Psychological Test and Assessment Modeling, Volume 55, 2013 (4), 361-382

Modeling responses and response times in personality tests with rating scales

Jochen Ranger[1]

Abstract

In this article several latent trait models for the joint distribution of the responses and response times in rating scales are compared. Among these models are two generalizations of established models for binary items, namely a generalization of the approach of Ferrando and Lorenzo-Seva (2007a) and a generalization of the approach of Ranger and Ortner (2011). Two new models and a variant of the hierarchical model of van der Linden (2007) are also considered. All these models combine the graded response model with a response time model based on the log-normal distribution. The models differ in the assumed relationship between the expected log response time and the underlying latent traits. Although the proposed models have different interpretations and implications they can all be calibrated within the same general framework using marginal maximum likelihood estimation and an application of the ECM-algorithm. The models are used for the analysis of an empirical data set. According to the AIC index, the generalization of the model of Ranger and Ortner (2011) can represent the data best.

Keywords: response time, log-normal distribution, inverted U-shaped relationship, rating scale

[1] *Correspondence concerning this article should be addressed to:* Jochen Ranger, PhD, Martin-Luther-Universität Halle-Wittenberg, Institut für Psychologie, Brandbergweg 23c, 06120 Halle (Saale), Germany; email: jochen.ranger@psych.uni-halle.de

The analysis of response times in psychological tests has a long history (Thurstone, 1937; Furneaux, 1952). The focus of psychological research has thereby been on achievement tests, and substantial progress has been made in this area. A major step forward was the development of latent trait models that can be used for the joint analysis of the responses and response times in a test. In comparison to the simple indices, which have formerly been used in order to combine the accuracy and speed of a respondent, the usage of a latent trait model has several advantages. First, some latent trait models have an epistemological foundation as these models can be derived from the principle of specific objectivity, which states that the result of a comparison of two individuals must not depend on the specific item used for the comparison (Fischer, 1989). And second, latent trait models allow for more sophisticated research questions and applications. Latent trait models have been used for item selection in adaptive testing (van der Linden, 2008), the detection of collusion between test takers (van der Linden, 2009a) and the detection of aberrant responses (van der Linden & van Krimpen-Stoop, 2003).

Much progress has been made in the field of achievement testing, where a large number of different models for responses and response times has been proposed. For an overview over the different models proposed so far see van der Linden (2009b) as well as Lee and Chen (2011). Some of these models explicitly refer to the concept of specific objectivity and have a sound measurement theoretic foundation. See for example Scheiblechner (1979) who proposed a response time model based on the exponential distribution. In this model the total test time has a similar function as the sum score in the Rasch model, being a sufficient statistic for the speed of a test taker. Other models have been derived from assumptions about the response process. The components of such process models are closely related to psychological concepts, such that these models go beyond the usual measurement models of item response theory (Tuerlinckx & De Boeck, 2005; van der Maas, Molenaar, Maris, Kievit, & Boorsboom, 2011; Vandekerckhove, Tuerlinckx, & Lee, 2011).

Less progress has been made in the area of attitudinal scales and personality tests. The latent trait models devised for the responses and response times in achievement tests can not simply be transferred to this area of application. This is due to a different relation between the time needed to give a response and the trait that is intended to be measured with the test. In personality and attitudinal scales the individuals located at either end of the trait continuum usually respond fast, a data pattern that is described as an inverted-U relationship (Kuiper, 1981; Ferrando, 2006; Akrami, Hedlund, & Ekehammar, 2007). This inverted-U relationship can not be found in achievement tests, where the average response time and the test score are related monotonously (Lavergne & Vigneau, 1997; MacLennan, Jackson, & Bellantino, 1988; Rafaeli & Tractinsky, 1991). Therefore, different latent trait models are needed for achievement tests and personality scales.

Models that are based on a nonmonotone relation between the time needed to give a response and the underlying latent trait are less numerous than models based on a monotone relation. A nonmonotone model for binary items has been proposed by Ferrando and Lorenzo-Seva (2007a) as well as by Ranger and Ortner (2011). Ferrando and Lorenzo-Seva (2007a) use the two-parameter logistic model for the responses and combine this model with a response time model based on the log-normal distribution. The re-

sponse time model relates the expectations of the log response times to two determining factors. The first factor is the distance between the item location given by the two-parameter logistic model and the trait level of an individual. The second factor consists in an additional latent trait that reflects the general work pace of an individual. The model of Ranger and Ortner (2011) is similar in spirit, likewise combining a two-parameter logistic model with a response time model based on the log-normal distribution. The major difference consists in the implementation of the inverted-U relationship. Instead of using the distance between the item location and the individual as a determining factor of the expected response times, the authors use the probability of the given response implied by the item response model. Which of the two models is the right one, or the better approximation to reality respectively, can not be assessed on basis of the available research findings, as the relative fit of both models has not been compared so far.

Personality scales usually do not consist of items with just two response categories representing the endorsement or rejection of a statement. It is more common to use a response format consisting of several ordered response categories, each accompanied by a label that defines the extend of endorsement represented by the category. Such items are sometimes denoted as Likert-type items (Baker, 1992, p.222) and the response format is called a rating scale (Seiwald, 2003). The response time models proposed by Ferrando and Lorenzo-Seva (2007a) as well as Ranger and Ortner (2011) require binary scaled items and can not be used for such tests. Nevertheless, the idea that some sort of distance between the item and the individual is a predictor of the expected response time can also be implemented in rating scales. One implementation of this idea is due to Ferrando and Lorenzo-Seva (2007b). Although their model is intended for scales consisting of multiple ordered response categories, the authors start from the standard factor model for continuous responses. The codes assigned to the response categories of the rating scale are then shrunk to the range of 0 and 1 by a linear transformation and the factor model is reparameterized correspondingly. The item location is defined as the trait level that is needed for an expected response of 0.5 in the reparameterized factor model. This definition of the item distance resembles the item location in the two-parameter logistic model. Having located the item on the scale, the distance of each individual to the item can be determined. Similar to the model of Ferrando and Lorenzo-Seva (2007a) for binary items, the distance is then used as a predictor of the expected log response time.

Although the model of Ferrando and Lorenzo-Seva (2007b) is a reasonable extension to scales with several response categories it has a drawback. In case of a continuous response model, such as the standard factor model, the support of the responses ranges from $-\infty$ to ∞ and can not be scaled to (0,1) by a linear transformation. In the case of rating scales with few numerically scored response options such a linear transformation exists. However, the numerical values assigned to the response categories are arbitrary and the actual range of agreement reflected by the response categories is unknown. Therefore, a transformation to (0,1) might only be possible under strong assumptions about the spreading of the response categories. A crucial question is the question whether the response categories can be assumed to be equidistant. This limits the applicability of the model.

