Fitting diffusion item response theory models for responses and response times using the R package diffIRT

Molenaar, D.; Tuerlinckx, F.; van der Maas, H.L.J.

DOI
10.18637/jss.v066.i04

Publication date
2015

Document Version
Final published version

Published in
Journal of Statistical Software

Journal of Statistical Software

August 2015, Volume 66, Issue 4. *http://www.jstatsoft.org/*

Fitting Diffusion Item Response Theory Models for Responses and Response Times Using the R Package diffIRT

Dylan Molenaar
University of Amsterdam

Francis Tuerlinckx
University of Leuven

Han L. J. van der Maas
University of Amsterdam

Abstract

In the psychometric literature, item response theory models have been proposed that explicitly take the decision process underlying the responses of subjects to psychometric test items into account. Application of these models is however hampered by the absence of general and flexible software to fit these models. In this paper, we present **diffIRT**, an R package that can be used to fit item response theory models that are based on a diffusion process. We discuss parameter estimation and model fit assessment, show the viability of the package in a simulation study, and illustrate the use of the package with two datasets pertaining to extraversion and mental rotation. In addition, we illustrate how the package can be used to fit the traditional diffusion model (as it has been originally developed in experimental psychology) to data.

Keywords: item response theory, diffusion model, psychometrics, mathematical psychology, R.

1. Introduction

In the behavioral sciences, inferences about traits such as motivation, extraversion, arithmetic ability, and attitudes require measurable indicators for these traits. Commonly, such indicators are responses to tests (e.g., intelligence tests like the Wechsler Adult Intelligence Scales; Wechsler 1997) and questionnaires (e.g., personality questionnaires like the Big-5 Personality Inventory; Digman 1990). To enable inferences about the traits underlying these observed data, the indicators are linked to the trait by specifying a measurement model. The exact mathematical form of the measurement model depends on the distribution of the observed data and the assumed distribution of the trait.

The family of item response theory (IRT) models includes popular measurement models like

the Rasch model (Rasch 1960), the 2-parameter logistic model (2PL; Birnbaum 1968), the graded response model (GRM; Samejima 1969), and the common factor model (Spearman 1904; Thurstone 1931). Traditionally, these measurement models have been formulated purely on basis of desirable statistical properties like sufficiency of the total score for the trait (in the case of the Rasch model, see Fischer 1995), model flexibility (in case of the 2PL and the GRM; see Glas 1999), and linearity between the indicators and the trait (in case of the common factor model, see Bollen 1989). These properties make these measurement models extremely useful for a wide range of applications, e.g., for investigating differential item functioning (Mellenbergh 1989; Meredith 1993), for testing group differences (Jöreskog 1971), for the construction of tests and questionnaires (Kline 1986), for investigating the structure of theoretical constructs (i.e., structural equation modeling; Bollen 1998), and for testing the effects of experimental manipulations (Wicherts, Dolan, and Hessen 2005). Despite these advantages, these measurement models are purely statistical. That is, the traits in these models have no direct connection to the psychological process that allows the respondent to make a certain response to the test or questionnaire items (Borsboom, Mellenbergh, and van Heerden 2004). As argued by van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011) this is problematic for a number of reasons. Specifically, (1) it hampers the investigation of test validity; (2) it obscures the connection between intra-individual differences and inter-individual differences; and (3) it makes the interpretation of the data in terms of substantive processes more ambiguous.

Effort has been devoted to formulate more substantively informed measurement models that explicitly incorporate the underlying psychological process that elicited the response to a given test item. These *process IRT models* mainly draw from process models that already exist in the field of mathematical psychology. For instance, Ranger and Kuhn (2014) proposed a model based on the proportional hazard model (see, e.g., Luce 1986), Tuerlinckx and De Boeck (2005) and Rouder, Province, Morey, Gomez, and Heathcote (2015) proposed extensions of the Race model (Audley and Pike 1965), and Tuerlinckx, Molenaar, and van der Maas (2016),Tuerlinckx and De Boeck (2005), and van der Maas *et al.* (2011) proposed extensions of the so called *diffusion model* (Ratcliff 1978).

The diffusion model is arguably the most popular model for decision making used in experimental psychology. It is a model appropriate for trials in which subjects need to decide between two answer options (e.g., "yes/no", "left/right", etc.). A typical example of a psychological experiment that is suitable for diffusion modeling is the lexical decision task. In this task, subjects have to decide as quickly as possible whether a sequence of letters, shown on a computer screen, form a word (e.g., in the case of "table") or a non-word (e.g., in the case of "telab"). In the diffusion model, it is assumed that once a subject has encoded the letters presented on the screen, information starts accumulating over time in favor of either a "word" response or a "non-word" response. Each response option is characterized by a boundary which quantifies the amount of information that is needed for that option to elicit a response by the subject in favor of that option. If the amount of information for a given option reaches the boundary of that option, the subject starts making a response. The response time is then a function of the average rate with which the information accumulated (drift rate, μ), the distance between the two boundaries (boundary separation, α), and the time needed for encoding of the letters on the screen and the physical responding (non-decision time, Ter), see Figure 1. Diffusion model applications to lexical decision experiments have shown for instance that older subjects have larger values for the non-decision time Ter (Ratcliff, Thapar, Gomez,

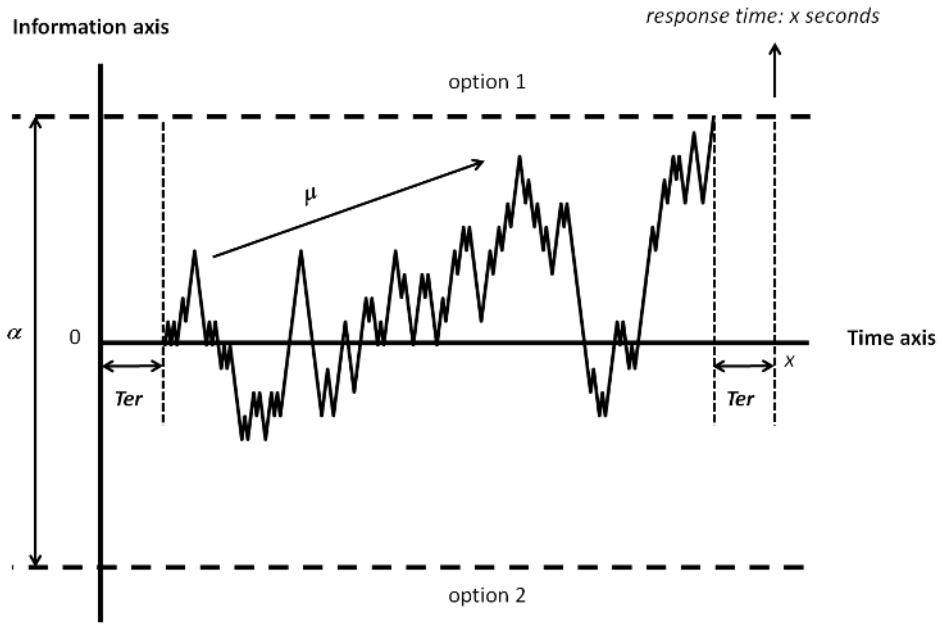

Figure 1: Graphical representation of the traditional diffusion model including the parameter non-decision time, *Ter*, boundary separation, α, and drift rate, μ.

and McKoon 2004) and that instructions that emphasize speed lead to closer boundaries α (Wagenmakers, Ratcliff, Gomez, and McKoon 2008). More extended versions of the diffusion model also include a bias parameter which models the tendency to favor one option over the other irrespective of the stimulus content. However, here we focus on the more basic model as described above.

On its own, the diffusion model is not a measurement model as it does not disentangle person and item characteristics which is the key property of a measurement model (Borsboom and Molenaar 2015). Van der Maas *et al.* (2011) however proposed such a decomposition for personality questionnaires and ability tests separately (see also Tuerlinckx *et al.* 2016). The resulting models are referred to as diffusion IRT models. At present, no general model fit routines are available to fit diffusion IRT models to real data. Van der Maas *et al.* (2011) used a Bayesian model implementation within the **WinBUGS** program (Lunn, Thomas, Best, and Spiegelhalter 2000). However this implementation is specific to the data in the van der Maas *et al.* (2011) study and could thus not straightforwardly be used for different datasets. In addition, it requires advanced programming knowledge to adapt the scripts as they make use of the **WinBUGS** add-on package **wbDEV** (Lunn 2003). Tuerlinckx and De Boeck (2005) and Tuerlinckx *et al.* (2016) used SAS (SAS Institute Inc. 2013) code to fit a specific instance of the diffusion IRT model (for personality data only, see below), which is inflexible as it needs adaptation when there are a different number of items or when there are parameter constraints. In addition, the procedure is slow and the assessment of absolute goodness of fit of the model is not readily possible. Finally, none of the above studies focused on parameter recovery of the different models to establish the feasibility of the estimation procedures.

In the present paper we present a general approach to fitting diffusion IRT models to real data using the package **diffIRT** (Molenaar 2014) within the statistical software environment

R (R Core Team 2015). Package **diffIRT** is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=diffIRT. As compared to the approaches above, the package is easy to use (no programming knowledge is needed besides some basic R understanding like reading in data), the package is reasonably fast, it contains tools to assess absolute goodness of fit, and it is flexible in specifying parameter constraints. In addition, we show in a simulation study that true model parameters are well recovered and that the proposed statistics to assess model fit follow their theoretical distribution. Finally, the package can also be used to fit the traditional diffusion model to data.

The outline of this paper is as follows. First, we introduce the general family of diffusion IRT models together with instances for personality questionnaires and ability tests. Next, we discuss how parameters are estimated and how model fit is assessed in the **diffIRT** package. Next, we present a simulation study to show the viability of the estimation procedure. We then illustrate the use of the package with a single simulated dataset, followed by two applications to real data pertaining to extraversion and mental rotation. In a third application, we illustrate how the **diffIRT** package can be used to fit the traditional diffusion model to data. We end with a discussion on limitations and future possibilities.

2. Diffusion IRT models

In the traditional diffusion model, a given response of subject p, x_{pi}, with response time t_{pi}, is assumed to be originating from the following joint distribution function:

$$h(x_{pi}, t_{pi}) = \frac{\pi}{\alpha^2} \exp[\alpha\mu(x_{pi} - \frac{1}{2}) - \frac{\mu^2}{2}(t_{pi} - Ter)] \times \sum_{k=1}^{\infty} k \sin(\frac{1}{2}\pi k) \exp(-\frac{\pi^2 k^2}{2\alpha^2}(t_{pi} - Ter)) \quad (1)$$

As can be seen, no separation between person and item parameters is involved. That is, in the majority of applications of the diffusion model, subjects are given a large number of items and boundary separation, α, drift rate, μ, and non-decision time, Ter, are estimated for each subject separately assuming that all items are interchangeable.

To make the diffusion model more appropriate as a measurement model, Tuerlinckx and De Boeck (2005) and van der Maas *et al.* (2011) introduced person and item characteristics by separating μ, and α, into a person specific part and an item specific part, i.e.,

$$\mu_{pi} = u(\theta_p, v_i),$$
$$\alpha_{pi} = w(\gamma_p, a_i),$$

where v_i and a_i are the item drift and item boundary parameter of item i, and θ_p and γ_p are the person drift and person boundary parameter of subject p. Choices for the functions $u(.)$ and $w(.)$ could be made on statistical grounds. For instance, Vandekerckhove, Tuerlinckx, and Lee (2011) proposed a multilevel version of the model above, implying additive functions for $u(.)$ and $w(.)$. However, as pointed out by van der Maas *et al.*, substantive considerations lead to different forms of $u(.)$ and $w(.)$ which are not necessarily linear. We will discuss these next.

2.1. D-diffusion

Tuerlinckx and De Boeck (2005) proposed the following functions to decompose drift rate and boundary separation into a person and item contribution:

$$\mu_{pi} = \theta_p - v_i \text{ with } \theta_p, v_i \in \mathbb{R}, \tag{2}$$

$$\alpha_{pi} = \frac{\gamma_p}{a_i} \text{ with } \gamma_p, a_i \in \mathbb{R}^+. \tag{3}$$

Substituting this in Equation 1 and integrating out t_{pi}, the probability of a correct response is given by

$$\mathsf{P}(x_{pi} = 1|\theta_p, \gamma_p) = \frac{\exp[\frac{\gamma_p}{a_i}(\theta_p - v_i)]}{1 + \exp[\frac{\gamma_p}{a_i}(\theta_p - v_i)]} \text{ with } \gamma_p, a_i \in \mathbb{R}^+ \text{and } \theta_p, v_i \in \mathbb{R}, \tag{4}$$

which is referred to as the D-diffusion IRT model (see also Tuerlinckx *et al.* 2016). In this model, the item boundary, a_i, is interpreted as time pressure, the item drift, v_i, as item difficulty, person boundary, γ_p, as response caution, and person drift, θ_p, is interpreted as the actual construct being measured. Application of the model requires both the response time data, t_{pi}, and the response data, x_{pi}. In the case of personality data (e.g., extraversion items "Do you like to meet new people?" with answer options 0: no and 1: yes), $x_{pi} = 0$ corresponds commonly to a response indicative for the lower end of the underlying personality construct, and $x_{pi} = 1$ is indicative for the upper end of the underlying personality dimension. That is a "yes" response, $x_{pi} = 1$, indicates that the subject is more extraverted and a "no" response, $x_{pi} = 0$, indicates that the subject is less extraverted. In the case of ability items however (e.g., an arithmetic problem) $x_{pi} = 0$ indicates an incorrect answer and $x_{pi} = 1$ indicates a correct answer.

