British Journal of Mathematical and Statistical Psychology (2017), 70, 209–224

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Paper

Analysing model fit of psychometric process models: An overview, a new test and an application to the diffusion model

Jochen Ranger[1]*, Jörg-Tobias Kuhn[2] and Carsten Szardenings[2]

[1]Martin Luther University Halle-Wittenberg, Germany

[2]University of Münster, Germany

Cognitive psychometric models embed cognitive process models into a latent trait framework in order to allow for individual differences. Due to their close relationship to the response process the models allow for profound conclusions about the test takers. However, before such a model can be used its fit has to be checked carefully. In this manuscript we give an overview over existing tests of model fit and show their relation to the generalized moment test of Newey (*Econometrica*, 53, 1985, 1047) and Tauchen (*J. Econometrics*, 30, 1985, 415). We also present a new test, the Hausman test of misspecification (Hausman, *Econometrica*, 46, 1978, 1251). The Hausman test consists of a comparison of two estimates of the same item parameters which should be similar if the model holds. The performance of the Hausman test is evaluated in a simulation study. In this study we illustrate its application to two popular models in cognitive psychometrics, the Q-diffusion model and the D-diffusion model (van der Maas, Molenaar, Maris, Kievit, & Boorsboom, *Psychol Rev.*, 118, 2011, 339; Molenaar, Tuerlinckx, & van der Maas, *J. Stat. Softw.*, 66, 2015, 1). We also compare the performance of the test to four alternative tests of model fit, namely the M_2 test (Molenaar *et al.*, *J. Stat. Softw.*, 66, 2015, 1), the moment test (Ranger *et al.*, *Br. J. Math. Stat. Psychol.*, 2016) and the test for binned time (Ranger & Kuhn, *Psychol. Test. Asess. Model.*, 56, 2014b, 370). The simulation study indicates that the Hausman test is superior to the latter tests. The test closely adheres to the nominal Type I error rate and has higher power in most simulation conditions.

1. Introduction

Cognitive psychometrics is a new paradigm that aims to apply cognitive process models to psychological assessment (Batchelder, 2007). This is approached by embedding a cognitive process model into a latent trait framework in order to account for individual differences. The resulting model then serves as a measurement model by which the responses and response times in a test can be analysed. Several such models have lately been proposed, among them latent trait versions of the race model (Ranger & Kuhn, 2014a; Ranger, Kuhn, & Gaviria, 2015; Rouder, Province, Morey, Gomez, & Heathcote, 2015; Tuerlinckx & De Boeck, 2005) and of the diffusion model (Molenaar, Tuerlinckx, &

*Correspondence should be addressed to Jochen Ranger, Martin-Luther-Universität Halle-Wittenberg, Institut für Psychologie, 06099 Halle, Germany (email: jochen.ranger@psych.uni-halle.de).

DOI:10.1111/bmsp.12082

van der Maas, 2015; Tuerlinckx & De Boeck, 2005; Tuerlinckx, Molenaar, & van der Maas, 2016; Vandekerckhove, Tuerlinckx, & Lee, 2011; van der Maas, Molenaar, Maris, Kievit, & Boorsboom, 2011).

Due to their close relation to the response process the psychometric process models allow for profound conclusions about the test takers. These conclusions, however, are only justified if the chosen model is able to represent the response processes adequately. According to Standard 3.9 of the Standards for Educational and Psychological Testing, one has to verify the appropriateness of a measurement model before applying it in psychological assessment (Sinharay & Haberman, 2014). This requires a thorough analysis of model fit with statistical tests that evaluate the chosen model in a rigorous manner. What is needed are counterparts to the tests of model fit and item fit that are available for item response models (Glas, 2016; Mavridis, Moustaki, & Knott, 2007; Maydeu-Olivares, 2013; Swaminathan, Hambleton, & Rogers, 2006). In this paper we describe several such tests. Some of the tests have been suggested before, while others are new. All are variants of the generalized moment test (Newey, 1985; Tauchen, 1985). The paper is organized as follows. In Section 2, the tests will be described. Although the tests are not coupled to a specific process model, only their application to the diffusion model is discussed in detail due to space restrictions. Applications to other process models are similar. In Section 3, we compare the tests in a simulation study. Again, the focus is on their performance in the diffusion model. Section 4 concludes.

2. Tests of model fit

In experimental psychology the fit of a process model can be evaluated in three ways. One can check whether predictions from the process model are roughly met by the data. The diffusion model, for example, implies a skewed response time distribution, a constant hazard function in the right tail, a functional relation between the mean and the standard deviation of the response times, and time differences between correct and incorrect responses (Ratcliff & McKoon, 2008; Wagenmakers, 2009). One can assess model fit with diagnostic plots such as quantile–quantile (Q–Q) or quantile–probability plots (Ratcliff & Smith, 2004). Or one can perform a statistical test of model fit such as the χ^2 test (Ratcliff & Smith, 2004) or Kolmogorov–Smirnov test (Voss & Voss, 2008). Unfortunately, these techniques are not directly applicable to psychometric process models for several reasons. In experimental psychology the fit of a model is usually analysed for each experimental condition separately. This is suboptimal in cognitive psychometrics where the data are multivariate and model fit includes the model's capability to account for dependencies between the responses and response times from different items. Furthermore, the common tests in experimental psychology assess the fit of a model at the subject level by exploiting the fact that each subject is observed under each experimental condition repeatedly. In cognitive psychometrics each test taker responds to each item just once. Hence, an analysis of model fit at the subject level requires an estimate of the test taker's latent trait and a statistical approach that accounts for the effect of trait estimation. Last, but not least, the statistical tests of model fit in experimental psychology require specific approaches to model estimation, such as the χ^2 approach (Ratcliff & Tuerlinckx, 2002) or the Kolmogorov–Smirnov approach (Voss & Voss, 2008), which are not used in cognitive psychometrics (Molenaar *et al.*, 2015; Ranger, Kuhn, & Szardenings, 2016).

