathbf{s}|\theta, \boldsymbol{\xi}), \tag{46}$$

where the summation $\mathbf{s} : u$ is over all possible score patterns $\mathbf{s}$ that lead to sum score u. In a regular IRT setting, i.e., with discrete scores, the Lord–Wingersky algorithm can be used (Lord & Wingersky, 1984; Kim, 2013), which can be used to determine the distribution of the sum of correct responses under the speed–accuracy model. Finding the fitted distribution of the sum score based on the speed–accuracy model is a bit more involved, because the score is continuous.

In general, exponential family results can be used in combination with the moment-generating function to find the exact density. However, saddlepoint approximations have proven to be convenient and accurate (Butler, 2007; Biehler, Holling, & Doebler, 2015). The saddlepoint density approximation to the density of the weighted sum score V is given by (Butler, 2007, Eq. 5.40)

$$\begin{aligned}\hat{f}(V = v|\theta) &= \frac{1}{\sqrt{2\pi I(\hat{\theta})}} \frac{L(\theta)}{L(\hat{\theta})} &(47)\\ &= \frac{1}{\sqrt{2\pi I(\hat{\theta})}} \frac{\exp((v\theta) - b(\theta))}{\exp((v\hat{\theta}) - b(\hat{\theta}))}, &(48)\end{aligned}$$

where $\hat{\theta}$ is the ML estimate of θ associated with v, $I(\theta)$ is the information function, and $L(\theta)$ is the likelihood. Note that the complicated parts of the likelihood cancel out. The marginal approximation to the density of the weighted sum score V is then given by

$$\hat{f}(V = v) = \int \frac{1}{\sqrt{2\pi I(\hat{\theta})}} \frac{\exp((v\theta) - b(\theta))}{\exp((v\hat{\theta}) - b(\hat{\theta}))} f(\theta) d(\theta), \tag{49}$$

where the integral can be approximated using Gauss–Hermite quadrature. Equation (49) could be simplified further using the Lugannini–Rice approximation, but this is not necessary for our present purposes. More results on saddlepoint approximations in the context of IRT such as cumulative probabilities are described in Biehler et al. (2015). Finally, we note that a generalized residual could be found for the difference between the observed and fitted distribution, but this is beyond our scope.

6. Simulation Study

We performed a relatively small simulation study to evaluate the estimation of model parameters and the item fit statistics. Since the model is an exponential family model and a well-established algorithm is used, an elaborate simulation study on parameter recovery is not deemed necessary. However, the simulation results on the fit statistics can be helpful for practitioners in order to assess how well they perform under different conditions (e.g., numbers of items and sample sizes). All analysis were performed with the computer program SARM (van Rijn & Ali, 2018).[1]

6.1. Generating Data

In order to generate data under the model, one can make use of the pseudo-time variable T_{ij}^* defined by (Maris & van der Maas, 2012, Eq. 27)

$$T_{ij}^* = \begin{cases} T_{ij}, & \text{if} \quad X_{ij} = 1, \\ d_j - T_{ij}, & \text{if} \quad X_{ij} = 0, \end{cases} \tag{50}$$

which is independent of response accuracy. Marsman (2014, Appendix D) showed that it is then straightforward to generate data for the SRT scoring rule. The pseudo-times are based, however, on symmetry properties of the response time density that hold only under the SRT scoring rule, i.e., $f(t|X_{ij} = 1, \eta_{ij}) = f(d_j - t|X_{ij} = 0, \eta_{ij}) = f(t|X_{ij} = 0, -\eta_{ij}) = f(d_j - t|X_{ij} = 1, -\eta_{ij})$. Since our scoring rule is more general (e.g., asymmetrical scoring of incorrect and correct responses is allowed), these properties no longer hold and we need to generalize the procedure. This can be done as follows: First, the response accuracy x_{ij} is generated, which is straightforward. Then, we can sample from the conditional response time density, which is given by

$$f(t|x, \theta_i) = \frac{w_{j0}(1 - x_{ij}) + w_{j1}x_{ij}\eta_{ij} \exp[w_{j0}(1 - x_{ij}) + w_{j1}x_{ij}(d_j - t)\eta_{ij}]}{\exp[w_{j0}(1 - x_{ij}) + w_{j1}x_{ij}d_j\eta_{ij}] - 1}. \tag{51}$$

In our simulations, rejection sampling was used to sample from this density. The R code for simulating data can be found in the additional materials online.

6.2. Simulation Design

The sample sizes were 500 and 3000, and the numbers of items were 25 and 50. The slopes α were (0.5, 0.75, 1, 1.25, 1.5), and the intercepts β were (− 2, − 1, 0, 1, 2). The vector of slopes was crossed with the vector of intercepts, so that each of the 25 items had a unique combination of parameters. In the 50-item condition, each parameter combination occurred twice. Ability θ was sampled from a standard normal distribution. The number of replications was set at 500.

