PSYCHOMETRIKA—VOL. 77, NO. 4, 615–633
OCTOBER 2012
DOI: 10.1007/S11336-012-9288-Y

SPEED-ACCURACY RESPONSE MODELS: SCORING RULES BASED ON RESPONSE TIME AND ACCURACY

GUNTER MARIS

CITO – UNIVERSITY OF AMSTERDAM

HAN VAN DER MAAS

UNIVERSITY OF AMSTERDAM

Starting from an explicit scoring rule for time limit tasks incorporating both response time and accuracy, and a definite trade-off between speed and accuracy, a response model is derived. Since the scoring rule is interpreted as a sufficient statistic, the model belongs to the exponential family. The various marginal and conditional distributions for response accuracy and response time are derived, and it is shown how the model parameters can be estimated. The model for response accuracy is found to be the two-parameter logistic model. It is found that the time limit determines the item discrimination, and this effect is illustrated with the Amsterdam Chess Test II.

Key words: item response theory, response times, two-parameter logistic model, scoring rule.

1. Introduction

In the recent literature (e.g., van der Linden, 2007; Tuerlinckx & De Boeck, 2005), models for both response accuracy and response time have become increasingly popular, mainly due to the increased availability of computers in classrooms. Models have been developed along different lines. The most prominent being the statistical approach of van der Linden (2007) and the psychological approach of Tuerlinckx and De Boeck (2005), and Van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011). In this paper, we address the issue of modeling response time and accuracy from a *measurement* point of view. Our approach is in line with the classical derivation of the Rasch (1960) model, which starts from an explicit scoring rule (i.e., the number correct score) and leads to the derivation of a model that conforms to this scoring rule. This automatically leads to two research questions. First, we need to formulate a scoring rule incorporating both accuracy and response time. Second, we need to derive a statistical model that conforms to the scoring rule. Possible scoring rules have been proposed by Van der Maas and Wagenmakers (2005) and Dennis and Evans (1996), for instance.

Van der Maas and Wagenmakers (2005) propose an explicit scoring rule based on both response accuracy *and* response time for the measurement of chess expertise. This scoring rule, called the *correct item summed residual time* (CISRT) scoring rule, assigns the following credit for an item response X_{pi}, where X_{pi} equals one for correct, and zero for incorrect responses, after T_{pi} time units when the time limit for responding is d:

$$X_{pi}(d - T_{pi}), \tag{1}$$

where the subscript i is used to index the items, and p to index the persons. The total score is simply the sum of the item scores. In other words, for a correct response, the student earns

Requests for reprints should be sent to Gunter Maris, Cito – University of Amsterdam, Arnhem, The Netherlands. E-mail: Gunter.Maris@cito.nl

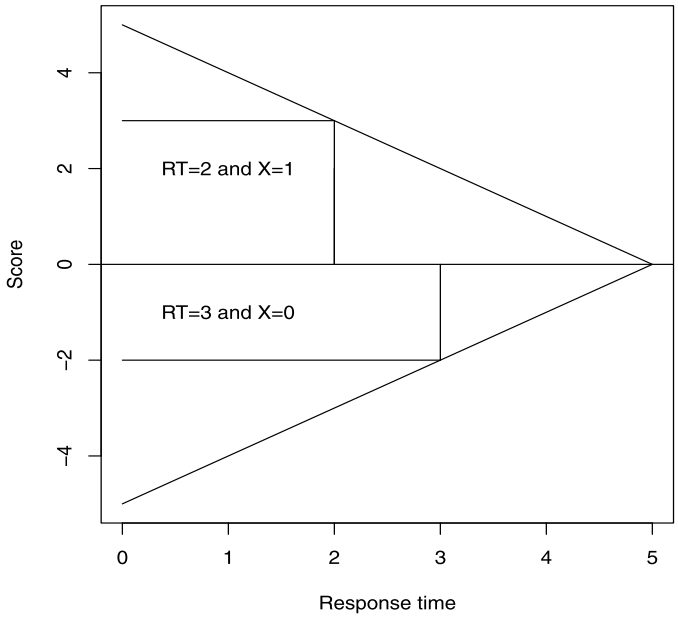

FIGURE 1.
SRT scoring rule for a time limit of 5.

the remaining time as credit, such that fast correct responses earn more credit than slow correct responses; and no credit is earned for incorrect responses.

An advantage of this scoring rule is that subjects know how speed and accuracy are weighted. This gives subjects the opportunity to choose an optimal speed accuracy trade-off (Wickelgren, 1977). Other scoring rules for accuracy and response time, such as the Ratio Index or the Log A Index (Dennis & Evans, 1996), only apply to test scores and cannot be applied to items.

One drawback of the use of the CISRT as a scoring rule is that it may promote guessing, which generally is undesirable. We consider 4 multiple choice questions with 4 alternatives to illustrate this point. The items all have difficulty relative to the ability of a student such that his expected score using the CISRT equals d. If this student would guess quickly, for the sake of the argument, let the guessing response be instantaneous ($T_{pi} = 0$); his expected score would also be d. It is clear that for students with a low ability level the expected score from guessing will be higher than the expected score from giving an honest answer. So, even if a model which conforms to the CISRT scoring rule holds, once the scoring rule becomes known people may well adjust their response behavior to maximize their expected score (and, hence, the model no longer describes the response behavior).

The alternative scoring rule we consider here is one in which fast incorrect responses are *penalized*. Specifically, we consider the following scoring rule:

$$(2X_{pi} - 1)(d - T_{pi}), \tag{2}$$

where for a correct response you earn the residual time as score, and for an incorrect response you lose the residual time as score. We refer to this scoring rule as the *Signed Residual Time* (SRT) scoring rule. Figure 1 gives a graphical illustration of the SRT scoring rule. As before, the total score is simply the sum of the item scores. If we reconsider our example, and assume that the items have difficulty relative to the ability of the student such that his expected score equals zero, we obtain that if this person guesses instantaneously the expected score would be equal to $-2d$. That is, for such a person it is not beneficial to guess.

The CISRT and the SRT scoring rules clearly differ in how speed and accuracy are weighted. For the CISRT, there are many persons (i.e., everybody with ability below the item difficulty) for whom pure fast guessing is the optimal strategy. That is, speed is favored over accuracy. For the SRT, this is not the case; persons need to be both fast and accurate to obtain a high score and, thereby, a high estimated ability.

This paper is organized as follows. In Section 2, we derive an IRT model for which the SRT scoring rule is the sufficient statistic for ability. In order to better understand the operating characteristics of this new IRT model, we derive its various marginal and conditional distributions. In Section 3, we deal with the problem of parameter estimation. In Section 4, we consider some ways in which this scoring rule can be generalized. In Section 5, an illustrative application is presented. The focus of the application is on checking qualitative predictions implied by the IRT model. The paper ends with some concluding remarks in Section 6.