In cognitive psychometrics a thorough evaluation of model fit comprises the assessment of the model's appropriateness for the joint distribution of the responses and the response

times in all items. This can be approached with multivariate generalizations of Q–Q plots (Dhar, Chakraborty, & Chaudhuri, 2014) or checks based on the Rosenblatt transform (Rosenblatt, 1952). Alternatively, one can assess the global fit of the model with statistical tests. An ideal basis for such tests is the generalized moment test proposed by Newey (1985) and Tauchen (1985). Let $\mathbf{x}$ and $\mathbf{t}$ denote the responses and the response times in the items of a psychological test, respectively, and let $\boldsymbol{\omega}$ denote the parameters of the model. Generalized moment tests are based on auxiliary functions $m_j(\mathbf{x}, \mathbf{t}, \boldsymbol{\omega})$ which are known to have an expectation of zero in the case of model fit. These functions are averaged $(1/N)\sum_{n=1}^{N} m_j(\mathbf{x}_n, \mathbf{t}_n, \hat{\boldsymbol{\omega}})$ over the N test takers of a calibration sample, which was previously drawn in order to determine the estimate $\hat{\boldsymbol{\omega}}$ of the model's parameters. If the assumed model is valid or provides a good approximation of the true data generation process the sample average should be close to zero. Several tests for the fit of psychometric process models – among them those of Molenaar *et al.* (2015), Ranger and Kuhn (2014b) and Ranger *et al.* (2016) – can be shown to be variants of the generalized moment test. In addition to these tests we propose a new one, the Hausman test of misspecification (Hausman, 1978). All four tests are described more thoroughly below.

2.1. The Molenaar et al. (2015) test

The Molenaar *et al.* (2015) test is an adaptation of the M_2 test proposed by Maydeu-Olivares and Joe (2005) for item response models and is limited to an analysis of the responses. The test assesses whether a model is able to account for the association of the responses in different items. The target quantities are the observed frequencies of correct responses in two items that are contained in the marginal cross-tabulations of order 2. These frequencies are compared to the corresponding expected frequencies that are implied by the process model and the maximum likelihood estimates $\hat{\boldsymbol{\omega}}_{\mathrm{ML}}$. Let $\mathbf{o}$ denote the observed frequencies in all item pairs that were stacked into a vector and let $\mathbf{e}(\hat{\boldsymbol{\omega}}_{\mathrm{ML}})$ denote the corresponding expected frequencies. The compatibility of both vectors can be tested via the test statistic

$$M_2 = \frac{1}{N}(\mathbf{o} - \mathbf{e}(\hat{\boldsymbol{\omega}}_{\mathrm{ML}}))^t \mathbf{C}(\mathbf{o} - \mathbf{e}(\hat{\boldsymbol{\omega}}_{\mathrm{ML}})), \tag{1}$$

where $\mathbf{C}$ is a matrix derived from the asymptotic covariance matrix of the residuals $\mathbf{o} - \mathbf{e}(\hat{\boldsymbol{\omega}}_{\mathrm{ML}})$ and an orthogonal complement of the gradient vector; see Maydeu-Olivares and Joe (2005) for more details. The distribution of M_2 converges to χ^2 with $s - m$ degrees of freedom, where s is the number of elements in $\mathbf{o}$ and m the number of the estimated item parameters (Maydeu-Olivares & Joe, 2005). The test was implemented by Molenaar *et al.* (2015) for the diffusion model and is part of the `diffIRT` package of the software environment R. Although one motivation for the test was the relation of the diffusion model to the two-parameter logistic model (Tuerlinckx & De Boeck, 2005), the test can be used for other process models also, as the test is a generalized moment test with auxiliary function $\mathrm{E}(\mathbf{o} - \mathbf{e}(\boldsymbol{\omega})) = \mathbf{0}$. The test evaluates the global fit of the model, but item-specific variants can also be implemented (Maydeu-Olivares, 2013).

2.2. The Ranger et al. (2016) test

In structural equation modelling the fit of a model is evaluated by comparing the observed and implied covariance matrix (Shapiro, 1986). This approach can be adapted to

psychometric process models. Denote by $\mathbf{s}$ the vector of stacked non-redundant elements of the covariance matrix of the responses and response times and by $\boldsymbol{\sigma}(\hat{\boldsymbol{\omega}}_{\text{ULS}})$ the corresponding vector which is implied by the model and the parameter estimates $\hat{\boldsymbol{\omega}}_{\text{ULS}}$. Here, least squares estimates (Ranger *et al.*, 2016) are used as this permits the application of the tests developed in structural equation modeling. Using the quantities defined above, the fit can be assessed via the residuals $\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\omega}}_{\text{ULS}})$, which have an expectation of zero if the model holds. This assumption can be tested with all the tests developed in structural equation modelling for non-normal data, such as the asymptotic distribution-free test of Browne (1984). Define the test statistic

$$M_{\text{ULS}} = (\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\omega}}_{\text{ULS}}))^t(\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\omega}}_{\text{ULS}})). \tag{2}$$

The test statistic is asymptotically distributed as a mixture of independent χ^2 distributions (Maydeu-Olivares, 2013). This follows from the fact that the residuals have a multivariate normal distribution as both the observed and implied elements of the covariance matrix are approximately normally distributed in large samples. The mixing proportions depend on the covariance matrix of the residuals which is a function of the expected fourth-order moments and some Jacobian matrix (Browne, 1984; Shapiro, 1986). The distribution of M_{ULS} can be approximated by a standard χ^2 distribution when correcting the degrees of freedom (Maydeu-Olivares, 2013; Yuan & Bentler, 2010). Similar to the M_2 test of Molenaar *et al.* (2015), the M_{ULS} test is a test of global fit. It improves upon the former test by considering response times in addition to the responses. The test was implemented for the diffusion model by Ranger *et al.* (2016).