Assuming the coding of the observed responses for x_{pi} as discussed above, van der Maas *et al.* (2011) discuss why the D-diffusion model is appropriate for personality data but not for ability data. Specifically the D-diffusion model in Equation 4 predicts:

1. Response times are slowest for $\mu = \theta_p - v_i = 0$ (i.e., no evidence accumulation) and get faster when $\theta_p - v_i$ deviates more from 0. That is, subjects with a low position on θ_p are as fast in responding as subjects high on θ_p.

2. When the item time pressure, a_i, decreases, α increases, which causes subjects on the lower end of the θ-range to have a smaller probability obtaining $x_{pi} = 1$. The subjects on the higher end of the θ-range have a larger probability of obtaining $x_{pi} = 1$.

Note that prediction 1 holds only for personality data: When $\theta_p \approx v_i$ the subject is near the middle of the extraversion continuum of that item. That is, on the question "Do you like to meet new people?", the subject tends a bit toward answering "no" or the subject tends a bit towards answering "yes". A response will take considerably longer as compared to extreme subjects as the subject in the middle needs to carefully consider the item and decide whether it is a "yes" or a "no". Extreme subjects, that is $\theta_p \gg v_i$ and $\theta_p \ll v_i$, are extremely extraverted and introverted respectively, so they will immediately know to respond "yes" or "no" respectively. Note that this is in line with prediction 1. For ability items, subjects with

$\theta_p \gg v_i$ (the high ability subjects), will also respond fast as these subjects have no difficulty solving the arithmetic problem. However, subjects with $\theta_p \ll v_i$ (the low ability subjects) will take considerably longer to answer the item as for them the item is challenging. Note that this is not in line with prediction 1.

In addition, prediction 2 also only holds for personality data: When the time limit of a test decreases, subjects have more time to think about their answer. This will result in more extraverted subjects choosing "yes" (i.e., $x_{pi} = 1$) and more introverted subjects choosing "no" (i.e., $x_{pi} = 0$). Note that this is in line with prediction 2. For ability items, decreasing the time limit will commonly result in a larger probability of correct (i.e., $x_{pi} = 1$) for *all* subjects. Note that this is not in line with prediction 2.

2.2. Q-diffusion

As the D-diffusion model is only suitable for personality data, a different parameterization of the diffusion IRT model is needed for ability tests. Taking into account the above, for ability data, the resulting model should predict: (1) increasing response times for decreasing θ_p (i.e., low ability subjects have more difficulty solving the items), and (2) increasing probability corrects for the whole θ_p-range for decreasing time pressure, a_i. As discussed by van der Maas *et al.* (2011) the following parameterization conforms these predictions:

$$\mu_{pi} = \frac{\theta_p}{v_i} \text{ with } \theta_p, v_i \in \mathbb{R}^+, \tag{5}$$

$$\alpha_{pi} = \frac{\gamma_p}{a_i} \text{ with } \gamma_p, a_i \in \mathbb{R}^+. \tag{6}$$

Substituting in Equation 1 and integrating out t_{pi}, the so-called Q-diffusion model is obtained, i.e.,

$$\mathsf{P}(x_{pi} = 1|\theta_p, \gamma_p) = \frac{\exp(\frac{\gamma_p}{a_i}\frac{\theta_p}{v_i})}{1 + \exp(\frac{\gamma_p}{a_i}\frac{\theta_p}{v_i})} \text{ with } \gamma_p, a_i, \theta_p, v_i \in \mathbb{R}^+, \tag{7}$$

in which the interpretation of the parameters is the same as in the D-diffusion model, that is, a_i, is the time pressure, v_i is the item difficulty, γ_p is the response caution, and θ_p is the actual ability being measured by the test. As a_i, v_i, θ_p, and γ_p are strictly positive, μ has a lower bound of 0 which ensures that response times are slowest for $\mu_p = 0$ and get faster for increasing μ, because of more difficult items (larger v_i) or higher ability (larger θ_p).

2.3. Relation to other models

The diffusion IRT model family has some interesting relations to existing models. Most notably, van der Maas *et al.* (2011) discuss how the parameters from the Q-diffusion model have a relation with the parameters from van der Linden's hierarchical model for responses and response times (van der Linden 2007). Specifically, the time intensity and speed parameter from the hierarchical model are related linearly to $\log \theta_p/\gamma_p$ and $\log v_i/a_i$ respectively for large $\mu \times \alpha$. In addition, the D-diffusion model can be applied to investigate person fit (Meijer and Sijtsma 2001). That is, the D-diffusion model in Equation 4 is equivalent to a two-parameter IRT model with a random discrimination parameter, a_p. Such models are commonly used in assessing person fit (e.g., Strandmark and Linn 1987). Next, a similar

model as the Q-diffusion model in Equation 7 was proposed by Ramsay (1989). Specifically, Ramsay proposed the so-called quotient IRT model which – for two-choice data – is given by: $\text{logit}[\mathsf{P}(x_{pi} = 1|\theta)] = \theta_p/\lambda_i$, where λ_i is an item parameter (see also van der Maas *et al.* 2011). Finally, Tuerlinckx *et al.* (2016) discuss how the D-diffusion model can be seen as a generalized linear latent variable model with two interacting latent variables. Specifically, if γ_p from Equation 4 is rewritten as a normally distributed zero-centered latent variable plus a mean, $\gamma_p^{'} + \kappa$, then it follows that

$$\text{logit}\mathsf{P}(x_{pi} = 1|\theta_p, \gamma_p^{'}) = \frac{\gamma_p^{'} + \kappa}{a_i} \times (\theta_p - v_i) = -\frac{\kappa}{a_i} v_i + \frac{\kappa}{a_i}\theta_p - \frac{v_i}{a_i}\gamma_p^{'} + \frac{1}{a_i}\gamma_p^{'}\theta_p, \tag{8}$$

which is a latent variable model with two latent variables, $\gamma_p^{'}$ and θ_p, and their interaction.

3. The diffusion IRT model package

Using the package **diffIRT** in the R statistical programming environment, the general model given by Equation 1 subject to Equations 2 and 3 for the D-diffusion model and subject to Equations 5 and 6 for the Q-diffusion model can be fitted to responses and response time data. In this section, we discuss the main modeling tools, i.e., the parameter estimation procedure and the assessment of model fit.

3.1. Parameter estimation

The likelihood function

Parameters of the diffusion IRT model are estimated using marginal maximum likelihood (MML; Bock and Aitkin 1981). In this likelihood-based procedure, the person parameters from a given statistical model (in this case θ_p and γ_p) are treated as nuisance parameters. That is, by assuming a distribution for these parameters, they can be integrated out of the likelihood equation. The resulting marginal likelihood is only a function of the item parameters and the population parameters that describe the distribution of the person parameters in the population. In present case, we choose a normal distribution for the person parameters that are in $\mathbb{R}$ (i.e., θ_p in the D-diffusion model, see Equation 2) and a log-normal distribution for the person parameters that are in $\mathbb{R}^+$ (i.e., γ_p and θ_p in case of the Q-diffusion model, see Equations 5 and 6; and γ_p in case of the D-diffusion model, see Equation 3). The log-normal distribution is chosen mainly for reasons of convenience, that is, the log-normal distribution is appropriate for the parameters at hand as this distribution is bounded by zero. In addition, a logarithmic transformation of a log-normal variate has a normal distribution which numerically facilitates the use of Gauss-Hermite quadrature approximation of the integrals in the likelihood function (see below). Note that van der Maas *et al.* (2011) and Tuerlinckx *et al.* (2016) also used a log-normal distribution for the strictly positive person parameters in the D- and Q-diffusion model.

Given the above, the marginal log likelihood of the $N \times k$ matrix with responses $\boldsymbol{X}$ and response times $\boldsymbol{T}$ with elements x_{pi} and t_{pi} respectively, is given for the Q-diffusion model by

$$\ell\left(\boldsymbol{\tau};\boldsymbol{X},\boldsymbol{T}\right)=\sum_{p=1}^{N}\log\iint\limits_{\mathbb{R}}\prod_{i=1}^{k}h\left[x_{pi},t_{pi}|\exp(\gamma_p^*),\exp(\theta_p^*)\right]\times f(\gamma_p^*;\omega_\gamma)\times g\left(\theta_p^*;\omega_\theta\right)d\gamma_p^*d\theta_p^* \quad (9)$$

In this equation, $h(.)$ is given by Equation 1 with μ and α given by Equation 5 and Equation 6 respectively. Functions $f(\gamma_p^*;\omega_\gamma)$ and $g(\theta_p^*;\omega_\theta)$ are normal density functions with mean 0 and standard deviation ω_γ and ω_θ respectively. Note that in the above we used $\theta_p=\exp(\theta_p^*)$ and $\gamma_p=\exp(\gamma_p^*)$ such that θ_p^* and γ_p^* are normal variates and θ_p and γ_p are log-normal variates. τ is a vector of free parameters in the model, i.e., $\tau=[a_1^*,\ldots,a_k^*,v_1^*,\ldots v_k^*,Ter_1^*,\ldots,Ter_k^*,\omega_\gamma^*,\omega_\theta^*]$ in which the star denotes that the corresponding parameter is log-transformed to parameter space $(-\infty,\infty)$ for numerical convenience.

The likelihood function in Equation 9 is given for the Q-diffusion model. However, the likelihood function for the D-diffusion model is straightforwardly obtained by first using Equation 2 for α and Equation 3 for μ in function $h(.)$, and then by dropping the transformation of θ_p, i.e., $\theta_p=\theta_p^*$, and by dropping the transformation of v_i, i.e., $v_i=v_i^*$. That is the likelihood function for the D-diffusion model is given by

$$\ell\left(\boldsymbol{\tau};\boldsymbol{X},\boldsymbol{T}\right)=\sum_{p=1}^{N}\log\iint\limits_{\mathbb{R}}\prod_{i=1}^{k}h\left[x_{pi},t_{pi}|\exp(\gamma_p^*),\theta_p\right]\times f(\gamma_p^*;\omega_\gamma)\times g\left(\theta_p;\omega_\theta\right)d\gamma_p^*d\theta_p \quad (10)$$

with $h(.)$ given by Equation 1 with μ and α given by Equation 2 and Equation 3 respectively.

Evaluation of function $h(.)$ in Equation 9 and Equation 10 can be numerically demanding due to the presence of the infinite sum (see Equation 1). Therefore, in the **diffIRT** package, we evaluate this function using the procedure outlined in Navarro and Fuss (2009). In short, this procedure uses two different infinite series expansions of $h(.)$, one which converges quickly for small response times, and one which converges quickly for larger response times. We refer to Navarro and Fuss (2009) for more details.

Approximation of the integrals

The marginal likelihood function above contains two integrals that do not have a closed form expressions. To enable optimization of this function, we approximate the integrals using Gauss-Hermite quadratures. Within IRT modeling, this is a commonly used approach, e.g., it is used in R package **ltm** (Rizopoulos 2006) and in the software packages Mplus (Muthén and Muthén 2007), **BILOG** (Mislevy and Bock 1990), and **MULTILOG** (Thissen 1991). Using Gauss-Hermite quadratures, the log likelihood function above can be written as

$$\ell\left(\boldsymbol{\tau};\boldsymbol{X},\boldsymbol{T}\right)\approx\sum_{p=1}^{N}\log\sum_{r=1}^{R}\sum_{s=1}^{S}W_{1r}^{'}W_{2s}^{'}\prod_{i=1}^{k}h\left(x_{pi},t_{pi}|N_{1r}^{'},N_{2s}^{'}\right) \quad (11)$$

where

$$W_{1r}^{'}=\frac{1}{\sqrt{\pi}}W_{1r}\quad\text{and}\quad W_{2r}^{'}=\frac{1}{\sqrt{\pi}}W_{2r}$$

and

$$N_{1r}^{'}=\sqrt{2}N_{1r}\quad\text{and}\quad N_{2s}^{'}=\sqrt{2}N_{2s}.$$

That is, the integrals are replaced by weighted sums. Within these summations, N_{1r} and N_{2s} are the so-called "nodes" which are positions on the dimensions of integration and W_{1r} and W_{2s} are the corresponding weights. In addition, R and S denote the number of nodes that are being used for the θ_p^* and γ_p^* dimension, respectively. The nodes and weights can be found in standard tables (Stroud and Secrest 1966). In **diffIRT**, we used the function `gauss.quad` from the R package **statmod** (Smyth, Hu, Dunn, Phipson, and Chen 2015) to obtain the weights and nodes. Generally, the more nodes are used, the better the approximation of the likelihood will be. However, computational demands increase rapidly as, for instance, with 10 nodes for each dimension, $10 \times 10 = 100$ evaluations of $h(.)$ are necessary for each response of a single subject. In the package **diffIRT**, the number of nodes can be set by the user, with default $K = S = 7$. This number of quadrature points showed satisfactory results during the package development. This is also demonstrated below in the simulation study. However, in practice it is advisable to try different numbers of nodes to check the stability of the results. In Equation 11, missing values (`NA`) are allowed in $\boldsymbol{X}$ and $\boldsymbol{T}$. If a given element in $\boldsymbol{X}$ is missing, the corresponding element in $\boldsymbol{T}$ is also assumed treated as missing and vice versa. For missing elements, calculation of the likelihood is skipped.

