Psychological Test and Assessment Modeling, Volume 54, 2012 (2), 128-148

# The case of dependency of responses and response times: A modeling approach based on standard latent trait models

*Jochen Ranger*[1] *& Tuulia Ortner*[2]

**Abstract**

When modeling responses and response times in tests with latent trait models, the assumption of conditional independence between responses and response times might be too strong in the case that both data are gained from reactions to the same item. In order to account for the possible dependency of responses and response times from the same item, a generalization of the model of van der Linden (2007) is proposed. The basic idea consists in the assumption of a latent continuous response that underlies the observed binary response. This latent response is assumed to be correlated with the corresponding response time. The main advantage of this approach consists in the fact that the marginal models for responses and response times follow well known, standard latent trait models. Model estimation can be accomplished by marginal maximum likelihood estimation. The adequacy of the estimation approach is demonstrated in a small scale simulation study. An empirical data application illustrates the practicability of the approach in practice.

Key words: item response theory, response time, log-normal distribution, conditional independence

[1] *Correspondence concerning this article should be addressed to:* Jochen Ranger, PhD, Martin-Luther-University Halle-Wittenberg, Institute for Psychology, Universitätsplatz 10, 06108 Halle, Germany; email: jochen.ranger@psych.uni-halle.de

[2] University of Salzburg, Austria

## 1. An approach to account for the dependency of responses and response times in tests

Due to the computerized application of tests, item response times are widely available today. Therefore, it is not surprising that studies on the meaning or utility of response times in tests have come into the focus of psychological research (Schnipke & Scrams, 2002; van der Linden, 2009). One field of research addresses the question, whether it is possible to extend latent trait models to both, the responses and the response times of individual test takers. Such models are attractive as they provide the opportunity to include response times into the measurement of individual characteristics (Ferrando & Lorenzo-Sevas, 2007; van der Linden, 2008; Ranger & Ortner, 2011). However, care must be exercised when formulating latent trait models for responses and response times. Standard latent trait models might not be appropriate for all fields of application. This is the case when responses and response times gained from the same item are more closely related than responses and response times from different items. As will be shown later, ignoring this extra association in the data can lead to noticeable distortions of certain parameter estimates. Referring to these considerations, the following manuscript proposes an approach to account and test for the possible dependency between responses and response times in the same item.

## 2. Approaches to modeling responses and response times in tests

In general, latent trait models intend to model dependencies between observable quantities (Bartholomew & Knott, 1999). These models are based on the assumption of latent traits that are supposed to represent the entirety of all systematic influencing factors of the observable quantities. In the framework of item response models, this reasoning leads to the local independence assumption, which states that responses from different items are independent when conditioning on the underlying latent traits.

When modeling the joint distribution of responses and response times in tests, it is tempting to assume local independence as well. This comprises four aspects of local independence: The local independence of the responses, the local independence of the response times, the local independence of the responses and response times from different items and the local independence of the response and response time in the same item. While conditional independence is reasonable for data from different items, it is rather controversial when it comes to the response and response time in the same item (Thissen, 1983). As both quantities can be seen as the result of the same response process, they might share common influences that have not been accounted for when using the latent traits as conditioning variables.

The possible causes for a violation of the fourth facet of local independence are numerous. In achievement tests it is well known that individuals can increase their speed of responding at the cost of response accuracy, a phenomenon called speed accuracy trade-off. Generally speaking, a speed accuracy trade-off can be seen as a negative

relation between the level of ability and work pace, at which the individual is able to operate. As Linden (2009) has pointed out, the existence of a speed accuracy trade-off does not contradict a latent trait model with constant latent traits, as long as individuals choose a level for ability and work pace before beginning the test and maintain this level throughout working. However, it seems possible that individuals do not rely totally on a stable choice during the test, but unsystematically fluctuate slightly around their chosen level over different items. This is a potential source of local dependency between item responses and response times. Additional sources of local dependency are random fluctuations of attention (Pieters & van der Ven, 1982). With reference to personality scales, research has revealed the average response latencies increase if a single item possesses an emotional evocative character that is independent from the trait measured (Temple & Geisinger, 1990; Tyron & Mulloy, 1993). It seems probable that such specific arousal evoked by single items also affects the test taker's response process for this particular item. Therefore, when modeling responses and response times, the used model should allow for violations of the local independence assumption in data from the same item.

## 2.1 A classification of responses and response time models

In the next paragraphs, different approaches to the joint distribution of responses and response times will be discussed. The following starting point will be used. Let the joint distribution of the responses and the response times in a test depend on two latent traits, namely ability $\theta$ and work pace $\omega$, and some item parameters $\gamma_g$. The item parameters of the item response model will be denoted as $\beta_g(\gamma_g)$ and the item parameters of the response time model as $\alpha_g(\gamma_g)$, however the dependency on $\gamma_g$ will only be stated when necessary. When assuming conditional independence of observations from different items, the joint distribution of the responses $\boldsymbol{x} = [x_1, \ldots, x_G]$ and response times $\boldsymbol{t} = [, t_1 \ldots, t_G]$ can be stated most generally for a test taker as

$$\mathrm{f}(\boldsymbol{x}, \boldsymbol{t} \mid \theta, \omega; \gamma) = \prod_{g=1}^{G} \mathrm{f}(x_g, t_g \mid \theta, \omega; \gamma_g), \tag{1}$$

where $G$ denotes the number of the items, $x_g$ is the response to item $g$, $t_g$ is the corresponding response time and vector $\gamma = [\gamma_1, \cdots, \gamma_G]$ represents the parameters of the different items.

Different approaches can be chosen in order to specify the joint distribution $\mathrm{f}(x_g, t_g \mid \theta, \omega; \gamma_g)$ of the response and the response time in a single item (Bloxom, 1985). In the simplest case, the assumption of conditional independence can be applied to single items. In this case one can factor $\mathrm{f}(x_g, t_g \mid \theta, \omega; \gamma_g)$ as $\mathrm{f}(x_g \mid \theta, \omega; \beta_g(\gamma_g))\mathrm{f}(t_g \mid \theta, \omega; \alpha_g(\gamma_g))$, such that standard latent trait models can be used for the responses and the response times. Of course, further simplifications like $\mathrm{f}(x_g \mid \theta; \beta_g(\gamma_g))\mathrm{f}(t_g \mid \omega; \alpha_g(\gamma_g))$ might be more reasonable if one assumes that the

responses and the response times depend on different latent traits. This approach has been advocated by Thissen (1983) and van der Linden (2007).

Alternatively, one can factor $f(x_g, t_g \mid \theta, \omega; \gamma_g)$ as $f(x_g \mid \theta, \omega; \beta_g(\gamma_g))f(t_g \mid x_g, \theta, \omega; \alpha_g(\gamma_g))$. Again simplifications like $f(x_g \mid \theta; \beta_g(\gamma_g))f(t_g \mid x_g, \omega; \alpha_g(\gamma_g))$ might be more reasonable. This approach has been investigated by van der Linden and Glas (2010) and revealed excellent power to detect even minor violations of conditional independence.