2. Derivation and Model Properties

Many item response theory (IRT) models can be derived from the *scoring rule* they imply, together with some auxiliary assumptions, such as conditional independence of the item responses. This approach to psychometric modeling originated with Georg Rasch in his classic derivation of the Rasch (1960) model. Some of the IRT models that can be derived in this way are the one-parameter logistic model (OPLM) (Verhelst & Glas, 1995), the nominal response model (NRM) (Bock, 1972), and its various special cases. The main idea is that the score of a student is considered to be the sufficient statistic for his ability. That is, we assume that the response $\mathbf{X}$ is independent of ability θ, given the scoring rule $S(\mathbf{X})$. By assuming that the score is a sufficient statistic, we immediately can formulate a measurement model that belongs to the exponential family of distributions:

$$P(\mathbf{X}=\mathbf{x}|\theta)=\frac{a(\mathbf{x})\exp(S(\mathbf{x})\theta)}{b(\theta)}, \tag{3}$$

where

$$b(\theta)=\int_R a(\mathbf{x})\exp\bigl(S(\mathbf{x})\theta\bigr)\,d\mathbf{x}. \tag{4}$$

All of the models developed in this fashion have in common that the scoring rule they employ only involves *response accuracy*. The same approach generalizes directly to scoring rules that depend both on response accuracy and response time.

In order to specify the model for the SRT scoring rule, we assume that responses to different items from the same person are independent:

$$\perp\!\!\!\perp_i (X_{pi}, T_{pi})|\theta_p \tag{5}$$

which is the traditional conditional independence assumption. That is, the responses to items i and j, say, which consist both of accuracy and response time, are independent given ability θ_p. Observe that this does not imply that the response accuracy and response time for the same item are independent given ability. We furthermore assume that the SRT score of a person is the sufficient statistic for his ability (denoted by θ_p):

$$\mathbf{X}_p, \mathbf{T}_p \perp\!\!\!\perp \theta_p | \sum_i (2X_{pi}-1)(d-T_{pi}). \tag{6}$$

Like persons, items differ from one another. Consider two items, one for which people tend to get a large positive SRT item score and the other for which people tend to get a large negative SRT

item score. Our model should account for such differences between items. One way to achieve this is by assuming that the total score of an item, i.e.,

$$\sum_p (2X_{pi} - 1)(d - T_{pi}) \tag{7}$$

is the sufficient statistic for an item *difficulty* parameter δ_i:

$$(\mathbf{X}_i, \mathbf{T}_i) \perp\!\!\!\perp \delta_i \,|\, \sum_p (2X_{pi} - 1)(d - T_{pi}). \tag{8}$$

Taken together, these assumptions imply that we may express the joint distribution of response time and response accuracy as follows:

$$f(\mathbf{x}_p, \mathbf{t}_p|\theta_p) = \prod_i f(x_{pi}, t_{pi}|\theta_p), \tag{9}$$

where

$$f(x_{pi}, t_{pi}|\theta_p) = \frac{1}{C_{pi}} \exp\big((2x_{pi} - 1)(d - t_{pi})(\theta_p - \delta_i)\big) \tag{10}$$

with C_{pi} representing a normalization factor to make the probability density function integrate to one, namely[1]

$$\begin{aligned} C_{pi} &= \sum_{j=0}^{1} \int_0^d \exp\big((2j - 1)(d - s)(\theta_p - \delta_i)\big)\, ds \\ &= \frac{\exp(d(\theta - \delta_i)) - 1}{\theta_p - \delta_i} + \frac{1 - \exp(-d(\theta_p - \delta_i))}{\theta_p - \delta_i}. \end{aligned} \tag{11}$$

In the following subsections, we derive the relevant marginal and conditional distributions corresponding to the distribution of a single observation X_i, T_i to gain further insight into the operating characteristics of this new measurement model. In order to simplify the derivations, we suppress the item difficulty parameter δ_i, and drop the subscripts p and i. All results in the rest of this section pertain to one response (accuracy and response time) from one person to one item. For all distributions, it holds that replacing θ with $\theta_p - \delta_i$ gives the corresponding distribution for the general case.

2.1. Item Response Function

The first thing we consider is the item response function (IRF), which gives the marginal probability with which a person with ability θ solves an item correctly. Direct integration of the joint distribution of response accuracy and response time with respect to response time yields the following expression for the IRF:

$$P(X = 1|\theta) = \frac{\frac{\exp(d\theta)-1}{\theta}}{\frac{\exp(d\theta)-1}{\theta} + \frac{1-\exp(-d\theta)}{\theta}} = \frac{\exp(d\theta)}{1 + \exp(d\theta)} \tag{12}$$

in which we recognize the two-parameter logistic model (2PL) (Birnbaum, 1968) with the item discrimination equal to the time limit d. Hence, we see that the model in this section has the 2PL (or the Rasch model if the time limit is the same for different items) as the marginal distribution for the response quality X, and that a decrease in time limit d corresponds to a decrease in item discrimination for the IRF. Conversely, as the time limit tends to infinity, the IRF tends to that

[1] L'Hôpital's rule needs to be used to evaluate limits for θ_p tending to δ_i.

of a Guttman item. Systematic variation of the time limit, for the same item, might provide a powerful test for this model based on (the IRF for) response quality.

The SRT scoring rule implies an interpretation of the discrimination parameter in the 2PL model as a time limit. This interpretation is similar to the interpretation derived from a diffusion model by Tuerlinckx and De Boeck (2005). Tuerlinckx and De Boeck (2005) show that the discrimination parameter corresponds to boundary separation in a drift diffusion model. Of course, both models are developed for very different paradigms. The SRT model is meant for tasks which have an explicit time limit, whereas the drift diffusion model does not impose any definite upper limit on the response time. On the other hand, *soft* manipulation of the time limit is a popular way to induce differences in the speed-accuracy trade-off in empirical research for which the drift diffusion model is deemed appropriate.

2.2. *Response Time Distribution*

From the joint distribution, we readily obtain the following marginal response time distribution by summing the joint density of response time and accuracy for correct and incorrect responses:

$$f(t|\theta) = \theta \frac{\exp((d-t)\theta) + \exp(-(d-t)\theta)}{\exp(d\theta) - \exp(-d\theta)}. \tag{13}$$

Application of L'Hôpital's rule shows that response times are uniformly distributed between 0 and d for θ equal to zero. We see that the response time distribution is symmetric in θ. That is, we obtain the same marginal response time distribution for a person with ability equal to θ and for a person with ability equal to $-\theta$. In other words, this means that response time does not tell us whether a person is more or less able than the item is difficult, but informs us about the distance between the ability of a person and the difficulty of the item, which is another property the SRT model shares with the drift diffusion model. Moreover, as the time limit increases without bound we obtain the following:

$$\lim_{d\to\infty} f(t|\theta) = |\theta| \exp\bigl(-|\theta|t\bigr). \tag{14}$$

That is, in the limit as d goes to ∞, response time is distributed as a negative exponential random variable with hazard rate equal to $|\theta|$. Put differently, the absolute value of ability $|\theta|$ may be interpreted as the hazard rate of response time when the time limit increases without bound.

Some algebra[2] gives the following expression for the expected response time:

$$\begin{aligned}
\mathcal{E}(T|\theta) &= \int_0^d t\theta \frac{\exp((d-t)\theta) + \exp(-(d-t)\theta)}{\exp(d\theta) - \exp(-d\theta)}\, dt \\
&= \frac{1}{\theta} \frac{\exp(d\theta) + \exp(-d\theta) - 2}{\exp(d\theta) - \exp(-d\theta)} \\
&= \frac{1}{\theta} \frac{\exp(d\theta) - 1}{\exp(d\theta) + 1} \overset{d\to\infty}{\to} \frac{1}{|\theta|}.
\end{aligned} \tag{15}$$

Figure 2 illustrates both the symmetry of (expected) response times with respect to ability, and the general feature that as ability becomes more extreme, response time decreases.