Optimization and starting values

As the likelihood function is now numerically tractable and efficiently formulated, we move on to discuss optimization of the function. Within the **diffIRT** package, we implemented $-2 \times \ell(\boldsymbol{\tau}; \boldsymbol{X}, \boldsymbol{T})$ for the D- and Q-diffusion model in C code to increase speed of computation. The C code is subsequently linked to R in which we minimize this function using the built-in R function `optim`. We use the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS; see, e.g., Nocedal and Wright 2006, p. 194) which is a quasi-Newton algorithm that uses function evaluations and first-order derivatives. We approximated the first-order derivatives of $-2 \times \ell(\boldsymbol{\tau}; \boldsymbol{X}, \boldsymbol{T})$ with respect to $\boldsymbol{\tau}$ by a finite difference approximation. The spacing of the finite difference between two function evaluations equals `1e-7` by default as this value turned out to be sufficiently small during the package development.

As the procedure might be sensitive to the starting values for the parameter vector $\boldsymbol{\tau}$, these values should be carefully chosen. In the **diffIRT** package, starting values can be chosen by the user. However, by default, starting values are calculated on basis of the EZ-diffusion moment estimators of the diffusion model (Wagenmakers, van der Maas, and Grasman 2007). In the EZ-diffusion approach, α, μ, and Ter from the traditional diffusion model in Equation 1 are calculated using closed form expressions for these parameters. The expressions are derived by equating the expected mean response time, the expected variance of the response times, and the expected proportion correct of the responses to the corresponding observed mean, variance and proportion correct. Here, these expressions are used in **diffIRT** to obtain starting values for $a_i^*, v_i^*, Ter_i^*, \omega_\theta^*$, and ω_γ^* by adapting the R code provided by Wagenmakers *et al.* (2007). That is, item parameters are calculated by applying the EZ-diffusion model on the responses and response times of each item separately. Starting values for ω_θ^* and ω_γ^* are obtained by applying the EZ-diffusion model on the item responses and response times for each subject separately and calculating ω_θ^* and ω_γ^* using the formula for the log-normal distribution (in case of the D-diffusion model, this is not necessary for ω_θ^* as in this model θ_p follows a normal distribution). The resulting values are biased by definition as for boundary separation, α, for instance, it holds that $\alpha = \gamma_p / a_i$ (Equations 3 and 6), i.e., the starting values for a_i are conflated by γ_p. We reduce this bias by multiplying the initial value for a_i by $\mathsf{E}(\gamma_p)$ and the

initial value for v_i by $\mathsf{E}(\theta_p)$ for the Q-diffusion model. In case of the Q-diffusion model, both $\mathsf{E}(\theta_p)$ and $\mathsf{E}(\gamma_p)$ can be calculated using the initial value for ω_γ. In case of the D-diffusion model, only $\mathsf{E}(\gamma_p)$ needs to be calculated from the initial value of ω_γ as $\mathsf{E}(\theta_p) = 0$. This method of correcting the starting values reduces bias but does not eliminate it totally. This is plausible as item and person effects cannot be fully disentangled using the EZ-diffusion model as it assumes interchangeable items and subjects. However, most importantly, the starting values obtained in this way showed good performance. This will be illustrated in the simulation study.

3.2. Assessment of model fit

Assessment of model fit is challenging in models such as the present as a general statistical method to assess the absolute goodness of fit on the responses and response times simultaneously does not exist. We therefore follow Tuerlinckx *et al.* (2016) and assess absolute goodness of fit on the responses and response times separately. In addition, we propose some model fit indices to assess comparative goodness of fit.

Absolute goodness of fit

Responses. For the responses we use the limited-information test for multivariate contingency tables proposed by Maydeu-Olivares and Joe (2005, 2006). This test consists of comparing the observed and expected joint moments of the binary data up to order r. As advocated in the papers by Maydeu-Olivares and Joe, the traditional full-information Pearson $\chi 2$ test statistic (which arises when r equals the total number of joint moments, i.e., when r is equal to k, the number of items), does not follow its hypothetical χ^2-distribution in the case of small cell values. In addition, as the χ^2-statistic needs all k joint moments, this involves calculations of $2^k - 1$ predicted score patterns, which quickly becomes numerically demanding. As Maydeu-Olivares and Joe show, in case of both limited sample sizes (100 subjects) and large sample sizes (2500 subjects), their test – using $r = 2$ or $r = 3$ – outperforms the traditional χ^2-statistic in terms of type I error rate for a two parameter IRT model. We therefore adopt this statistic in the **diffIRT** package. The statistic, M_r, is given by

$$M_r = N(\boldsymbol{p}_r - \boldsymbol{\pi}_r)^\top \boldsymbol{C}_r (\boldsymbol{p}_r - \boldsymbol{\pi}_r) \tag{12}$$

where $\boldsymbol{p}_r$ is the vector of observed joint moments up to order r, e.g., for 2 items and $r = 2$, this vector contains $[p_{x1=1}, p_{x2=1}, p_{x1=1,x2=1}]$, $\boldsymbol{\pi}_r$ is a vector containing the corresponding joint moments predicted by the statistical model given the parameter estimates (obtained with maximum likelihood estimation or another minimum variance estimator), and $\boldsymbol{C}_r$ is given by

$$\boldsymbol{C}_r = \boldsymbol{\Omega}_r^{-1} - \boldsymbol{\Omega}_r^{-1}\boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r^\top \boldsymbol{\Omega}_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r^\top \boldsymbol{\Omega}_r^{-1} \tag{13}$$

where $\boldsymbol{\Omega}_r$ is the covariance matrix of the residuals (i.e., $\boldsymbol{p}_r - \boldsymbol{\pi}_r$ from Equation 12), and $\boldsymbol{\Delta}_r$ is a matrix containing the first-order derivatives of $\boldsymbol{\pi}_r$ with respect to the model parameters. The test statistic M_r in Equation 12 has an asymptotic χ^2-distribution with degrees of freedom equal to

$$df = \sum_{q=1}^{r} \binom{n}{q} - s,$$

where the summation gives the number of observed proportions (i.e., the observed joint moments up to order r) on which the statistic is based, and s will generally be equal to the number of free parameters in the statistical model. See for more details Maydeu-Olivares and Joe (2005). In the **diffIRT** package, the procedure as outlined above is implemented using for $\boldsymbol{\pi}_r$ the proportions of correct responses as predicted by the D- or Q-diffusion model given the MML parameter estimates. It is important to note that the statistic M_r is based on the response data only which implies that s above will not be equal to $3 \times k + 2$ (which is the total number of parameters in the full diffusion IRT model). At the level of the response data, for the Q-diffusion model only, $s = k + r + 2$ parameters are identified, and for the D-diffusion $s = 2 \times k + r + 2$ parameters are identified. Note that this implies that the M_r test is not possible for $r = 1$ as the degrees of freedom will be smaller than 0 (i.e., in that case, we have more parameters than observed statistics). In the simulation study below we show that using the correction above, the M_r statistic follows its theoretical distribution. In calculating M_r within the **diffIRT** package, the order r can be inputted by the user with default $r = 2$ as this number showed satisfactory results in Maydeu-Olivares and Joe (2005).

Response times. For the response times, we use QQ-plots to investigate goodness of fit. That is, the quantiles of the observed response time distribution for a given item, $q_{1i}, \ldots, q_{zi}$, are plotted against the predicted quantiles given the parameter estimates, $\tilde{q}_{1i}, \ldots, \tilde{q}_{zi}$. If the model fits the data, the observed and predicted quantiles are on a straight line. The probability between successive quantiles is given by

$$\frac{1}{z} = \int_{\tilde{q}_{(l-1)i}}^{\tilde{q}_{li}} \left\{ \mathsf{E}(x_{pi}) \times \tilde{h}(x_{pi} = 1, t_{pi}) + [1 - \mathsf{E}(x_{pi})] \times \tilde{h}(x_{pi} = 0, t_{pi}) \right\} dt_{pi}, \tag{14}$$

where $\tilde{h}(.)$ is the joint density of the diffusion model (Equation 1) evaluated at the estimated parameters. By approximating the integral in Equation 14 using the R function `integrate`, $\tilde{q}_{li}$ can be solved for all l and all i by using the R function `uniroot`. This procedure requires an approximate interval which covers the predicted quantiles $(0, c_i)$, where c_i is the upper bound of this interval for item i. As the range of the predicted quantiles it commonly not exactly known, c_i could be loosely chosen. However, choosing a very large value for c_i will generally result in a successful solution by `uniroot`, but will be computationally demanding. Choosing a very small value for c_i will be less computationally demanding, but if c_i is too small, $(0, c_i)$ will not contain the solution and `uniroot` will fail. In **diffIRT**, the upper bound of the interval c_i equals $2 \times \text{MAX}(t_{pi})$ by default. This default was chosen more or less intuitively (i.e., the upper bound should be large enough to cover the solution, but not too large to avoid heavy computational burden) and worked well in the testing stage of the package. If the value is not large enough, an error message will be produced. In this case c_i could be manually raised.

Comparative goodness of fit

As we use MML estimation, the value of the likelihood function at its maximum can be used to calculate various comparative goodness of fit indices. In the **diffIRT** package we include the Akaike's information criterion (AIC; Akaike 1974), the Bayesian information criterion (BIC; Schwarz 1978), the sample size adjusted BIC (sBIC; Sclove 1987), and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, and Linde 2002), see Neale, Boker, Xie, and Maes (2006, p. 93) for the mathematical expressions for these indices. The indices can be used to compare different models in terms of goodness of fit, where it holds for all

indices that a lower value indicates a better model fit. Indices above can be used to compare the diffusion IRT model to different non-nested models. In addition, standard output of **diffIRT** contains $-2 \times \ell(\hat{\boldsymbol{\tau}}; \boldsymbol{X}, \boldsymbol{T})$ evaluated at the estimated parameters, $\hat{\boldsymbol{\tau}}$. This value can be used to conduct likelihood ratio tests between two nested diffusion IRT models. For instance, a full model with estimated parameter vector, $\hat{\boldsymbol{\tau}}_A$, could be compared to a model in which the a_i parameters are constrained to be equal with estimated parameter vector $\hat{\boldsymbol{\tau}}_0$. To do so, $-2 \times (\ell(\hat{\boldsymbol{\tau}}_A; \boldsymbol{X}, \boldsymbol{T}) - \ell(\hat{\boldsymbol{\tau}}_0; \boldsymbol{X}, \boldsymbol{T}))$ needs to be calculated which is asymptotically χ^2-distributed with degrees of freedom that are equal to the difference in the number of free parameters between both models. A significant likelihood ratio test indicates that the parameter constraints in the restricted model are not tenable. As the likelihood ratio test is known to be sensitive to large sample size (see Schermelleh-Engel, Moosbrugger, and Müller 2003, p. 34), it is advisable to consider other fit indices as well.

4. Simulation study

To show that the models discussed in this paper are feasible, we conducted a simulation study to demonstrate (1) that true parameter values are adequately recovered; and (2) that the test statistics proposed above follow their theoretical distributions. To do so, we simulated data according to the D-diffusion and the Q-diffusion IRT model with the function `simdiff` from the **diffIRT** package. This function uses the rejection algorithm described in Tuerlinckx, Maris, Ratcliff, and De Boeck (2001) with the appropriate Q- and D-diffusion decompositions for boundary, α, and drift, μ.

4.1. Design

We used the item parameter setup in Table 1. In addition, the population parameters were chosen to equal $\omega_\gamma = 0.3$ and $\omega_\theta = 0.3$ for the Q-diffusion model, and $\omega_\gamma = 0.3$ and $\omega_\theta = 1$ for the D-diffusion model. As can be seen in the table, we systematically varied the item parameters across items resulting in different expected values for the responses, x_{pi}, and the response times, t_{pi}. These expected values are calculated using the expression for the mean, variance, and probability correct in Wagenmakers *et al.* (2007), together with the appropriate decomposition of α and μ (Equations 2, 3, 5, and 6) and integrating out the person variables γ_p and θ_p.