2.3. The Ranger and Kuhn (2014b) test

Ranger and Kuhn (2014b) proposed a general test of model fit for response and response time models. The test is similar to the χ^2 test of Ratcliff and Smith (2004) and requires binned data. It differs from the latter test by using marginal maximum likelihood estimates of the item parameters, a fact that complicates the analysis (Chernoff & Lehmann, 1954). For the test, the response times in each item g have to be categorized into intervals defined by the prespecified cut points $\mathbf{c}_g = (c_{0g} = 0, c_{1g}, \ldots, c_{Kg} = \infty)$. Then, for each time interval defined by two consecutive cut points the number of correct and incorrect responses within the interval is counted. The counts are stacked into the vector $\mathbf{o}_g(\mathbf{c}_g)$. In addition, the corresponding expected numbers of responses $\mathbf{e}_g(\mathbf{c}_g, \hat{\boldsymbol{\omega}}_{\text{ML}})$ are determined from the marginal cumulative distribution function of the model using the marginal maximum likelihood estimates $\hat{\boldsymbol{\omega}}_{\text{ML}}$ of the model parameters. The observed and expected frequencies in all items are stacked into the vector $\mathbf{o}(\mathbf{c})$ and the vector $\mathbf{e}(\mathbf{c}, \hat{\boldsymbol{\omega}}_{\text{ML}})$. Model fit can be evaluated with the elements of the residual vector $\mathbf{o}(\mathbf{c}) - \mathbf{e}(\mathbf{c}, \hat{\boldsymbol{\omega}}_{\text{ML}})$. These elements are normally distributed random variates with expectation zero in the case of model fit (Ranger & Kuhn, 2014b). Define the test statistic

$$M_{\text{B}} = (\mathbf{o}(\mathbf{c}) - \mathbf{e}(\mathbf{c}, \hat{\boldsymbol{\omega}}_{\text{ML}}))^t(\mathbf{o}(\mathbf{c}) - \mathbf{e}(\mathbf{c}, \hat{\boldsymbol{\omega}}_{\text{ML}})), \tag{3}$$

where the redundant elements of the residual vector have been removed in order to avoid linear dependencies. The test statistic is a quadratic form, such that its asymptotic distribution can be approximated by the χ^2 distribution. One has to adjust the degrees of freedom in order to account for the effect of estimating the item parameters though (Ranger & Kuhn, 2014b). The M_{B} test is a test of global fit. The placement of the cut points

determines the power for detecting different forms of misfit. Using roughly equiprobable intervals is known to have some optimum properties (Tauchen, 1985). Alternatively, one can test for deviations in tail areas of the response time distribution. Tests of item fit can be derived by using just the residuals from a single item. The test has not been used for the diffusion model so far; for applications to race models, see Ranger and Kuhn (2014a) and Ranger *et al.* (2015).

2.4. The Hausman (1978) test

The Hausman test of misspecification in based on a comparison of two different estimates of the same item parameters. The first estimates $\hat{\boldsymbol{\omega}}_1$ are provided by an estimator that is consistent and efficient, that is, attains the Cramer–Rao bound. An example of such an estimator is the marginal maximum likelihood estimator. The second estimates $\hat{\boldsymbol{\omega}}_2$ are derived with an estimator that is consistent, but not efficient. Examples of such an estimator are the least squares estimator and the χ^2 estimator. Hausman (1978) suggested using the difference of the two estimates as a measure of model misspecification and proposed the test statistic

$$M_{\mathrm{H}} = (\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1)^t \boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1} (\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1), \tag{4}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1}$ denotes the asymptotic covariance matrix of the differences. As both estimators are normally distributed in large samples and the difference is a linear combination with expectation zero in the case of model fit, the test statistic M_{H} follows approximately a χ^2 distribution. The covariance matrix of the differences $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1}$ is the difference of the covariance matrix of the inefficient and the efficient estimator $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\omega}}_2 - \hat{\boldsymbol{\omega}}_1} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\omega}}_2} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\omega}}_1}$. This follows from the fact that the efficient estimator does not correlate with the difference, as otherwise a linear combination of the efficient estimator and the difference would yield a consistent estimator with lower asymptotic variance (Hausman, 1978; White, 1982). The Hausman test requires that the two estimators converge to different limits under misspecification. The marginal maximum likelihood estimator converges to the values that best approximate the data in the Kullback–Leibler sense. If model misspecification affects the inefficient estimator differently, the Hausman test is consistent. The Hausman test can be interpreted as a generalized moment test. It has been shown by Newey (1985) that for each Hausman test there exists a generalized moment test that is asymptotically equivalent under certain conditions. The Hausman test is a test of global fit, but can be used as a test of item fit when considering the parameters in single items. Up to now the Hausman test has not been used in psychometrics.

3. Performance of the tests: Illustration for the diffusion model

In order to compare the performance of the tests a simulation study was conducted. Due to space restrictions the simulation study is limited to the diffusion model. The diffusion model is the most popular psychometric process model for two alternative choice tasks in the context of time. It is implemented in standard software and has often been used successfully (Schubert, Hagemann, Voss, Schankin, & Gergmann, 2015; Vandekerckhove, 2014; White, Ratcliff, Vasey, & McKoon, 2009, 2010). The model is mathematically demanding and provides a good baseline for the tests. First, we briefly review the diffusion model with a special emphasis on two specific variants, the D-diffusion and the Q-diffusion

model (van der Maas *et al.*, 2011) in the version implemented by Molenaar *et al.* (2015). Then the tests, the simulation conditions and the results are described.