It is instructive to compare the expected response time according to the IRT model derived from the SRT to the expected response time derived from the drift diffusion model by Tuerlinckx

[2]The main result needed is the following indefinite integral $\int xa\exp(ax)\,dx = \frac{ax-1}{a^2}\exp(ax)$. Observe that this, and all other integrals, can be checked via tools for symbolic mathematics, such as Wolfram|Alpha (http://www.wolframalpha.com).

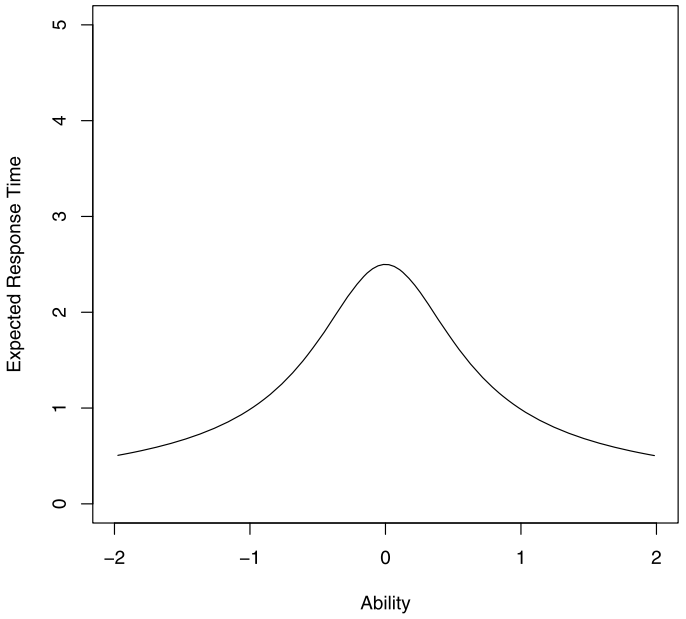

FIGURE 2.
Expected response time as a function of ability for a time limit equal to 5.

and De Boeck (2005). In our notation, the expected response time according to the drift diffusion model can be written as

$$\mathcal{E}(T|\theta) = \frac{a}{2\theta}\frac{\exp(a\theta)-1}{\exp(a\theta)+1}, \tag{16}$$

where a is the boundary separation chosen by a subject.

Similar to what we found for the probability to give the correct response, the expressions for the expected response time are strikingly similar. According to the SRT model, expected response time increases with d; whereas, according to the drift diffusion model, it increases with a. However, according to the SRT model, this increase is not without bound, whereas it is according to the drift diffusion model. The difference is, of course, reasonable as the SRT model presupposes a time limit for responding; whereas the diffusion model allows response time to become arbitrarily large, and, hence, is less suited for tasks with a time limit. The effect of increasing $|\theta|$, however, is the same according to both models.

The SRT model implies that response time tends to decrease as ability increases. This results in instantaneous responses for people of very extreme ability. Clearly, even very (un)able people need some time to read the item, make the actual response, and so on. If we assume that response time consists of a decision time T_d and a residual time T_r, and we adjust the actual time limit d for this residual time (i.e., $d^* = d - T_r$) we obtain that $d - T = d^* - T_d$. That is, the actual score of a person is not affected by the magnitude of the residual time. Put differently, whether the SRT model is used for decision times or response times does not make a difference. The value of the discrimination parameter, however, is affected by the magnitude of the residual time, via the effective reduction of the time limit. As long as the residual time is a constant, independent of both the item and the person, this causes little difficulties. If the residual time, however, differs between items, we see that these differences induce differences in item discriminatory power. We return to this issue in the discussion section.

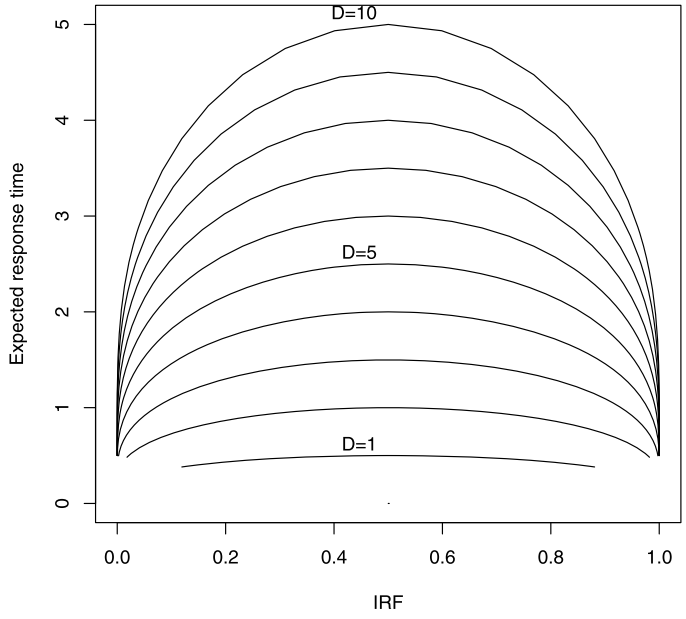

FIGURE 3.
SATF for different values of the time limit. Each line corresponds to a time limit in the range from 1 to 10.

2.3. *Speed-Accuracy Trade-Off*

In order to gain more insight in the relation between response time and response accuracy implied by the model, we consider how both response time and response accuracy relate to each other, to the time limit d and to ability θ. The relation between the item response function and the marginal expected response time that follows from Equations (12) and (15) is the following linear function:

$$P(X=1|\theta)=\frac{1}{2}\theta\mathcal{E}(T|\theta)+\frac{1}{2} \tag{17}$$

or put differently

$$\mathcal{E}(T|\theta)=\frac{2P(X=1|\theta)-1}{\theta}. \tag{18}$$

As a function of the time limit d, these curves specify a speed-accuracy trade-off function (SATF) for every ability θ; or conversely, as a function of ability θ, these curves specify a SATF for every time limit d. The SATF as a function of the time limit, for different values of ability, is given in Figure 3. We see in Figure 3 that increasing the time limit leads to larger increases in the expected response time for people with a probability of giving a correct response close to $1/2$, compared to people for whom the probability of giving a correct response is closer to either zero or one. Figure 4 gives the SATF as a function of ability, for different time limits. Figure 4 shows that the SATF is a linear decreasing function for people with negative ability values, and a linear increasing function for people with positive ability values. If ability is assumed to be fixed, the trade-off between speed and accuracy is completely determined by the time limit, and so considering the SATF as a function of ability may seem misleading. Nevertheless, it is instructive to see how, according to the SRT model, persons of different ability levels *trade-off* speed and accuracy.