For both the Q- and D-diffusion model we simulated 100 datasets for $N = 100$ and $N = 200$. Note that we chose not to study the asymptotic behavior of the model (i.e., by taking a sample size of for instance 10000) as we are mainly interested in establishing whether the parameter recovery is acceptable in more realistic sample sizes. However, given the results as presented below, we have no reason to doubt the asymptotic properties of the model. To the data we fitted four models: (1) the full model; (2) a model with a_1 to a_4 equal, a_5 to a_8 equal, and a_9 to a_{12} equal; 3) a model with v_1, v_5, v_9 equal, v_2, v_6, v_{10} equal, v_3, v_7, v_{11} equal, and v_4, v_8, v_{12} equal; 4) a model with Ter_1 to Ter_6 equal and Ter_7 to Ter_{12} equal. Note that all these equality constraints hold (see Table 1). In each replication we conducted a likelihood ratio test between a model with and without the above constraints. In addition, we conducted the M_r test on the responses using $r = 2$. All settings not discussed (e.g., number of quadrature points, etc.) equaled the default values of the **diffIRT** package.

Par	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12
						Q-diffusion model						
a_i^*	0.37	0.37	0.37	0.37	0.47	0.47	0.47	0.47	0.61	0.61	0.61	0.61
v_i^*	1	1.22	1.49	1.82	1	1.22	1.49	1.82	1	1.22	1.49	1.82
Ter_i	2	2	2	2	2	2	3	3	3	3	3	3
$\mathsf{E}(x_{pi})$	0.92	0.89	0.85	0.82	0.88	0.85	0.81	0.77	0.84	0.8	0.76	0.72
$\mathsf{E}(t_{pi})$	3.26	3.42	3.57	3.71	2.88	2.98	4.06	4.13	3.61	3.65	3.69	3.72
$\mathsf{VAR}(t_{pi})$	0.96	1.33	1.75	2.2	0.53	0.69	0.85	1.02	0.27	0.33	0.39	0.44
						D-diffusion model						
a_i^*	0.37	0.37	0.37	0.37	0.47	0.47	0.47	0.47	0.61	0.61	0.61	0.61
v_i^*	−1	−0.5	0.5	1	−1	−0.5	0.5	1	−1	−0.5	0.5	1
Ter_i	2	2	2	2	2	2	3	3	3	3	3	3
$\mathsf{E}(x_{pi})$	0.8	0.66	0.34	0.2	0.78	0.65	0.35	0.22	0.75	0.63	0.37	0.25
$\mathsf{E}(t_{pi})$	3.3	3.47	3.47	3.3	2.88	2.98	3.98	3.88	3.59	3.64	3.64	3.59
$\mathsf{VAR}(t_{pi})$	1.47	1.79	1.79	1.47	0.66	0.81	0.81	0.66	0.29	0.35	0.35	0.29

Table 1: True parameter values used in the simulation study, with model implied expected marginal response time, $\mathsf{E}(t_{pi})$, response time variance, $\mathsf{VAR}(t_{pi})$, and proportion correct $\mathsf{E}(x_{pi})$ for each item. The expected values for x_{pi} and t_{pi} are calculated given $\omega_\gamma = 0.3$ and $\omega_\theta = 0.3$ for the Q-diffusion model, and $\omega_\gamma = 0.3$ and $\omega_\theta = 1$ for the D-diffusion model.

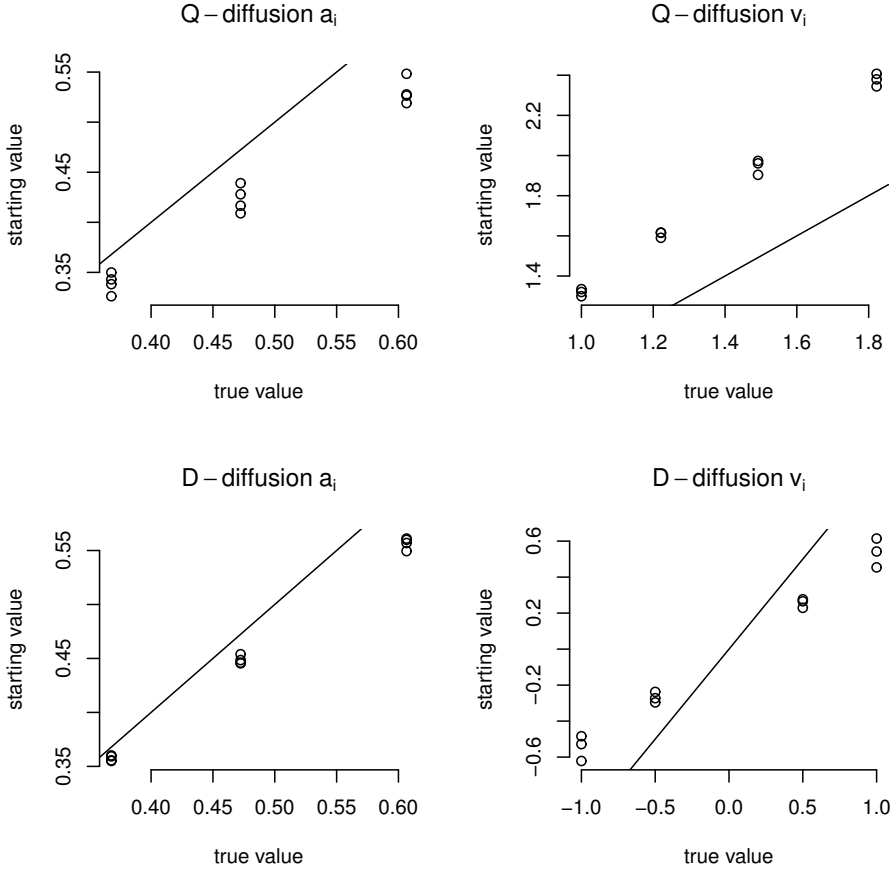

Figure 2: Average starting values of a_i and v_i calculated by the `diffIRT` function plotted against the true parameter values in the simulation study for $N = 200$. The straight line denotes a one-to-one correspondence.

Par		Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12
a_i	true	0.37	0.37	0.37	0.37	0.47	0.47	0.47	0.47	0.61	0.61	0.61	0.61
	$N = 100$	0.37	0.37	0.37	0.37	0.47	0.49	0.48	0.47	0.61	0.62	0.61	0.61
		(0.03)	(0.03)	(0.03)	(0.02)	(0.05)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.04)	(0.04)
	$N = 200$	0.37	0.37	0.37	0.37	0.48	0.48	0.47	0.47	0.61	0.61	0.61	0.61
		(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
v_i	true	1.00	1.22	1.49	1.82	1.00	1.22	1.49	1.82	1.00	1.22	1.49	1.82
	$N = 100$	1.01	1.21	1.49	1.83	1.00	1.27	1.56	1.87	1.01	1.25	1.55	1.90
		(0.11)	(0.17)	(0.22)	(0.37)	(0.14)	(0.20)	(0.28)	(0.40)	(0.16)	(0.18)	(0.32)	(0.50)
	$N = 200$	1.00	1.23	1.52	1.85	0.99	1.22	1.52	1.84	1.01	1.24	1.47	1.81
		(0.08)	(0.11)	(0.16)	(0.22)	(0.10)	(0.13)	(0.17)	(0.22)	(0.12)	(0.15)	(0.18)	(0.23)
Ter_i	true	2.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
	$N = 100$	2.00	2.01	2.02	2.01	1.98	2.01	3.01	3.00	3.00	3.01	3.00	3.00
		(0.03)	(0.04)	(0.03)	(0.03)	(0.20)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	$N = 200$	2.00	2.01	2.01	2.00	2.01	2.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)

Table 2: Mean (and standard deviation) of the item parameter estimates over the replications in the simulation study for the Q-diffusion model. True value for ω_γ equaled 0.3 which was estimated to be 0.29 (0.02) for $N = 100$ and 0.29 (0.02) for $N = 200$. True value for ω_θ equaled 0.3 which was estimated to be 0.29 (0.07) for $N = 100$ and 0.30 (0.05) for $N = 200$.

Par		Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12
a_i	true	0.37	0.37	0.37	0.37	0.47	0.47	0.47	0.47	0.61	0.61	0.61	0.61
	$N = 100$	0.37	0.37	0.37	0.37	0.48	0.47	0.48	0.48	0.60	0.61	0.61	0.62
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	$N = 200$	0.37	0.37	0.37	0.38	0.48	0.48	0.48	0.48	0.62	0.62	0.62	0.62
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
v_i	true	−1.00	−0.50	0.50	1.00	−1.00	−0.50	0.50	1.00	−1.00	−0.50	0.50	1.00
	$N = 100$	−0.96	−0.48	0.52	1.02	−0.94	−0.48	0.52	1.03	−0.98	−0.47	0.49	1.02
		(0.21)	(0.22)	(0.23)	(0.19)	(0.23)	(0.22)	(0.22)	(0.20)	(0.24)	(0.24)	(0.22)	(0.23)
	$N = 200$	−0.97	−0.47	0.52	1.00	−0.97	−0.48	0.51	1.01	−0.96	−0.46	0.51	1.01
		(0.17)	(0.17)	(0.18)	(0.17)	(0.17)	(0.16)	(0.18)	(0.16)	(0.18)	(0.19)	(0.19)	(0.19)
Ter_i	true	2.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
	$N = 100$	2.01	2.01	2.01	2.01	2.01	2.00	3.01	3.01	3.00	3.01	3.00	3.01
		(0.03)	(0.04)	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	$N = 200$	2.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table 3: Mean (and standard deviation) of the item parameter estimates over the replications in the simulation study for the D-diffusion model. True value for ω_γ equaled 0.3 which was estimated to be 0.30 (0.02) for $N = 100$ and 0.29 (0.02) for $N = 200$. True value for ω_θ equaled 1.0 which was estimated to be 0.87 (0.08) for $N = 100$ and 0.85 (0.05) for $N = 200$.

4.2. Results

In Figure 2, the true parameter values are plotted against the starting values for a_i and v_i for $N = 200$ for both the D-diffusion and Q-diffusion model. As can be seen, the starting values are well correlated to the true values, but biased. As discussed above, this bias is due to the fact that γ_p cannot be disentangled from a_i, and θ_p cannot be disentangled from v_i using the EZ-diffusion model. However, as starting values, these values have utility as will appear below.

Results concerning the item parameter recovery are displayed in Table 2 for the Q-diffusion model and in Table 3 for the D-diffusion model. For both models, all parameters are recovered well with the v_i estimates having somewhat more variability as compared to the a_i and Ter_i estimates. As can be seen in the footnotes of the tables, the population parameters, ω_γ and ω_θ, are also recovered well, with ω_θ slightly underestimated in case of the D-diffusion model

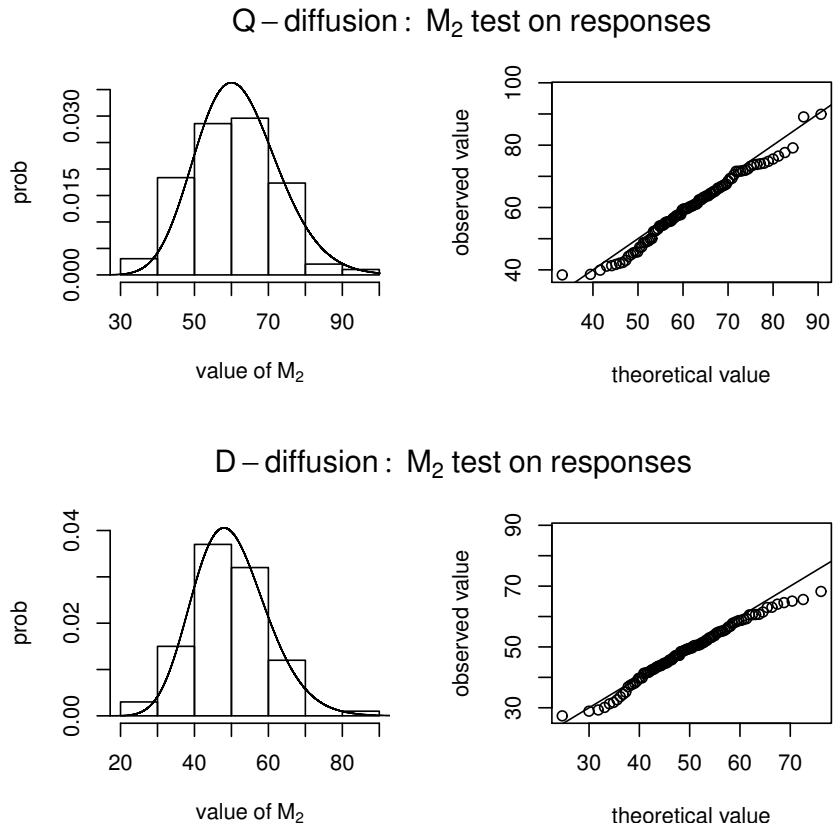

Figure 3: Histogram (left column) and corresponding QQ-plot (right column) of the distribution of the M_2 statistic calculated on the responses in the simulation study for $N = 200$. The theoretical distributions are $\chi^2(62)$ for the Q-diffusion model and $\chi^2(50)$ for the D-diffusion model.