3.1. The diffusion model and its psychometric variants

The diffusion model (Ratcliff, 1978; Wagenmakers, 2009) is a sequential sampling model for responses and response times in binary decision tasks. The model describes the development of the momentary preference for one response option over the other which is represented by a single value $X(t)$. Positive (negative) values indicate preference for the first (second) response option. The momentary preference is a stochastic process, whose instantaneous changes can be described with the stochastic differential equation

$$\mathrm{d}X(t) = v\mathrm{d}t + \mathrm{d}B(t), \tag{5}$$

where $\mathrm{d}B(t)$ is a Brownian increment with infinitesimal variance $1^2\mathrm{d}t$ and v is the drift rate. Equation 5 describes how individuals change their momentary preference from one instant to the next by integrating small amounts of new information. The process of preference formation fluctuates around the expected preference $\mathrm{E}(X(t)) = vt + X(0)$ at time t according to a normal distribution. A response occurs as soon as the level of preference attains an upper or lower boundary, whereby reaching the upper (lower) boundary elicits the first (second) response. This moment, the so-called first hitting time, corresponds to the response time. The expected response time depends on the drift rate v and on the distance of the boundaries to the start point $X(0)$. Usually, a quantity d called non-decision time is added to the first hitting time, which accounts for delays not related to information processing.

In cognitive psychometrics the parameters of the model are related to characteristics of the test taker and to features of the items via a latent trait model. Thereby, the effect of an item is modelled by a fixed intercept and the effects of the test taker's characteristics by latent traits. Such latent trait versions of the diffusion model have been suggested by Tuerlinckx and De Boeck (2005), Vandekerckhove *et al.* (2011) and van der Maas *et al.* (2011); see Tuerlinckx *et al.* (2016) for an overview. The models differ, *inter alia*, in the way the model parameters are related to the latent traits. Here, we focus on two popular variants, the D-diffusion model and the Q-diffusion model (Tuerlinckx *et al.*, 2016; van der Maas *et al.*, 2011) in the version implemented by Molenaar *et al.* (2015). In both variants the test takers are unbiased, that is, start in a neutral state $X(0) = 0$ with equidistant boundaries. (This simplification is typically made in cognitive psychometrics. It is motivated by the difficulty of calibrating the model when it is too complex. Unbiasedness can also be justified theoretically by the fact that in psychological assessment the base rates are unknown and the same response is scored differently over the items. Nevertheless it might be worth studying the utility of asymmetric boundaries for modelling response styles like endorsement tendencies.) Two latent traits are assumed. The first trait θ_i represents the information processing capability of test taker i and is related to the drift rate. The second trait τ_i is related to the boundary separation and reflects the test taker's response cautiousness or conscientiousness.

In the D-diffusion model, originally proposed by Tuerlinckx and De Boeck (2005) and slightly modified by van der Maas *et al.* (2011), the item-specific drift rate v_{ig} and the distance between the two boundaries, the boundary separation a_{ig}, are modelled as

$$\begin{aligned} v_{ig}(\theta_i) &= \theta_i - \beta_g, \\ a_{ig}(\tau_i) &= \tau_i / \alpha_g, \end{aligned} \tag{6}$$

where β_g and α_g account for all attributes of item g such as its difficulty. The non-decision time d_g is considered to be a fixed item parameter. The quantities τ_i and α_g must be positive as the boundary separation cannot be negative. The drift rate can assume negative values such that the probability of choosing the first response ranges from 0 to 1. The D-diffusion model should be used for personality tests or attitudinal scales (van der Maas *et al.*, 2011).

The Q-diffusion model of van der Maas *et al.* (2011) decomposes the diffusion model parameters differently, namely as

$$\begin{aligned} v_{ig}(\theta_i) &= \theta_i / \beta_g, \\ a_{ig}(\tau_i) &= \tau_i / \alpha_g. \end{aligned} \tag{7}$$

The latent traits and item parameters have a similar interpretation to the D-diffusion model. Here all quantities must be positive, which implies that the probability of choosing the first response is at least .5. The Q-diffusion model was proposed for achievement tests (van der Maas *et al.*, 2011); for more thorough discussions of the models and their differences see Molenaar *et al.* (2015) and van der Maas *et al.* (2011).

3.2. Simulation conditions

The performance (size, power) of the different tests in the two variants of the diffusion model was assessed in a simulation study. The simulation study was based on a test of 12 items. Different simulation conditions were defined by systematically crossing two sample sizes (500 and 1,000 subjects) and five different data generation processes. The data generation processes differed with respect to model misspecification as follows:

3.2.1. Correctly specified model

In the first simulation setting the data were generated according to one of the two diffusion models. For the D-diffusion model the item parameters were set by fully crossing three values of the boundary parameter ($\alpha_1 = 0.37, \alpha_2 = 0.47, \alpha_3 = 0.61$) with four values of the drift parameter ($\beta_1 = -1.00$, $\beta_2 = -.50$, $\beta_3 = 0.50$, $\beta_4 = 1.00$). The non-decision time was $d_1 = 2$ in the first six items and $d_2 = 3$ in the last six items. The latent speed θ was normally distributed with standard deviation of $\sigma_\theta = 1$ and the latent response caution τ was log-normally distributed with scale parameter $\sigma_\tau = 0.3$. These parameter values implied average response times between 2.90 and 3.90, variances between 0.30 and 2.50 and solution probabilities between .25 and .80 in the 12 items. For the Q-diffusion model the parameters of the items were determined by fully crossing three values of the boundary parameter ($\alpha_1 = 0.37$, $\alpha_2 = 0.47$, $\alpha_3 = 0.61$) with four values of the drift parameter ($\beta_1 = 1.00$, $\beta_2 = 1.22$, $\beta_3 = 1.49$, $\beta_4 = 1.82$). The non-decision times were again set to $d_1 = 2$ in the first six items and to $d_2 = 3$ in the last six items. The latent traits of the test were log-normally distributed with scale parameters of $\sigma_\theta = 0.3$ and $\sigma_\tau = 0.3$. These values implied average response times between 2.9 and 4.1, variances between 0.3

and 2.8 and solution probabilities between .75 and .95. The first simulation setting resembled the simulation study of Molenaar *et al.* (2015) and served to assess the size of the tests.