We also estimate the item parameters using response accuracy only, which is equivalent to fitting a 2PLM. Hence, we will compare the estimation of the 2PLM and the 2P-SARM. The 2PLM was estimated using multidimensional IRT software developed by Haberman (2013). The software employs a Newton–Raphson algorithm to perform marginal maximum likelihood estimation. We developed our own software to estimate the 2P-SARM (van Rijn & Ali, 2017). It is emphasized that the accuracy data for both models are the same.

To evaluate the estimation, we compute the bias and root-mean-squared error (RMSE) of the estimated parameters across replications. In addition, we compare the estimated standard errors with empirical standard errors, which are defined by the standard deviation of estimated parameters across replications. Although our software computes all three sets of standard errors (see Eqs. 34–36), we only evaluated the first set based on the observed information matrix. Furthermore, we compute the bias and RMSE of the estimated person parameters. Finally, we compared the quantiles of the generalized residuals of the ISFs to their theoretical quantities (i.e., those of a standard normal distribution).

[1] The executable and software manual are freely available for noncommercial use at *sarm@ets.org*.

TABLE 1.
Mean bias and RMSE for estimated item parameters under 2PLM and 2P-SARM ($N = 500$, $n = 25$).

α	β	2PLM				2P-SARM			
		Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Bias $\hat{\beta}$	RMSE $\hat{\beta}$	Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Bias $\hat{\beta}$	RMSE $\hat{\beta}$
0.5	− 2	0.000	0.168	− 0.013	0.157	− 0.003	0.112	− 0.004	0.114
0.75	− 2	0.007	0.166	− 0.008	0.160	0.000	0.125	0.000	0.115
1	− 2	0.024	0.186	− 0.020	0.183	0.002	0.127	0.000	0.130
1.25	− 2	− 0.004	0.201	− 0.001	0.186	− 0.008	0.144	0.001	0.135
1.5	− 2	0.025	0.228	− 0.023	0.191	0.006	0.151	− 0.007	0.136
0.5	− 1	− 0.001	0.122	− 0.003	0.104	− 0.003	0.096	− 0.004	0.091
0.75	− 1	0.004	0.137	− 0.005	0.114	0.000	0.107	− 0.004	0.096
1	− 1	0.014	0.167	− 0.003	0.129	0.001	0.119	0.001	0.108
1.25	− 1	0.020	0.184	− 0.009	0.133	− 0.004	0.134	0.002	0.109
1.5	− 1	0.018	0.209	− 0.010	0.147	− 0.003	0.134	0.000	0.119
0.5	0	0.004	0.110	0.006	0.095	0.000	0.090	0.004	0.080
0.75	0	0.011	0.124	− 0.001	0.101	0.003	0.099	0.002	0.086
1	0	0.005	0.139	0.011	0.108	0.000	0.108	0.009	0.093
1.25	0	0.009	0.163	− 0.002	0.114	0.002	0.123	− 0.003	0.102
1.5	0	0.011	0.198	0.002	0.124	− 0.003	0.146	0.000	0.110
0.5	1	0.010	0.127	0.003	0.104	− 0.001	0.096	− 0.004	0.083
0.75	1	0.003	0.141	0.016	0.116	0.004	0.107	0.011	0.096
1	1	0.014	0.159	0.019	0.129	0.004	0.122	0.012	0.109
1.25	1	0.023	0.170	0.008	0.133	0.016	0.127	0.003	0.110
1.5	1	0.012	0.197	0.004	0.149	0.000	0.139	0.004	0.119
0.5	2	− 0.007	0.148	0.001	0.149	− 0.004	0.111	− 0.004	0.114
0.75	2	0.009	0.168	0.021	0.168	0.002	0.119	0.009	0.116
1	2	0.011	0.172	0.015	0.166	− 0.006	0.121	0.003	0.119
1.25	2	0.027	0.207	0.042	0.203	− 0.001	0.131	0.014	0.132
1.5	2	0.022	0.219	0.029	0.200	0.004	0.151	0.012	0.139

6.3. Simulation Results

6.3.1. Item Parameter Estimation Since the results with respect to item parameter estimation for all four conditions were highly similar, we only discuss the outcomes for the 500-25 condition. The results for the other three conditions can be found in the additional materials.

Table 1 shows the results of the simulations with respect to mean bias and RMSE of the estimated item parameters using the 2PLM and 2P-SARM. It is quite clear that the estimated parameters are better for the 2P-SARM than for the 2PLM: The estimates generally show less bias and smaller RMSE. More specifically, the estimates of the intercept parameter β show more bias in the 2PLM than in the 2P-SARM, especially for more extreme values (e.g., − 2 and 2). For both models, it holds that if the slope parameter α increases, then the RMSE increases for both the slope and the intercept parameter estimates.

Table 2 displays the mean estimated SEs and the empirical standard errors of the item parameter estimates using the 2PLM and 2P-SARM. The empirical standard errors are simply the standard deviation of the estimated item parameters over replications. Although there are slight differences in some cases, they are generally quite close for both the 2PLM and 2P-SARM parameter estimates. In addition, the standard errors for the 2P-SARM are considerable smaller than those for the 2PLM.