(0.87 for $N = 100$ and 0.85 for $N = 200$, where the true value equals 1).

In Figure 3, the theoretical and observed distributions of the M_r statistic are depicted for the M_2 test on the responses in the case of $N = 200$. As argued above, the theoretical distribution is expected to be χ^2 with degrees of freedom equal to $df = \sum_{q=1}^{r} \binom{n}{q} - s$, which gives 62 for the Q-diffusion model and 50 for the D-diffusion model. Note that s needs to be adjusted as described above. As can be seen in the figure, there is no reason to suspect systematic departures of M_2 from these theoretical distributions. Figures 4 and 5 contain similar graphs for the likelihood ratio test statistics under H_0 in the case that respectively a_i and v_i are constrained to the equality pattern of their true values (graphs for Ter_i are not given to save space, but results for this parameter are the same). In case of a_i, the full model containing 12 a_i parameters, is compared to a model with only 3 a_i parameters (i.e., $a_1^{'} = a_1 = a_2 = a_3 = a_4; a_2^{'} = a_5 = a_6 = a_7 = a_8$; and $a_3^{'} = a_9 = a_{10} = a_{11} = a_{12}$). Thus, a likelihood ratio test between these models will result in a $\chi^2(9)$ distribution under H_0. In case of v_i, the constrained model contains only 4 v_i parameters (i.e., $v_1^{'} = v_1 = v_5 = v_9; v_2^{'} = v_2 = v_6 = v_{10}; v_3^{'} = v_3 = v_7 = v_{11}; v_4^{'} = v_4 = v_8 = v_{12}$). Thus, a likelihood ratio test between these models will result in a $\chi^2(8)$ test. As can be seen in the figure, there are no systematic departures from these theoretical distributions.

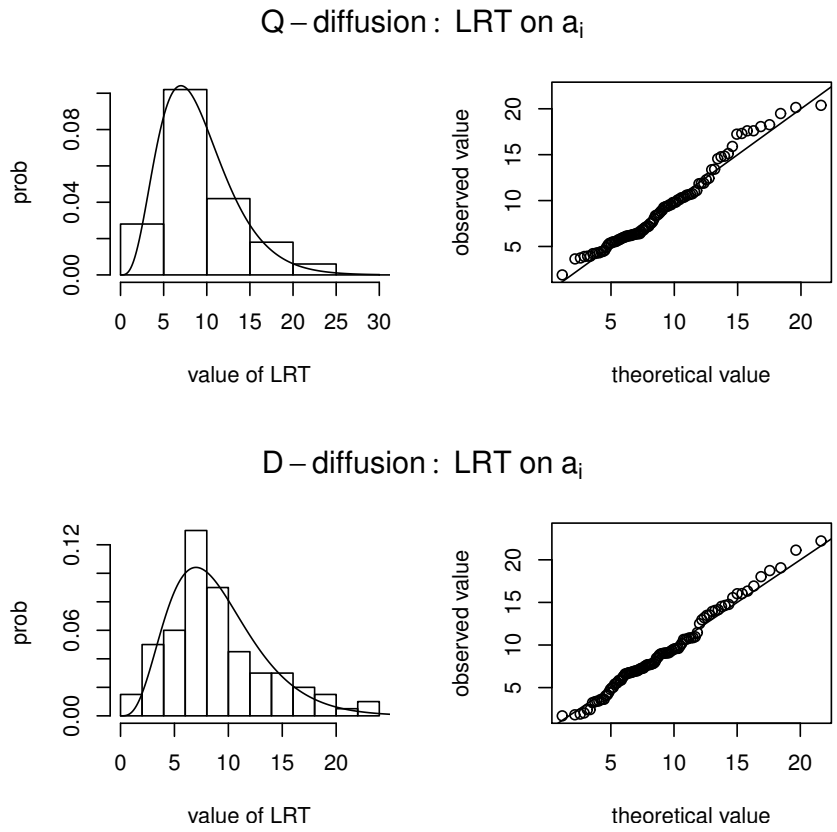

Figure 4: Histogram (left column) and corresponding QQ-plot (right column) of the distribution of the likelihood ratio statistic under H_0 comparing the full model with a constrained model subject to $a_1^{'} = a_1 = a_2 = a_3 = a_4; a_2^{'} = a_5 = a_6 = a_7 = a_8$; and $a_3^{'} = a_9 = a_{10} = a_{11} = a_{12}$ in the simulation study for $N = 200$. The theoretical distributions are $\chi^2(9)$ for both the Q-diffusion model and for the D-diffusion model.

5. Package description

In this section we describe and illustrate the basic functions of the **diffIRT** package. For an overview of all the functions, see Table 4.

To start, the function `simdiff` can be used to simulate data according to either the D-diffusion model or the Q-diffusion model. Here we simulate data for 100 subjects and 10 items that follow a Q-diffusion model:

```
R> set.seed(1310)
R> data <- simdiff(100, 10, model = "Q")
```

True values are randomly chosen by the function. It is also possible to provide user-specified true values to the function. We fit the Q-diffusion model to these simulated data using:

```
R> out <- diffIRT(data$rt, data$x, model = "Q", se = TRUE)
```

As can be seen, we explicitly requested standard errors of the parameter estimates. By default these are not calculated as it will increase estimation time. We can now check the results:

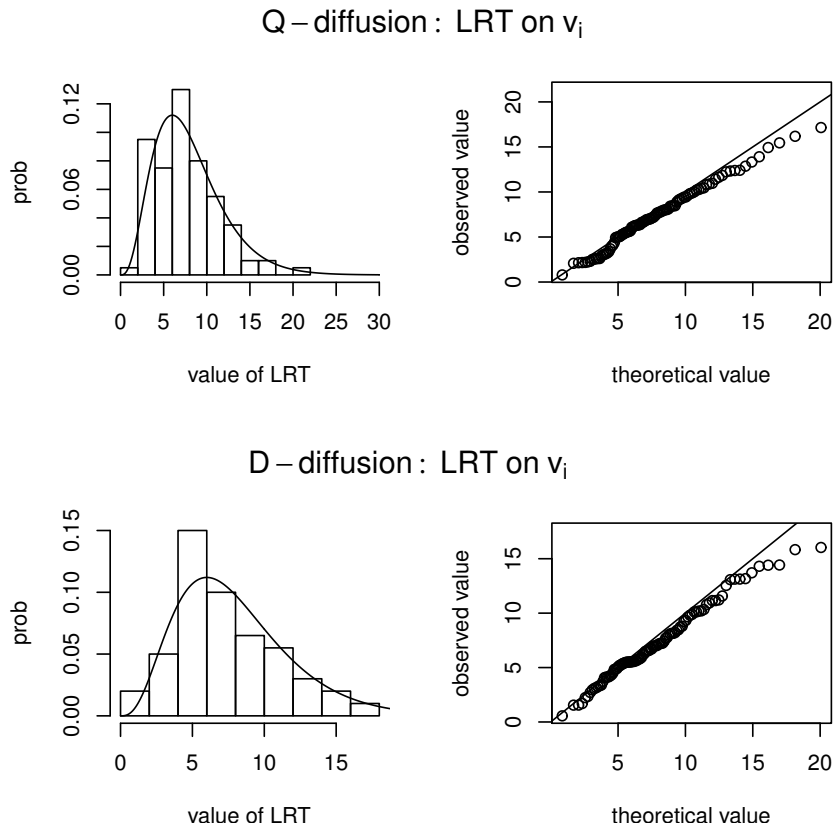

Figure 5: Histogram (left column) and corresponding QQ-plot (right column) of the distribution of the likelihood ratio statistic under H_0 comparing the full model with a constrained model subject to $v_1^{'} = v_1 = v_5 = v_9; v_2^{'} = v_2 = v_6 = v_10; v_3^{'} = v_3 = v_7 = v_{11}; v_4^{'} = v_4 = v_8 = v_{12}$ in the simulation study for $N = 200$. The theoretical distributions are $\chi^2(8)$ for both the Q-diffusion model and for the D-diffusion model.

```
R> summary(out)

        RESULTS
        -------
         a[i] se.a[i]  v[i] se.v[i] Ter[i] se.Ter[i]
item 1  0.728   0.040 1.794   0.469  0.843     0.011
item 2  1.071   0.055 4.448   3.932  0.622     0.004
item 3  0.433   0.029 1.348   0.203  1.383     0.028
item 4  0.348   0.027 1.106   0.143  0.708     0.041
item 5  0.762   0.043 1.426   0.317  0.885     0.010
item 6  0.603   0.037 1.163   0.193  0.914     0.013
item 7  0.365   0.026 1.178   0.158  0.690     0.039
item 8  0.772   0.044 1.656   0.415  1.450     0.010
item 9  0.523   0.031 1.541   0.280  0.544     0.020
item 10 0.419   0.024 2.706   0.637  0.915     0.035

omega[gamma]: 0.248
```

Name	Example call	Description
`diffIRT`	`out <- diffIRT(T, X, model = "Q")`	Fits the Q-diffusion model (`model = "Q"`) or the D-diffusion model (`model = "D"`, default) to the response time data in matrix `T` and response matrix `X`.
`RespFit`	`RespFit(out, 2)`	Calculates the M_2 statistic from the modeling results in object `out`.
`QQdiff`	`QQdiff(out, item = 1:4)`	Plots a histogram and corresponding QQ-plot of the predicted and observed response time distribution for items 1 to 4.
`simdiff`	`data <- simdiff(100, 10, model = "D")`	Simulates data according to the Q-diffusion model (`model = "Q"`) or the D-diffusion model (`model = "D"`, default) for 100 subjects and 10 items.
`factest`	`factest(out)`	Estimates the factor scores of γ_p and θ_p for all subjects.
`anova`	`anova(out1, out2)`	Conducts a likelihood ratio test between two nested models.
`coef`	`coef(out)`	Returns the estimated parameters in object `out`.
`summary`	`summary(out)`	Returns a summary of the model fitting results in object `out` including parameter estimates and fit statistics.
`simdiffT`	`data <- simdiff(1000, 2, 1, .3, 3)`	Simulates data according to the traditional diffusion model for a single subject and 1000 trials.

Table 4: Overview of the functions in the **diffIRT** package with example call and description.

```
std. err:     0.065
omega[theta]: 0.402
std. err:     0.169

---------------------------

--POPULATION DESCRIPTIVES--
Person Boundary (lognormal): Var(gamma):  0.067
Person Drift (lognormal):    Var(theta):  0.206
---------------------------

---MODEL FIT STATISTICS---
-2 x logLikelihood: 2053.414

no. of parameters: 32
AIC:  2117.414
BIC:  2200.779
sBIC: 2099.715
DIC:  2142.601
---------------------------
```

Note that `omega[gamma]` and `omega[theta]` refer to the estimates of ω_γ and ω_θ respectively in Equation 9. To see whether the parameters are adequately recovered, we plot the true

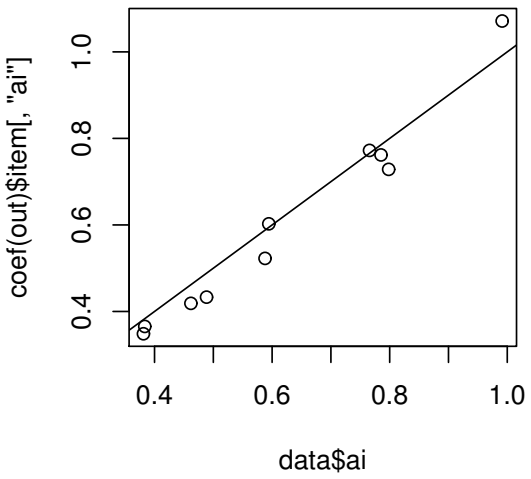

Figure 6: Plot from the illustration: Estimated a_i parameters against the true a_i parameters. The straight line denotes a one-to-one correspondence.

values (which are stored in object `data`) against the estimated parameters (that are in object `out`):

```
R> plot(data$ai, coef(out)$item[, "ai"])
R> abline(0, 1)
```

The resulting plot is in Figure 6. As can be seen, the a_i parameters are adequately recovered. Next, we study the model fit. First, we use the M_r-test of order 2 on the responses using:

```
R> resp_out <- RespFit(out, 2)
R> resp_out

Q-diffusion Model Fit of Responses
----------------------------------
Maydeu-Olivares & Joe Test of Order 2

Overall Test Statistic
Mr = 27.874   df= 41   p =  0.941
```

The M_r statistic is non-significant indicating that the model fits. In case of a significant M_r statistics, residuals could be examined using `resp_out$Z`, which gives the standardized residual proportions for each score pattern of the first r moments. Next we examine the model fit on the response times using the function `QQdiff`:

```
R> QQdiff(out, item = 1:3)
```

which produces the plot in Figure 7. It turns out that for the first three items, the predicted and observed distributions appear to coincide, with some minor misfit in the tails as is common in QQ-plots due to the relatively few observations in this region of the distribution.

