A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times

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

DOI
10.1080/00273171.2014.962684

Publication date
2015

Document Version
Final published version

Published in
Multivariate Behavioral Research

Multivariate Behavioral Research, 50:56–74, 2015

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.962684

Routledge
Taylor & Francis Group

A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times

Dylan Molenaar
Department of Psychology, University of Amsterdam

Francis Tuerlinckx
Quantitative Psychology and Individual Differences, University of Leuven

Han L. J. van der Maas
Department of Psychology, University of Amsterdam

A generalized linear modeling framework to the analysis of responses and response times is outlined. In this framework, referred to as bivariate generalized linear item response theory (B-GLIRT), separate generalized linear measurement models are specified for the responses and the response times that are subsequently linked by cross-relations. The cross-relations can take various forms. Here, we focus on cross-relations with a linear or interaction term for ability tests, and cross-relations with a curvilinear term for personality tests. In addition, we discuss how popular existing models from the psychometric literature are special cases in the B-GLIRT framework depending on restrictions in the cross-relation. This allows us to compare existing models conceptually and empirically. We discuss various extensions of the traditional models motivated by practical problems. We also illustrate the applicability of our approach using various real data examples, including data on personality and cognitive ability.

Latent variable analysis is concerned with the specification of appropriate psychometric measurement models to link observed item scores to the underlying latent variable that the test purports to measure. In the case of a continuously distributed latent variable, various models have been proposed—for example, the Rasch model (Rasch, 1960) and the 2-parameter model (2PM; Birnbaum, 1968) for dichotomous items, the graded response model (Samejima, 1969) for ordinal items, and the linear factor model (LFM; Spearman, 1904; Thurstone, 1947) for continuously scored items. Although most of these models have been developed independently, it is well known that most measurement models are special cases of a general class of models commonly referred to by generalized linear item response theory (GLIRT; Bartholomew, Knott, & Moustaki, 2011; Mellenbergh, 1994; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004). In addition to the models mentioned above, GLIRT contains the nominal response model (Bock, 1972), the partial credit model (Masters, 1982), and the nonlinear factor model (McDonald, 1962). For an overview of all models in GLIRT, we refer to Table 1 of Mellenbergh (1994).

All models within GLIRT have been developed with the specific aim of analyzing item responses gathered using traditional paper and pencil tests. However, nowadays, item responses are increasingly collected using computerized tests, resulting in the availability of item response times in addition to item scores. Within latent variable analysis, this begs the question of how to incorporate this additional source of information in the measurement model. Different approaches have been taken, such as hierarchical modeling of the relation between responses and response times (van der Linden, 2007; see also Fox, Klein Entink, & van der Linden, 2007; Glas & van der Linden, 2010; Klein Entink, Fox, & van der Linden, 2009;), linear (Furneaux, 1961; Thissen, 1983) and nonlinear (Ferrando & Lorenzo-Seva, 2007a; 2007b) regressions of the IRT model parameters on the response times, and IRT modeling of the categorized response times (De Boeck

Correspondence concerning this article should be addressed to Dylan Molenaar, Psychological Methods, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, The Netherlands. E-mail: D.Molenaar@uva.nl

& Partchev, 2012; Ranger & Kuhn, 2012; Partchev & De Boeck, 2012).

As these approaches have been developed independently and on different substantive and/or statistical grounds, it is currently unclear how they are related, which approach should be taken for a given dataset, and how they can be compared in terms of model fit. In addition, the absence of flexible fit routines hampers the application of the above approaches in many situations. Most approaches above are only suitable for unidimensional latent variables and are only implemented for the 2PM.

In this article, we formulate a generalized linear latent variable modeling approach for the analysis of responses and response times. This approach will include all models above as special cases. The key idea is to formulate a GLIRT measurement model linking the responses to the latent ability variable and a separate GLIRT measurement model linking the response times to a latent speed variable. Subsequently, the two measurement models are connected by specifying cross-relations between them. We will therefore refer to this approach as bivariate generalized linear item response theory (B-GLIRT).