TABLE 2.
Mean SEs and empirical SEs (SDs) of item parameter estimates under 2PLM and 2P-SARM ($N = 500$, $n = 25$).

α	β	2PLM				2P-SARM			
		$\bar{\text{SE}}(\hat{\alpha})$	$\text{SD}(\hat{\alpha})$	$\bar{\text{SE}}(\hat{\beta})$	$\text{SD}(\hat{\beta})$	$\bar{\text{SE}}(\hat{\alpha})$	$\text{SD}(\hat{\alpha})$	$\bar{\text{SE}}(\hat{\beta})$	$\text{SD}(\hat{\beta})$
0.5	−2	0.159	0.168	0.150	0.156	0.117	0.112	0.112	0.114
0.75	−2	0.166	0.167	0.159	0.160	0.121	0.125	0.118	0.115
1	−2	0.180	0.184	0.173	0.182	0.127	0.128	0.124	0.130
1.25	−2	0.195	0.201	0.183	0.186	0.135	0.144	0.132	0.135
1.5	−2	0.222	0.227	0.200	0.190	0.146	0.151	0.140	0.136
0.5	−1	0.122	0.122	0.108	0.104	0.098	0.096	0.090	0.091
0.75	−1	0.134	0.137	0.115	0.114	0.105	0.107	0.096	0.096
1	−1	0.151	0.166	0.124	0.129	0.115	0.119	0.103	0.108
1.25	−1	0.172	0.183	0.135	0.133	0.126	0.134	0.111	0.109
1.5	−1	0.197	0.209	0.146	0.147	0.138	0.134	0.119	0.119
0.5	0	0.111	0.110	0.095	0.094	0.091	0.090	0.083	0.080
0.75	0	0.125	0.123	0.101	0.101	0.100	0.099	0.088	0.086
1	0	0.142	0.139	0.109	0.108	0.111	0.109	0.095	0.093
1.25	0	0.163	0.163	0.118	0.114	0.123	0.123	0.103	0.102
1.5	0	0.189	0.198	0.127	0.124	0.136	0.146	0.112	0.111
0.5	1	0.123	0.127	0.108	0.104	0.098	0.096	0.090	0.083
0.75	1	0.134	0.141	0.115	0.115	0.105	0.107	0.096	0.095
1	1	0.151	0.159	0.125	0.128	0.115	0.122	0.103	0.108
1.25	1	0.172	0.169	0.135	0.133	0.127	0.126	0.111	0.110
1.5	1	0.196	0.197	0.145	0.149	0.139	0.139	0.119	0.119
0.5	2	0.158	0.148	0.149	0.149	0.117	0.111	0.112	0.114
0.75	2	0.167	0.168	0.160	0.167	0.121	0.119	0.118	0.116
1	2	0.179	0.172	0.172	0.166	0.127	0.121	0.124	0.120
1.25	2	0.200	0.205	0.188	0.199	0.136	0.131	0.132	0.132
1.5	2	0.222	0.218	0.201	0.198	0.146	0.151	0.140	0.139

TABLE 3.
Mean bias and RMSE of EAP estimates under 2PLM and 2P-SARM.

Sample size	Number of items	2PLM		2P-SARM	
		Bias θ	RMSE θ	Bias θ	RMSE θ
500	25	−0.001	0.444	−0.001	0.376
500	50	0.001	0.334	0.001	0.279
3000	25	0.001	0.439	0.001	0.372
3000	50	0.000	0.327	0.000	0.274

6.3.2. Person Parameter Estimation Table 3 shows the mean bias and RMSE of the EAP estimates of θ under the 2PLM and 2P-SARM for the four simulation conditions. The mean bias is negligible for both models, but, as expected, the RMSEs for the 2P-SARM are generally smaller than those for the 2PLM.

6.3.3. Generalized Residuals of ISFs Figure 4 shows the type I error of the generalized residuals of the ISFs for a nominal level of .05 as a function of θ for the four simulation conditions. It can be seen that the type I errors are approaching the nominal level with increasing sample size, but also that they are larger for more extreme values of θ. Furthermore, they are generally larger in the 50-item condition than in the 25-item condition.