And finally, one can factor $f(x_g, t_g \mid \theta, \omega; \gamma_g)$ as $f(x_g \mid t_g, \theta, \omega; \beta_g(\gamma_g))f(t_g \mid \theta, \omega; \alpha_g(\gamma_g))$ with the possible simplification of $f(x_g \mid t_g, \theta; \beta_g(\gamma_g))f(t_g \mid \omega; \alpha_g(\gamma_g))$. This approach has been proposed by van Breukelen (1991) and Verhelst, N. (1997).

The question addressing the adequate strategy for response time modeling should be answered empirically in each case, depending on the characteristics of the data. Nevertheless, some of the proposed models might be more preferable as a first choice from a theoretical point of view. The available publications show that responses from tests can be modeled with standard item response models when ignoring response time. This implies that the marginal response distribution

$$f(x_g \mid \theta, \omega; \beta_g(\gamma_g)) = \int f(x_g, t_g \mid \theta, \omega; \gamma_g) dt_g \tag{2}$$

should follow (a potentially bidimensional version of) a standard item response model. Likewise, response times in tests have been modeled with standard latent traits models for a long time (Scheiblechner, 1979; van der Linden, 2006). Therefore, the marginal response time distribution

$$f(t_g \mid \theta, \omega; \alpha_g(\gamma_g)) = \sum_{x_g} f(x_g, t_g \mid \theta, \omega; \gamma_g) \tag{3}$$

should also be a standard response time model. So, when setting up a new model, it would be desirable that the corresponding marginal distributions of responses and response times are known latent trait models, as this is what we would expect from empirical findings.

In fact, a similar claim has already been made by Ip (2002) for item response models that account for the dependency between responses. Referring to the different approaches described above, only the model of Verhelst et al. (1997) fulfills this claim. The model of Verhelst et al. (1997) however assumes exponentially distributed response times. The exponential distribution implies a constant hazard rate and possesses the memoryless property and therefore might be a rather unrealistic model for data sets. As a possible solution, we propose an alternative model that is based on the log-normal distribution. The log-normal distribution is known to fit real data remarkably well (van der Linden, 2009). In the following paragraphs we will introduce this model that can be regarded as a generalization of the approach of van der Linden (2007), with the slight modification that we use the two-parameter probit model whereas van der Linden (2007) used the three-parameter probit model for the responses.

## 3. A model for the joint distribution of responses and response times

The model for the joint distribution of responses and response times in a test is introduced in two steps. First, it is described how responses and response times are distributed in a single item when conditioning on the latent traits $\theta$ and $\omega$. At this level of the model, no assumption of conditional independence will be made. Second, the joint distribution of the responses and response times from different items will be derived.

### 3.1 The distribution of responses and response times in a single item

A standard model for binary responses in tests is the two-parameter probit model (Lord & Novick, 1968, p. 365). This model can be derived from the assumption that the binary response to an item rests on a continuous but unobservable response (Baker, 1992, p. 8). Let $\theta$ be the ability level of an individual and let the unobserved latent response $z_g$ to item $g$ depend on $\theta$ only. More specifically, it has to be assumed that conditionally on ability $\theta$ the latent response $z_g$ is normally distributed with expected value

$$\mathrm{E}(z_g \mid \theta; \beta_{0g}, \beta_{1g}) = \beta_{0g} + \beta_{1g}\theta \tag{4}$$

and variance $\sigma^2_{z_g} = 1$. The quantities $\beta_{0g}$ and $\beta_{1g}$ are item parameters, $\beta_{0g}$ reflecting the difficulty of an item and $\beta_{1g}$ being the item discrimination. Whenever the latent response $z_g$ exceeds the threshold zero, the observable item response is positive, otherwise it is negative. Or more formally, $x_g = 1$ when $z_g \geq 0$ and $x_g = 0$ when $z_g < 0$. In this case, the distribution of the observed response $x_g$ can be derived as a binomial distribution with success probability

$$\mathrm{P}(x_g = 1 \mid \theta; \beta_{0g}, \beta_{1g}) = \int_0^\infty \mathrm{f}(z_g \mid \theta; \beta_{0g}, \beta_{1g})\mathrm{d}z_g = \Phi\left(\beta_{0g} + \beta_{1g}\theta\right), \tag{5}$$

where $\Phi(x)$ denotes the distribution function of the standard normal distribution.

The second component of the model describes the distribution of the response times and is based on the log-normal distribution. Log-normal models have been used successfully for response times in tests (van der Linden, 2009). Such a response time model follows from the assumption that conditionally on work pace $\omega$ the logarithm of the response time $t_g' = \log(t_g)$ is normally distributed with expected value

$$\mathrm{E}(t_g' \mid \omega; \alpha_{0g}, \alpha_{1g}) = \alpha_{0g} + \alpha_{1g}\omega \tag{6}$$

and the variance $\sigma^2_{t_g'} = \alpha_{2g}$ independent of the test taker's characteristics. Again, $\alpha_{0g}$ and $\alpha_{1g}$ are item parameters, $\alpha_{0g}$ reflecting the general response time level of an item and $\alpha_{1g}$ accounting for the strength of the relationship between work pace and the response time.

Within this framework, the assumption of conditional independence within an item can easily be abandoned by allowing for a correlation between the latent response $z_g$ and the log response time $t'_g$. In this case, the distribution of $z_g$ and $t'_g$ follows a bivariate normal distribution with expected values according to Equation (4) and Equation (6) and correlation $\rho_g$, which accounts for the dependency of responses and response times in a single item. As a consequence, the joint distribution of the observable response $x_g$ and the log response time $t'_g$ is

$$\mathrm{f}(x_g, t'_g \mid \theta, \omega; \alpha_g, \beta_g, \rho_g) = \int_{-\infty}^{\infty} \mathrm{I}(z_g, x_g)\mathrm{f}(z_g, t'_g; \mu(\theta, \omega, \alpha_g, \beta_g), \Sigma(\alpha_g, \rho_g))\mathrm{d}z_g. \qquad (7)$$

In Equation (7), function $\mathrm{I}(z_g, x_g)$ is an indicator function with $\mathrm{I}(z_g, 1) = 1$ when $z_g > 0$, $\mathrm{I}(z_g, 0) = 1$ when $z_g < 0$ and zero elsewhere. Function $\mathrm{f}(z, t'; \mu, \Sigma)$ is a bivariate normal distribution with mean vector $\mu = \mu(\theta, \omega, \alpha_g, \beta_g)$ given by Equation (4) and Equation (6) and covariance matrix $\Sigma = \Sigma(\alpha_g, \rho_g)$ with diagonal elements $\Sigma_{11} = 1$, $\Sigma_{22} = \alpha_{2g}$ and off-diagonal elements $\rho_g \sqrt{\alpha_{2g}}$. Equation (7) can easily be generalized to polytomous items and the graded response model by slightly modifying the indicator function.