There are several advantages to this approach. First, B-GLIRT contains commonly used models for responses and response times from the psychometric literature as special cases. This is advantageous because it may enhance the understanding of the similarities and differences between the various models, aiding researchers in choosing the appropriate model for a given problem. Additionally, as B-GLIRT models are generalized linear latent variable models (Bartholomew et al., 2011; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004), one can profit from all well-developed and flexible modeling tools and extensions that exist within this framework. Below is a (non-exhaustive) list of the possibilities that the generalized linear modeling framework offers to the B-GLIRT models (some extensions will be discussed in this article):

(1) B-GLIRT models can be fit with standard latent variable modeling software like Mplus (Muthén & Muthén, 2007), Lisrel (Jöreskog & Sörbom, 1993), Amos (Arbuckle, 1997), Mx (Neale, Boker, Xie, & Maes, 2006), SAS (SAS Institute, 2011), EQS (Bentler, 2006), and OpenMX (Boker et al., 2010).
(2) One is not limited to one particular measurement model, such as the 2PM in case of Thissen (1983) and the 2PM and 3PM in case of van der Linden (2007) or the Rasch model in case of Partchev and De Boeck (2012). The measurement model for the response data can be any model of choice, as long as it is a special case of GLIRT.
(3) The model could be extended to include multiple latent variables to account for such phenomena as multidimensionality in case of, for instance, an intelligence test battery, or a test battery with item bundles that share a common property (i.e., testlets).
(4) It is straightforward to incorporate multilevel and multigroup components (as in Klein Entink et al., 2009).
(5) It enables structural modeling of the latent speed and or latent ability variables in a traditional factor analytic way.
(6) Time limits can be modeled using truncation or censoring of the response time component of the model (Dolan, van der Maas, & Molenaar, 2002) as these techniques are available within generalized linear latent variable modeling (see Skrondal & Rabe-Hesketh, 2004, p. 35).
(7) Various types of well-developed model selection tools become available, such as likelihood-ratio tests, power analysis, modification indices, model fit statistics (as also argued by Glas & van der Linden, 2010), and bootstrapping procedures.
(8) Measurement invariance can be investigated easily on both the responses and the response times.

In addition to the above statistical motivations, B-GLIRT may provide a flexible modeling approach to various substantive applications related to responses and response times. To illustrate, B-GLIRT models might be suitable for detecting faking on psychopathology questionnaires (Holden & Kroner, 1992), for investigating between different cognitive strategies to solve test items (Van der Maas & Jansen, 2003), for testing for differential speediness in multistage testing (van der Linden, Breithaupt, Chuah, & Zhang, 2007), for improving item selection in computerized adaptive testing (van der Linden, 2008), for testing hypotheses about the cognitive processes underlying ability test performance (Klein Entink, Kuhn, Hornke, & Fox, 2009), and for investigating the claim that slow responses are better indicators of intelligence as compared to fast responses ("the worst performance rule"; Coyle, 2003). For these applications, existing models can be used. However, each application requires a different approach. As all of these applications involve one or more of the modeling possibilities discussed above, B-GLIRT unifies these models into a single framework. One advantage of this is that it allows for a direct comparison of the different models. We therefore think that the present framework is a valuable tool for both statistical and substantive applications.

This article consists of five sections. In the first section, we present the general formulation of B-GLIRT and discuss identification and parameter estimation. In the second section, we discuss different forms of the cross-relation between the speed and ability measurement model and show how different instances of B-GLIRT arise. We propose new models motivated by practical problems that could arise in analyzing responses and response times but that could not readily be addressed using the existing models. In the third section,

we present four real data analyses. Specifically, in the first illustration we compare various models in terms of model fit and predictive validity using a chess ability dataset. This illustration is intended to show that different response time models are needed in different cases. In the second application we illustrate the applicability of the present approach to test for measurement invariance. In the third application we test for local independence and we model multidimensionality. This application is intended to show how violations of local dependence can be detected and subsequently taken into account in the statistical model. In the fourth application we illustrate how a five-point Likert scale can be accommodated in the measurement model for the responses. Finally, we discuss limitations and future directions.

BIVARIATE GENERALIZED LINEAR ITEM RESPONSE THEORY

Let X_{pi} denote the response of subject p to item i. In the traditional GLIRT framework, a monotone transformation of $\mathrm{E}(X_{pi})$ is modeled as a linear function of the underlying latent ability variable, θ_p. The specific GLIRT model depends on the transformation used in the link function $g_X(.)$, the exact form of the linear function, and the scales of the item scores and latent variable (nominal, ordinal or continuous). Consider for instance:

$$\mathrm{g_X}\left[\mathrm{E}\left(X_{pi}\right)\right] = \alpha_i\theta_p + \beta_i, \tag{1}$$

where X_{pi} is a binary scored item (1 for correct and 0 for incorrect), α_i is the item discrimination parameter, and β_i is the item difficulty. In this example, using the probit function $\Phi^{-1}(.)$, for $g_X(.)$ results in the normal ogive version of the 2PM (Lord, 1952):

$$\mathrm{P}\left(X_{pi} = 1 \middle| \theta_p\right) = \Phi\left(\alpha_i\,\theta_p + \beta_i\right), \tag{2}$$

because the expected value equals the probability of a correct response for this model. Other popular latent trait models may be specified in a similar manner. For example, when X_{pi} has a continuous scale and $g_X(.)$ is the identity link, the resulting GLIRT model is equivalent to the linear factor model. Similarly, when X_{pi} has an ordinal scale and $g_\mathrm{X}(.)$ is the cumulative logit function, Samejima's graded response model (1969) is obtained.