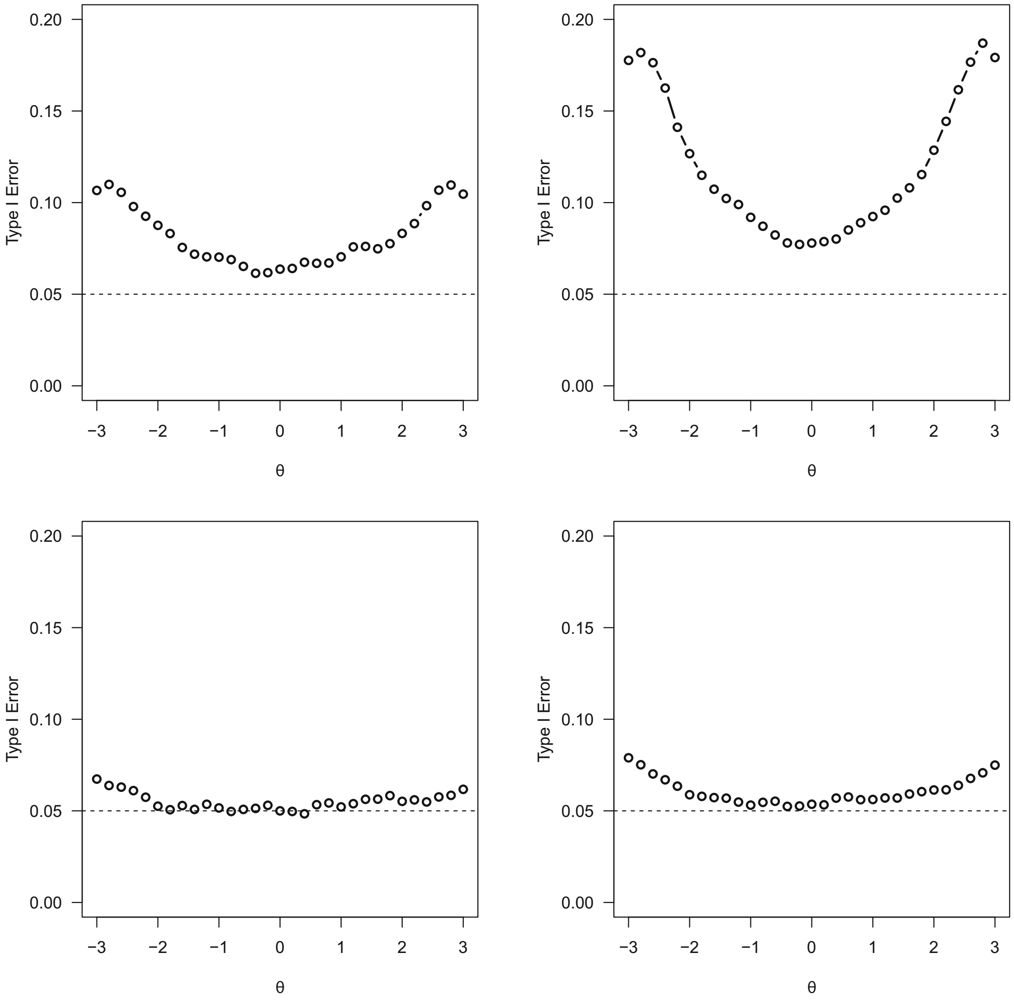

FIGURE 4.
Type I errors of generalized residuals of ISFs for nominal level of .05 (top left: $N = 500$, $n = 25$; top right: $N = 500$, $n = 50$; bottom left: $N = 3000$, $n = 25$; bottom right: $N = 3000$, $n = 50$).

7. Applications to Real Data

7.1. Application 1: Mathematics

We analyzed the responses of 1079 students to 40 multiple-choice and numeric-entry items from an eighth-grade mathematics assessment. The items were dichotomously scored and focused on basic topics in number, measurement, geometry, data analysis and algebra. Because no item-specific time limits were imposed in collecting the data, we initially selected a lenient time limit based on the observed response time distributions for each item: the 99th percentile. The mean and median of those time limits are 205 and 191 s, the standard deviation is 77 s, and the range is 83–425 s. We rescaled the time limits and the response time data so that all limits d_j are equal to one and all items have equal weight. Table 4 shows descriptive statistics of the accuracy and speed of the responses as well as the resulting score using the SRT scoring rule. The correlations

TABLE 4.
Descriptive statistics for different scores of mathematics items.

Description	Variable	Range	Min	Max	Mean	SD	Cronbach's α
Accuracy	X	0–40	3.00	40.00	23.35	8.03	0.89
Speed	T	0–40	2.77	21.61	11.85	2.79	0.85
SRT score	S	− 40–40	− 27.22	31.10	4.93	11.85	0.90

TABLE 5.
Relative model fit of different IRT and speed–accuracy response models.

Model (variable)	Parameters	Log-likelihood	AIC	BIC	Reliability (θ)
1PLM (X)	41	− 23,790.4	47,663	47,867	0.883
2PLM (X)	80	− 23,448.6	47,057	47,456	0.891
1P-SARM (S)	41	− 18,892.6	37,867	38,071	0.927
2P-SARM (S)	80	− 18,247.9	36,656	37,055	0.933

between the observed sum scores are -0.09 for X and T, 0.99 for X and S, and -0.10 for T and S.