3.2.2. Multidimensionality

In the second simulation setting we assessed whether the tests are able to detect multidimensionality. The data were simulated as in the first simulation setting, with one exception. The latent speed and the latent response cautiousness in the first half of the test (items 1–6) were different from the latent speed and the latent response cautiousness in the second half of the test (items 7–12). A correlation of $\rho = .5$ was assumed between the corresponding quantities of the two test halves. Such a violation was supposed to occur in items that tap different, but related content.

3.2.3. Shifted response times

In the third simulation setting the distribution of the response times was misspecified locally. In a first step the data were generated as in the first simulation setting. Then the response times in items 1, 4, 7 and 10 were shifted by adding a delay to the original response times in the incorrect responses. The delay was drawn from an exponential distribution with rate $\lambda = 0.5$. The third simulation setting dealt with the power of the tests to detect response time differences between correct and incorrect responses. This is a standard question when assessing model fit (van der Linden & Glas, 2009).

3.2.4. Complete model misspecification

The fourth simulation setting dealt with complete model misspecification. In this setting the data were generated according to the hierarchical model of van der Linden (2007). This model combines a unidimensional standard factor model for the log response times with a unidimensional three-parameter logistic model for the responses. The latent traits of both models are assumed to be correlated. The item parameters of the model were chosen in such a way that the model mimicked the data from the diffusion models as closely as possible. Therefore, a data set was generated as in the first simulation setting. The data set was analysed with the model of van der Linden (2007), whereby the guessing parameter was set to zero. The resulting item parameters were then used in order to generate the simulation data sets with the model of van der Linden (2007). The fourth simulation setting addressed the question whether the tests can distinguish between theoretically different, but empirically similar models.

3.2.5. Contaminated data

In the fifth simulation setting the effect of rapid guessing was simulated. Data were generated as in the first simulation setting. Then, for a random sample of 5% of the subjects, the original response times were replaced by random draws from a shifted exponential distribution with rate $\lambda = 2$ and shift parameter $d = 1$. The responses were replaced by random draws from the binomial distribution with parameter $\pi = 0.5$. Hence, the subjects in the subsample responded fast, with little variance and were correct on chance level.

Fully crossing the experimental factors generated $2 \times 2 \times 5$ simulation conditions defined by one of two variants of the diffusion model, one of two sample sizes and one of five data generating processes. For each simulation condition 250 data sets were generated. The simulation was conducted with the statistical software R.

3.3. Data analysis

The simulation samples were analysed as follows. First, the diffusion model was fitted to the data. As the different tests are based on different estimators, the item parameters were estimated three times: with marginal maximum likelihood estimation (Molenaar *et al.*, 2015) and with unweighted and diagonally weighted least squares estimation (Ranger *et al.*, 2016). For results concerning parameter recovery, see the simulation study of Ranger *et al.* (2016). Then, the tests of model fit were performed.

The first test was the M_2 test of Molenaar *et al.* (2015) which is implemented in the `RespFit` routine of the `diffIRT` package. The second test was the M_{ULS} test of Ranger *et al.* (2016) given in equation 2. The next tests were different implementations of the M_{B} test. For a first variant of the test no distinction was made between the response times in correct and incorrect responses. Time intervals were defined by four cut points that corresponded approximately to the sample quantiles. For these intervals, the expected and observed frequencies were compared according to equation 3. A global test ($M_{\mathrm{B\cdot GT}}$ test) considered the frequencies in all items jointly. Additionally, we implemented an item-specific version ($M_{\mathrm{B\cdot IT}}$ test) that compared the frequencies in each item separately. As the two tests focus on the response times we conjectured that they would be able to detect misspecifications of the response time distribution with high power. For a second variant of the M_{B} test the number of responses between fixed cut points was determined for correct and incorrect responses separately. Therefore, the response times were binned at three cut points. The correspondence between expected and observed frequencies was tested jointly for all items ($M_{\mathrm{B\cdot GXT}}$ test) and for each item separately ($M_{\mathrm{B\cdot IXT}}$ test). This variant of the test checks the model's adequacy for the joint distribution of the responses and the response times. For all tests the expected frequencies were determined with the package `RWiener` (Wabersich, 2014).

Finally, the marginal maximum likelihood and the diagonally weighted least squares estimates were compared with the Hausman test. The diagonally weighted least squares estimator was used as this estimator was superior to unweighted least squares estimation with respect to efficiency and convergence to the normal distribution. Three versions of the Hausman test were implemented. A global test of model fit ($M_{\mathrm{H\cdot G}}$ test) was implemented that compared the two estimates of the boundary separation parameter α_g in all items jointly. The global test was complemented by a test of item fit ($M_{\mathrm{H\cdot I1}}$ test) that tested the equality in each item separately. In addition, a second test of item fit was considered. This test ($M_{\mathrm{H\cdot I2}}$ test) was based on a comparison of the two estimates of the non-decision time d_g. The other item parameters were excluded as the distribution of their estimates was rather far from normality. This made them inappropriate for the Hausman test which depends crucially on the approximate normality of the parameter estimates.

An overview of the different tests is given in Table 1. Not all data sets could be analysed with all these tests. The Hausman tests and the M_2 test were not applicable to a minority of the data sets where the information matrix of the marginal maximum likelihood estimator became singular.