Next, we conduct a likelihood ratio test to see whether the v_i parameters are equal across items. To do so we fit a Q-diffusion model subject to the constraint that $v_1 = v_2 = \ldots = v_{10}$:

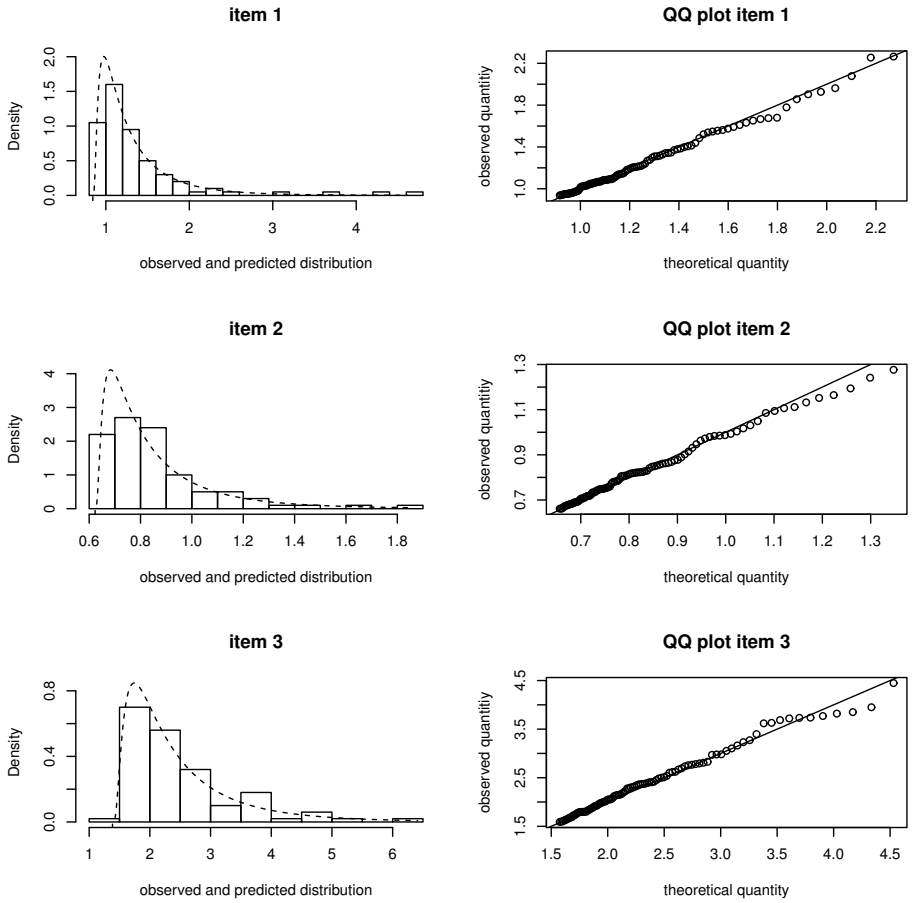

Figure 7: Plot from the illustration: Histogram with corresponding QQ-plot of the predicted and observed response time distribution for the Q-diffusion model.

```
R> out_vi <- diffIRT(data$rt, data$x, model = "Q",
+    constrain = c(1:10, rep(11,10), 12:21, 22, 23))
```

As can be seen, the constraint is introduced by providing a vector of parameter numbers. Each parameter should have a unique number. The order in which the parameters should be labeled is $a_1, \ldots, a_{10}$, $v_1, \ldots, v_{10}$, $Ter_1, \ldots, Ter_{10}$, ω_γ, and ω_θ, i.e., in the same order as the parameter vector $\boldsymbol{\tau}$ in Equations 9 and 11. Since we only have one v_i parameter in the present case (as we want to constrain all v_i to be equal across items), we label all v_i parameters using the same number, that is 11 in this case. The following code will do exactly the same:

```
R> out_vi_alt <- diffIRT(data$rt, data$x, model = "Q",
+    constrain = "vi.equal")
```

This code makes use of the pre-programmed constraint option, which is more easy to use but less flexible. Other pre-programmed constraint arguments are `"ai.equal"`, `"ter.equal"`, and `"scale.equal"`. In which the latter means that ω_γ and ω_θ are fixed to be equal in the model. For the model with equal v_i in object `out_vi`, the output (not printed here) displays fit indices AIC, BIC, sBIC, and DIC which are all larger for this constraint model as compared

to the full model in `out` except for BIC. To conduct a likelihood ratio test between the two models we use:

```
R> anova(out_vi, out)

 Likelihood Ratio Table
        AIC  BIC sBIC  DIC  log.Lik   LRT df p.value
out_vi 2127 2187 2114 2145 2081.171
out    2117 2201 2100 2143 2053.414 27.76  9   0.001
```

i.e., at a 0.05 level of significance, we reach the same conclusion as compared to the AIC, sBIC, and DIC, that is, the restrictions in the model in object `out_vi` are not tenable.

We now illustrate the use of fixed parameter constraints. As an example, we fix all a_i parameters in the simulated dataset to be equal to 0.5 as follows:

```
R> out_fix <= diffIRT(data$rt, data$x, model = "Q", constrain = c(rep(0,
+    nit), 1:10, 11:20, 21, 22), start = c(rep(.5, nit), rep(NA, 22)))
```

As can be seen, in the `constrain` argument we assigned all a_i parameters a number of `0` denoting that these parameters are fixed. In addition, we assigned the value `0.5` to the corresponding elements in the `start` argument, leaving the other elements `NA`. Requesting output using the `print` comment gives:

```
R> out_fix

RESULTS Q-DIFFUSION IRT ANALYSES

Item parameter estimates
        -------
        a[i]  v[i] Ter[i]
item 1   0.5 1.324  0.770
item 2   0.5 2.674  0.450
item 3   0.5 1.805  1.423
item 4   0.5 1.737  0.805
item 5   0.5 1.042  0.801
item 6   0.5 1.126  0.888
item 7   0.5 1.788  0.770
item 8   0.5 1.307  1.359
item 9   0.5 1.640  0.538
item 10  0.5 3.623  0.979

Population parameter estimates
        -------
omega[gamma]: 0.242
omega[theta]: 0.489
---------------------------
```

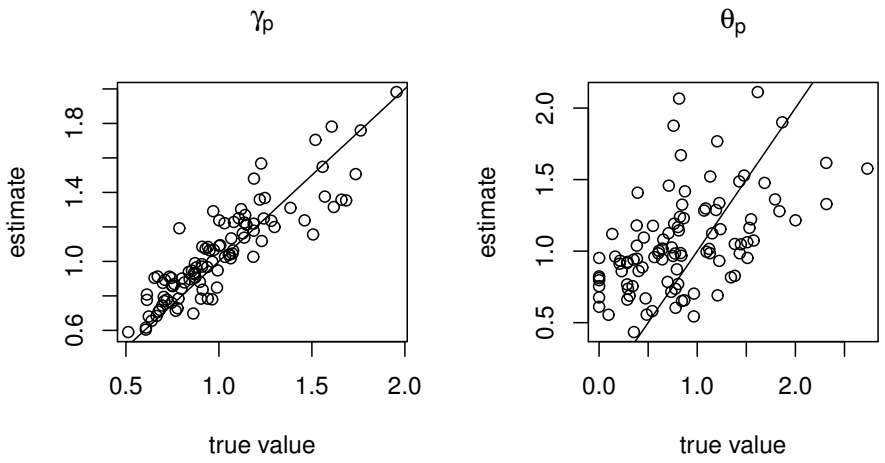

Figure 8: Plot from the illustration: Estimated γ_p and θ_p parameters against the true parameter values. The straight line denotes a one-to-one correspondence.

All a_i are fixed to 0.5. We could again conduct a likelihood ratio test, which has 10 degrees of freedom this time:

```
R> anova(out_fix, out)
```

```
 Likelihood Ratio Table
         AIC  BIC sBIC  DIC  log.Lik    LRT df p.value
out_fix 2508 2566 2496 2526 2464.194
out     2117 2201 2100 2143 2053.414 410.78 10  <0.001
```

Not surprisingly, the test is significant. Finally, we estimate the factor scores, i.e., we obtain estimates for γ_p and θ_p for each subject in the sample. To do so we use:

```
R> fs <- factest(out)
```

Note that we use the full model object, as this was the best fitting model. Matrix `fs` now contains estimates of γ_p and θ_p in the first and second columns, respectively. We plot these estimates against the true values in Figure 8. As can be seen, parameters are adequately recovered, where the estimates of γ_p are subject to more variability as compared to estimates of θ_p.

6. Applications

In this section we present three applications of the models presented in this paper to real data. In the first applications we fit the D-diffusion IRT model to an extraversion dataset. In the second application we fit the Q-diffusion IRT model to data pertaining to mental rotation. In the third application we illustrate how the package can be used to fit the traditional diffusion model to experimental data.

6.1. Application I: D-diffusion modeling of extraversion

The first application concerns an analysis of unpublished data comprising scores of 146 subjects on 10 items purported to measure extraversion. Each item consists of a particular description related to extraverted behavior, e.g., *active* or *noisy*. Subjects were asked to indicate whether (yes/no) these descriptions are applicable to their personalities. Both responses and response times were recorded. The data is available within the **diffIRT** package. The first 10 columns contain the responses, the next 10 columns contain the response times in seconds. Below we analyze the extraversion data using a D-diffusion model. First, we open the data and fit a D-diffusion model:

```
R> data("extraversion", package = "diffIRT")
R> x <- extraversion[, 1:10]
R> rt <- extraversion[, 11:20]
R> res1 <- diffIRT(rt, x, "D", se = TRUE)
```

In Table 5, the results from object `res1` are depicted (i.e., parameter estimates and standard errors) together with item content and results from a traditional two parameter model (obtained using R package **ltm**; Rizopoulos 2006). As can be seen some differences are present in the ordering of the items on basis of the difficulty parameters from the two models, that is, v_i in the D-diffusion model and β_i in the two parameter model. For instance, in the D-diffusion model *eupeptic* is the least difficult item while in the two parameter model the least difficult item is *jovial*. In addition, in the D-diffusion model *noisy* is the most difficult item, while in the two parameter model *impulsive* is the most difficult item. Standard errors of the difficulty parameter in the D-diffusion model, v_i are smaller as compared to those from the two parameter model, β_i. This is due to the additional information in the response times that is used in fitting the D-diffusion model. Note that the standard errors of the other parameters cannot be straightforwardly compared as these are on a different scale. Next an M_2 test is conducted:

Item	a_i	v_i	Ter_i	λ_i	β_i
1. 'active'	0.490 (0.023)	−0.711 (0.108)	0.570 (0.015)	0.541 (0.282)	−2.067 (1.026)
2. 'noisy'	0.510 (0.023)	−0.171 (0.108)	0.471 (0.013)	0.618 (0.276)	−0.269 (0.313)
3. 'energetic'	0.560 (0.029)	−1.309 (0.131)	0.499 (0.011)	2.675 (1.014)	−1.211 (0.217)
4. 'enthusiastic'	0.549 (0.035)	−1.768 (0.150)	0.456 (0.015)	2.709 (1.025)	−1.648 (0.280)
5. 'impulsive'	0.523 (0.024)	−0.228 (0.111)	0.459 (0.012)	0.724 (0.300)	−0.235 (0.271)
6. 'jovial'	0.475 (0.027)	−1.361 (0.128)	0.495 (0.015)	0.918 (0.406)	−2.761 (1.004)
7. 'viable'	0.467 (0.031)	−1.736 (0.146)	0.507 (0.016)	1.759 (0.659)	−2.179 (0.494)
8. 'eupeptic'	0.428 (0.031)	−1.986 (0.152)	0.427 (0.019)	2.879 (1.295)	−2.050 (0.372)
9. 'chatty'	0.393 (0.020)	−0.883 (0.107)	0.606 (0.019)	0.901 (0.368)	−1.950 (0.676)
10. 'spontaneous'	0.598 (0.032)	−1.492 (0.141)	0.458 (0.011)	2.024 (0.633)	−1.422 (0.265)

Table 5: Parameter estimates (standard errors) and item content of the extraversion data in application I for the D-diffusion model and the traditional two parameter model. λ_i and β_i are respectively estimates of the discrimination parameter and difficulty parameter from a traditional two parameter model.