Table 5 shows relative model fit statistics for different IRT and speed–accuracy response models. The table shows the log-likelihood, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC). As a reference, we fitted the one-parameter logistic model (1PLM) and 2PLM to the response accuracies using marginal maximum likelihood estimation as implemented in IRT software developed by Haberman (2013). It can be seen that the 2PLM has better fit statistics than the 1PLM. The same holds for the 2P-SARM, although the relative difference is substantially larger than the difference for the IRT models. In addition, the model-based reliabilities of the speed–accuracy models are considerably larger than those of the IRT models.

Figures 5 and 6 show fit plots of the item score functions for a selection of items for the 1P-SARM and 2P-SARM, respectively. The solid lines indicate the fitted ISF, and the dashed lines indicate a 95% confidence interval for the observed ISF. For the 1P-SARM (Fig. 5), there is considerable misfit, but most of it seems related to the slope. It can be seen in Fig. 6 that the fit improves considerably, which is in line with the relative model fit statistics. Most of the remaining misfit for the 2P-SARM seems to occur for very low values of θ. Note that type I errors in this range were larger than the nominal level in the simulations (see Fig. 4).

Figure 7 shows the observed density and the saddlepoint approximation to the fitted density of the sum score for the 1P-SARM (left) and the observed density and the saddlepoint approximation to the fitted density of the weighted sum score for the 2P-SARM. It can be observed that both models predict the observed density quite well.

Finally, we fitted the 2P-SARM using a number of different item-level time limits based on the observed response time distributions. We used the following percentiles: 99th, 95th, 90th, 75th and 50th. Again, the time limits and the response time data were rescaled, so that all limits d_j are equal to one and all items have equal weight.

Table 6 shows the results of fitting the 2P-SARM with different time limits. It can be seen that with shorter time limits the reliabilities decrease. This is not at all surprising because the number of zero scores increases with decreasing time limits. That is, if a response is outside the time limit, the scoring rule does not discriminate between incorrect and correct responses. This results in information loss, and hence, the reliabilities drop.

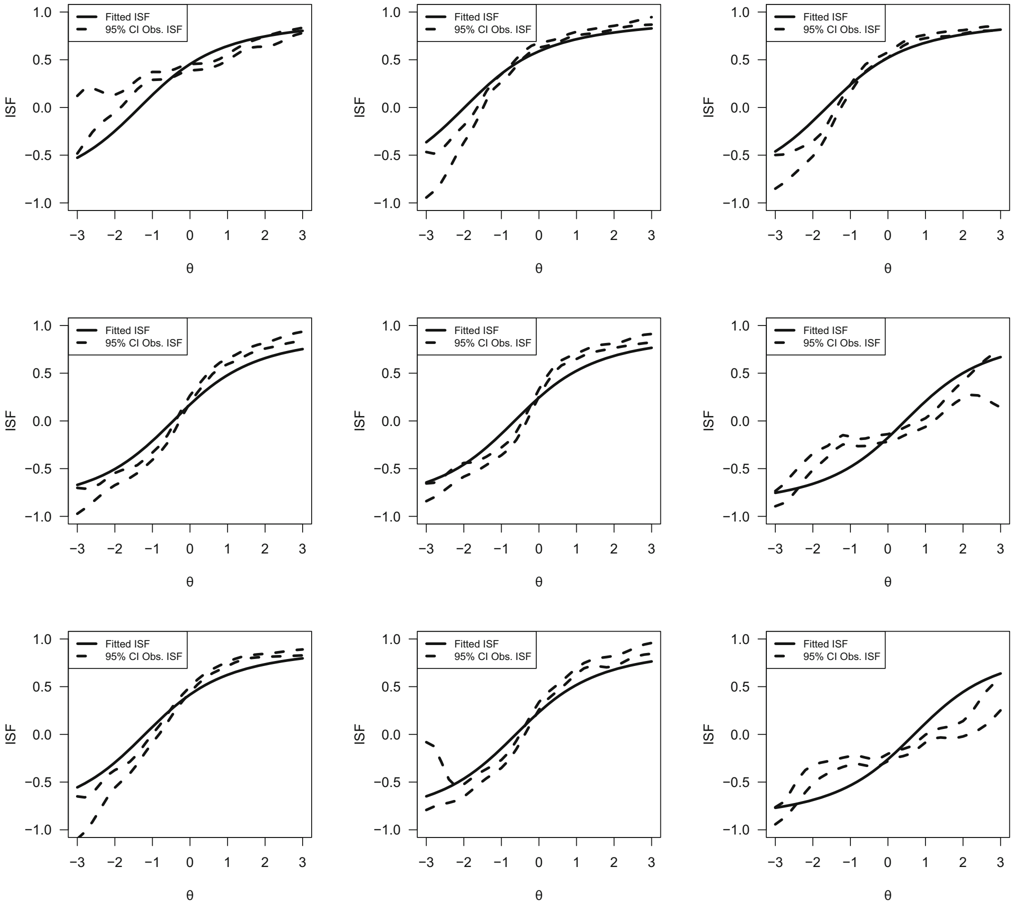

FIGURE 5.
Item fit plots for ISF for 9 items of math data showing misfit with 1P-SARM.