Table 1. Overview of the different tests used in the simulation study

Test	Variant	Focus	Target quantity
M_2	–	Global	Responses in cross-tabulations
M_{ULS}	–	Global	Elements of covariance matrix
M_{B}	$M_{\text{B·GT}}$	Global	Frequencies of binned times
	$M_{\text{B·IT}}$	Item	Frequencies of binned times
	$M_{\text{B·GXT}}$	Global	Frequencies of binned times/responses
	$M_{\text{B·IXT}}$	Item	Frequencies of binned times/responses
M_{H}	$M_{\text{H·G}}$	Global	Boundary separation estimates
	$M_{\text{H·I1}}$	Item	Boundary separation estimates
	$M_{\text{H·I2}}$	Item	Non-decision time estimates

3.4. Simulation results

The tests were evaluated with respect to their size and power. The results for the first simulation setting can be found in Table 2. This table contains the empirical rejection rates of the tests for different nominal Type I error rates. Note that the empirical rejection rates should be close to the nominal Type I error rates as no misspecification was present. In Table 2 the results of the item-specific tests have been averaged over the items.

None of the tests of global fit attains the nominal Type I error rate. The only test that is close to it is the M_{ULS} test in samples of 1,000 subjects. The performance of this test, however, declines in samples with 500 subjects, especially in the D-diffusion model. The M_2 test is slightly too conservative, while the $M_{\text{H·G}}$ test is slightly too liberal. The $M_{\text{B·GT}}$ test and the $M_{\text{B·GXT}}$ test are far too liberal, especially in the D-diffusion model, where the test should not be used. The problematic behaviour of the tests for binned time might be caused by the slow convergence of the maximum likelihood estimator to the normal distribution. Non-normality has a strong effect on these tests as they depend on all item parameters and their asymptotic distribution via the expected frequencies. The tests of

Table 2. Empirical rejection rates of the tests for different nominal Type I error rates α and two sample sizes in the D-diffusion and Q-diffusion model in the absence of model misspecification

	D-diffusion model						Q-diffusion model					
Model	1,000			500			1,000			500		
Sample												
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
Global fit												
M_2	0.07	0.03	0.00	0.06	0.03	0.01	0.07	0.04	0.01	0.05	0.03	0.00
M_{ULS}	0.09	0.05	0.02	0.04	0.02	0.00	0.09	0.05	0.02	0.08	0.05	0.02
$M_{\text{H·G}}$	0.14	0.09	0.03	0.14	0.09	0.02	0.12	0.08	0.02	0.10	0.05	0.02
$M_{\text{B·GT}}$	0.32	0.21	0.06	0.22	0.11	0.04	0.14	0.08	0.02	0.16	0.09	0.02
$M_{\text{B·GXT}}$	0.80	0.75	0.60	0.58	0.48	0.32	0.11	0.05	0.01	0.12	0.06	0.02
Item fit												
$M_{\text{H·I1}}$	0.10	0.06	0.02	0.10	0.06	0.02	0.10	0.05	0.01	0.09	0.05	0.01
$M_{\text{H·I2}}$	0.08	0.04	0.01	0.08	0.04	0.01	0.10	0.05	0.01	0.08	0.04	0.01
$M_{\text{B·IT}}$	0.15	0.08	0.02	0.12	0.07	0.02	0.11	0.05	0.01	0.11	0.05	0.02
$M_{\text{B·IXT}}$	0.35	0.22	0.08	0.25	0.16	0.05	0.11	0.06	0.01	0.10	0.05	0.01

Notes. Results based on 250 simulation samples. Results for test of item fit are averaged over items.

item fit perform better. In general, they are close to the nominal Type I error rate. The only exceptions are the $M_{B\cdot G}$ test and the $M_{B\cdot IXT}$ test in the D-diffusion model. In summary, the results indicate that with respect to the Type I error rate the M_{ULS} test, the M_2 test and the different versions of the Hausman test are unproblematic, while the tests for binned time do not work as expected.

In addition to the size of the tests, we investigated their power. Therefore, the empirical rejection rates in the remaining simulation settings were determined. The results for the D-diffusion and the Q-diffusion model can be found in Tables 3 and 4, respectively. The results concerning the $M_{B\cdot GT}$ test should be interpreted with care as this test has an elevated Type I error rate. Furthermore, no results are reported in Table 3 for the $M_{B\cdot GXT}$ test due to its erratic behaviour in the first simulation setting in the D-diffusion model. The rejection rates of the tests of item fit are reported separately for items with and without misfit in the simulation setting with shifted response times where this distinction could be made. The results for the tests of item fit have been averaged over the items.

The power of the tests depends strongly on the simulation setting. Multidimensionality can be detected by all tests of global fit, with the exception of the M_B tests for binned time. The M_{ULS} test performs best, followed by the Hausman test $M_{H\cdot G}$. The M_2 test is only powerful in the case of the D-diffusion model. The item-specific tests generally have lower power than their global counterparts. Complete model misspecification is revealed by the different versions of the Hausman test with high probability. The other tests clearly fall behind. The M_2 test even lacks any power. This is hardly surprising, as the diffusion model and the van der Linden (2007) model imply the same distribution for the responses. A shift in the response time distribution is hard to detect. The test for binned time $M_{B\cdot GXT}$ is most powerful but can only be used for the Q-diffusion model due to the Type I error inflation in the D-diffusion model. The global version of the Hausman test $M_{H\cdot G}$ performs well. The other tests have little power. If only some items are affected by the shift, the item-specific versions of the Hausman test ($M_{H\cdot I1}/M_{H\cdot I2}$) are able to detect the affected items with moderate power. This does not come at the price of a high false-alarm rate which is near the nominal Type I error rate in the $M_{H\cdot I2}$ test. For contaminated data all tests indicate model misspecification, with the exception of the M_{ULS} test. This has a simple explanation. Contamination affected the marginal maximum likelihood estimates tremendously. As the contaminated response times were relatively fast, the estimated non-decision times were highly distorted, and this also affected the other item parameters. The weighted least squares estimator was less compromised by contamination.