General Formulation

In the analysis of item responses and item response times, we have two observations per item. Thus, contrary to GLIRT, where a subject receives one score per item, we are interested in modeling the bivariate distribution of the responses and response times, X_{pi} and T_{pi} respectively. In B-GLIRT this is accomplished by formulating a GLIRT measurement model for the responses that includes the latent ability variable, θ_p, and a separate GLIRT measurement model for the response times including the latent speed variable, τ_p. The two measurement models are then connected by specifying a cross-relation between them. Specifically, in the case of dichotomous or continuous X_{pi} and T_{pi} the general model is given by:

$$\begin{aligned} \mathrm{E}(Z_{pi}) &= g_X\left[\mathrm{E}\left(X'_{pi}\right)\right] = \alpha_i\theta_p + \beta_i \\ \text{with } \mathrm{var}\left(Z_{pi}\right) &= \sigma^2_{\varepsilon i} \end{aligned} \tag{3}$$

$$\begin{aligned} \mathrm{E}\left(W_{pi}\right) &= g_T\left[\mathrm{E}\left(T'_{pi}\right)\right] = \varphi_i\tau_p + \lambda_i + f(\theta_p;\boldsymbol{\rho}) \\ \text{with } \mathrm{var}\left(W_{pi}\right) &= \sigma^2_{\omega i} \end{aligned} \tag{4}$$

where ϕ_i is a time discrimination parameter, λ_i is a time intensity parameter, and $f(.)$ is the cross-relation function with cross-relation parameter vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \ldots]$ which is invariant across subjects. The prime symbol in $X_{pi}{}'$ and $T_{pi}{}'$ in Equations (3) and (4) leaves open the possibility of model transformations of X_{pi} and T_{pi}, which might be desirable in case of (approximately) continuous responses, such as responses to a line segment or response times. In the case of categorical responses and response times, $X_{pi}{}' = X_{pi}$ and $T_{pi}{}' = T_{pi}$ will suffice. We elaborate on this later. In addition, Z_{pi} and W_{pi} denote the responses and response time variables after applying the link function. Note that for discrete data the Z_{pi} and W_{pi} can be seen as continuous variables underlying the discrete responses and response times respectively (Takane & De Leeuw, 1987; Wirth & Edwards, 2007). In case of an identity link function for $\mathrm{g_X}(.)$ or $\mathrm{g_T}(.)$, the Z_{pi} and W_{pi} variables coincide with the observed variables, such that $\mathrm{E}(Z_{pi}) = \mathrm{E}(X_{pi}{}')$ and $\mathrm{E}(W_{pi}) = \mathrm{E}(T_{pi}{}')$.

Distributions for the Variables and Parameters

In the general B-GLIRT framework one can assume different distributions for the responses, response times, and parameters. For the observed data, possible distributions include discrete distributions like the Bernoulli distribution (logit or probit link) and the Poisson distribution (logarithmic link), and continuous distributions like the normal distribution (identity link) and the exponential distribution (reciprocal link). For dichotomous distributions for X_{pi} and T_{pi} (i.e., $X_{pi}{}' = X_{pi}$ and $T_{pi}{}' = T_{pi}$), the residual variances, $\sigma_{\varepsilon i}{}^2$ and $\sigma_{\omega i}{}^2$ are a deterministic function of $\mathrm{E}(X_{pi})$ and $\mathrm{E}(T_{pi})$ respectively, such as $\sigma_{\varepsilon i}{}^2 = \mathrm{E}(.) \times [1\text{-}\mathrm{E}(.)]$ for the Bernoulli distribution and $\sigma_{\varepsilon i}{}^2 = \mathrm{E}(.)$ for the Poisson distribution. For multinomial responses and/or ordinal response times, category parameters β_{ic} and/or λ_{ic} replace the item parameters β_i and λ_i. In the case of nominal responses, the model additionally incorporates category specific discrimination parameters, α_{ic} instead of the item discrimination parameters.