7.2. *Application 2: Chess*

As a second application, we discuss data from the two main subtests of the Amsterdam Chess Test (van der Maas & Wagenmakers, 2005) in which item-specific time limits were used.[2] These tests are so-called choose-a-move tests in which test takers choose the move which they think is best from a given chess position. The data consist of responses to 80 items (40 from test A and 40 from test B) from 259 chess players, who participated in the 1998 open Dutch championship. Both the test and the data are publicly available.[3] A subset of test takers had an official Elo chess rating, which was used as an external criterion to evaluate validity. The item-specific time limit was set at 30 s for both tests.

Table 7 shows the relative model fit statistics for the chess data. The sample size for Choose-a-move A was 256 and the sample size for Choose-a-move B was 251. For both tests, the AIC and BIC are better for the 2P-SARM than for the 1P-SARM. However, it seems that adding response time to the scoring only works for the Choose-a-move B test: The correlation of $\bar{\theta}$ (EAP) with the

[2] This data set was suggested by an anonymous reviewer.

[3] See http://hvandermaas.socsci.uva.nl/Homepage_Han_van_der_Maas/Chess_Psychology.html.

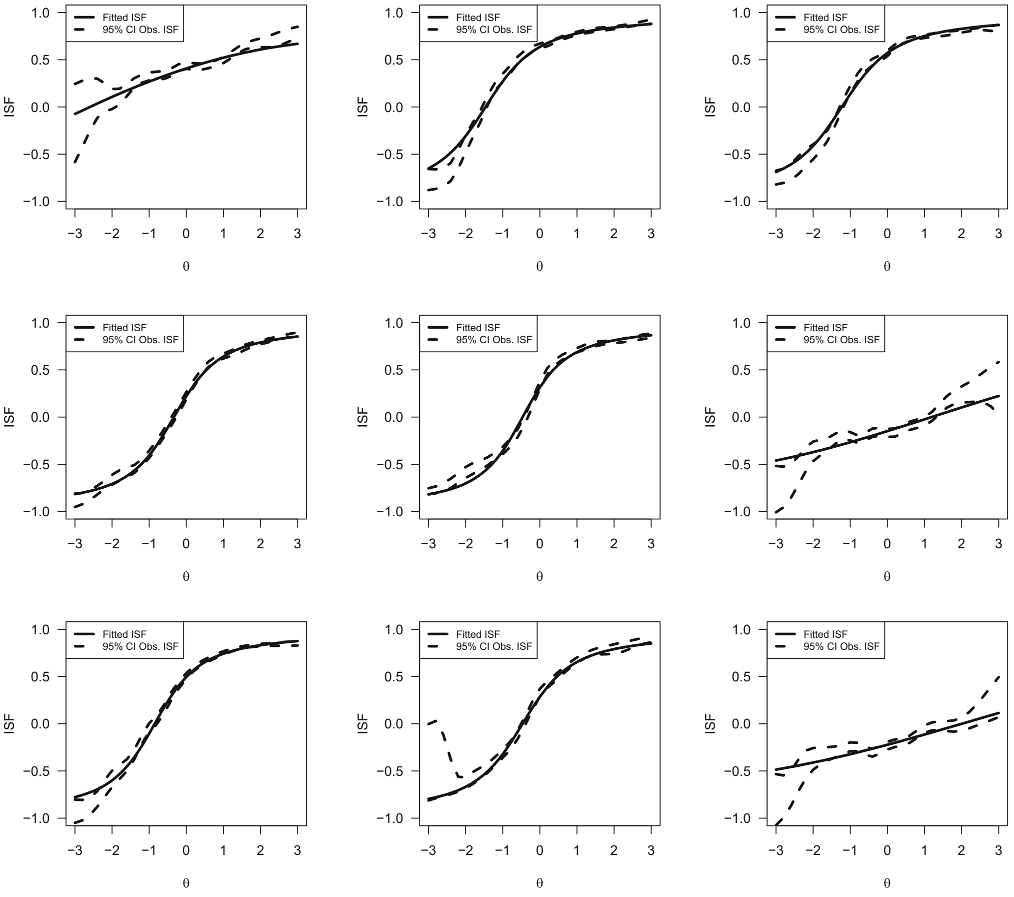

FIGURE 6.
Item fit plots for ISF for 9 items of math data showing improved fit with 2P-SARM.