The results can be summarized as follows. The global tests based on responses and response times usually perform better than the local tests or the tests based on the responses. The Hausman test seems to be a good all-purpose test as it usually has high power. However, it is very sensitive to outliers, a property that might not always be desirable.

4. Discussion

Cognitive psychometric models merge cognitive process models and latent trait models into measurement models that allow for the assessment of individual differences. Due to their close relation to the response process the models allow for profound conclusions about the test takers. This is certainly desirable, but comes at the price of restrictive assumptions about the response process. Although cognitive psychometric models are based on models that have been used successfully in cognitive psychology for many years,

Table 3. Empirical rejection rates of the tests for different nominal Type I error rates α and two sample sizes in the D-diffusion model and the conditions with model misspecification

Condition	Multidimensionality				van der Linden model				Shifted RT				Contamination			
Sample	1,000		500		1,000		500		1,000		500		1,000		500	
α	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Global fit																
M_2	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.08	0.04	0.07	0.04	0.97	0.96	0.87	0.80
M_{ULS}	1.00	1.00	1.00	1.00	0.62	0.52	0.34	0.25	0.40	0.26	0.17	0.10	0.18	0.12	0.04	0.02
$M_{\text{H·G}}$	0.98	0.97	0.88	0.80	0.78	0.70	0.63	0.57	1.00	1.00	0.99	0.98	1.00	1.00	1.00	1.00
$M_{\text{B·GT}}$	0.58	0.43	0.34	0.22	0.58	0.43	0.34	0.24	0.97	0.93	0.73	0.62	1.00	1.00	1.00	1.00
Item fit: misfitting items																
$M_{\text{H·I1}}$	0.66	0.57	0.43	0.33	0.57	0.42	0.30	0.20	0.64	0.60	0.57	0.53	1.00	1.00	0.99	0.99
$M_{\text{H·I2}}$	0.54	0.44	0.29	0.19	0.83	0.76	0.61	0.53	0.59	0.54	0.54	0.51	1.00	1.00	1.00	1.00
$M_{\text{B·IT}}$	0.20	0.12	0.15	0.08	0.21	0.12	0.15	0.08	0.50	0.41	0.37	0.30	1.00	1.00	1.00	1.00
Item fit: fitting items																
$M_{\text{H·I1}}$	–	–	–	–	–	–	–	–	0.27	0.19	0.18	0.12	–	–	–	–
$M_{\text{H·I2}}$	–	–	–	–	–	–	–	–	0.13	0.07	0.07	0.03	–	–	–	–
$M_{\text{B·IT}}$	–	–	–	–	–	–	–	–	0.17	0.11	0.13	0.07	–	–	–	–

Notes. Results based on 250 simulation samples. Results for item-specific tests are averaged over items.

Table 4. Empirical rejection rates of the tests for different nominal Type I error rates α and two sample sizes in the Q-diffusion model and the conditions with model misspecification

Condition	Multidimensionality				van der Linden model				Shifted RT				Contamination			
Sample	1,000		500		1,000		500		1,000		500		1,000		500	
α	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
Global fit																
M_2	0.23	0.14	0.08	0.05	0.06	0.03	0.04	0.02	0.09	0.05	0.10	0.05	0.98	0.98	0.99	0.98
M_{ULS}	1.00	1.00	0.94	0.92	0.32	0.24	0.19	0.12	0.31	0.22	0.14	0.09	0.49	0.35	0.10	0.05
$M_{H\cdot G}$	0.60	0.51	0.42	0.33	0.90	0.86	0.75	0.68	0.62	0.49	0.40	0.29	1.00	1.00	1.00	1.00
$M_{B\cdot GT}$	0.26	0.15	0.17	0.08	0.27	0.15	0.22	0.14	0.17	0.10	0.16	0.07	1.00	1.00	1.00	1.00
$M_{B\cdot GXT}$	0.24	0.16	0.18	0.08	0.26	0.16	0.23	0.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Item fit: misfitting items																
$M_{H\cdot I1}$	0.24	0.17	0.17	0.11	0.57	0.47	0.33	0.24	0.38	0.27	0.25	0.15	1.00	1.00	1.00	1.00
$M_{H\cdot I2}$	0.19	0.12	0.11	0.05	0.43	0.31	0.20	0.12	0.42	0.33	0.26	0.18	1.00	1.00	1.00	1.00
$M_{B\cdot IT}$	0.13	0.07	0.12	0.06	0.13	0.08	0.13	0.08	0.12	0.06	0.12	0.06	1.00	1.00	1.00	1.00
$M_{B\cdot IXT}$	0.14	0.07	0.10	0.06	0.14	0.08	0.13	0.07	0.87	0.82	0.74	0.66	1.00	1.00	1.00	1.00
Item fit: fitting items																
$M_{H\cdot I1}$	–	–	–	–	–	–	–	–	0.11	0.07	0.12	0.07	–	–	–	–
$M_{H\cdot I2}$	–	–	–	–	–	–	–	–	0.10	0.05	0.09	0.05	–	–	–	–
$M_{B\cdot IT}$	–	–	–	–	–	–	–	–	0.11	0.06	0.11	0.06	–	–	–	–
$M_{B\cdot IXT}$	–	–	–	–	–	–	–	–	0.11	0.06	0.10	0.05	–	–	–	–

Notes. Results based on 250 simulation samples. Results for item-specific tests are averaged over items.

it is far from clear whether they can be generalized beyond simple perceptual decision tasks. Some of the assumptions inherent in cognitive process models might not be reasonable for non-transparent problems where the influx of information is not constant but dependent on the solution process itself. Such doubts can only be diminished by careful model checks.