A second limitation of the current state of research is the focus on single models. There are hardly any model comparison studies, which are essential for identifying the more fruitful approaches to data analysis. This is especially important as different models have different interpretations. The mechanisms behind the inverted-U relationship are far from clear. When responses to test items are given, individuals have to assess information, generate an internal response and map the internal response to the response options offered by the rating scale. Which stage of the response process is accelerated by extreme trait levels is unknown. The distance to the item thresholds is only relevant for the last response stage, the mapping of the internal response to the given response options. Therefore, the usage of a distance measure as a predictor of the expected response time can only account for effects on the last response stage. One could also assume that it is the overall trait level that affects the response time, as information processing is facilitated in individuals with extreme trait levels. This is a more schema based interpretation of the inverted-U relationship: Strong self schemata enhance information processing. Summing up, it would be interesting to compare different models in order to locate the mechanism behind the inverted-U relationship within the response process.

In this manuscript I compare several models for responses and response times in items with ordered categorical response format. All models respect the discrete nature of the responses by using the graded response model. The different response time models considered in this manuscript are based on the log-normal distribution. The models differ in the way they relate the expected log response time to the trait level the test is supposed to measure. One possibility is to use the distance between an individual and the item as defined in Ferrando and Lorenzo-Seva (2007b). This replicates the model of Ferrando and Lorenzo-Seva (2007b) with a graded response model instead of the standard factor model. Alternative approaches are the usage of the minimal distance of an individual to one of the response thresholds or the usage of the probability of the uttered response. All these new alternatives are considered in the manuscript. The manuscript is organized as follows. First, the different models are described. Then, a general framework for model calibration is proposed. And finally, the models are applied to a real data set. Models are compared according to Akaike's information criterion (AIC) (Akaike, 1992).

Response time modeling in rating scales

The agreement with an item is a continuous quantity that has to be assigned to a discrete response category when the response is given on a rating scale. This process of transforming a continuous quantity into a discrete response can be modeled with the graded response model, which can be derived from the following assumptions. Let the continuous agreement to item g follow a standard factor model, such that the agreement y_g^* depends on the trait level θ according to the linear model $y_g^* = \alpha_{0g}^{'} + \alpha_{1g}^{'}\theta + r_g$. The intercept parameter $\alpha_{0g}^{'}$ determines the expected agreement in individuals with trait level $\theta = 0$ and the regression coefficient $\alpha_{1g}^{'}$ regulates the strength of the relation between θ and the level of agreement y_g^*. The residual term r_g is assumed to be distributed according to a normal distribution with an expectation of zero and a variance of $\sigma_{r_g}^2$.

Each of the $k = 1,\ldots,K$ response categories of the rating scale corresponds to a bandwidth on the agreement continuum. These bandwidths are defined by the $K-1$ thresholds c_{1g}, $\ldots$, $c_{(K-1)g}$. The first category is chosen in case the agreement y_g^* is lower than the first threshold c_{1g}, the second category is chosen in case the agreement y_g^* falls between the first threshold c_{1g} and the second threshold c_{2g} and the last category K is chosen in case the agreement y_g^* is greater than the last threshold $c_{(K-1)g}$. Instead of observing the continuous agreement y_g^* one can only register the chosen category, which will be denoted as y_g in the following. The probability that a respondent with trait level θ selects category k or lower follows from the assumption of the linear model $y_g^* = \alpha_{0g}^{'} + \alpha_{1g}^{'}\theta + r_g$ as

$$\begin{aligned} \mathrm{P}\left(y_g \le k \mid \theta\right) &= \mathrm{P}\left(\alpha_{0g}^{'} + \alpha_{1g}^{'}\theta + r_g \le c_{kg}\right) = \mathrm{P}\left(\frac{r_g}{\sigma_{r_g}} \le \frac{c_{kg} - \alpha_{0g}^{'}}{\sigma_{r_g}} - \frac{\alpha_{1g}^{'}}{\sigma_{r_g}}\theta\right) \\ &= \Phi\left(\alpha_{1g}\left(\theta - \alpha_{0kg}\right)\right), \end{aligned} \tag{1}$$

when using the reparameterization $\alpha_{1g} = -\alpha_{1g}^{'} / \sigma_{r_g}$ and $\alpha_{0kg} = \left(c_{kg} - \alpha_{0g}^{'}\right) / \alpha_{1g}^{'}$. Function $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. A logistic version of the graded response model can be derived from the assumption that residual r_g is distributed according to the logistic distribution. Alternative models for ordered multicategorical data are described in Kubinger (1989) and Baker (1992).

The inverted-U relationship suggests that the response times are related to the latent trait θ. However, it is obvious that θ is not the only systematic influence on the time needed to give the response. Response times also depend on further individual characteristics like reading speed, mental speed or impulsivity. Subsuming all these additional influences under the second latent trait ω that represents general work pace, a log-normal model can be formulated that relates the log response time in item g to the latent trait θ and the work pace ω via

$$\log\left(t_g\right) = \beta_{0g} + \beta_{1g} \cdot \mathrm{f}_g\left(\theta\right) + \beta_{2g} \cdot \omega + e_g\,, \tag{2}$$

where residual e_g is a normally distributed random variate with expectation of zero and variance of $\sigma_{e_g}^2$. By using the log transformation of the response times, the model amounts to an accelerated failure time model, a popular approach to the analysis of event times (Wei, 1992). For a more theoretical justification of the log transformation see van der Linden (2009a). In the accelerated failure time model the parameters can be interpreted as follows. Intercept β_{0g} determines the general time demand of an item and depends on the item length, wording and other characteristics of the item. The regression coefficients β_{1g} and β_{2g} control the strength of the relation between the latent traits and the response times. These parameters can be interpreted as accelerating factors, accelerating the time scale by the factor $\exp\left(\beta_{1g} \cdot \mathrm{f}_g\left(\theta\right) + \beta_{2g} \cdot \omega\right)$. The crucial component in Equation 2 is function $\mathrm{f}_g\left(\theta\right)$ that underlies the inverted-U relationship. Different assumptions about the form of $\mathrm{f}_g\left(\theta\right)$ can be made, leading to different models. In the

following, four forms of $f_g(\theta)$ will be considered. These forms can be classified into two classes, namely item-person distance models and scale location models.

Item-person distance models

Findings from experimental psychology, see for example Maddox, Ashby, and Gottlob (1998) as well as Ashby and Maddox (1994), suggest that a decision is hard for an individual in case the latent agreement y_g^* is close to the decision criterion employed by the individual. This observation motivates the item-person distance models, which are based on the assumption that $f_g(\theta)$ is some sort of distance measure between the location of the item and the location of the individual on the latent trait continuum. The inverted-U relationship is thereby due to the formatting of the internal, continuous response to the response format imposed by the item. Different definitions of the distance measure $f_g(\theta)$ can be used:

The approach of Ferrando and Lorenzo-Seva (2007b): Ferrando and Lorenzo-Seva (2007b) locate the item in the point of the latent trait continuum that corresponds to an average response, that is, to that value $\theta_{0.5}$ of the latent trait θ that corresponds to an expected response of 0.5 when the responses options are transformed to the range (0,1). The predictor of the log response times is $f_g(\theta) = |\theta - \theta_{0.5}|$. As pointed out in the introduction, the approach requires that the thresholds of the rating scales are equidistant.