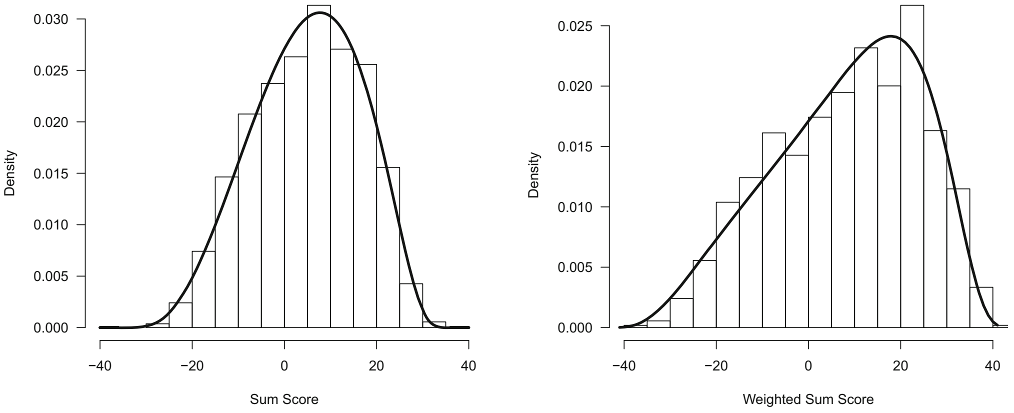

FIGURE 7.
Observed and fitted distributions for sum score with 1P-SARM (left) and weighted sum score with 2P-SARM (right).

TABLE 6.
Relative model fit of 2P-SARM under different item time limits.

RT percentile	Mean time limit (SD)	Log-likelihood	AIC	BIC	Reliability (θ)
99	202 (77)	– 18,247.9	36,656	37,055	0.933
95	136 (55)	– 22,740.2	45,640	46,039	0.897
90	110 (47)	– 25,041.6	50,243	50,642	0.858
75	77 (36)	– 27,981.6	56,123	56,522	0.718
50	53 (27)	– 29,537.8	59,236	59,634	0.225

TABLE 7.
Relative model fit of different speed–accuracy response models for chess data.

Test	Model	Parameters	Log-likelihood	AIC	BIC	Reliability (θ)	Correlation of $\bar{\theta}$ with Elo
Choose-a-move A	1PLM	41	– 4366.2	8814	8960	.882	.765
	2PLM	80	– 4231.4	8623	8906	.903	.773
	1P-SARM	41	– 4069.8	8221	8336	.850	.755
	2P-SARM	80	– 3821.8	7804	8087	.890	.746
Choose-a-move B	1PLM	41	– 4316.2	8714	8759	.905	.817
	2PLM	80	– 4157.7	8475	8859	.924	.822
	1P-SARM	41	– 3000.0	6082	6226	.915	.816
	2P-SARM	80	– 2605.5	5371	5653	.933	.842

Elo rating is lower for the 1P- and 2P-SARM than for the 1PLM and 2PLM for Choose-a-move A, whereas this correlation is .842 for the 2P-SARM for Choose-a-move B and 0.822 for the 2PLM.

8. Discussion

We presented an extension to the approach of developing psychometric models using scoring rules that include both speed and accuracy of item responses introduced by Maris and van der Maas (2012). We derived a generalized speed–accuracy response model and algorithms to perform maximum marginal likelihood estimation of item parameters. Our simulations indicated that parameter recovery was adequate and standard errors were accurate. In addition, we developed two approaches to evaluate absolute model fit. The simulations showed that the type I error rates for the generalized residuals for ISFs converged to their nominal level with increasing sample size.

In two applications, the modeling approach was demonstrated: one without item-specific time limits and one with time limits. In both applications, the 2P-SARM showed better fit than the 1P-SARM, which leads us to an affirmative answer to our motivating question: Do item scores differ in discriminating power? In addition, the reliabilities of the ability estimates (EAPs) for the speed–accuracy models were considerably larger than those for IRT models, which indicate that including RTs can be used to obtain higher measurement precision. In terms of validity, it was found in the chess application that the correlation of the ability estimate of the 2P-SARM with an external measure (the Elo rating) can exceed the correlation for the 2PLM (although it can also be lower).

Recently, van Rijn and Ali (2017) compared the present modeling approach with two other modeling approaches for item responses and response times in the context of adaptive testing. They compared with the hierarchical modeling approach developed by van der Linden (2007) and with the diffusion modeling approach described by Tuerlinckx, Molenaar, and van der Maas (2016). In application to data from a math test and spelling test, they found an improvement in model precision when response time was included through the scoring rule framework, but not through the hierarchical framework. However, a limitation was that the data were not collected under item-level time limits.

As noted, Goldhammer (2015) advocated the use of item-level time limits. Although he argued that test takers should be aware of this, we found in the mathematics application that meaningful results can be obtained if no time limits are used in the administration of the test. However, it is wise in this case to explore several time limits based on the observed response time distributions (e.g., De Boeck, Chen, & Davison, 2017). We emphasize that the modeling approach would have to be re-evaluated if scoring rules including both accuracy and speed are to be used in practice, because response behavior is likely to be affected when test takers are informed about the particular scoring rule that is used.

A potential area for further investigation is model fit. For example, the generalized residual approach could also be developed for the ITF of Eq. 7. In addition, the model-based density of the sum of the item response times can be determined to evaluate fit. This is less straightforward than the model-based density of the sum of item response accuracies for which the Lord–Wingersky algorithm can be used, and the model-based density of the sum of item scores for which we developed a saddlepoint approximation. Finally, as noted, generalized residuals could be developed for the difference between observed and fitted distributions.