In this paper we gave an overview of tests of model fit. We also proposed a new test, the Hausman test of model misspecification (Hausman, 1978).This test is little known in psychometrics. In the limited simulation study the test was superior to its rivals. It adhered to the nominal Type I error rate reasonably well and had high power in all simulation settings. One drawback of the Hausman test is the need for two estimates. However, very often these are readily available and the test can be implemented easily. Another drawback is its sensitivity to outliers, which is partly due to the lack of robustness of the maximum likelihood estimator. Although the ability to detect misfit in the form of outliers is not a defect, it might not always be desired. We did not pursue alternative tests from the generalized moment testing framework such as the information matrix test (White, 1982). A generalized χ^2 test with random cell boundaries (Moore & Spruill,1975) might also be an interesting candidate. Alternatively, one could implement a score test by embedding the psychometric process model into a more flexible model.

This paper was written from the perspective of a psychometrician. The focus was on tests that parallel those typically used in item response modelling. These tests assess the congruence of the model and the data internally and help to identify model violations. The results can then be analysed further in order to assess their practical relevance (Sinharay & Haberman, 2014).One could alternatively assess the validity of the model via an external validation. In experimental psychology it is common practice to assess whether experimental manipulations have the predicted effects on the model's parameters (Voss, Rothermund, & Voss, 2004). One could also try to connect parts of the model to neural activity (Purcell, Heitz, Cohen, Schall, Logan, & Palmeri, 2010). Such investigations ideally complement the proposed tests.

Acknowledgements

We would like to thank the editor and two referees for their constructive comments, which helped to improve the first version of this paper. We also thank for the invitation to provide a manuscript to the special issue on response time modeling.

References

Batchelder, W. (2007). *Cognitive psychometrics: Combining two psychological traditions*. Amsterdam, the Netherlands: CSCA Lecture.

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x

Chernoff, H., & Lehmann, E. (1954). The use of the maximum likelihood estimates in χ^2-tests for goodness of fit. *Annals of Mathematical Statistics*, *25*, 579–586. doi:10.1214/aoms/1177728726

Dhar, S., Chakraborty, B., & Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile-quantile plots and related tests. *Bernoulli*, *20*, 1484–1506. doi:10.3150/13-BEJ530

Glas, C. (2016). Frequentist model-fit tests. In W. van der Linden (Ed.), *Handbook of item response theory: Vol. 2. Statistical tools* (pp. 343–361). Boca Raton, FL: Chapman and Hall/CRC Press.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, *46*, 1251–1271. doi:10.2307/1913827

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 135–161). Amsterdam, the Netherlands: Elsevier.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*, 71–101. doi:10.1080/15366367.2013.831680

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*, 1009–1020. doi:10.1198/016214504000002069

Molenaar, D., Tuerlinckx, F., & van der Maas, H. (2015). Fitting diffusion item response theory models for responses and response times using the R package diffIRT. *Journal of Statistical Software*, *66*, 1–34. doi:10.18637/jss.v066.i04

Moore, D., & Spruill, M. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics*, *3*, 599–616. doi:10.1214/aos/1176343125

Newey, W. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica*, *53*, 1047–1070. doi:10.2307/1911011

Purcell, B., Heitz, R., Cohen, J., Schall, J., Logan, G., & Palmeri, T. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143. doi:10.1037/a0020311

Ranger, J., & Kuhn, J. (2014a). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *67*, 388–407. doi:10.1111/bmsp.12025

Ranger, J., & Kuhn, J. (2014b). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling*, *56*, 370–392.

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

Ranger, J., Kuhn, J., & Szardenings, C. (2016). Limited information estimation of the diffusion based item response theory model for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *69*(2), 122–138. doi:10.1111/bmsp.12064

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333–367. doi:10.1037/0033-295X.111.2.333

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481. doi:10.3758/BF03196302

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, *23*, 470–472. doi:10.1214/aoms/1177729394

Rouder, J., Province, J., Morey, R., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 491–513. doi:10.1007/s11336-013-9396-3

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Gergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence*, *51*, 28–46. doi:10.1016/j.intell.2015.05.002

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, *81*, 142–149. doi:10.1080/01621459.1986.10478251

Sinharay, S., & Haberman, S. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*, 23–35. doi:10.1111/emip.12024

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 683–718). Amsterdam, the Netherlands: Elsevier.

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, *30*, 415–443. doi:10.1016/0304-4076(85)90149-6

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

Tuerlinckx, F., Molenaar, D., & van der Maas, H. (2016). Diffusion-based item response theory modeling. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory*. Boca Raton, FL: Chapman and Hall/CRC Press.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. doi:10.1016/j.jmp.2014.06.004

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62. doi:10.1037/a0021765

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

van der Linden, W., & Glas, C. (2009). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139. doi:10.1007/s11336-009-9129-9

van der Maas, H., Molenaar, D., Maris, G., Kievit, R., & Boorsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356. doi:10.1037/a0022749

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220. doi:10.3758/BF03196893

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*, 1–9. doi:10.1016/j.jmp.2007.09.005

Wabersich, D. (2014). *RWiener: Wiener process distribution functions [Computer software manual]*. (R package version 1.2-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://sourceforge.net/projects/rwiener/

Wagenmakers, E. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671. doi:10.1080/09541440802205067

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25. doi:10.2307/1912526

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion*, *23*, 181–205. doi:10.1080/02699930801976770

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*, 39–52. doi:10.1016/j.jmp.2010.01.004

Yuan, K.-H., & Bentler, P. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology*, *63*, 273–291. doi:10.1348/000711009X449771

Received 24 March 2016; revised version received 9 September 2016