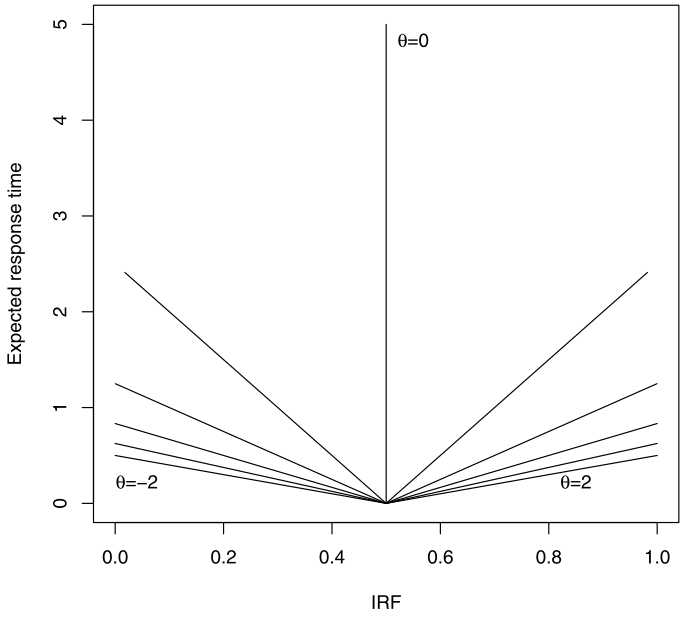

FIGURE 4.
SATF for different values of θ. Each line corresponds to a θ-value in the range from -2 to 2.

Observe that from the SATF we may infer a characterization of θ in terms of the expectation of X and T (where we suppress in our notation their dependence on θ, and treat them as quantities that are, in principle, observable):

$$\theta = \frac{2\mathcal{E}(X) - 1}{\mathcal{E}(T)}. \tag{19}$$

Because the numerator tends to plus or minus one as the time limit tends to infinity, we may interpret θ as velocity, and $|\theta|$ as speed. That is, the unit of measurement of θ is one over time. It is important to remember, at this point that θ here refers to the *difference* between the ability of the student and the difficulty of the item. However, if we multiply this quantity by two, the (expected) response time gets divided by two. That is, the difference between ability and item difficulty is measured on a ratio scale.

2.4. Conditional Accuracy Function

A different perspective on the relation between response time and accuracy is obtained from the *conditional accuracy function* (CAF) corresponding to this model. The CAF gives the probability of a correct response conditionally on the response time:

$$P(X = 1|T = t, \theta) = \frac{\exp(2(d - t)\theta)}{1 + \exp(2(d - t)\theta)}. \tag{20}$$

Figure 5 gives the CAF for different values of ability. Figure 5 shows that if θ is larger than zero, the CAF decreases from

$$\frac{\exp(2d\theta)}{1 + \exp(2d\theta)} > 1/2, \tag{21}$$

when t equals zero to $1/2$ when t equals d. If θ is smaller than zero, the CAF increases from

$$\frac{\exp(2d\theta)}{1 + \exp(2d\theta)} < 1/2, \tag{22}$$

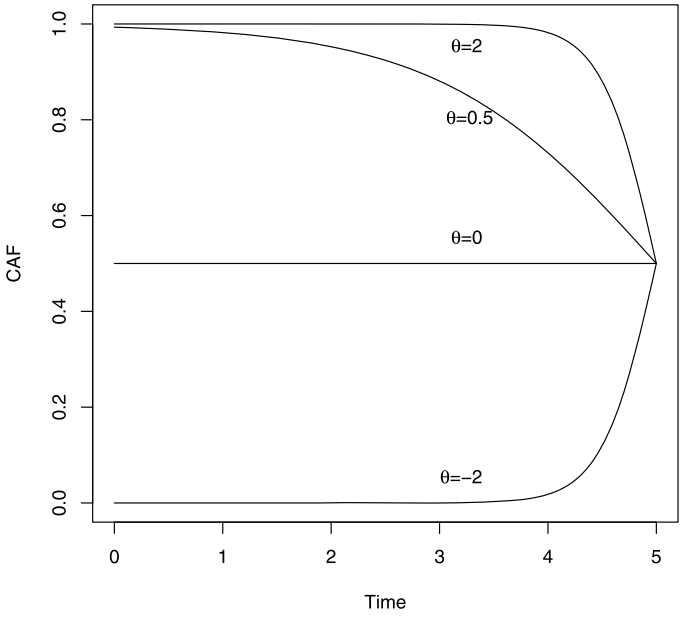

FIGURE 5.
CAF for different values of ability, with a time limit of 5.

when t equals zero to $1/2$ when t equals d. In words, this means that if the ability of a person is above the item difficulty, fast responses are more likely to be correct, whereas slow responses are more likely to be incorrect. Similarly, for a person with ability below the item difficulty, the reverse holds. Here, the SRT model differs from the (unbiased) drift diffusion model, which implies that response time and accuracy are independent (Tuerlinckx & De Boeck, 2005). After we have considered the conditional response time distributions corresponding to the CAF (via Bayes' theorem), we will have more to say about the meaning of the CAF.

2.5. *Conditional Response Time Distributions*

Corresponding to the CAF considered above there are conditional response time distributions for correct and incorrect responses, which are related to the CAF via Bayes' theorem. If we consider the response time distribution conditionally on response quality X_i, we obtain that

$$f(t|X_i = x, \theta) = \frac{\theta \exp((2x-1)(d-t)\theta)}{(2x-1)[\exp((2x-1)d\theta)-1]} \quad (0 \le t \le d). \tag{23}$$

The conditional expectations corresponding to these densities are

$$\mathcal{E}(T|X_i = 1, \theta) = \frac{1-(d\theta+1)\exp(-d\theta)}{\theta(1-\exp(-d\theta))} \tag{24}$$

and

$$\mathcal{E}(T|X_i = 0, \theta) = \frac{1+(d\theta-1)\exp(d\theta)}{\theta(\exp(d\theta)-1)}. \tag{25}$$

Observe that

$$(T|X=1,\theta) \underset{st}{=} (d-T|X=0,\theta) \underset{st}{=} (T|X=0,-\theta) \underset{st}{=} (d-T|X=1,-\theta). \tag{26}$$

From this chain of (stochastic) equalities, we obtain that there is a qualitative difference between able ($\theta > 0$) and unable ($\theta < 0$) persons. Able persons are those whose correct responses are fast,

and whose errors are slow; whereas unable persons are those whose correct responses are slow, and whose errors are fast. How fast and how slow depends on the absolute magnitude of ability $|\theta|$ (that is, on speed).

From $(T|X=1,\theta) \underset{st}{=} (d-T|X=0,\theta)$, we obtain that a new variable T^*, defined as follows:

$$T^* = \begin{cases} T & \text{if } X=1 \\ d-T & \text{if } X=0 \end{cases} \sim (T|X=1,\theta) \tag{27}$$

is independent of response accuracy ($X \perp\!\!\!\perp T^*$). Furthermore, both the distribution of X and the distribution of T^* belong to the exponential family of distributions. Both the sum, across items, of the item responses X and the sum of the *pseudo* times T^* are sufficient for θ.