The presented modeling approach seems to be most promising for tests that focus on fluency (e.g., basic arithmetic or basic grammar tasks) in low-stake conditions (e.g., formative assessment). For such tests, it can be problematic to estimate ability with a reasonable precision, because these tests typically have fewer items than high-stakes tests and can be too easy for a substantial proportion of test takers. However, an interesting topic for future research is to combine different model types for different sections of a test. For example, a between-item two-dimensional model can be developed where the first dimension is a 2PLM and the second dimension is a 2P-SARM. This would give the possibility to jointly model and study lower-order and higher-order skills.

Acknowledgments

Funding was provided by Educational Testing Service. The authors would like to thank Rebecca Zwick, Yi-Hsuan Lee, Fred Robin, and three anonymous reviewers for their comments on earlier drafts of the paper.

References

Andersen, E. B. (1973). Conditional inference and multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44. https://doi.org/10.1111/j.2044-8317.1973.tb00504.x.

Biehler, M., Holling, H., & Doebler, P. (2015). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika*, *80*, 665–688. https://doi.org/10.1007/s11336-014-9405-1.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge: Cambridge University Press.

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*, 225–237. https://doi.org/10.1111/bmsp.12094.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd edn). Berlin: Springer. https://doi.org/10.1007/978-1-4757-3454-6.

Goldhammer, F. (2015). Measuring, ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement*, *13*, 133–164. https://doi.org/10.1080/15366367.2015.1100020.

Haberman, S. J. (2006). *Joint and conditional estimation for implicit models for tests with polytomous item scores* (ETS Research Report RR-06-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02009.x.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (ETS research report RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x.

Haberman, S. J. (2016). Exponential family distributions relevant to IRT. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: Statistical tools* (pp. 47–70). Boca Raton, FL: CRC Press.

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, *108*, 1435–1444. https://doi.org/10.1080/01621459.2013.835660.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417–440. https://doi.org/10.1007/s11336-012-9305-1.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, *74*, 419–442. https://doi.org/10.1007/S11336-009-9111-6.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*, 153–162. https://doi.org/10.1007/s11336-011-9238-0.

Kim, S. (2013). Generalization of the Lord–Wingersky algorithm to computing the distributions of summed test scores based on real-number item scores. *Journal of Educational Measurement*, *50*, 381–389.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *3*, 359–379.

Lord, F. M. (1975). Formula scoring and number right scoring. *Journal of Educational Measurement*, *12*, 7–11. https://doi.org/10.1111/j.1745-3984.1975.tb01003.x.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of “IRT” true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, *8*, 453–461.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*, 226–233. https://doi.org/10.2307/2345828.

Luce, R. D. (1986). *Response times*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195070019.001.0001.

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633. https://doi.org/10.1007/s11336-012-9288-y.

Marsman, M. (2014). *Plausible values in statistical inference*. Doctoral dissertation, University of Twente, Enschede.

Meng, X. L., & Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for efficient computation of posterior distributions. *Applied Statistics*, *31*, 214–225. https://doi.org/10.2307/2347995.

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*, 31–47. https://doi.org/10.1007/s11336-011-9231-7.

Ranger, J., Kuhn, J. T., & Gaviria, J. L. (2015). A race model for responses and response times in tests. *Psychometrika*, *80*, 791–810. https://doi.org/10.1007/s11336-014-9427-8.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogike Institut.

Roskam, E. E. (1997). Models for speed and time-limit tests. In R. K. Hambleton & W. J. van der Linden (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.

Spearman, C. (1927). *The abilities of men*. London: MacMillan.

Thurstone, L. L. (1919). A scoring method for mental tests. *Psychological Bulletin*, *16*, 235–240.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, *2*, 249–254.

Tuerlinckx, F., & de Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*, 629–650. https://doi.org/10.1007/s11336-000-0810-3.

Tuerlinckx, F., Molenaar, D., & van der Maas, H. L. J. (2016). Diffusion-based item response modeling. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 283–302). Boca Raton, FL: Chapman & Hall/CRC Press.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20. https://doi.org/10.3102/1076998607302626.

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

van der Maas, H., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*, 29–60.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 317–345. https://doi.org/10.1111/bmsp.12101.

van Rijn, P. W., & Ali, U. S. (2018, in press). *SARM: A computer program for estimating speed-accuracy response models* (ETS Research Report). Princeton, NJ: Educational Testing Service.

van Rijn, P. W., & Rijmen, F. (2015). On the explaining-away phenomenon in multivariate latent variable models. *British Journal of Mathematical and Statistical Psychology*, *68*, 1–22. https://doi.org/10.1111/bmsp.12046.

Yuan, K. H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*, 232–254. https://doi.org/10.1007/S11336-013-9334-4.

Manuscript Received: 15 OCT 2015
Published Online Date: 21 NOV 2017