The proposed model in Equation (7) is a variant of the model of van der Linden (2007). In the model of van der Linden (2007), the probability of a correct solution is given by a three parameter logistic model $\mathrm{P}(x_g = 1) = c_g + (1 - c_g)\Phi(\beta_{1g}(\theta - \beta_{0g}))$ and the response times are distributed according to a log-normal distribution with $\mathrm{E}(t'_g \mid \omega; \alpha_{0g}) = \alpha_{0g} - \omega$. Contrary to the present model (see Equation (6)), the model of Linden (2007) contains a restriction of the different $\alpha_{1g}$ parameters to the same value. The present model avoids this assumption as such constraints are unusual in factor analysis, but it always is possible to implement it in case it is justified by the data set. The major difference between the two models is the assumption of conditional independence between the response and the response time in the same item, which is made by van der Linden (2007) but not in the present model.

There are several reasons to follow the approach proposed in this manuscript. The structure of Equation (7) accounts for the fact that responses and response times can often be modeled separately by unidimensional standard latent trait models. The applicability of unidimensional models to responses and response times clearly excludes the existence of additional, neglected common traits that have not been taken into consideration and that are responsible for remaining associations between responses and response times. However, even though one can exclude the presence of traits that influence all items, one still can assume specific factors that influence the response and response time in just one item, thereby causing a correlation between $z_g$ and $t'_g$. The presence of such specific factors leaves the validity of the unidimensional response model and the unidimensional response time model unaffected. The assumption of specific factors resembles models for testlets, where the dependency of items based on the same content is similarly modeled by assuming a testlet specific factor (Wainer, Bradlow, & Wang, 2007; Li, Bolt, & Fu, 2006). Contrary to the present model however, the specific influences in the testlet model affect items measuring the same trait.

The psychological interpretation of the item specific factor depends on the kind of the test and the context of testing. As outlined in the introduction, it could represent random fluctuations in the speed accuracy level of an individual, account for the effects of isolated random guessing or be the consequence of the specific arousal evoked by a single item.

### 3.2 The distribution of responses and response times in a test

Responses and response times in a single item are based at least in part on the same cognitive process. As a consequence, the assumption of conditional independence seems not realistic. Reactions to different items however do not share the same response process. Therefore, the assumption of conditional independence is plausible for responses and response times from different items. Let $f(x_g, t'_g \mid \theta, \omega; \alpha_g, \beta_g, \rho_g)$ be the distribution of the response and the response time of a test taker in item $g$. According to the conditional independence assumption, the joint distribution of the latent traits and the responses and response times in the $G$ items of the test can be stated as

$$f(\boldsymbol{x}, \boldsymbol{t}', \theta, \omega; \gamma, \rho_{\theta\omega}) = \prod_G f(x_g, t'_g \mid \theta, \omega; \alpha_g, \beta_g, \rho_g) f(\theta, \omega; \rho_{\theta\omega}). \quad (8)$$

In Equation (8), the distribution $f(\theta, \omega; \rho_{\theta\omega})$ is the distribution of the latent traits in the population of the potential test takers. As in the original model of van der Linden (2007), this is a bivariate normal distribution with zero means, unit variances and coefficient of correlation $\rho_{\theta\omega}$. In this aspect the model resembles an oblique factor model.

## 4. Estimating item parameters

Having observed the responses and response times of $N$ test takers, the unknown item parameters can be estimated according to the marginal maximum likelihood approach or the limited information approach. In limited information estimation one first estimates the tetrachoric correlation matrix between the responses, the correlation matrix between the response times and the biserial correlation matrix between responses and response times. Using these correlation matrices, the model parameters can be estimated with standard software for structural equation models by allowing for correlated residuals in responses and response times from the same item. However, as limited information estimates are not efficient, marginal maximum likelihood estimation is preferred in this manuscript. We therefore propose an algorithm that can generally be used for response and response time modeling and might be useful for other response time models as well.

Marginal maximum likelihood estimates can be found by an application of the expectation maximization (EM) algorithm (Rubin, 1976; McLachlan & Krishnan, 1997). It is well known that the unknown latent traits can be considered as missing data. This idea is the basis for the application of the EM algorithm to the estimation of item parameters in the two parameter logistic model (Bock & Aitkin, 1981) or to the

estimation of the loadings in the linear factor model (Rubin & Thayer, 1982). However, to estimate the parameters of the proposed model, it is advantageous to introduce another type of missing data. The item characteristic curve of the two-parameter probit model can be justified by the assumption of a latent continuous response $z_g$, which underlies the observed binary response $x_g$, see the explanations above. In fact, the exact value of the latent response $z_g$ cannot be observed as it only is known whether this variable exceeds the threshold zero or not. Therefore, this latent response can also be considered as missing data. Although not immediately apparent, the estimation of the item parameters can be simplified by pretending that the latent responses to the different items are known. In fact, this approach is similar to the technique of data augmentation, which has been applied in Markov Chain Monte Carlo estimation of item response models (Albert, 1992).

Let $\boldsymbol{z}_i = [z_{i1},\ldots,z_{iG}]$ be the latent responses and $\boldsymbol{t}'_i = [\log(t_{i1}),\ldots,\log(t_{iG})]$ be the log response times of the $i$-th individual from altogether $N$ test takers. Simplifying the notation slightly as $\mu_{ig} = \mu(\theta_i,\omega_i,\alpha_g,\beta_g)$ and $\Sigma_g = \Sigma(\alpha_g,\rho_g)$ and using the nomenclature $\boldsymbol{y}_{ig} = [z_{ig},\log(t_{ig})]'$ and $\lambda_i = [\theta_i,\omega_i]'$, the relevant kernel of the complete log-likelihood function can be written as

$$\mathrm{LL} = \sum_{i=1}^{N}\left[\sum_{g=1}^{G}\log\left[\frac{1}{|\Sigma_g|^{1/2}}\right] - \frac{1}{2}[(\boldsymbol{y}_{ig}-\mu_{ig})'\Sigma_g^{-1}(\boldsymbol{y}_{ig}-\mu_{ig})] + \log\left[\frac{1}{|\Sigma_{\theta\omega}|^{1/2}}\right] - \frac{1}{2}[\lambda_i'\Sigma_{\theta\omega}^{-1}\lambda_i]\right], \quad (9)$$

where $\Sigma_{\theta\omega}$ denotes the variance covariance matrix of the latent traits. The complete log-likelihood function is a function of the unknown item parameters and of sufficient statistics of the missing data, that is, the latent traits and the latent responses. A more accessible version of the complete log-likelihood function as well as a list of the unknown sufficient statistics is given in the Appendix. In case of known latent observations, the maximization of the complete log-likelihood function would be straightforward. However, as the latent variables and likewise the sufficient statistics are not known, the complete log-likelihood function can not directly be used to estimate the item parameters.