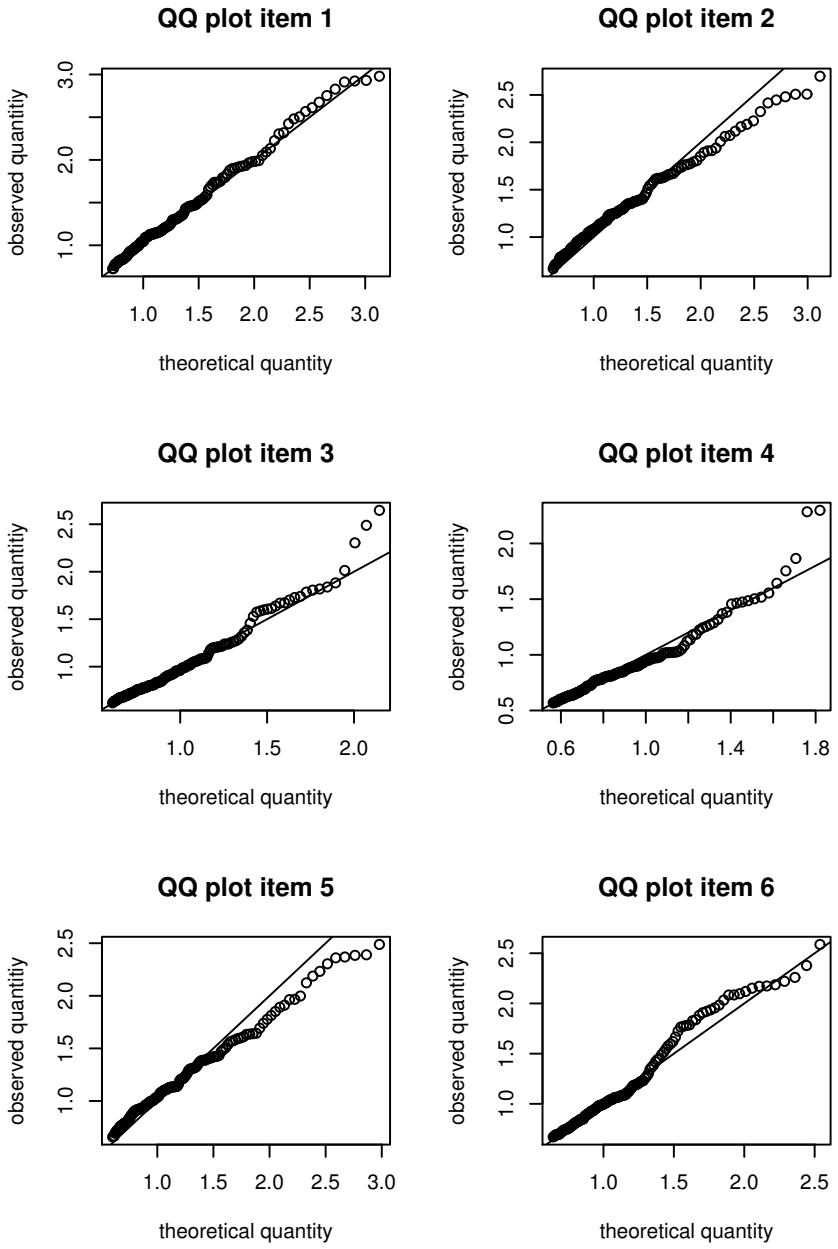

Figure 9: QQ-plots for the first 6 items of the extraversion data in application I.

```
R> out_resp <- RespFit(res1, 2)

D-diffusion Model Fit of Responses
----------------------------------
Maydeu-Olivares & Joe Test of Order 2

Overall Test Statistic
Mr = 37.794   df= 31   p =  0.187
```

As can be seen, the M_2 test indicates that the model fits to the responses. QQ-plots for the

first 6 items are obtained by using:

```
R> QQdiff(res1, item = 1:6, plot = 1)
```

which produces the plots in Figure 9. As can be seen, the plots also indicate that the model fits the observed response times.

6.2. Application II: Q-diffusion modeling of mental rotation data

The data comprises scores of 121 subjects on 10 items purported to measure mental rotation. These data are taken from a larger database published in Kievit (2010); see also Borst, Kievit, Thompson, and Kosslyn (2011). The 10 selected items are part of the article by van der Maas *et al.* (2011). van der Maas *et al.* analyzed these data using a Bayesian version of the Q-diffusion model including random *Ter* parameters. Here, we re-analyze the data (with fixed *Ter*) and conduct various goodness of fit tests that are not straightforward in the Bayesian framework employed by van der Maas *et al.* (2011). Each item consists of a graphical display of two 3-dimensional objects. The second object was either a rotated version of the first object, or a rotated version of a different object. Subjects were asked whether the second object was the same as the first object (yes/no). The degree of rotation of the second object was either 50, 100, or 150 degrees. Answers are coded to be correct (1) or false (0). Response times were recorded in seconds. The data are available in the package **diffIRT**. The first 10 columns contain the responses, the next columns contain the response times in seconds. From the original data we omitted four response times (i.e., we replaced these by `NA`) that were smaller than 0.3s, as these suspiciously fast response times are likely to be invalid and could cause problems in estimating Ter_i. Below we report a Q-diffusion model analysis of the resulting dataset. To start, we open the data and fit a Q-diffusion IRT model to it using:

```
R> data("rotation", package = "diffIRT")
R> x <- rotation[, 1:10]
R> rt <- rotation[, 11:20]
R> res <- diffIRT(rt, x, "Q", se = TRUE)
```

Table 6 contains the results in object `res` (i.e., parameter estimates and standard errors) together with the degree of rotation of the object in each item. As appears from the estimates, the effect rotation degree is most notable in the Ter_i parameters. That is, the items with a more rotated object are associated with a higher Ter_i estimate. This indicates that items with a higher degree of rotation need more stimulus encoding time than items with a smaller degree of rotation. To investigate model fit, we conduct an M_2 test using

```
R> RespFit(res, 2)

Q-diffusion Model Fit of Responses
---------------------------------
Maydeu-Olivares & Joe Test of Order 2

Overall Test Statistic
Mr = 144.865   df= 41   p =  0
```

Item	a_i	v_i	Ter_i
1: 150°	0.249 (0.016)	1.484 (0.174)	0.880 (0.036)
2: 50°	0.230 (0.018)	1.231 (0.135)	0.538 (0.060)
3: 100°	0.219 (0.014)	1.424 (0.160)	0.778 (0.038)
4: 150°	0.248 (0.015)	2.345 (0.323)	0.999 (0.046)
5: 50°	0.238 (0.020)	0.884 (0.100)	0.824 (0.043)
6: 100°	0.267 (0.018)	1.210 (0.130)	0.878 (0.036)
7: 150°	0.190 (0.015)	1.155 (0.128)	0.580 (0.036)
8: 50°	0.214 (0.021)	0.764 (0.086)	0.617 (0.050)
9: 150°	0.231 (0.016)	1.191 (0.126)	0.794 (0.040)
10: 100°	0.202 (0.013)	1.814 (0.209)	0.801 (0.041)

Table 6: Parameter estimates (standard errors) and degree of rotation of the mental rotation items in application II. λ_i and β_i are respectively estimates of the discrimination parameter and difficulty parameter from a traditional two parameter model.

Model	−2LL	LRT	df	AIC	BIC	sBIC	DIC
Full	4355.78			4420	4509	4408	4451
a_i equal	4378.58	22.80	9	4425	4489	4416	4447
v_i equal	4437.01	81.23	9	4483	4547	4475	4505
ter_i equal	4439.43	83.65	9	4485	4550	4477	4508
a_i rotate	4378.12	22.34	7	4428	4498	4419	4452
v_i rotate	4406.95	51.17	7	4457	4527	4448	4481
ter_i rotate	4417.80	62.02	7	4468	4538	4459	4492

Table 7: Model fit indices of various Q-diffusion models on the mental rotation data in application II. 'rotate' denotes a model in which the corresponding item parameters are constrained equal for the items that have the same amount of rotation, see Table 6. All LRTs are against the full model.

Thus, according to the M_2 test, the model did not fit well. QQ-plots (not displayed), however, looked reasonable. Next we fitted four constraint models: (1) a model with equal a_i; (2) a model with equal v_i; (3) a model with equal v_i for items that have the same amount of rotation; and (4) a model with equal Ter_i. This required the following R code:

```
R> res_ai_equal <- diffIRT(rt, x, model = "Q", constrain = "ai.equal")
R> res_vi_equal <- diffIRT(rt, x, model = "Q", constrain = "vi.equal")
R> res_vi_rotation <- diffIRT(rt, x, model = "Q", constrain =
+    c(1:10, c(11, 12, 13, 11, 12, 13, 11, 12, 11, 13), 14:23, 24, 25))
R> res_ter_equal <- diffIRT(rt, x, model = "Q", constrain = "ter.equal")
```

In Table 7, output is summarized for the full model (i.e., the model we fit above, see Table 6) and the four constraint models. As we are interested in making an inference about the tenability of the equality constraints introduced in the constraint models, the full model – with all parameter unconstrained –, serves as a baseline model here. As can be seen from the table, the model with equal a_i parameters fits best according to the fit indices, i.e., AIC, BIC, sBIC, and DIC are smallest for this model. In addition, the likelihood ratio test between the full model and this model is insignificant as appears from

```
R>  anova(res_ai_equal,res)

 Likelihood Ratio Table
              AIC  BIC sBIC  DIC  log.Lik  LRT df p.value
res_ai_equal 4307 4370 4298 4329 4261.354
res          4316 4404 4303 4346 4252.032 9.32  9   0.408
```

which indicates that the equality restrictions on a_i are tenable.

6.3. Application III: Fitting the traditional diffusion model

As the diffusion IRT model is in essence a traditional diffusion model with random effects on the drift rate and boundary separation parameters, the **diffIRT** package can also be used to fit the traditional diffusion model to data. Specifically, the traditional diffusion model can be seen as a restricted D-diffusion IRT model without random γ_p and with random θ_p over trial (instead of random over subjects). To specify this, we will refer to traditional diffusion model parameters with a superscript *trad* for sake of clarity. In addition, to denote that the parameters can be different across experimental conditions, we use subscript c. As a result, in the traditional diffusion model we have, α_c^{trad}, Ter_c^{trad}, μ_c^{trad}, and σ_c^{trad} where c runs from 1 to the number of conditions, μ_c^{trad} refers to the mean drift rate over trials, and ω_c^{trad} refers to the inter-trial standard deviation of the drift rate (commonly used in experimental applications of the traditional diffusion model). To specify the traditional diffusion model, in the D-diffusion model we set $\omega_\gamma = 0$ such that γ_p is constant across subjects and equal to $\exp(0) = 1$. Then, from Equation 3 it follows that

$$\alpha_c^{trad} = \exp(0)/a_i = 1/a_i \tag{15}$$

and from Equation 2 it follows that

$$\mu_c^{trad} = \mathsf{E}(\theta_p - v_i) = -v_i, \tag{16}$$

as θ_p is normally distributed with mean 0 in the D-diffusion model. By equating $c = i$ it can be seen that the items in the D-diffusion model correspond to the conditions in the traditional diffusion model. In addition, subjects in the D-diffusion model are the trials in the traditional model. Thus, the response and the response time data matrices of a given subject need to be arranged in such a way that trials are the rows, and the conditions are the columns. Then, a D-diffusion model can be fitted to these matrices and the parameters can be converted to the traditional diffusion model parameters using Equations 15 and 16 together with $Ter_c^{trad} = Ter_i$ and $\sigma_c^{trad} = \omega_\theta$. This can be illustrated by applying the restricted D-diffusion model to a simulated data set. To simulate data according to the traditional diffusion model, we use the function `simdiffT` from the **diffIRT** package, that is:

```
R> set.seed(1310)
R> alpha <- 2
R> mu <- 1
R> ter <- 2
R> sdv <- .3
R> N <- 10000
R> data <- simdiffT(N, alpha, mu, sdv, ter)
```

We fit the traditional diffusion model by using

```
R> res <- diffIRT(data$rt, data$x, model = "D",
+    constrain = c(1, 2, 3, 0, 4), start = c(rep(NA, 3), 0, NA))

RESULTS D-DIFFUSION IRT ANALYSES

Item parameter estimates
        -------
        a[i]   v[i] Ter[i]
item 1 0.498 -0.997  2.001

Population parameter estimates
        -------
omega[gamma]: 0
omega[theta]: 0.319
---------------------------
```

From the output one can determine $1/a_i$ which gives 2.008 and $-1 \times v_i$ which gives 0.997. Both are close to the true values of respectively α^{trad} and μ^{trad}. In addition, estimates for ω_θ and Ter_i are close to the true values of σ^{trad} and Ter^{trad}. Note that the diffusion model fitted in this way does not include a random *Ter* or a (random) starting point as is the case in the software package fast-dm for instance (Voss and Voss 2007).

Multiple experimental conditions

Here we illustrate the application of the traditional diffusion model on a simulated dataset with multiple experimental conditions. The data are simulated according to a design similar to that of the real brightness discrimination experiment by Ratcliff and Rouder (1998). In this experiment, a subject had to decide for a number of trials whether the brightness of a stimulus (a randomly generated array of pixels displayed on a computer screen) was either "high" or "low". The true brightness of the stimuli were manipulated into a number of levels and administered with a speed instruction ("respond as fast as possible") and with an accuracy instruction ("respond as accurate as possible"). Here we simulated data for a single subject using function `simdiff` for a design with 6 different brightness levels and 2 speed instructions resulting in $6 \times 2 = 12$ conditions. The data can be obtained using `data("brightness", package = "diffIRT")` where the first 12 columns are the responses and the next 12 columns are the response times.