The distribution of Z_{pi} and W_{pi} can be inferred from the imposed observed data distribution and the link function. For instance, in the case of a Bernoulli distribution for X_{pi}, a probit link implies a normal distribution for Z_{pi} and a logit link implies a logistic distribution for Z_{pi}.

In the case of a normal distribution for the observed data, the residual variances $\sigma_{\varepsilon i}{}^2$ and $\sigma_{\omega i}{}^2$ are free parameters and assumed to be homoscedastic. In some cases, this might require a transformation of X_{pi} and T_{pi} resulting in $X_{pi}{}'$ and $T_{pi}{}'$. For responses to a line segment (which are bounded from above and below) appropriate transformations might be a (scaled) probit or logit function that maps the responses onto a $(-\infty, \infty)$ domain (implying a uniform distribution for the raw responses). For response times (which are bounded by zero and skewed; see Luce, 1986) appropriate transformations might be logarithmic or square root transformation (implying respectively a lognormal and a chi-square distribution for the raw response times).

On the parameter side of B-GLIRT, the item and/or person parameters can be considered fixed effects or random effects (e.g., following a normal distribution). This enables, for instance, the possibility of imposing latent classes or mixtures on the person parameters θ_p and τ_p and including random effects for the item parameters (see De Boeck, 2008).

In this article we focus on the case in which θ_p and τ_p are normally distributed random variables and the item parameters are fixed effect parameters. We mainly consider the common setting in which the responses are dichotomous (Bernoulli distributed, e.g., correct/incorrect) and in which the log transformed response times ($T_{pi}{}' = \ln T_{pi}$) can be considered normally distributed. However, we also discuss the case of ordinal response times. In addition, we will also propose models for ordinal responses (multinomially distributed, e.g., Likert scales).

In the case of dichotomous responses and normal log response times, Equations (3) and (4) simplify to:

$$\mathrm{E}\left(Z_{pi}\right) = \Phi^{-1}\left[\mathrm{E}\left(X_{pi}\right)\right] = \alpha_i\theta_p + \beta_i \quad (5)$$

$$\mathrm{E}\left(\ln T_{pi}\right) = \varphi_i\tau_p + \lambda_i + f\left(\theta_p; \boldsymbol{\rho}\right)$$
$$\text{with var}\left(\ln T_{pi}\right) = \sigma^2_{\omega i}. \quad (6)$$

We thus use a logarithmic transformation for T_{pi} which is common practice in response time modeling (e.g., Ferrando & Lorenzo-Seva, 2007a; van der Linden, 2007; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). As already noted, other transformations are allowed, such as square root (see Rummel, 1970 for more options). Note that we omitted the W_{pi} variable in the response time model because of the identity link, W_{pi} coincides with $\ln T_{pi}$.

The Cross-Relation Function $f(.)$

The function $f(.)$ in Equation (6) appears in the measurement model of the response times. The use of such a function was previously proposed by Ranger (2013) to model responses and response times on personality questionnaire items. There are two reasons to incorporate the cross-relation in the response time model only. The first reason is that we are primarily concerned with measuring the latent ability (θ_p). By collecting response times in addition to the responses, we hope to increase the measurement precision of θ_p. Thus, we leave the measurement model for the responses intact, and we model the information about θ_p that is available in the response times (if any). This requires a cross-relation function in the measurement model for the response times, but not in the measurement model for the responses. If the cross-relation function would have been incorporated into the measurement model of the responses, τ_p would account for the shared speed variance in the responses and the response times and θ_p would account for the unique ability variance in the responses. This approach would have been interesting if our aim was to partial out any speed effects in θ_p. However, as our objective is to increase measurement precision of θ_p using the response time information, we include the cross-relation function in the measurement model for the response times. In that case, θ_p accounts for the shared ability variance in the responses and the response times and τ_p accounts for the unique speed variance in the response times.

Our second reason for doing so is that by using a cross-relation function in the response time measurement model, we are able to specify various popular models from the literature as special cases (e.g., Thissen, 1983; Ferrando & Lorenzo-Seva, 2007a; 2007b; Ranger & Kuhn, 2012), which is one of the main objectives of present undertaking. In the section "other models"we shortly discuss a case by Roskam (1987) in which there is a cross-relation function in the measurement model of the responses.

Note that by specifying the cross-relation function in the measurement model of the response times, the interpretation of θ_p remains the same irrespective of the choice of $f(.)$, while it leaves open the exact interpretation of the speed factor τ_p. In principle, it is a latent variable for which higher levels are associated with larger responses times. However, depending on $f(.)$, the interpretation may range from a substantive speed factor to a method factor accounting for method variance. We will illustrate this in the application section.