One possible solution to this problem consists in the iterated replacement of the unknown sufficient statistics by preliminary values. These values are determined as follows. First, provisional item parameters have to be chosen for the items. With these item parameters it is possible to calculate the conditional expectation of the unobserved sufficient statistics when conditioning on the observed data, that is, on the responses $\boldsymbol{x}_i$ and log response times $\boldsymbol{t}'_i$ of the test takers. For example, the conditional expectation of $\sum_{i=1}^{N} z_{ig}$ can be calculated as

$$\sum_{i=1}^{N}\mathrm{E}(z_{ig}\,|\,\boldsymbol{x}_i,\boldsymbol{t}'_i) = \sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_{ig}\;\mathrm{f}(z_{ig}\,|\,\theta,\omega,\boldsymbol{x}_i,\boldsymbol{t}'_i)\mathrm{f}(\theta,\omega\,|\,\boldsymbol{x}_i,\boldsymbol{t}'_i)\mathrm{d}z_{ig}\,\mathrm{d}\theta\mathrm{d}\omega. \quad (10)$$

Although not explicitly stressed, the different distributions in Equation (10) depend on the provisional item parameters. The inner most integral over $z_{ig}$ is the expectation of a truncated normal distribution that can be stated in closed form. Therefore the triple

integral simplifies to a twofold integration problem. The integral over $f(\theta,\omega|\boldsymbol{x}_i,\boldsymbol{t}'_i)$ can be approximated with Gauss Hermite Quadrature (Stroud, 1971). A more thorough description of the algorithm is given in the Appendix.

After all the unobserved quantities in the complete log-likelihood function have been replaced by their conditional expectation, the resulting equation is a function of the item parameters alone that can be maximized easily. Maximization for the item parameters is not computationally intensive because the time-demanding calculations have been made when calculating the conditional expectations. Having found the maximum, one can use the corresponding item parameter estimates as new provisional values for determining the updated conditional expectations of the unknown sufficient statistics. This sequence of calculating expected statistics and maximization is consecutively iterated until parameter estimates converge.

## 5. Simulation study

In order to test the practicability of the proposed approach, we performed a simulation study. With this simulation two intentions were pursued. First, to demonstrate the applicability of the estimation method. And second, to investigate whether a maximum likelihood ratio test comparing the proposed model with a model assuming independence has proper Type I error rates and power.

### 5.1 Estimation of item parameter

Model estimation was demonstrated with a test of 20 items for samples of 500 and 1000 subjects. This range was supposed to cover the sample sizes reported in empirical applications. Ability and work pace were sampled from a standard bivariate normal distribution. Thereby, a correlation of $\rho_{\theta\omega} = 0.3$ was assumed between ability and work pace. Such correlations between ability and work pace have been reported for achievement tests (van der Linden, 2009). Responses and response times were generated according to the proposed model. The employed item parameters are given in Table 1. The chosen item parameters resembled more or less values that had been found in previous studies.

Four different levels of correlation between the responses and response times were considered, ranging from $\rho_g = 0.0$ for the first five items to $\rho_g = 0.3$ for the last five items. This increase was thought to be a realistic pattern as correlations between responses and response times might increase during the test due to effects of test speededness.

Altogether 500 datasets were generated for every sample size. Preliminary item parameters for the response model were estimated by fitting a probit model to the responses alone. Additionally, preliminary item parameters for the response time model were estimated by factor analyzing the logarithmized response times. These estimates were used as starting values for the EM algorithm. The starting values for the correlations between the responses and the response times were set to zero.

**Table 1:**
True item parameter of the simulated items

| Item | $\beta_0$ | $\beta_1$ | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\rho_g$ |
|---|---|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 3.00 | 0.40 | 0.36 | 0.00 |
| 2 | 1.00 | 1.00 | 3.00 | 0.40 | 0.36 | 0.00 |
| 3 | 1.00 | 1.00 | 3.00 | 0.40 | 0.36 | 0.00 |
| 4 | 1.00 | 1.00 | 3.00 | 0.40 | 0.36 | 0.00 |
| 5 | 0.50 | 1.00 | 3.50 | 0.40 | 0.36 | 0.00 |
| 6 | 0.50 | 1.00 | 3.50 | 0.40 | 0.36 | 0.10 |
| 7 | 0.50 | 1.00 | 3.50 | 0.40 | 0.36 | 0.10 |
| 8 | 0.50 | 1.00 | 3.50 | 0.40 | 0.36 | 0.10 |
| 9 | 0.00 | 1.00 | 4.00 | 0.40 | 0.36 | 0.10 |
| 10 | 0.00 | 1.00 | 4.00 | 0.40 | 0.36 | 0.10 |
| 11 | 0.00 | 1.00 | 4.00 | 0.40 | 0.36 | 0.20 |
| 12 | 0.00 | 1.00 | 4.00 | 0.40 | 0.36 | 0.20 |
| 13 | -0.50 | 1.00 | 4.50 | 0.40 | 0.36 | 0.20 |
| 14 | -0.50 | 1.00 | 4.50 | 0.40 | 0.36 | 0.20 |
| 15 | -0.50 | 1.00 | 4.50 | 0.40 | 0.36 | 0.20 |
| 16 | -0.50 | 1.00 | 4.50 | 0.40 | 0.36 | 0.30 |
| 17 | -1.00 | 1.00 | 5.00 | 0.40 | 0.36 | 0.30 |
| 18 | -1.00 | 1.00 | 5.00 | 0.40 | 0.36 | 0.30 |
| 19 | -1.00 | 1.00 | 5.00 | 0.40 | 0.36 | 0.30 |
| 20 | -1.00 | 1.00 | 5.00 | 0.40 | 0.36 | 0.30 |

The EM algorithm was implemented in R (R Development Core Team, 2009). The integrals in the E-Step were approximated with Gauss Hermite Quadratur and 20 nodes per dimension. The expected log-likelihood function was maximized with the package optim. Note that although the true discrimination coefficients of the items were the same, their estimates were not restricted to the same value. The EM algorithm was ended when item parameter values did not change for more than $0.0008$. The code can be obtained from the authors on request.

Altogether, the EM algorithm worked well as it converged in every sample. On the whole, the true item parameters could be recovered well without bias. The mean and the standard deviation of the estimates are given in Figure 1 for the item correlation parameter $\rho_g$. Results for the remaining parameters can be obtained from the authors.