2.6. Item Score Distribution

Because the sufficient statistic for ability (i.e., the item score) carries all the information about the value of the latent trait, we derive its distribution here. The distribution of the item score $S=(2X-1)(d-T)$ is found to be

$$P(S \le s|\theta) = \begin{cases} P(T \le d+s|X=0,\theta)P(X=0|\theta) & \text{if } s<0, \\ P(X=0|\theta) + P(T > d-s|X=1,\theta)P(X=1|\theta) & \text{if } s \ge 0, \end{cases} \tag{28}$$

with corresponding density:

$$f(S=s|\theta) = \begin{cases} f(d+s|X=0,\theta)P(X=0|\theta) & \text{if } s<0, \\ f(d-s|X=1,\theta)P(X=1|\theta) & \text{if } s \ge 0, \end{cases} \tag{29}$$

which simplifies as follows:

$$f(S=s|\theta) = \frac{\theta \exp(s\theta)}{\exp(d\theta)-1} \frac{\exp(d\theta)}{1+\exp(d\theta)} = \frac{\theta \exp((s+d)\theta)}{\exp(2d\theta)-1} = \frac{\exp(s\theta)}{\frac{\exp(d\theta)-\exp(-d\theta)}{\theta}}. \tag{30}$$

Hence, the cumulative distribution may be written as follows:

$$P(S \le s|\theta) = \frac{\exp((s+d)\theta)-1}{\exp(2d\theta)-1} \tag{31}$$

with expectation equal to

$$\mathcal{E}(S|\theta) = d\left(2\frac{\exp(2d\theta)}{\exp(2d\theta)-1} - 1\right) - \frac{1}{\theta} \tag{32}$$

and in general with cumulant generating function:

$$\ln \mathcal{E}\big(\exp(Sx)\big) = \ln\left(\frac{\exp(d\theta)}{\exp(2d\theta)-1}\frac{\theta}{\theta+x}\big[\exp\big(d(\theta+x)\big) - \exp\big(-d(\theta+x)\big)\big]\right). \tag{33}$$

Observe that the mean satisfies the following nice symmetry $\mathcal{E}(S|\theta) = -\mathcal{E}(S|-\theta)$, which means that it is an odd function of θ. Some rewriting gives the following expression:

$$\mathcal{E}(S|\theta) = d\frac{\exp(2d\theta)+1}{\exp(2d\theta)-1} - \frac{1}{\theta}. \tag{34}$$

As for every exponential family model, the variance of the sufficient statistic is the Fisher information:

$$\mathcal{I}(\theta) = \mathcal{V}(S|\theta) = \mathcal{E}\big(S^2|\theta\big) - \big(\mathcal{E}(S|\theta)\big)^2, \tag{35}$$

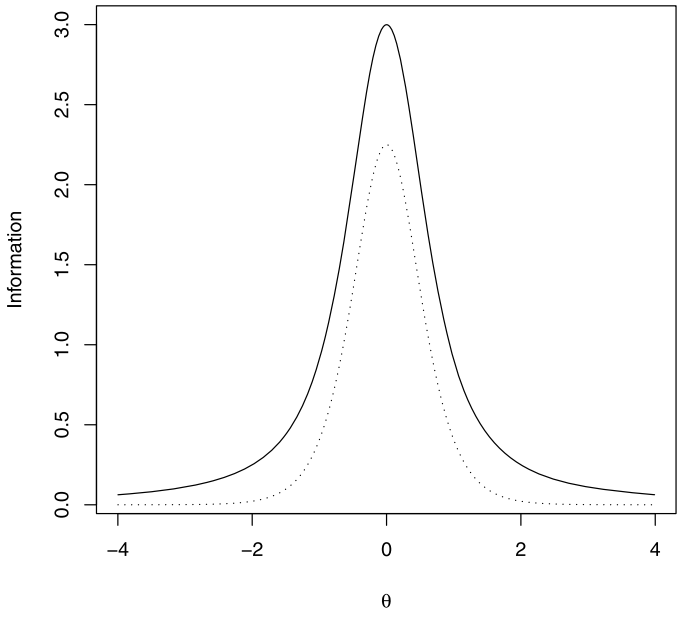

FIGURE 6.
Information function for the SRT model (*solid line*) and the 2PL model for accuracy (*dotted line*) as a function of ability for a time limit of 3.

where $\mathcal{E}(S^2|\theta)$ is readily found to be:[3]

$$\mathcal{E}\left(S^2|\theta\right) = \frac{\exp(d\theta)\frac{d^2\theta^2+2-2d\theta}{\theta^2} - \exp(-d\theta)\frac{d^2\theta^2+2+2d\theta}{\theta^2}}{\exp(d\theta) - \exp(-d\theta)}. \tag{36}$$

The mathematical expression for the information is not very instructive in its own right, but it is informative to compare it to the information derived from accuracy data alone (i.e., the information according to a 2PL model):

$$\mathcal{I}^{\text{2PL}}(\theta) = d^2\frac{\exp(d\theta)}{(1+\exp(d\theta))^2}. \tag{37}$$

Figure 6 shows both information functions for a time limit of 3. It is seen that throughout the ability range the use of response times adds to the information. Another way of comparing both information functions is to look at the effect of time limit. For the SRT, we see in Figure 7 that information is an increasing function of the time limit (referred to as "deadline" in Figure 7), whereas this does not hold true for the 2PL (unless ability is equal to zero).

A different perspective on the relation between information from accuracy alone and that from both accuracy and response time follows from the alternative formulation of the model presented in the previous section. Because X and T^* are independent, they contribute independently to the total information. Put differently, the total information is the information in X plus that in T^*. The information in T^* is readily found to be:

$$\mathcal{I}^{T^*}(\theta) = \frac{\exp(2d\theta) - (2+d^2\theta^2)\exp(d\theta) + 1}{\theta^2(\exp(d\theta)-1)^2}. \tag{38}$$

[3]The main result needed for this derivation is the following indefinite integral $\int x^2 a\exp(ax)\,dx = \frac{a^2x^2-2ax+2}{a^2}\exp(ax)$.

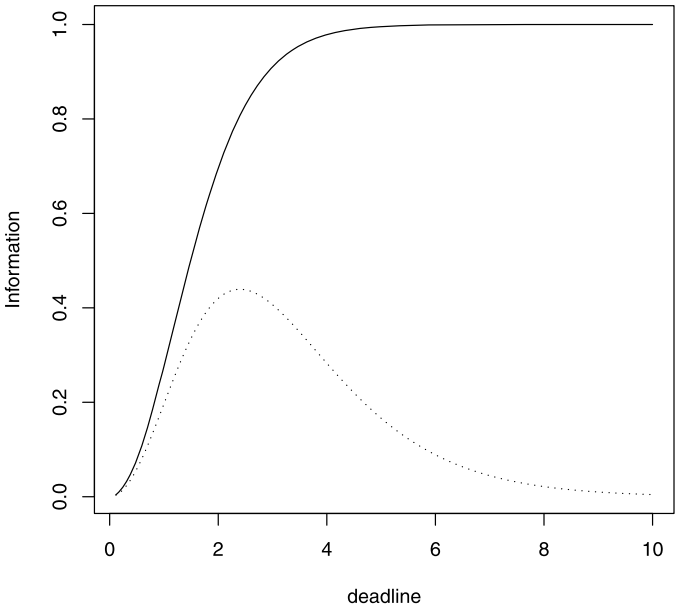

FIGURE 7.
Information function for the SRT model (*solid line*) and the 2PL model for accuracy (*dotted line*) as a function of time limit, for θ equal to 1.

We find that the information in the pseudo response times T^* is at least one-third of the information in the accuracy alone:

$$\mathcal{I}^{T^*}(\theta) \geq \frac{\mathcal{I}^{2PL}(\theta)}{3}. \tag{39}$$

If ability is equal to zero, the information in the pseudo response time is exactly one third of the information in the accuracy alone. In Figure 8, we see both the information function for accuracy and that for the pseudo response time. Observe that for values of the ability sufficiently different from zero, there is more information about ability in the pseudo response time than there is in the accuracy.