The approach based on the item thresholds c_{kg} : In the latent agreement interpretation of the graded response model, the distance between the threshold c_{kg} and the systematic component of the latent agreement $\alpha_{0g}' + \alpha_{1g}'\theta$ is $\left|c_{kg} - \left(\alpha_{0g}' + \alpha_{1g}'\theta\right)\right|$. This quantity is proportional to $\left|\alpha_{0kg} - \theta\right|$, such that the distance can alternatively be defined with the reparameterized model parameters. It is straightforward to use the minimal distance as a predictor of the log response time, such that in this case $f_g(\theta) = \min_k\left(\left|\alpha_{0kg} - \theta\right|\right)$.

Scale location models

It is well known that the existence of strong self-schemata can facilitate the processing of information about the self. The effect of an articulated self-schema for response times in personality scales was demonstrated by Holden, Fekken, and Cotton (1991), who found that mean decision times for endorsed items were negatively correlated with relevant self-report scale scores, whereas mean decision times for rejected items were correlated positively with the corresponding self-report scale scores. This data pattern suggests that it is the absolute location of the individual on the trait continuum that determines the response time: Individuals with a strong self-schema are able to process self-information faster. Models based on this assumption will be denoted as scale location models. Two versions of such models are considered, namely:

The approach of Ranger and Ortner (2011): Based on the findings of Holden et al. (1991), Ranger and Ortner (2011) developed a model that relates the expected log response time to the probability of the given response. Although the original model was supposed for binary items it is straightforward to generalize the model to rating scales with several response options. This can be implemented by assuming that $\mathrm{f}_g(\theta) = \mathrm{P}(\mathrm{y_g}|\theta)$, where $\mathrm{P}(\mathrm{y_g}|\theta)$ is the probability of the given response, which is determined by the graded response model in Equation 1.

The approach based on a quadratic relationship: As an alternative to the approach of Ranger and Ortner (2011) one can simply assume that the expected response time declines at both ends of the trait continuum. Such a relationship can be modeled with the quadratic function $\mathrm{f}_g(\theta) = \lambda_{2g}\theta^2 + \lambda_{1g}\theta + \lambda_{0g}$. The parameters λ_{2g} to λ_{0g} are ordinary regression coefficients. Note that when using the quadratic relationship the parameter β_{1g} in Equation 2 is not needed any longer. Although the model is similar to the approach of Ferrando and Lorenzo-Seva (2007b), there is an important difference. The quadratic relationship is more flexible in choosing the location of the trait level θ with the longest response time.

Comparison of the models

Two aspects of the different models have to be discussed. First, their ability to account for the inverted-U relationship and second their relevance for psychological assessment. Both aspects will be addressed in the following.

The relation between θ and $\mathrm{f}_g(\theta)$ is depicted exemplary in Figure 1 for an item with five response options and four thresholds located at $c_{1g} = -2.0$, $c_{2g} = -1.0$, $c_{3g} = 0.0$ and $c_{4g} = 1.0$ and the four different models. The item thresholds are equidistant as it is assumed in the model of Ferrando and Lorenzo-Seva (2007b). Note that the functions $\mathrm{f}_g(\theta)$ in the plots are not the original functions, but were linearly transformed in order to have approximately the same range. This is possible as the response time model in Equation 2 contains an intercept (β_{0g}) and a regression parameter (β_{1g}) that can account for such a transformation. All models can reproduce the inverted-U relationship in case the regression coefficient β_{1g} is negative. Overall, the predictions of the different models are similar. The quadratic model is the most general model. The location of the minimal value can be chosen freely in this model. This is different to the alternative models where the location is determined by the item parameters of the graded response model. This flexibility however is bought with additional parameters.

All models are relevant for psychological assessment. As the response times depend on the latent trait θ, this relationship can be used in order to infer the latent trait from the response times. Thereby, the response times provide additional information about the latent trait value; see the empirical application described later for a demonstration of this effect. The response times also support the calibration of the graded response model. In two models, namely the threshold model and the model of Ranger and Ortner (2011), the response time model shares parameters with the graded response model. Thereby, when calibrating the graded response model and the response time model jointly, one can

reduce the standard error of estimation in the parameters of the graded response model; see the empirical application below for an example with real data.

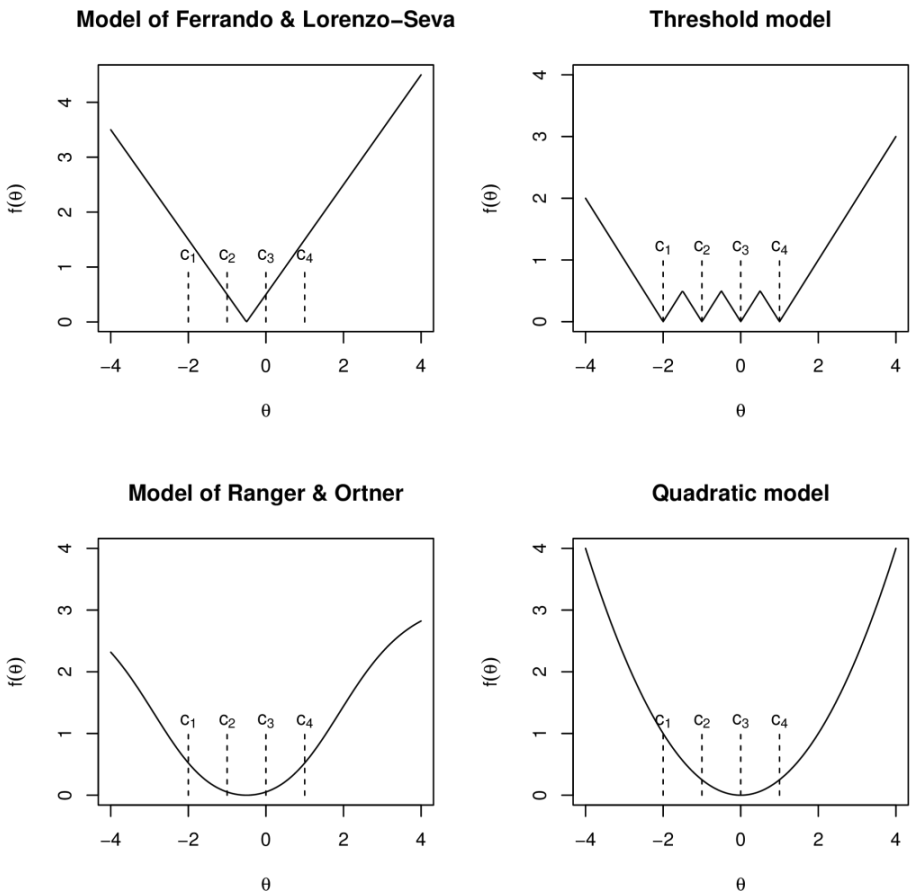

Figure 1:
Plot of the different functions $f_g(\theta)$ representing the inverted-U relationship in the different models for different values of the latent trait θ and an item with four thresholds located at $c_1 = -2.0$, $c_2 = -1.0$, $c_3 = 0.0$ and $c_4 = 1.0$. Note that the range of the functions was rescaled to approximately the same span.