Additionally, the item parameters were estimated according to the limited information approach. As not all programs for structural equation modeling can handle mixtures of

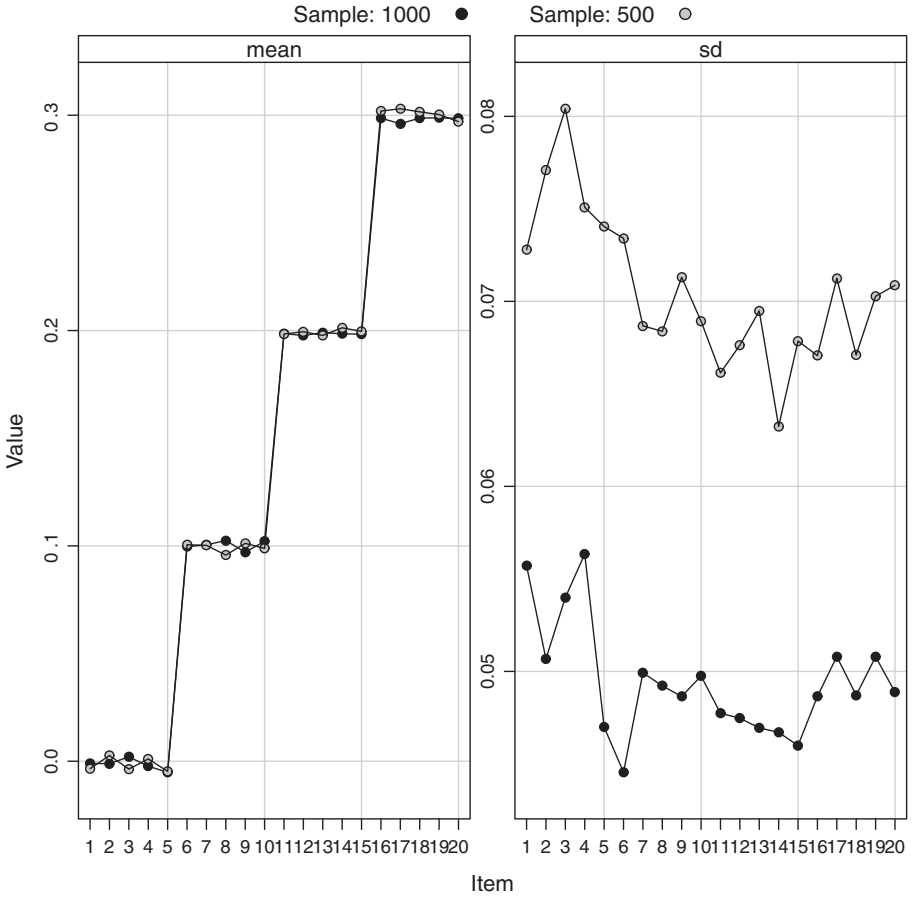


**Figure 1:**
Estimation results for correlation parameter $\rho_g$

continuous and discrete responses, the response times were dichotomized and the item parameters were estimated according to Muthen (1978). However, instead of the weighted least squares approach proposed by Muthen (1978), the item parameters were estimated with unweighted least squares. Item parameter estimates were unbiased, but not as efficient as the corresponding maximum likelihood estimates. The standard deviations of the correlation coefficients for example were about twice as large as the corresponding standard deviations of the maximum likelihood estimates. Therefore, as it is well known, maximum likelihood estimation is the first choice when one is interested in precise estimates.

## 5.2 Testing for independence

The proposed approach offers a framework for testing whether the responses and response times in the same item are independent or not. A first test of this hypothesis is the likelihood ratio test that compares the proposed model with a version where $\rho_g$ is set to zero. As second test may serve a $z$-test that evaluates whether the correlation parameters $\rho_g$ deviate from zero. This test employs Wald's second partial deviations for variance estimation. Both tests were evaluated with respect to power and Type I error rate in a simulation study.

Altogether, three scenarios were investigated. In all scenarios, the responses and response times were generated for 20 items. In the first scenario, the Type I error rate of the tests was investigated. The data was generated according to the item parameters given in Table 1 with the exception that there was no correlation between the responses and response times in a single item. In the second and third scenario, the power of the tests was the quantity of interest. In the second scenario, the data was generated according to the proposed model using the parameter values in Table 1. Local independence is violated as the correlations $\rho_g$ are not zero any more. In the third scenario, a different violation of the conditional independence assumption was considered. This time, the data was generated according to the model of van der Linden and Glas (2010). In this model, the response times are distributed log-normally with expected value according to Equation (6) and the modification that the intercept term $\alpha_{0g}$ is different for positive and negative responses. As a consequence, the response times are distributed differently for positive and negative responses. The motivation for this approach is the observation that wrong answers sometimes take longer than right answers (Thissen, 1983). Contrary to the original version of the model of van der Linden and Glas (2010), the two-parameter probit model was used for the responses instead of the three-parameter probit model.

For every scenario, 500 simulation samples with a size of 500 and 1000 subjects were generated. Two different models were calibrated with marginal maximum likelihood estimation: The proposed model with the correlation parameters $\rho_g$ estimated freely and the restricted version with all correlations $\rho_g$ set to zero. The validity of this restriction was tested with the proposed likelihood ratio test. The model was also calibrated with limited information estimation. The results were used for a $z$-test of the hypothesis that $\rho_g = 0$ for all items. Empirical rejection rates are given in Table 2 for the three scenarios and the two tests. Note that the first two lines of Table 1 (Scenario 1) contain the empirical Type I error rate of the tests whereas the remaining lines contain the power.

As can be seen, both the likelihood ratio test and the $z$-test adhere to the nominal Type I error rate well. The likelihood ratio test however seems to have a slightly reduced nominal Type I error rate in small samples. The power of both tests is excellent. With sample sizes that are usually used for item response models both tests can detect small deviations from the independence assumption with a high probability. This is especially remarkable for the $z$-test that is based on dichotomous responses such that some information of the response times is lost. Interestingly, both tests also can verify model violations when the true model is the model of van der Linden and Glas (2010).

**Table 2:**
Empirical Type I error rates and empirical power

| Scenario | Sample | Test: LR-Test | | | Test: $z$-Test | | |
|---|---|---|---|---|---|---|---|
| | | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| 1 - Independence | 500 | 0.074 | 0.032 | 0.008 | 0.104 | 0.066 | 0.010 |
| | 1000 | 0.104 | 0.046 | 0.010 | 0.126 | 0.052 | 0.012 |
| 2 - Correlation | 500 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.996 |
| | 1000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 - Intercept | 500 | 0.974 | 0.960 | 0.888 | 0.782 | 0.666 | 0.428 |
| | 1000 | 1.000 | 1.000 | 1.000 | 0.974 | 0.960 | 0.866 |

However, in this case the likelihood ratio test is clearly superior to the $z$-test. Therefore, although the implementation of the $z$-test in standard software for structural equation models offers a quick check for model violations, the proposed maximum likelihood ratio test might be preferable.