3. Parameter Estimation

A straightforward and robust approach to statistical inference is to use response accuracy for estimating the model parameters, which is a standard problem, and to only use response time for model validation purposes. Even though statistically inefficient, this approach is very powerful when it comes to evaluation of the tenability of the measurement model.

Here, however, a more efficient approach will be developed that uses both response accuracy and response time for estimating the model parameters. In particular, a computationally attractive variant of the EM algorithm (Dempster, Laird, & Rubin, 1977) for computing marginal maximum likelihood (MML) estimates is developed.

Assuming that ability is normally distributed with an unknown expectation μ and unknown variance σ^2, we obtain the following marginal likelihood function:

$$p(\mathbf{s}|\boldsymbol{\delta},\mu,\sigma) = \prod_p \int_{-\infty}^{\infty} \prod_i \frac{\exp(s_{pi}(\theta_p-\delta_i))}{\frac{\exp(d(\theta_p-\delta_i))-\exp(-d(\theta_p-\delta_i))}{\theta_p-\delta_i}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta_p-\mu)^2}{2\sigma^2}\right) d\theta_p. \tag{40}$$

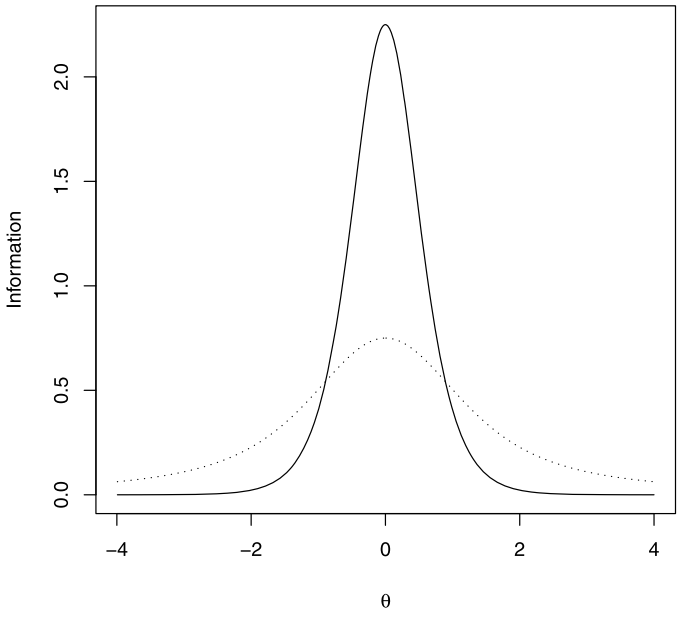

FIGURE 8.
Information function for the 2PL model for accuracy (*solid line*) and the pseudo response time T^* (*dotted line*) as a function of θ for a time limit of 3.

An EM algorithm may be used for estimating $\boldsymbol{\delta}$, μ, and σ. In order to apply the EM algorithm, we need to specify the so-called "Q-function," which for our model looks as follows:

$$
\begin{aligned}
&Q_{\mathbf{s}}(\boldsymbol{\delta}, \mu, \sigma \,|\, \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma}) \\
&\quad = \sum_p \int_{-\infty}^{\infty} \Bigg(\sum_i \big(s_{pi} (\theta_p - \delta_i) \big) - \ln \Bigg(\frac{\exp(d(\theta_p - \delta_i)) - \exp(-d(\theta_p - \delta_i))}{\theta_p - \delta_i} \Bigg) \\
&\qquad - \ln(\sqrt{2\pi}\sigma) - \frac{(\theta_p - \mu)^2}{2\sigma^2} \Bigg) f(\theta_p | \mathbf{s}; \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma}) \, d\theta_p .
\end{aligned} \tag{41}
$$

Even though with an EM algorithm for the problem at hand we reduce a complex multidimensional estimation problem to a sequence of unidimensional problems, the M step of the algorithm does not have an analytical solution. As a consequence, a not necessarily monotonic iterative approach, such as Newton–Raphson, has to be used here. The complications arise from the presence of the following term in the Q-function:

$$
\ln \Bigg(\frac{\exp(d(\theta_p - \delta_i)) - \exp(-d(\theta_p - \delta_i))}{\theta_p - \delta_i} \Bigg). \tag{42}
$$

We develop an alternative estimation method that is based on the idea of minorization (de Leeuw, 1994; Hunter & Lange, 2004; de Leeuw, 2006). This alternative estimation method shares with the EM algorithm the desirable property of monotonic convergence, but allows for an analytical solution to its M step.

The minorization approach for finding the maximum of the log-likelihood function $l_{\mathbf{y}}(\boldsymbol{\omega}) = \ln P(\mathbf{y}|\boldsymbol{\omega})$ rests on the following chain of inequalities:

$$
l_{\mathbf{y}}(\boldsymbol{\omega}) = M_{\mathbf{y}}(\boldsymbol{\omega}, \boldsymbol{\omega}) \leq M_{\mathbf{y}}(\widehat{\boldsymbol{\omega}}, \boldsymbol{\omega}) \leq l_{\mathbf{y}}(\widehat{\boldsymbol{\omega}}) \tag{43}
$$

known as the *sandwich* inequality. That is, we need a function M such that

$$M_{\mathbf{y}}(\widehat{\boldsymbol{\omega}}, \boldsymbol{\omega}) \begin{cases} = l_{\mathbf{y}}(\boldsymbol{\omega}) & \text{if } \widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}, \\ \leq l_{\mathbf{y}}(\boldsymbol{\omega}) & \text{for all } \widehat{\boldsymbol{\omega}} \text{ and } \boldsymbol{\omega}. \end{cases} \tag{44}$$

van Ruitenburg (2005) considers a quadratic minorization of part of the conditional log-likelihood for the Rasch model. The key property used to find a quadratic minorization is that the function to be minorized is odd (van Ruitenburg, 2005, p. 13). It is readily seen that the *complicated* part in the Q-function also is an odd function, and hence, admits a quadratic minorization:

$$f(x) = -\ln\left(\frac{\exp(dx) - \exp(-dx)}{x}\right) \geq a(\widehat{x})x^2 + b(\widehat{x}), \tag{45}$$

where

$$a(\widehat{x}) = \frac{\frac{d}{dx} f(x)|_{x=\widehat{x}}}{2\widehat{x}}, \tag{46}$$

$$b(\widehat{x}) = f(\widehat{x}) - a\widehat{x}^2 \tag{47}$$

and

$$\frac{d}{dx} f(x) = -\left(d\frac{\exp(dx) + \exp(-dx)}{\exp(dx) - \exp(-dx)} - \frac{1}{x}\right). \tag{48}$$

It will become clear that using this minorization in the Q-function of an EM algorithm not only gives us a monotonically convergent estimation algorithm, but also one in which the M step has an analytical solution.