A model for the joint distribution of the responses and response times in tests

A model for the joint distribution of the responses and response times in a test consisting of *G* items follows from the assumption of conditional independence typically made in latent trait models. This assumption comprises two aspects. First, the assertion that conditional on the two latent traits θ and ω the responses and response times from different items are independent. And second, the assertion that the responses and response times in the same item are independent. On the whole, these assumptions imply that the joint distribution of the responses $\mathbf{y}' = [y_1,...,y_G]$ and the response times $\mathbf{t}' = [t_1,...,t_G]$ of an individual in a test consisting of *G* items follows as the product of the single distributions

$$\mathrm{f}(\mathbf{y},\mathbf{t} \mid \theta,\omega) = \prod_{g=1}^{G} \mathrm{P}(y_g \mid \theta)\, \mathrm{f}(t_g \mid \theta,\omega), \tag{3}$$

where $P(y_g \mid \theta)$ is the response probability defined by the graded response model given in Equation 1 and $\mathrm{f}(t_g \mid \theta,\omega)$ is the density of the log-normal distribution corresponding to Equation 2.

Model estimation

In latent trait models not only the item parameters are unknown but also the latent traits of the individuals. The joint estimation of both quantities, item parameters and latent traits, is not feasible, as the number of unknown latent traits grows with the sample size, a fact that invalidates the consistency of standard maximum likelihood estimators. The problem of inconsistent estimators can be resolved with conditional maximum likelihood estimation. In conditional maximum likelihood estimation the dependency of the likelihood function on the unknown latent traits is removed by conditioning on a sufficient statistic for the latent trait. The resulting conditional distribution of the data does not depend on the latent traits anymore and can be used for maximum likelihood estimation of the item parameters. This approach is denominated as conditional maximum likelihood estimation. Conditional maximum likelihood estimation is possible in the Rasch model, where the required sufficient statistic is the sum score of the test (Kubinger, 1989). The response time model of Scheiblechner (1979) also allows for conditional maximum likelihood estimation when conditioning on the total test time. Unfortunately, not all latent trait models can be estimated by conditional maximum likelihood estimation. The approach requires the existence of an observable sufficient statistic, that is, a sufficient statistic that does not depend on the item parameters itself. In case of the graded response model (and the proposed response time model given in Equation 3), no such observable quantity exists, such that conditional maximum likelihood estimation is not possible.

One solution to this problem is the strategy to use the marginal response and response time distribution. The marginal distribution of the responses and the response times follows from Equation 3 after integrating the conditional distribution $\mathrm{f}(\mathbf{y},\mathbf{t}|\theta,\omega)$ over

the distribution $f(\theta,\omega)$, which describes the distribution of the latent traits in the population of the test takers. Marginal maximum likelihood estimation is based on this marginal distribution; those parameter values are chosen that maximize the likelihood function corresponding to the marginal distribution. One method to find this maximum is the so called expectation conditional maximization (ECM) algorithm of Meng and Rubin (1993). The implementation of the ECM-algorithm proposed in this manuscript closely follows the estimation approach that was suggested by Ranger and Kuhn (2012) for the model of Ferrando and Lorenzo-Seva (2007a). As the general approach has already been described in Ranger and Kuhn (2012) its implementation will only be sketched here. Like the ordinary EM-algorithm, the ECM-algorithm consists of two steps, the E-step and the M-step. The M-step can be divided into several substeps, where the item parameters are updated sequentially. This is the difference to the standard EM-algorithm.

E-step

In the E-step of the EMC-algorithm one determines the conditional expectation of the log-likelihood function when conditioning on the observed responses and response times. First, one has to specify the distribution $f(\theta,\omega)$ of the latent traits in the population of the potential test takers. In the following this distribution is assumed to be a bivariate standard normal distribution with coefficient of correlation ρ. And second, one needs preliminary values for the unknown parameters, that is, for the parameters of the graded response model, for the parameters of the response time model and for the correlation coefficient ρ.

Let $\boldsymbol{\gamma}$ represent the vector of the unknown parameters and denote by $\boldsymbol{\gamma}'$ some preliminary values. Having observed the responses $\mathbf{y}_i$ and the response times $\mathbf{t}_i$, $i = 1,\ldots,N$, of the N individuals in a test of G items, the conditional expectation of the log-likelihood function follows from Equation 3 as

$$\sum_{i=1}^{N} \mathrm{E}\left[\log\left(f\left(\mathbf{y}_i,\mathbf{t}_i|\theta,\omega;\boldsymbol{\gamma}\right)\right)|\mathbf{y}_i,\mathbf{t}_i;\boldsymbol{\gamma}'\right] = \sum_{i=1}^{N}\int\int\left[\log\left(f\left(\mathbf{y}_i,\mathbf{t}_i|\theta,\omega;\boldsymbol{\gamma}\right)\right)\right] f\left(\theta,\omega|\mathbf{y}_i,\mathbf{t}_i;\boldsymbol{\gamma}'\right)\partial\omega\partial\theta \quad (4)$$

The distribution $f\left(\mathbf{y}_i,\mathbf{t}_i\,|\,\theta,\omega;\boldsymbol{\gamma}\right)$ depends on the specific version of the response time model of course. However, irrespective of the exact model, the integral can be simplified because the inner integral over ω has a closed form solution. The components $\log\left(f\left(\mathbf{y}_i,\mathbf{t}_i|\theta,\omega;\boldsymbol{\gamma}\right)\right)$ of the log-likelihood function are linear functions of ω and ω^2. This follows from the fact that the conditional distribution $f\left(\log(t_{gi})\,|\,\theta,\omega\right)$ of each logarithmized response time is just a normal distribution and all proposed models are linear in ω. As a consequence, the conditional expectation of the log-likelihood function is a function of $\sum_{i=1}^{N}\mathrm{E}\left(\omega\,|\,\mathbf{y}_i,\mathbf{t}_i;\boldsymbol{\gamma}'\right)$ and $\sum_{i=1}^{N}\mathrm{E}\left(\omega^2\,|\,\mathbf{y}_i,\mathbf{t}_i;\boldsymbol{\gamma}'\right)$. When conditioning on the latent trait θ, the joint distribution of the response times $\mathbf{t}_i$ and the second latent trait ω follow a multivariate normal distribution. Therefore, one can first determine the condi-

tional expectations $\sum_{i=1}^{N} \mathrm{E}\left(\omega \mid \theta, \mathbf{y}_i, \mathbf{t}_i; \boldsymbol{\gamma}'\right)$ and $\sum_{i=1}^{N} \mathrm{E}\left(\omega^2 \mid \theta, \mathbf{y}_i, \mathbf{t}_i; \boldsymbol{\gamma}'\right)$, which can be given in closed form due to well known properties of the multivariate normal distribution. And then, one integrates over the latent trait θ. No closed form solution exists for this integral but it can be approximated numerically by Gauss Hermite quadrature (Stroud, 1971). The strategy of replacing ω and ω^2 by their conditional expectation reduces the two dimensional integral over ω and θ to a one dimensional integral over θ. This saves a large amount of computing time.