### 5.3 Consequences of model violations

Testing the local independence assumption is only necessary when parameter estimates are highly distorted in the case of unaccounted dependency. Therefore, in a further simulation study the effects of a misspecified model were investigated. Thereby, response patterns were generated for 10000 subjects and six different tests. The first two tests consisted of 20 items with item parameters as given in Table 1. However, for the first test the correlation between the response and the response time in the same item was set to $\rho_g = 0.50$ for all items, whereas for the second test this correlation was set to $\rho_g = 0.25$. The third and fourth test were generated by using every second item of the first and second test, thus reducing test length from 20 items to 10 items. The last two tests were generated similarly by choosing only every fourth item of the first two tests. In all conditions the correlation between ability and work pace $\rho_{\theta\omega}$ was set to zero.

Having generated the responses and response times, the proposed model was fit to the data with the correlation parameters $\rho_g$ restricted to zero, thus ignoring the extra association between the responses and response times in the same item. Despite fitting the wrong model, the item parameters of the item response and the response time model could be recovered without serious bias. The parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ of the independence model differed maximally by $0.03$ from the corresponding estimates of the full model. However, the estimates of the correlation between ability and work pace were distorted, ranging from $0.018$ up to $0.148$. The exact results are given in Table 3.

**Table 3:**
Estimated $\hat{\rho}_{\theta\omega}$ in a misspecified model depending on the length of the test and the correlation between the response and the response time

| Items | 5 | | 10 | | 20 | |
|---|---|---|---|---|---|---|
| $\rho_g$ | 0.250 | 0.500 | 0.250 | 0.500 | 0.250 | 0.500 |
| $\hat{\rho}_{\theta\omega}$ | 0.075 | 0.148 | 0.036 | 0.075 | 0.018 | 0.035 |

Two effects can be noted. First, the misspecification of the model affects mostly the correlation of the latent traits. This is due to the structure of the model. As the responses and the response times depend on different latent traits, the only way to allow for an association between the responses and response times in the reduced model consists in the admission of a positive correlation between the latent traits. Therefore, it is expectable that the effect of the misspecification is mostly reflected in a distortion of $\hat{\rho}_{\theta\omega}$. Note that a special feature of the proposed model is the fact that the responses and response times follow standard latent trait models when considered separately. Second, the effect diminishes with a growing number of items. This is due to the fact that the number of misspecified associations grows more slowly than the number of correctly specified associations. In a test of $G$ items, only $2G$ associations, the associations between responses and response times in the same items, are misspecified, while the remaining $2G \times 2G - 2G$ associations are correctly specified. This is similar to the concept of essential independence, which is present when the average covariance tends to zero (Junker, 1991). However, as Junker (1991) pointed out, even though the effects on consistency might not be large, they can be considerable on the standard errors of estimates.

From a practitioner's perspective, the correlation between ability and work pace is a key quantity. It can be shown that the accuracy of ability estimates can be improved by jointly considering responses and response times (van der Linden, Klein Entink, & Fox, 2010). The actual gain however depends on the amount of correlation between ability and work pace in the population of the test takers. Therefore, without checking the independence assumption one risks the overestimation of the benefits of response time modeling.

## 6. Empirical data application

To investigate the applicability of the proposed approach to real data, the model was used for data from an application of the German pre-version of the Eysenck Personality Profiler. The German Eysenck Personality Profiler has been published with reduced amount of items by Eysenck, Wilson and Jackson (1998). The Eysenck Personality Profiler measures 21 traits of personality which are consistent with the three major dimensions of personality as defined by Eysenck. In line with the original form of the

questionnaire, three response options were offered including the 'don't know' option. Responses were dichotomized by scoring 'don't know' answers as rejections. Data was collected by Ortner (2008) and consisted of 171 men called up for military service. If they agreed (about 80%), they were tested after the standardized psychological testing conducted by the Psychological Service of the Austrian Armed Forces. Persons were only included if they were evaluated as being motivated by the conductor and if no language problems were known. To reduce faking, the conductor pointed out that all results are handled anonymously and are not evaluated to determine the military appropriateness of the persons. Nevertheless one individual had to be excluded due to unusual short response times. Here only results for the anxious scale will be presented. Although originally the scale consists of 15 items one item had to be excluded as it was rejected by almost all subjects.

First, the response times were logarithmized. Normal Q-Q plots revealed that this transformation was capable of normalizing the data, see Figure 2 for an example.