Plugging this minorization into the Q-function gives the following function:

$$\begin{aligned} Q^*_{\mathbf{s}}(\boldsymbol{\delta}, \mu, \sigma | \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma}) = \sum_p \int_{-\infty}^{\infty} \Bigg(& \sum_i \big(s_{pi}(\theta_p - \delta_i)\big) + a(\theta_p - \widehat{\delta}_i)(\theta_p - \delta_i)^2 + b(\theta_p - \widehat{\delta}_i) \\ & - \ln(\sqrt{2\pi}\sigma) - \frac{(\theta_p - \mu)^2}{2\sigma^2} \Bigg) f(\theta_p | \mathbf{s}; \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma})\, d\theta_p \end{aligned} \tag{49}$$

the derivatives of which, with respect to the δ_i's:

$$\frac{\partial}{\partial \delta_i} Q^*_{\mathbf{s}}(\boldsymbol{\delta}, \mu, \sigma | \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma}) = \sum_p \int_{-\infty}^{\infty} \big(-s_{pi} - 2a(\theta_p - \widehat{\delta}_i)(\theta_p - \delta_i)\big) f(\theta_p | \mathbf{s}; \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma})\, d\theta_p \tag{50}$$

admit the following closed form solution when equated to zero:

$$\delta_i = \frac{\sum_p s_{pi} + 2\sum_p \int_{-\infty}^{\infty} a(\theta_p - \widehat{\delta}_i)\theta_p f(\theta_p | \mathbf{s}; \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma})\, d\theta_p}{2\sum_p \int_{-\infty}^{\infty} a(\theta_p - \widehat{\delta}_i) f(\theta_p | \mathbf{s}; \widehat{\boldsymbol{\delta}}, \widehat{\mu}, \widehat{\sigma})\, d\theta_p}. \tag{51}$$

4. Model Extensions

There are different directions into which we can fruitfully extend the SRT model. Here, one particular model extension and its implications will be discussed. First, we consider how the extended SRT scoring rule may be used for questions with more than two response alternatives. Second, we consider how the extended SRT scoring rule may be used to reflect different trade-offs between speed and accuracy.

We consider a situation where a person can choose between multiple decisions. We assume that every response alternative earns a certain amount of credit and that the item score equals the

credit times the residual time. That is, if Y_{pij} equals one if person p chooses alternative j for item i, we may denote the weighted SRT (WSRT) as follows:

$$\sum_j Y_{pij} a_{ij} (d - T_{pi}). \tag{52}$$

The IRT model in which this is the sufficient statistic is straightforward to derive, as are the various marginal and conditional distributions.

Two properties of this model deserve special attention. First, contrary to the SRT model, the resulting model for accuracy data alone does *not* necessarily (i.e., for all values of the weights a_{ij}) belong to the exponential family of distributions

$$P(Y_{pi0} = 1|\theta_p) = \frac{\frac{\exp(a_{i0} d\theta_p) - 1}{a_{i0}}}{\sum_j \frac{\exp(a_{ij} d\theta_p) - 1}{a_{ij}}}. \tag{53}$$

Second, if for every option of an item the credit is positive (or negative), the probability of a correct response does *not* go to zero as ability decreases, but rather tends to a non-zero lower asymptote that depends on the credit assigned to options as follows:

$$\text{If } \forall j : a_{ij} > 0 \quad \text{then} \lim_{\theta_p \to -\infty} P(Y_{pi0} = 1|\theta_p) = \frac{\frac{1}{a_{i0}}}{\sum_j \frac{1}{a_{ij}}}. \tag{54}$$

Qualitatively, this model closely resembles the three parameter logistic model (3PL) (Birnbaum, 1968). The IRT model derived from the SRT scoring rule implies that as ability decreases, the probability of an incorrect response increases. While this may be plausible for open ended questions, it is a questionable assumption for closed form questions with a limited number of response alternatives (e.g., Van der Maas et al., 2011; Tuerlinckx & De Boeck, 2005). The IRT model derived from the WSRT does not necessarily make this questionable assumption.

If we reconsider the CISRT and the SRT scoring rules, we see that they are both instances of the WSRT in which the weights for correct and incorrect responses are equal to one and zero, respectively, for the CISRT scoring rule, and one and minus one for the SRT scoring rule. In general, we can consider the following family of scoring rules for two-choice decisions:

$$\sum_i \big(C_i X_{pi} - P_i (1 - X_{pi})\big)(d - T_{pi}), \tag{55}$$

where credit C_i is earned for a correct response to item i and punishment P_i is earned for an incorrect response.

5. Illustration: The Measurement of Chess Ability

The SRT model was applied to data from the computerized adaptive Amsterdam Chess Test II[4] (ACT-II), collected during the Corus Chess Tournament 2008 in Wijk aan Zee in the Netherlands. A total of 295 participants were tested, but 34 participants are removed from the analysis because they did not complete the test, and 4 because the computer system failed during test administration, leaving a total of 257 participants. The computerized adaptive Amsterdam Chess Test II consists of 100 choose-a-move items that are administered with a time limit of either 20 or 40 seconds. The test is administered as a computerized adaptive test where item

[4]We thank Daan Zult for collecting the data used in the example.

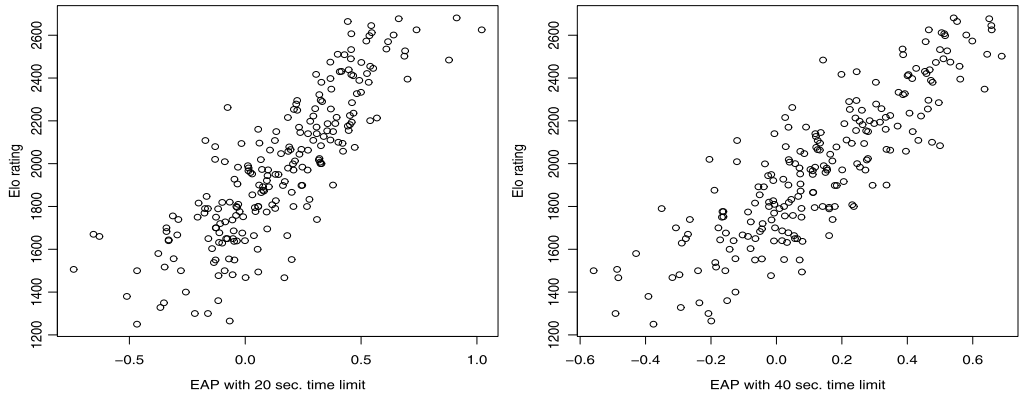

FIGURE 9.
Scatterplot of EAP for the 20-second (*left*) and 40-second (*right*) time limit conditions versus Elo rating.

difficulties are estimated during test administration. For the purpose of this illustration, however, the exact details of how items were assigned to participants is of no importance.

One of the core predictions of the SRT model is that the time limit is equal to the item discrimination parameter. Because the same items are administered with two different time limits, the ACT-II allows for a direct test of this hypothesis. An important advantage of this illustrative application is that for chess expertise an established external criterion is available in the form of the Elo ratings of chess players, which has high predictive power for game results.

5.1. Relation to Elo ratings

In order to validate the ACT-II as well as the SRT model, we estimated the expected a posteriori (EAP) estimates for all participants from the SRT model both for the 20- and 40-second time limit conditions, separately, using both response time and accuracy. For both conditions, a normal distribution of ability was assumed of which the mean was set equal to zero, in order to identify the parameters, and of which the variance was estimated. The item discrimination was set equal to 2 for the 20-second time limit condition and 4 for the 40-second time limit condition, in line with the assumption that the item discrimination parameter is equal to the time limit.

For those 217 participants for whom a reliable Elo rating was available, we correlated the EAP estimates with their Elo ratings. The results are given in Figure 9. The correlation between EAP estimates and Elo ratings is equal to 0.837 and 0.858 for the 20-second and 40-second time limit conditions, respectively. We see that the ACT-II has good predictive validity for the actual Elo ratings of the chess players involved. More importantly, we find a (slightly) higher correlation for the 40-second time limit condition compared to the 20-second time limit condition, in line with the assumption that the time limit modulates the discriminatory power of the items.