The brightness data are prepared such that the 12 conditions are in the columns and the 800 trials are in the rows. The data is arranged in such a way that the first 6 conditions are the speed instructed stimuli and the next 6 conditions are the corresponding accuracy instructed versions of these stimuli. As the trials are random, each trial is assigned to a separate row with the response time of that trial in the corresponding column and `NA`'s on the remaining columns. Similarly for the responses. Note that the response and the response time matrices thus have $800 \times 12 = 9600$ rows and 12 columns.

We now fit the hypothesized traditional diffusion model to these data using:

```
R> data("brightness", package = "diffIRT")
R> x <- brightness[, 1:12]
R> rt <- brightness[, 13:24]
R> res <- diffIRT(rt, x, model = "D", constrain = c(rep(1, 6), rep(2, 6),
+    3:8, 3:8, rep(9, 12), 0, 10), start = c(rep(NA, 36), 0, NA))
R> res

RESULTS D-DIFFUSION IRT ANALYSES

Item parameter estimates
        -------
         a[i]  v[i] Ter[i]
item 1  0.392 3.584  0.298
item 2  0.392 3.012  0.298
item 3  0.392 2.521  0.298
item 4  0.392 2.030  0.298
item 5  0.392 1.526  0.298
item 6  0.392 1.014  0.298
item 7  0.652 3.584  0.298
item 8  0.652 3.012  0.298
item 9  0.652 2.521  0.298
item 10 0.652 2.030  0.298
item 11 0.652 1.526  0.298
item 12 0.652 1.014  0.298

Population parameter estimates
        -------
omega[gamma]: 0
omega[theta]: 0.806
---------------------------
```

In the `constrain` argument we fixed the a_i parameters for the first 6 stimuli to be equal, and the a_i parameters for the next 6 stimuli to be equal. In addition, we fixed the v_i for the same stimuli (i.e., v_i of the tasks in column 1 of the data is fixed to equal v_i of column 7, etc.), and we fixed all Ter_i parameters to be equal. In addition, ω_γ is fixed to equal 0 to omit this dimension from the model. As discussed above, from the output, traditional diffusion model estimates can be obtained by $1/a_i$ and $-1 \times v_i$. Ter_i and ω_θ are already equal to the traditional parameters.

7. Discussion

Process IRT models incorporating separate person and item parameters have the potential to bridge the gap between the statistically oriented psychometric measurement models and the theoretically oriented mathematical process models. Model estimation of such process IRT models is however challenging due to the presence of the random person effects and due to the relatively complex forms of the process models like the diffusion model. With

the **diffIRT** package we provide a useful tool to fit a diffusion IRT model to data obtained from multiple subjects answering multiple items. Challenges for future developments within this line of research remain. For instance, at present, as the diffusion model is a model for two-choice data, the package can only handle binary item data. It would be interesting to consider possibilities for multiple choice data, including extensions of the diffusion models discussed in this paper and other process models, e.g., the linear ballistic accumulator model (Brown and Heathcote 2005) and the race model (Audley and Pike 1965; see also Tuerlinckx and De Boeck 2005). Other challenges are: development of models for multiple processes (e.g., an arithmetic process and a reading process for worded arithmetic items), multi-group approaches (e.g., to test gender differences in person boundary) and latent regression of the person drift and boundary parameters (e.g., to test for the effects of age on these parameters). These options might be considered in future developments of the **diffIRT** package.

References

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Audley RJ, Pike AR (1965). "Some Alternative Stochastic Models of Choice." *British Journal of Mathematical and Statistical Psychology*, **18**(2), 207–225.

Birnbaum A (1968). "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In EM Lord, MR Novick (eds.), *Statistical Theories of Mental Test Scores*, pp. 397–479. Addison Wesley, Reading.

Bock RD, Aitkin M (1981). "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika*, **46**(4), 443–459.

Bollen KA (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Bollen KA (1998). *Structural Equation Models*. John Wiley & Sons.

Borsboom D, Mellenbergh GJ, van Heerden J (2004). "The Concept of Validity." *Psychological Review*, **111**(4), 1061–1071.

Borsboom D, Molenaar D (2015). "Psychometrics." In JD Wright (ed.), *International Encyclopedia of the Social & Behavioral Sciences*, pp. 418–422. Elsevier.

Borst G, Kievit RA, Thompson WL, Kosslyn SM (2011). "Mental Rotation Is Not Easily Cognitively Penetrable." *Journal of Cognitive Psychology*, **23**(1), 60–75.

Brown S, Heathcote A (2005). "A Ballistic Model of Choice Response Time." *Psychological Review*, **112**(1), 117–128.

Digman JM (1990). "Personality Structure: Emergence of the Five-Factor Model." *Annual Review of Psychology*, **41**(1), 417–440.

Fischer GH (1995). "Derivations of the Rasch Model." In GH Fisher, IW Molenaar (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*. Springer-Verlag.

Glas CA (1999). "Modification Indices for the 2-PL and the Nominal Response Model." *Psychometrika*, **64**(3), 273–294.

Jöreskog KG (1971). "Simultaneous Factor Analysis in Several Populations." *Psychometrika*, **36**(4), 409–426.

Kievit RA (2010). "Representational Inertia: The Influence of Associative Knowledge on 3D Mental Transformations." *Technical report*, University of Amsterdam.

Kline P (1986). *A Handbook of Test Construction: Introduction to Psychometric Design*. Methuen.

Luce RD (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*, volume 8. Oxford University Press.

Lunn D (2003). "**WinBUGS** Development Interface (**WBDev**)." *ISBA Bulletin*, **10**(3), 10–11.

Lunn D, Thomas A, Best NG, Spiegelhalter DJ (2000). "**WinBUGS** – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing*, **10**(4), 325–337.

Maydeu-Olivares A, Joe H (2005). "Limited-and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework." *Journal of the American Statistical Association*, **100**(471), 1009–1020.

Maydeu-Olivares A, Joe H (2006). "Limited Information Goodness-of-Fit Testing in Multidimensional Contingency Tables." *Psychometrika*, **71**(4), 713–732.

Meijer RR, Sijtsma K (2001). "Methodology Review: Evaluating Person Fit." *Applied Psychological Measurement*, **25**(2), 107–135.

Mellenbergh GJ (1989). "Item Bias and Item Response Theory." *International Journal of Educational Research*, **13**(2), 127–143.

Meredith W (1993). "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika*, **58**(4), 525–543.

Mislevy RJ, Bock RD (1990). ***BILOG**: Item Analysis and Test Scoring with Binary Logistic Models*. Scientific Software International Inc.

Molenaar D (2014). ***diffIRT**: Diffusion IRT Models for Response and Response Time Data*. R package version 1.4, URL http://CRAN.R-project.org/package=diffIRT.

Muthén LK, Muthén BO (2007). ***Mplus** User's Guide*. 5th edition. Muthén & Muthén, Los Angeles.

Navarro DJ, Fuss IG (2009). "Fast and Accurate Calculations for First-Passage Times in Wiener Diffusion Models." *Journal of Mathematical Psychology*, **53**(4), 222–230.

Neale MC, Boker SM, Xie G, Maes HH (2006). ***Mx**: Statistical Modeling*. 7th edition. Department of Psychiatry, VCU Box 900126, Richmond, VA 23298.

Nocedal J, Wright S (2006). *Numerical Optimization*. Series in Operations Research and Financial Engineering. Springer-Verlag.

Ramsay JO (1989). "A Comparison of Three Simple Test Theory Models." *Psychometrika*, **54**(3), 487–499.

Ranger J, Kuhn JT (2014). "An Accumulator Model for Responses and Response Times in Tests Based on the Proportional Hazards Model." *British Journal of Mathematical and Statistical Psychology*, **67**(3), 388–407.

Rasch G (1960). "Probabilistic Models for Some Intelligence and Attainment Tests." *Technical report*, Danmarks Paedogogiske Institut, Copenhagen, Denmark.

Ratcliff R (1978). "A Theory of Memory Retrieval." *Psychological Review*, **85**(2), 59.

Ratcliff R, Rouder JN (1998). "Modeling Response Times for Two-Choice Decisions." *Psychological Science*, **9**(5), 347–356.

Ratcliff R, Thapar A, Gomez P, McKoon G (2004). "A Diffusion Model Analysis of the Effects of Aging in the Lexical-Decision Task." *Psychology and Aging*, **19**(2), 278–289.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rizopoulos D (2006). "**ltm**: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software*, **17**(5), 1–25. URL http://www.jstatsoft.org/v17/i05/.

Rouder JN, Province JM, Morey RD, Gomez P, Heathcote A (2015). "The Lognormal Race: A Cognitive-Process Model of Choice and Latency with Desirable Psychometric Properties." *Psychometrika*, **80**(2), 491–513.

Samejima F (1969). "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph Supplement, No 17*.

SAS Institute Inc (2013). *SAS/STAT Software, Version 13.2*. SAS Institute Inc., Cary. URL http://www.sas.com/.

Schermelleh-Engel K, Moosbrugger H, Müller H (2003). "Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures." *Methods of Psychological Research Online*, **8**(2), 23–74.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464.

Sclove SL (1987). "Application of Model-Selection Criteria to Some Problems in Multivariate Analysis." *Psychometrika*, **52**(3), 333–343.

Smyth G, Hu Y, Dunn P, Phipson B, Chen Y (2015). ***statmod**: Statistical Modeling*. R package version 1.4.21, URL http://CRAN.R-project.org/package=statmod.

Spearman C (1904). ""General Intelligence", Objectively Determined and Measured." *The American Journal of Psychology*, **15**(2), 201–292.

Spiegelhalter DJ, Best NG, Carlin BP, Linde AVD (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society B*, **64**(4), 583–639.

Strandmark NL, Linn RL (1987). "A Generalized Logistic Item Response Model Parameterizing Test Score Inappropriateness." *Applied Psychological Measurement*, **11**(4), 355–370.

Stroud AH, Secrest D (1966). *Gaussian Quadrature Formulas*. Prentice-Hall Series in Automatic Computation. Prentice-Hall.

Thissen D (1991). ***MULTILOG*** *User's Guide: Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory*. Scientific Software International.

Thurstone LL (1931). "Multiple Factor Analysis." *Psychological Review*, **38**(5), 406–427.

Tuerlinckx F, De Boeck P (2005). "Two Interpretations of the Discrimination Parameter." *Psychometrika*, **70**(4), 629–650.

Tuerlinckx F, Maris E, Ratcliff R, De Boeck P (2001). "A Comparison of Four Methods for Simulating the Diffusion Process." *Behavior Research Methods, Instruments, & Computers*, **33**(4), 443–456.

Tuerlinckx F, Molenaar D, van der Maas HLJ (2016). "Diffusion-Based Item Response Modeling." In WJ van der Linden, RK Hambleton (eds.), *Handbook of Modern Item Response Theory*. Chapman & Hall/CRC.

van der Linden WJ (2007). "A Hierarchical Framework for Modeling Speed and Accuracy on Test Items." *Psychometrika*, **72**(3), 287–308.

van der Maas HLJ, Molenaar D, Maris G, Kievit RA, Borsboom D (2011). "Cognitive Psychology Meets Psychometric Theory: On the Relation Between Process Models for Decision Making and Latent Variable Models for Individual Differences." *Psychological Review*, **118**(2), 339–356.

Vandekerckhove J, Tuerlinckx F, Lee MD (2011). "Hierarchical Diffusion Models for Two-Choice Response Times." *Psychological Methods*, **16**(1), 44.

Voss A, Voss J (2007). "**fast-dm**: A Free Program for Efficient Diffusion Model Analysis." *Behavior Research Methods*, **39**(4), 767–775.

Wagenmakers EJ, Ratcliff R, Gomez P, McKoon G (2008). "A Diffusion Model Account of Criterion Shifts in the Lexical Decision Task." *Journal of Memory and Language*, **58**(1), 140–159.

Wagenmakers EJ, van der Maas HLJ, Grasman RPPP (2007). "An EZ-Diffusion Model for Response Time and Accuracy." *Psychonomic Bulletin & Review*, **14**(1), 3–22.

Wechsler D (1997). *WAIS-III: Wechsler Adult Intelligence Scale*. Psychological Corporation San Antonio.

Wicherts JM, Dolan CV, Hessen DJ (2005). "Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance." *Journal of Personality and Social Psychology*, **89**(5), 696–716.

Affiliation:

Dylan Molenaar
Psychological Methods
Department of Psychology
University of Amsterdam
1018 XA, Amsterdam, The Netherlands
Telephone: +31/205256584
E-mail: D.Molenaar@uva.nl

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Issue 4 *Submitted:* 2013-11-07
August 2015 *Accepted:* 2014-10-20