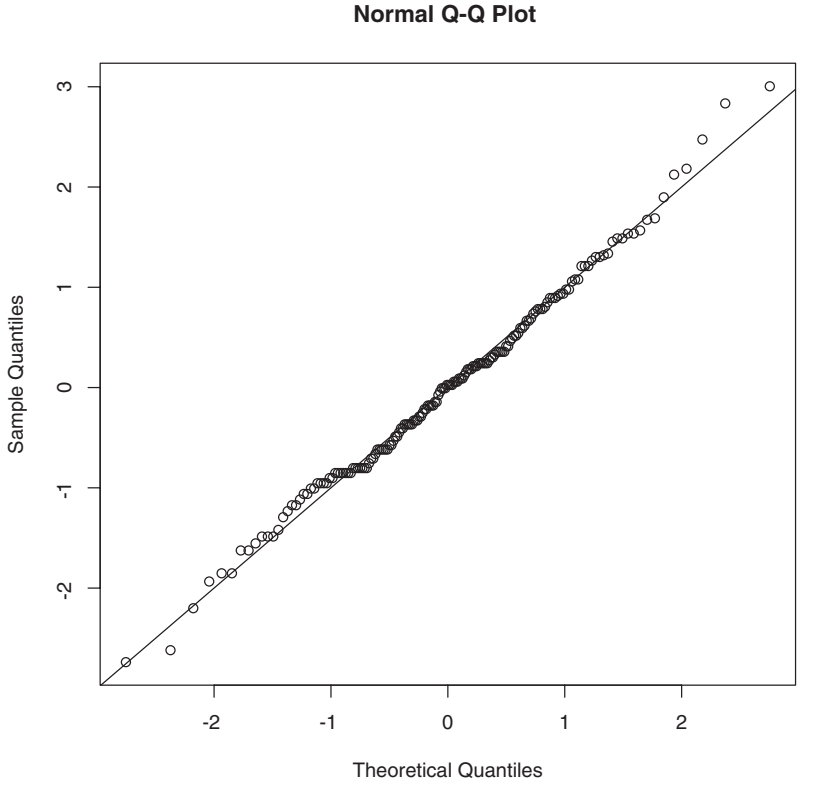


**Figure 2:**
Normal Q-Q plot: Fit of the log-normal distribution to response times of item 13 of the anxious scale

Then, the responses and log response times were analyzed separately. The motivation behind this step was the generation of starting values for the EM algorithm and the assessment of model fit for the marginal models. Responses were analyzed with R and the ltm package (Rizopoulos, 2006). As the probit model is not implemented in the ltm package, a one-parameter logit model was used instead. When appropriately transforming the parameters, logit models are virtually identical to probit models. Assessing model fit via a parametric bootstrap test based on Pearson's chi-squared statistic did not show any evidence for model violations ( $p = 0.44$ ). Therefore, the one-parameter probit model was used for the subsequent analysis because parsimonious models are preferable, especially in samples of moderate size. Replacing the two-parameter probit model with the one-parameter probit model is a change with little implications for the proposed estimation approach. Then, the log response times were analyzed by maximum likelihood factor analysis. A one dimensional model was sufficient to account for the dependence between the log response times ( $p = 0.41$ ). Results of the factor analysis did not change when extreme observations were truncated, such that no truncation of the data seemed necessary. Finally, standardized residuals were calculated and plotted against the estimated factor scores (Bollen & Arminger, 1991). These plots did not show any systematic violation of the assumption of linearity and variance homogeneity. Summing up, it seemed reasonable to use the one-parameter probit model and the linear factor model for the marginal distribution of responses and log response times.

In the next step, the item parameters of the joint distribution of responses and log response times were estimated using the proposed EM algorithm. The resulting parameter estimates are given in Table 4. Ability and work pace were correlated with $\widehat{\rho}_{\theta\omega} = 0.246$. When the original model of van der Linden (2007) was used, almost the same estimates resulted. The estimate of the correlation between the latent traits slightly increased to $\widehat{\rho}_{\theta\omega} = 0.262$. The remaining estimates (the coefficients of correlation $\rho_g$ excluded) were identical up to the second decimal place. This indicates that no extra association of the responses and response times in the same item has to be considered.

In order to test for $\rho_g = 0$, a likelihood ratio test was used. In this test the effect of restricting every parameter $\widehat{\rho}_g$ to zero was evaluated. This test yielded a non-significant test statistic of $\chi^2 = 11.49$ ( $\mathrm{df} = 14$, $p = 0.65$ ). Therefore, the results indicate that the assumption of conditional independence can be extended to observations from the same item.

Finally, the independence model, the model with correlated responses and response times and a version of the model of van der Linden and Glas (2010) based on the one-parameter probit model were compared with respect to AIC. This comparison yielded values of $4628.50$ for the independence model, of $4645.00$ for the proposed model and of $4646.61$ for the model of van der Linden and Glas (2010). Again findings suggest the independence model can describe the data best. The actual model and the model of van der Linden and Glas (2010) revealed similar model fit with a slight advantage for the actual model.

## 7. Discussion

Due to the popularity of computer administered tests, interest in item response times and their possible applications is growing. Whereas not for all applications joint models for responses and response times are needed, some applications depend on them crucially. This is always the case when response times are incorporated into the estimation of the unknown trait level, see for example van der Linden et al. (2010).

When jointly modeling responses and response times Achilles' heel is the question whether responses and response times in the same item can be considered as independent when conditioning on the latent traits. Although there is evidence for the independence (van der Linden & Glas, 2010) this assumption could be too strong for some tests. As it is well known that ignoring association in the data can distort confidence intervals (Ip, 2002) and parameter estimates (Wang, Cheng, & Wilson, 2005), it is a wise choice to check this assumption and to account for the dependency when it exists.

In the present article a new method was proposed that can account for the dependency between responses and response times in the same item. This can be done by only slightly generalizing the model of Linden (2007). The first advantage of this approach consists in the fact that the resulting marginal models are standard latent trait models. This is especially advantageous as the marginal response and response time distributions have been analyzed routinely with these models. A second advantage is the possibility to implement the actual approach in standard software for structural equation models. Although in this case, suboptimal limited information estimation has to be used, the results might be good enough for practical applications. Only when the exact amount of extra correlation between responses and response times has to be determined or one is interested in very high power, more complex estimation routines are recommended.

An empirical data application demonstrated the usefulness of the proposed approach in practice. Although most applications of response time models can be found in the field of achievement tests, in this study the applicability of the model to a personality test was shown. Therefore, the presented findings might have more implications than the mere checkout of a new model. In fact, it was shown that the model of van der Linden (2007) can be used for data from personality tests, such that it is not limited to the field of achievement tests. And second, equally to results from van der Linden and Glas (2010) findings indicate that the assumption of conditional independence in tests might not be totally unjustified in some cases. These findings might increase the popularity of response time modeling in the future.

# 8. Appendix

## 8.1 Sufficient statistics of the complete log-likelihood function

The relevant kernel of the complete log-likelihood function is given in Equation (9). Using Equation (4) and Equation (6), Equation (9) can be written as

$$\begin{aligned} \mathrm{LL} = & \sum_{i=1}^{N}\sum_{g=1}^{G}\left[ \log\left[ (\alpha_{2g}(1-\rho_g^2))^{-\frac{1}{2}} \right] - \frac{1}{2(1-\rho_g^2)}\left[ \frac{(z_{ig}-\beta_{0g}-\beta_{1g}\theta_i)^2}{1} \right.\right. \\ & \left.\left. + \frac{(t'_{ig}-\alpha_{0g}-\alpha_{1g}\omega_i)^2}{\alpha_{2g}} - 2\rho_g \frac{(z_{ig}-\beta_{0g}-\beta_{1g}\theta_i)(t'_{ig}-\alpha_{0g}-\alpha_{1g}\omega_i)}{\sqrt{\alpha_{2g}}} \right] \right] \\ & + \sum_{i=1}^{N}\left[ \log\left[ (1-\rho_{\theta\omega}^2)^{-\frac{1}{2}} \right] - \frac{1}{2(1-\rho_{\theta\omega}^2)}(\theta_i^2+\omega_i^2-2\rho_{\theta\omega}\theta_i\omega_i) \right]. \end{aligned} \tag{11}$$