5.2. Time Limit and Discrimination

We now test the crucial hypothesis that time limit selectively influences item discrimination, and *not* item difficulty, in a more formal way. Because the number of observations per item is limited, we confine attention to only two possible values for the item discrimination parameters, in line with the (very restrictive) hypothesis that item discriminatory power is *completely* explained by the imposed time limit. We assume that the item discrimination equals 1 for all items in the 40-second time limit condition, and equals either 1 or 2 in the 20-second time limit condition. The assumption that the item discrimination is twice as large in the 20-second condition compared to the 40-second condition means that the item discrimination does *not* depend on the

TABLE 1.
AIC information values for $\mathcal{H}_0$, $\mathcal{H}_1$, and $\mathcal{H}_2$.

	−2 log-likelihood	Number of parameters	AIC
$\mathcal{H}_0$	140.70	101	342.70
$\mathcal{H}_1$	139.42	201	541.42
$\mathcal{H}_2$	179.84	201	581.84

time limit. In terms of the WSRT, we assume that reward and punishment are equal (i.e., a_{i0} equals a_{i1}), but consider different ways in which the responses from both time limit conditions contribute to the final sufficient statistic:

$$\mathcal{H}_0: \quad \forall i: \quad a_{i1}^{(20)} = a_{i1}^{(40)} \quad \text{and} \quad \delta_i^{(20)} = \delta_i^{(40)}, \tag{56}$$

$$\mathcal{H}_1: \quad \forall i: \quad a_{i1}^{(20)} = a_{i1}^{(40)} \quad \text{and} \quad \delta_i^{(20)} \neq \delta_i^{(40)}, \tag{57}$$

$$\mathcal{H}_2: \quad \forall i: \quad a_{i1}^{(20)} = 2a_{i1}^{(40)} \quad \text{and} \quad \delta_i^{(20)} \neq \delta_i^{(40)}. \tag{58}$$

Specifically, $\mathcal{H}_0$ and $\mathcal{H}_1$ share the same sufficient statistic for ability:

$$\sum_{i:d_i=20} (2X_{pi} - 1)(d_i - T_{pi}) + \sum_{i:d_i=40} (2X_{pi} - 1)(d_i - T_{pi}) \tag{59}$$

and differ in the sufficient statistics for the item difficulties, whereas $\mathcal{H}_2$ assumes a different sufficient statistic for ability (where twice the remaining time is earned or lost for a correct or incorrect response, in the 20-second time limit condition):

$$\sum_{i:d_i=20} 2(2X_{pi} - 1)(d_i - T_{pi}) + \sum_{i:d_i=40} (2X_{pi} - 1)(d_i - T_{pi}). \tag{60}$$

Another way to look at these different hypotheses relates to the different models for accuracy they imply. Both $\mathcal{H}_0$ and $\mathcal{H}_1$ imply that the item discrimination will depend on the imposed time limit, whereas $\mathcal{H}_2$ does not.

The results of the analyses are summarized in Table 1. We see that, in terms of the AIC, the data provide strong evidence in favor of the conclusion that the time limit selectively influences the item discrimination and not the item difficulty.

6. Discussion

In this paper, a new measurement model was derived from an explicit scoring rule that involves both response time and accuracy for tasks that prescribe an explicit time limit for responding. This new model implies the 2PL model for response accuracy alone. The time limit was found to have an effect on the item discrimination parameter in the resulting 2PL model for accuracy. Application of the new model to the ACT-II lead us to conclude that item discrimination indeed depends on the time limit, whereas item difficulty remains unaffected. The relevant marginal and conditional distributions have been derived, and an estimation algorithm was proposed.

This paper provides an illustration of how new measurement models can be derived from explicit scoring rules. It should be clear that there are many possible variations on this theme, as many as there are scoring rules. If only response accuracy is considered, the number of different scoring rules that can be formulated is quite limited. If both response time and accuracy are considered, however, the number of substantively different rules is much larger, as was illustrated with the WSRT scoring rule.

From the SATF, we found a characterization of the difference between ability and difficulty ($\theta_p - \delta_i$) in terms of the expected response time and accuracy. As the time limit increases without bound, the expected response tends to zero or one, depending on whether the ability of the student is above or below the difficulty of the item. Since (expected) response time is measured on a ratio scale, it follows that also the difference between ability and difficulty is on a ratio scale.

The simple model as it was derived here provides a clear and elegant interpretation for the discrimination parameter in the 2PL model. However, the SRT model does not necessarily imply that the item discrimination parameter only depends on the imposed time limit. The model predicts that as we increase the time limit, the item discrimination parameter should increase as well. It is not necessary, however, that for given fixed time limits the item discrimination parameters should all be equal. In order to test this hypothesis, data are needed where the same item is administered with different time limits and only then can it be established whether or not the discrimination depends on the time limit. One reason why items may differ in discrimination is found in differences in the non-decision time T_r.

The SRT model, with equal credit for correct and punishment for incorrect responses, was proposed because the unequal balancing of the CISRT might promote guessing. That is, the way correct and incorrect responses are balanced may influence the behavior of students. Of course, for such an influence to take effect, feedback must be provided to students regarding their score, and the scoring rule must be known to them. An interesting question that follows from this observation is whether we may in fact use a scoring rule to induce behavior. For instance, if we consider the WSRT for correct and incorrect decisions with reward equal to one and punishment equal to 100, it is a safe bet that students will only answer when they are absolutely certain that their answer will be correct. If, however, reward and punishment are reversed, it is clearly beneficial for students to guess quickly. Further research is needed to establish the extent to which explicit feedback based on an explicit scoring rule balancing speed and accuracy forces (or entices) students to adopt a particular response model. The approach to item response modeling that derives from this can be formulated as follows: For a given response model, can we *train* students to comply to it, as opposed to the traditional point of view: For given responses of students to questions, can we *infer* the response model.

Observe that with the SRT scoring rule students may choose to not answer a question, that is, he may just wait until all time elapses. To counter such behavior, it may we advisable to adapt the difficulty level of the items to the ability level of the student, as in a computerized adaptive test. Since tests where not only response accuracy but also response time is registered are usually computerized tests anyhow, such behavior is easily prevented.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

de Leeuw, J. (1994). Block-relaxation algorithms in statistics. In H.H. Bock, W. Lenski, & M.M. Richter (Eds.), *Information systems and data analysis* (pp. 308–325). Berlin: Springer.

de Leeuw (J.2006). *Some majorization techniques* (Tech. Rep. No. 2006032401). Department of statistics, UCLA.

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

Dennis, I., & Evans, J. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology*, *87*, 105–129.

Hunter, D., & Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, *58*(1), 30–37.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press).

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650.

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

Van der Maas, H.L., Molenaar, D., Maris, G., Kievit, R.A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356.

Van der Maas, H.L., & Wagenmakers, E.J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*(1), 29–60.

van Ruitenburg, J. (2005). *Algorithms for parameter estimation in the Rasch model*. Unpublished master's thesis, Erasmus University.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model: OPLM. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments and applications* (pp. 215–238). New York: Springer.

Wickelgren, W. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Manuscript Received: 20 DEC 2010
Final Version Received: 25 NOV 2011
Published Online Date: 18 SEP 2012