M-step

The M-step is based on the conditional expectation of the log-likelihood function given in Equation 4. As the latent trait values have been integrated out, the conditional expectation of the log-likelihood function is a function of the item parameters $\boldsymbol{\gamma}$ only. In the M-step one improves the parameter estimates by setting them to the values that maximize the expected log-likelihood function. One advantage of the EM-algorithm is that these values can be determined itemwise, such that a time consuming search over a high-dimensional parameter space is avoided. Contrary to the standard EM-algorithm, the M-step of the ECM-algorithms consists of several substeps. In each substep, only a subset of the parameters of each item is improved while the remaining parameters of the item are set to fixed values. That is the reason why these substeps are denominated as conditional maximization (CM) substeps.

The exact implementation of the M-step depends on the variant of the response time model, or to be more specific, on the question whether some parameters are shared by the graded response model and the response time model. I suggest the following updating scheme. In a first substep, the expected log-likelihood function is maximized over the coefficient of correlation ρ, the parameters of the graded response model that are not shared by the response time model and the parameters of the response time model that are not shared by the graded response model. The unknown item parameters that appear in both models are set to the preliminary values that were used during the E-step. Having improved these parameters in the first CM-substep, the expected log-likelihood function is maximized over the remaining item parameters that are shared by the graded response model and the response time model. During this second substep, the improved values are used for those parameters, which have already been updated in the first CM-substep. After the two CM-substeps the M-step is completed.

The improved parameters are used for another E-step, where they replace the preliminary values. Then another M-step follows in order to improve the parameter estimates further. This sequence of E-steps and M-steps is repeated until the parameter estimates do not change considerably anymore. As has been shown by Meng and Rubin (1993), the ECM-algorithm converges to a local maximum of the marginal likelihood function under similar conditions as the ordinary EM-algorithm.

The ECM-algorithm was implemented within the statistical software environment R (R Development Core Team, 2009) for all model variants as suggested. The codes are available on request from the author. The performance of the algorithm was explored in a simulation study. Simulation samples were generated as follows. First, values of the latent traits were drawn from a bivariate normal distribution for fictitious test taker. Then, the responses and response times in a fictitious test of 20 items were generated according to one of the four models proposed before. The item parameters of the different models were chosen in order to mimic the data typically observed in personality questionnaires. Observations were simulated for samples consisting of 1000 and 500 subjects. Altogether 250 simulation samples were generated for each condition (4 models × 2 sample sizes). Each data set was analyzed according to the underlying response time model used for generating the data. The marginal maximum likelihood estimator was implemented as described before. Having estimated the item parameters in each simulation sample, the estimates were inspected with respect to bias and root mean squared error of estimation.

The algorithm was stable and converged in all data sets. Parameter recovery was good for most coefficients. All estimators were virtually unbiased. However, some of the estimators had a rather large standard error. Extreme item thresholds c_k corresponding to very infrequent response categories could not be estimated with high precision, a finding that is well known for the graded response model. Parameter recovery of the parameters corresponding to the response time part of the model was usually good. One exception was the estimator for β_{1g} in the model of Ranger and Ortner (2011), which had a rather large standard error of estimation. See the empirical application, in particular Table 2, for a similar result. Due to space limitations, a more thorough description of further results can not be given. More information, as well as the R codes used for the simulation study can be obtained from the author on request.

Empirical data application

A real data set was analyzed with the four models proposed before in order to compare their model fit under realistic conditions. The data set consisted of responses to the Neuroticism scale of the Spanish version of the Five-Factor Personality Inventory (Rodríguez-Fornells, Lorenzo-Seva, & Andrés-Pueyo, 2001) and of the corresponding response times. The data set has already been analyzed by Ferrando and Lorenzo-Seva (2007b), where a detailed description of the data and the process of data collection can be found. In short, the Neuroticism scale was composed of 20 items, which had to be rated on a scale with five ordered response options. The scale was running from “not at all applicable” to “entirely applicable”. The responses were given by 262 undergraduate students from a Spanish university. Data was collected via a computerized test. Although the original data set also included the responses to an Extraversion scale, the analysis was limited to the Neuroticism data. As in the Extraversion scale the extreme response categories of the rating scale were hardly used by the respondents, it was impossible to fit the graded response model with the necessary precision.

First, the data was preprocessed. Two items (item 3 and item 19) had to be removed as a preliminary analysis with the graded response model had revealed very low coefficients of discrimination (α_{1g}) in the two items. This indicates that the items do not measure the same construct as the remaining items. Then the response times were screened for unusual observations. Using boxplots, 17 subjects could be identified that had unusually large response times. As the extreme responses of these individuals would have had a large impact on the results of the data analysis, they were removed, reducing the sample size to 245 individuals. No further cleaning of the data was undertaken.

Subsequently the graded response model was fitted to the responses, using the `LTM` package of the software environment R (Rizopoulos, 2006). Model fit was evaluated using the marginal tables up to order two, comparing expected and observed frequencies. This analysis did not indicate any evidence for model misfit, such that the graded response model seems to be capable of representing the associations between the responses. Finally, the proposed models for the responses and response times were fitted to the data with the ECM-algorithm described before. In addition to the four models, a variant of the model of van der Linden (2007) was considered. The model of van der Linden (2007) is probably the most popular model for responses and response times in achievement tests. In its original formulation, the model combines a linear factor model for the log response times with the three-parameter logistic model for the responses. Although on first sight the model of van der Linden (2007) is different to the approaches proposed in the manuscript, it can be integrated within the present framework after a slight modification. One simply has to replace the three-parameter logistic model with the graded response model and delete the summand $\beta_{1g}\mathrm{f}_g(\theta)$ in Equation 2. This simplifies the response time model to the standard factor model $\log(t_g) = \beta_{0g} + \beta_{2g}\omega + e_g$. Hence, the model of van der Linden (2007) can serve as a benchmark and allows the decision whether response times in personality tests are related to the target trait θ via the assumed inverted-U relationship. This is a crucial question as it amounts to the question whether response times in personality tests possess construct validity.

After model calibration, the fit of the models was evaluated with QQ-plots. These plots revealed that the log-transformation was not capable of normalizing the response times satisfactorily. Therefore, the reciprocal of the response times was used instead of the log response times. In fact, this transformation had already been utilized by Ferrando and Lorenzo-Seva (2007b) when analyzing the data set. This change improved the fit considerably. Note that the reciprocal transformation inverts the order of the data. Hence, the inverted-U relationship requires positive β_{1g} coefficients.