Expanding Equation (11) and summing over the $N$ test takers reveals that the complete log-likelihood function is a function of the following unobserved sufficient statistics:

$$\sum_{i=1}^{N} z_{ig} \qquad \sum_{i=1}^{N} z_{ig}^2$$

$$\sum_{i=1}^{N} \theta_i \qquad \sum_{i=1}^{N} \theta_i^2$$

$$\sum_{i=1}^{N} \omega_i \qquad \sum_{i=1}^{N} \omega_i^2$$

$$\sum_{i=1}^{N} z_{ig} t_{ig} \qquad \sum_{i=1}^{N} \theta_i \omega_i$$

$$\sum_{i=1}^{N} \theta_i z_{ig} \qquad \sum_{i=1}^{N} \theta_i t'_{ig}$$

$$\sum_{i=1}^{N} \omega_i z_{ig} \qquad \sum_{i=1}^{N} \omega_i t'_{ig}$$

## 8.2 Calculation of conditional expectation of sufficient statistics

The complete log-likelihood function depends on unknown sufficient statistics, which are replaced by their conditional expectation during the E-Step. First, provisional values $\lambda^*$

have to be chosen for the unknown item parameters. Given these preliminary values for the item parameters, the conditional expectation of $\sum_{i=1}^{N} z_{ig}$ is

$$\sum_{i=1}^{N} \mathrm{E}(z_{ig} \mid \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) = \sum_{i=1}^{N} \iiint z_{ig} \, \mathrm{f}(z_{ig} \mid \theta, \omega, \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) \mathrm{f}(\theta, \omega \mid \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) \mathrm{d}z_{ig} \, \mathrm{d}\theta \mathrm{d}\omega. \quad (11)$$

The inner most integral over $z_{ig}$ can be given in closed form. Conditional on the latent traits, the latent response in item $g$ is independent of the responses and the response times from different items, such that $\mathrm{f}(z_{ig} \mid \theta, \omega, \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*)$ can be simplified to $\mathrm{f}(z_{ig} \mid \theta, \omega, x_{ig}, t'_{ig}; \gamma^*_g)$. Conditional on $\theta$ and $\omega$, the joint distribution of $z_{ig}$ and $t'_{ig}$ is a bivariate normal distribution, see Equation (7) for details. Therefore, the conditional distribution $\mathrm{f}(z_{ig} \mid \theta, \omega, t'_{ig}; \gamma^*_g)$ is a normal distribution, with expected value

$$\mathrm{E}\left(z_{ig} \mid \theta, \omega, t'_{ig}; \gamma^*_g\right) = (\beta^*_{0g} + \beta^*_{1g}\theta) + \frac{\rho^*_g}{\sqrt{\alpha^*_{2g}}} \cdot (t'_{ig} - (\alpha^*_{0g} + \alpha^*_{1g}\omega)) \quad (12)$$

and conditional variance $1 - \rho^{*2}_g$. Conditioning finally on the observed response $x_{ig}$ yields a truncated normal distribution with corresponding expected value and variance. The expectation of a truncated normal distribution can be given in closed form. Let $\mathrm{E}(z_{ig} \mid \theta, \omega, t'_{ig}, x_{ig}; \gamma^*_g)$ be the expectation of the truncated normal distribution implied by Equation (12). Using this expectation in case of $z_{ig}$, one can simplify Equation (11) to

$$\sum_{i=1}^{N} \mathrm{E}(z_{ig} \mid \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) = \sum_{i=1}^{N} \iint \mathrm{E}(z_{ig} \mid \theta, \omega, t'_{ig}, x_{ig}; \gamma^*_g) \mathrm{f}(\theta, \omega \mid \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) \mathrm{d}\theta \mathrm{d}\omega. \quad (13)$$

The solution of the two-fold integral can not be given in closed form. However, it can be approximated by Gauss Hermite quadrature. Using cartesian quadrature rules, Equation (13) can be approximated by

$$\sum_{i=1}^{N} \mathrm{E}(z_{ig} \mid \boldsymbol{x}_i, \boldsymbol{t}'_i; \gamma^*) = \sum_{i=1}^{N} \sum_{q_1=1}^{Q_1} \sum_{q_2=1}^{Q_2} \frac{\mathrm{E}(z_{ig} \mid \theta_{q_1}, \omega_{q_2}, t'_{ig}, x_{ig}; \gamma^*_g) \mathrm{f}(\boldsymbol{x}_i, \boldsymbol{t}'_i \mid \theta_{q_1}, \omega_{q_2}; \gamma^*) w_{q_1} w_{q_2}}{\sum_{q_1=1}^{Q_1} \sum_{q_2=1}^{Q_2} \mathrm{f}(\boldsymbol{x}_i, \boldsymbol{t}'_i \mid \theta_{q_1}, \omega_{q_2}; \gamma^*) w_{q_1} w_{q_2}} \quad (14)$$

where summation is over quadrature points $\theta_{q_1}$ and $\omega_{q_2}$ with corresponding weights $w_{q_1}$ and $w_{q_2}$. The remaining sufficient statistics are calculated alike.

## Acknowledgement

We would like to thank the editor and two referees for their helpful and constructive comments, which led to many improvements.

## References


Albert, J. H. (1992). Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling. *Journal of Educational Statistics, 17,* 251-269.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

Bartholomew, D., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika, 50,* 383-397.

Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology, 21,* 235-262.

Eysenck, H., Wilson, C., & Jackson, C. (1998). *Eysenck Personality Profiler (EPP-D).* Frankfurt: Swets.

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

Ip, E. H. (2002). Locally dependent latent trait model and the dutch identity revisited. *Psychometrika, 67,* 367-386.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255-278.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30,* 2-21.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Muthen, B. (1978). Contributions tofactor analysis of dichotomous variables. *Psychometrika, 43,* 551-560.

Ortner, T. (2008). Effects of changed item order: A cautionary note to practioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16,* 249-257.

Pieters, J. P. M., & van der Ven, A. H. G. S. (1982). Precision, Speed, and Distraction in Time-Limit Tests. *Applied Psychological Measurement, 6,* 93-109.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ranger, J., & Ortner, T. M. (2011). Assessing Personality Traits through Response Latencies using item response theory. *Educational and Psychological Measurement, 71,* 389-406.

Rizopoulos, D. (1976). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17,* 1-25.

Rubin, D. B. Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D. B., & Thayer, D. (1982). EM Algorithms for ML Factor Analysis. *Psychometrika, 47,* 69-76.

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

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analysis. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (eds.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-267). Mahwah: Lawrence Erlbaum.

Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals*. Englewood Cliffs: Prentice-Hall.

Temple, D. E., & Geisinger, K. F. (1990). Response latency to computer-administered inventory items as an indicator of emotional arousal. *Journal of Personality Assessment, 54,* 289-297.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Tyron, W. W., & Mulloy, J. M. (1993). Further Validation of Computer-Assessed Response Time to Emotionally Evocative Stimuli. *Journal of Personality Assessment*, *61,* 231-236.

van Breukelen, G., & Roskam, E. (1991). A Rasch model for the speed-accuracy trade of in time-limited tests. In J. Doignon & J. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp.251-271). NewYork: Springer.

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

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

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

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

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

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement, 34,* 327-347.

Verhelst, N., Verstralen, H., & Jansen, M. (1997). A logistic model for time-limit tests. In W. van der Linden & R. Hambelton (Eds.), *Handbook of modern item response theory* (pp. 169-185). NewYork: Springer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65,* 5-27.