In general, the model estimates were reasonable, with coefficient β_{1g} being positive in most items, irrespectively of the model considered. This is further support for the existence of the inverted-U relationship in personality scales. In order to identify the best fitting model, the models were compared with respect to Akaike's Information Criterion (AIC). The results can be found in Table 1. Note that smaller values correspond to a better fit of the model.

Table 1:
Akaike Information Criterion (AIC) for Five Different Models in the Neuroticism Data

Model	AIC
Model of Ranger and Ortner (2011)	16160.12
Quadratic model	16179.36
Model of Ferrando and Lorenzo-Seva (2007b)	16185.19
Threshold model	16210.10
Model of van der Linden (2007)	16230.46

All proposed models have a lower AIC than the model of van der Linden (2007), which was included as a benchmark. This is not surprising. The model of van der Linden (2007) was developed for achievement tests and assumes a monotone relation between the expected response time and the latent trait of an individual. The poor performance of the model clearly demonstrates that accounting for the inverted-U relation improves the representation of the data. Among the four proposed model variants that account for the inverted-U relation, the model of Ranger and Ortner (2011) performs best, although the difference to the quadratic model is not tremendous. This slightly suggests that it is the trait value itself that is responsible for the inverted-U relationship and not the item-person distance. However, one must not over-interpret the results. The model of Ferrando and Lorenzo-Seva (2007b) requires equidistant item thresholds, an assumption that clearly was violated in the data set. This misfit of the response model might alone be responsible for the somewhat elevated AIC. The threshold model had the highest AIC. This casts the utility of the model into doubt.

The estimates of the item parameters of the model of Ranger and Ortner (2011) are given in Table 2, as well as their standard errors of estimation. The coefficient of correlation was set to $\rho = 0$ because the two latent traits were almost independent. The standard errors of estimation were calculated by a Monte Carlo method using 500 simulation samples. Thus, the item parameters were estimated for 500 samples generated according to the model of Ranger and Ortner (2011) with the item parameters given in Table 2. The standard errors of estimation reported in Table 2 are just the standard errors of these estimates. A Monte Carlo approach was chosen because it was not clear whether the asymptotic distribution theory of maximum likelihood estimation already works satisfactorily in samples of 245 individuals.

The Monte Carlo simulation did not only provide estimates of the standard error of estimation, but also shed light on the utility of response time modeling. Even though one is not interested in the response times at all it can be beneficial to model responses and response times jointly. This finding was revealed when comparing two estimators of the parameters of the graded response model. First, the marginal maximum likelihood estimator based on the responses only, which is the standard approach to calibrating the graded response model. And second, the marginal maximum likelihood estimator based

Table 2:
Estimates (and Standard Error of Estimation) for the 18 Items in the Neuroticism Scale and the Model of Ranger and Ortner (2011).

Item	α_{01}	α_{02}	α_{03}	α_{04}	α_1	β_0	β_1	β_2	σ_e^2
1	-1.07 (.17)	1.05 (.17)	2.50 (.24)	4.84 (.61)	1.11 (.16)	1.82 (.15)	1.33 (.41)	0.65 (.07)	0.74 (.07)
2	-3.69 (.39)	-0.84 (.23)	1.45 (.23)	3.59 (.35)	2.19 (.25)	2.49 (.18)	1.04 (.41)	0.67 (.07)	0.88 (.09)
3	-2.61 (.30)	0.32 (.22)	2.36 (.27)	4.05 (.38)	2.17 (.25)	2.71 (.15)	0.33 (.31)	0.71 (.07)	0.69 (.07)
4	-2.75 (.25)	-1.29 (.16)	-0.12 (.14)	1.41 (.15)	1.00 (.14)	2.29 (.18)	2.12 (.62)	0.60 (.07)	0.92 (.09)
5	-2.67 (.25)	-0.87 (.15)	0.89 (.16)	2.62 (.25)	0.77 (.13)	2.74 (.24)	0.80 (.76)	0.64 (.09)	1.39 (.12)
6	-0.72 (.24)	1.28 (.26)	2.88 (.35)	4.36 (.44)	2.49 (.30)	2.54 (.14)	1.51 (.26)	0.69 (.07)	1.00 (.10)
7	-1.73 (.25)	0.49 (.22)	2.40 (.27)	3.95 (.38)	2.12 (.24)	2.74 (.14)	0.56 (.31)	0.66 (.07)	0.71 (.07)
8	-4.17 (.44)	-0.98 (.19)	1.99 (.24)	4.55 (.48)	1.55 (.19)	2.19 (.16)	0.97 (.34)	0.55 (.06)	0.70 (.07)
9	-2.59 (.26)	-0.97 (.16)	0.64 (.15)	2.61 (.27)	0.75 (.14)	2.24 (.12)	-0.31 (.41)	0.55 (.05)	0.36 (.04)
10	-1.10 (.16)	0.02 (.14)	1.17 (.16)	2.16 (.21)	1.06 (.15)	3.20 (.19)	2.54 (.63)	0.85 (.09)	1.35 (.13)
11	-2.14 (.27)	0.54 (.21)	2.76 (.31)	5.02 (.52)	2.17 (.25)	2.96 (.18)	0.92 (.41)	0.62 (.08)	1.14 (.11)
12	-2.87 (.26)	-1.26 (.17)	0.61 (.15)	2.25 (.22)	0.82 (.15)	2.92 (.16)	-0.02 (.47)	0.66 (.06)	0.65 (.06)
13	-2.24 (.22)	-0.14 (.16)	1.71 (.19)	3.45 (.37)	0.80 (.14)	3.23 (.20)	0.05 (.57)	0.68 (.08)	0.97 (.10)
14	-0.96 (.14)	0.58 (.13)	2.30 (.23)	3.75 (.46)	0.48 (.13)	2.52 (.22)	1.74 (.71)	0.68 (.07)	0.93 (.09)
15	-0.37 (.17)	1.03 (.18)	2.27 (.23)	4.32 (.50)	1.27 (.17)	2.80 (.13)	0.78 (.30)	0.57 (.07)	0.86 (.08)
16	-3.16 (.30)	-0.84 (.17)	1.09 (.17)	2.67 (.26)	1.21 (.16)	3.09 (.18)	0.56 (.51)	0.76 (.07)	0.94 (.09)
17	-1.21 (.17)	0.48 (.16)	1.89 (.20)	3.07 (.32)	0.81 (.15)	3.61 (.21)	0.06 (.64)	0.88 (.09)	1.19 (.11)
18	-2.89 (.28)	-0.64 (.19)	1.57 (.20)	3.40 (.32)	1.56 (.16)	2.63 (.18)	1.22 (.48)	0.43 (.07)	0.90 (.08)

Note. Calculation of the standard errors of estimation is based on a Monte Carlo simulation with 500 simulation samples. The parameter estimates for the graded response model are corresponding to the linear parameterization $\alpha_1\theta + \alpha_{0\kappa}$ of the model. Note that the correlation ρ between the two latent traits was set to zero.

on the responses and response times, which was described before. Both estimators were used for every simulated sample of the Monte Carlo study. Comparing the root mean squared errors of both estimators revealed that this quantity was about 6.7% lower for the new estimator. This is equivalent to a reduction of the sample size of about 13%. The ratio of the root mean squared error of the new estimator incorporating response time and the root mean squared error of the standard estimator ignoring response time is depicted in Figure 2 for the different parameters. Note that 1 is a reference value: A value larger

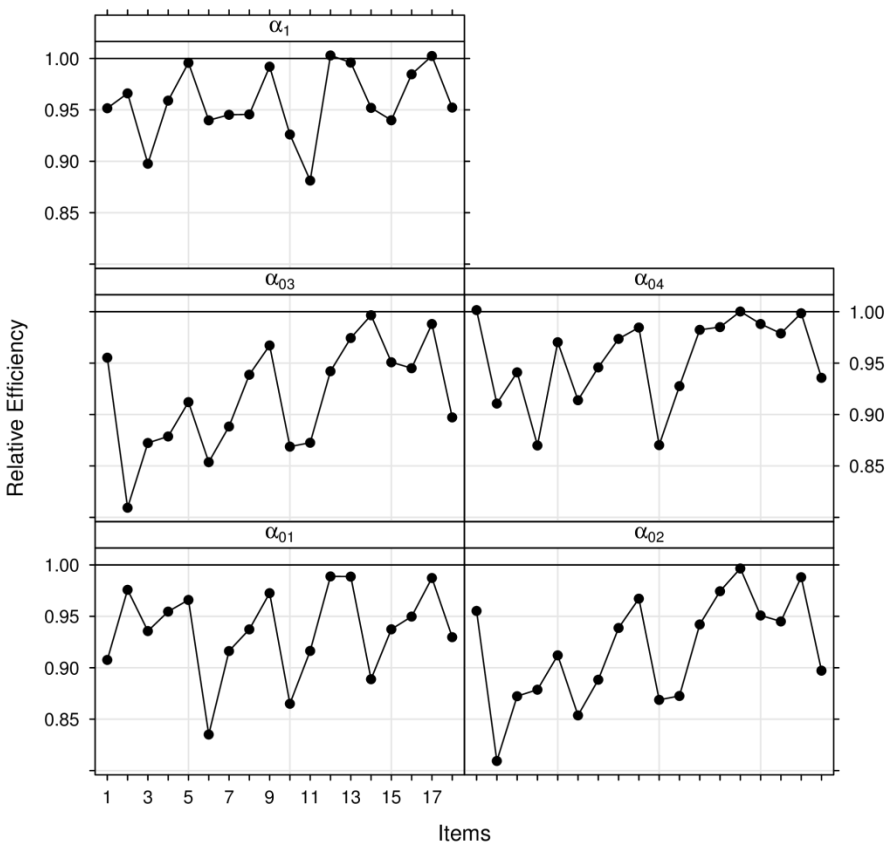

Figure 2:
Ratio of the root mean squared error of two estimators for the parameters of the graded response model in the simulation study. Data was generated according to the model of Ranger and Ortner (2011) and the parameter values as given in Table 2. The ratio represents the relative size of the root mean squared error of the new estimator incorporating response time versus the root mean squared error of the standard estimator ignoring response time. Note that values lower than one denote higher efficiency of the new estimator.

than 1 indicates that the inclusion of response time does not improve the precision, while a value lower than 1 indicates the converse. As can be seen, jointly modeling responses and response times generally improves parameter estimation in the graded response model.

Having estimated the item parameters, the proposed models can be used as measurement models in order to estimate the latent traits. As can be seen from Equation 2, not only the responses are related to the latent trait θ, but also the response times. This means that the response times provide additional information about θ by increasing the test information beyond the information provided by the responses, that is, the test information according to the graded response model. This reduces the variance when estimating θ. The benefit of using the new model for psychological assessment was explored in a simulation study. In this study, response pattern were generated for a test consisting of 18 items with item parameters as given in Table 2, using the model of Ranger and Ortner (2011). The data was generated for ten different levels of θ, equally spaced between $\theta = -3$ and $\theta = 3$. The second latent trait ω was randomly drawn from the standard normal distribution. Having generated the data, the two latent traits were determined

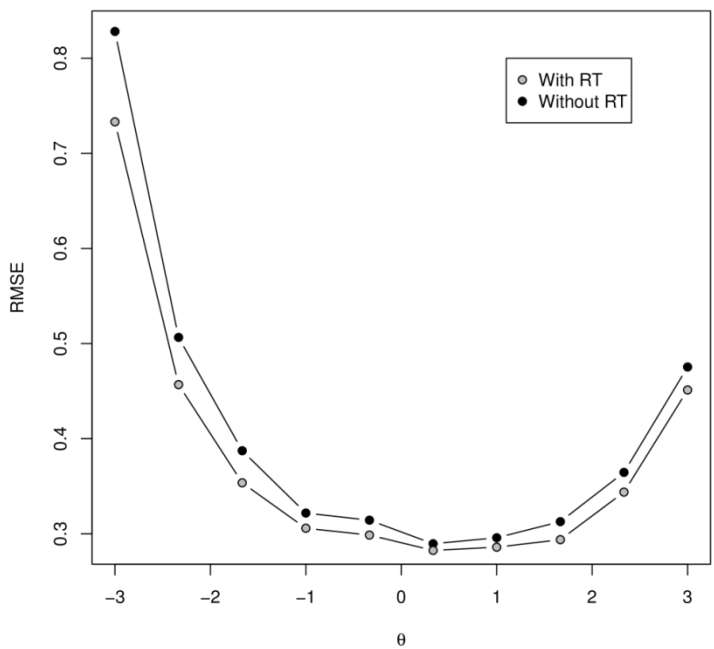

Figure 3:
Root mean squared error for the estimator of the latent trait θ when employing the neuroticism test with the item parameters given in Table 2. With RT denotes the estimator when using the model of Ranger and Ortner (2011). Without RT denotes the estimator when estimation is based on the responses using the graded response model.

with maximum likelihood estimation. A first estimator was based on the responses only, using the graded response model. The second estimator considered responses and response times, using the model of Ranger and Ortner (2011). The item parameters were regarded as known. Data analysis was repeated 1000 times for each trait level. The root mean squared error of the two estimators is given in Figure 3. As can be seen, the consideration of response times always reduces the root mean squared error. The reduction is highest for very low trait levels.

Discussion

There is a large body of evidence that the time needed to respond to items of a personality questionnaire is related to the trait level on an individual: Average response times are shorter for more extreme items (Dunn, Lushene, & O'Neil, 1972; Hanley, 1962; Kuncel, 1973; Rogers, 1973) and more extreme responses (Casey & Tryon, 2001). When factor analyzing questionnaire data, one can observe an inverted quadratic relationship between the factor score of a respondent and the average response time needed by the respondent for responding to indicator items of the corresponding factor (Akrami et al., 2007). Findings of Ferrando (2006) suggest that the distance between the respondent and the item on the underlying trait continuum is correlated with the time needed to respond to the item. These findings are usually subsumed under the label inverted-U relationship.

It is obvious, that any systematic relationship between an observable feature of the respondent behavior and the latent trait can be used for diagnostic purposes by devising a measurement model based on this relationship. Popular measurement models are models from item response theory, which represent the relationship between the given response and the underlying latent trait. However, the inverted-U relationship suggests that not only the responses, but also the response times possess construct validity. One only needs a measurement model relating the response times to the latent trait in order to exploit this extra source of information about the latent trait. Unfortunately, such models barely exist for multi-categorical rating scales. The sole model proposed up to now is the approach of Ferrando and Lorenzo-Seva (2007b). However, a good principle of data analysis is never to fall in love with just one model (McCullagh & Nelder, 1983). Alternative models might be equally useful and equally supported by the data. This is especially true as the psychological mechanism behind the inverted-U relationship is unclear. Summing up, it is far too early to restrict the attention to just one model. What is needed are studies that compare different approaches, as this might help in understanding the true nature of the inverted-U relationship.

In the actual manuscript several models were compared, among them a version of the original model of Ferrando and Lorenzo-Seva (2007b), but also alternative models with different assumptions about the relation between the latent trait and the expected response time. All models are intended to account for the inverted-U relation somehow. The different models can crudely be classified into two classes, item-person distance models and scale location models, although one should not over-interpret the meaning

suggested by this distinction. The empirical application clearly demonstrates the superiority of models that account for the inverted-U relationship. The question about the best model can not be answered as unambiguously. In the actual study, the best model was the model of Ranger and Ortner (2011), which relates the probability of the given response to the expected response time. Probable responses are given faster than unexpected (improbable) ones. As unusual response times and outliers have been removed from the data set before analyzing the response times, the data pattern can not result from distortions like rapid guessing or daydreaming. It seems that extra cognitive effort is needed to overcome typical reactions. However, it is wise to interpret the results from a single study with care. More research is needed in order to assess the generalizability of the present findings.

As has been shown in the manuscript, modeling responses and response times has several benefits. Considering response times improves the estimation of the item parameters of the graded response model and the estimation of the latent trait. The model might also be useful for other purposes like the identification of unusual responses, inappropriate response styles or untraited individuals. These applications will be addressed in future research.

Acknowledgment

The author would like to thank Pere J. Ferrando for the kindness of providing the data set.

References

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in statistics, Volume I* (pp. 599–624). New York: Springer.

Akrami, N., Hedlund, L., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-U effect revisited. *Personality and Individual Differences*, *43*, 611–618.

Ashby, F., & Maddox, W. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*, 423–466.

Baker, F. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Casey, M., & Tyron, W. (2001). Validating a double-press method for computer administration of personality inventory items. *Psychological Assessment*, *13*, 521–530.

Dunn, T., Lushene, R., & O´Neil, H. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology*, *39*, 381–387.

Ferrando, P. (2006). Person-item distance and response time: An empirical study in personality measurement. *Psicologica, 27*, 137–148.

Ferrando, P., & Lorenzo-Seva, U. (2007a). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543.

Ferrando, P., & Lorenzo-Seva, U. (2007b). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research, 42*, 675–706.

Fischer, G. (1989). Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells [Specific objectivity: An epistemological foundation of the Rasch model]. In K. Kubinger (Ed.), *Moderne Test Theorie [Modern test theory]* (2nd ed., pp. 87–111). Weinheim: Beltz.

Furneaux, W. (1952). Some speed, error and difficulty relationships within a problem-solving situation. *Nature, 170*, 37–38.

Hanley, C. (1962). The "difficulty" of a personality inventory item. *Educational and Psychological Measurement, 22*, 577–584.

Holden, R., Fekken, G., & Cotton, D. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment, 3*, 111–118.

Kubinger, K. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Current status and critical acclaim of probabilistic test theory]. In K. Kubinger (Ed.), *Moderne Test Theorie [Modern test theory]* (2nd ed., pp. 19–85). Weinheim: Beltz.

Kuiper, N. (1981). Convergent evidence for the self as a prototype: The "Inverted-U RT Effect" for self and other judgments. *Personality and Social Psychology Bulletin, 7*, 438–443.

Kuncel, R. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement, 33*, 545–563.

Lavergne, C., & Vigneau, F. (1997). Response speed on aptitude tests as an index of intellectual performance: A developmental perspective. *Personality and Individual Differences, 23*, 283–290.

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

MacLennan, R., Jackson, D., & Bellantino, N. (1988). Response latencies and the computerized assessment of intelligence. *Personality and Individual Differences, 9*, 811–816.

Maddox, W., Ashby, F., & Gottlob, L. (1998). Response time distributions in multidimensional perceptual categorization. *Perception and Psychophysics, 60*, 620–637.

McCullagh, P., & Nelder, J. (1983). *Generalized linear models*. New York: Chapman and Hall.

Meng, X., & Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*, 267–278.

R development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria: Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Rafaeli, S., & Tractinsky, N. (1991). Time in computerized tests: A multitrait, multimethod investigation of general-knowledge and mathematical-reasoning on-line examinations. *Computers in Human Behavior*, *7*, 215–225.

Ranger, J., & Kuhn, J.-T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*, 214–231.

Ranger, J., & Ortner, T. (2011). Assessing personality traits through response latencies using IRT. *Educational and Psychological Measurement*, *71*, 389–406.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Rodríguez-Fornells, A., Lorenzo-Seva, U., & Andrés-Pueyo, A. (2001). Psychometric properties of the Spanish adaptation of the Five Factor Personality Inventory. *European Journal of Psychological Assessment*, *17*, 133–145.

Rogers, T. (1973). Toward a definition of the difficulty of a personality item. *Psychological Reports*, *33*, 159–166.

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

Seiwald, B. (2003). Antwortformat [response format]. In K. Kubinger & R. Jäger (Eds.), *Schlüsselbegriffe der psychologischen Diagnostik [key terms in psychological assessment]* (pp. 23–28). Weinheim: Beltz.

Stroud, A. (1971). *Approximate calculation of multiple integrals*. Englewood Cliffs: Prentice-Hall.

Thurstone, 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.

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

van der Linden, W. (2008). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics*, *31*, 5–20.

van der Linden, W. (2009a). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*, 378–394.

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

van der Linden, W., & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251-265.

van der Maas, H., Molenaar, D., Maris, G., Kievit, R., & Boorsboom, 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*, 339–356.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62.

Wei, L. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, *11*, 1871–1879.