The cross-relation function $f(.)$ is required to have such a form that it retains the generalized linear nature of the response time measurement model in Equation (4). Note that this allows functions of the form $\mathrm{f}(\theta_p;\ \rho) = \rho_1\theta_p$, and $\mathrm{f}(\theta_\mathrm{p};\ \rho) = \rho_1\theta_p + \rho_2{\theta_p}^2$ but it excludes functions like $\mathrm{f}(\theta_\mathrm{p};\ \rho) = \exp(\rho_1\theta_p + \rho_2{\theta_p}^2)$. When the function in $f(.)$ conforms to this requirement, B-GLIRT is part of the generalized linear latent variable modeling framework (Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004). That is, the model is a two-factor model (Figure 1). The generalized linear latent variable family has the advantage that it includes all well-developed modeling tools that exist within this framework. In addition, standard latent variable software can be used to fit the model. The only requirement is that the software enable the specification of the (non)linear constraints in $f(.)$. Throughout this article, we will illustrate various choices for $f(.)$.

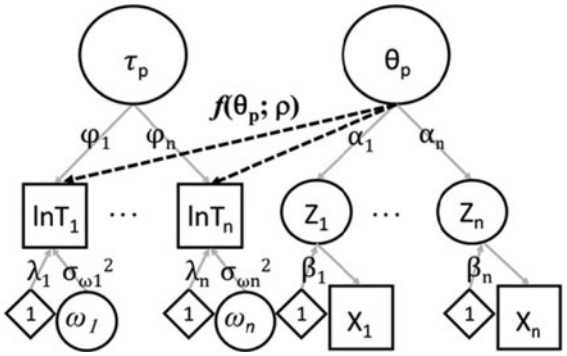

FIGURE 1 Schematic display of the B-GLIRT as a generalized linear latent variable model in the case of categorical responses and continuous response times [Equations (5) and (6)]. Note that in this figure, the cross-relation $f(.)$ is depicted with a dashed arrow, which indicates that this relation is not necessarily linear.

Identification and Parameter Estimation

To identify the model in Equations (5) and (6), standard constraints can be imposed (e.g., see Bollen, 1989, p. 238). That is, for both measurement models either a discrimination parameter is fixed for an arbitrary variable (e.g., $\alpha_i = 1$ and $\phi_j = 1$ for some i and j), or the variance of the latent variable is fixed (e.g., ${\sigma_\theta}^2 = {\sigma_\tau}^2 = 1$). Next, the latent variable means are fixed to 0 to ensure identification of the time intensity and item difficulty parameters (i.e., $\mu_\theta = \mu_\tau = 0$). As the relation between τ_p and θ_p is modelled in the cross-relation $f(.)$, the correlation between τ_p and θ_p needs to be fixed to 0. [1]

In the case of continuous responses and categorical response times, two popular estimation procedures can be used to fit the B-GLIRT to responses and response times. These are methods based on weighted least squares (WLS; e.g., diagonally weighted least squares; Jöreskog & Sörbom, 2001; robust weighted least squares; Muthén, du Toit, & Spisic, 1997; and "fully" weighted least squares; Muthén & Satorra, 1995) and marginal maximum likelihood (MML; Bock & Aitkin, 1981). First, WLS-based methods are attractive as they offer various absolute goodness of fit measures including the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). See Schermelleh-Engel, Moosbrugger, and Müller (2003) for an overview. We will refer to these fit measures as "absolute" fit measures, as they can be interpreted on their own. For instance, a value for the RMSEA smaller than 0.05 is commonly taken as an indication of good model fit. Using a WLS-based method is thus advantageous as the assessment of absolute model fit has been challenging for categorical response models. However, a disadvantage of the WLS method is that it can only be used when $f(.)$ has a linear form. An alternative estimation procedure is MML, which can be used to estimate B-GLIRT models incorporating various forms for $f(.)$ including linear and nonlinear forms. However, MML does not offer absolute fit measures like RMSEA and CFI. Fit measures that can be calculated when using MML are Akaike's information criterion (AIC; Akaike, 1974) , the bayesian information criterion (BIC; Schwarz, 1978), and the sample size adjusted BIC (sBIC; Sclove, 1987). These fit measures are comparative indices, which means that they are only interpretable when compared to the indices of another model. For all of these fit indices, a lower value indicates a better model fit. A disadvantage of MML is that it becomes computationally infeasible when the number of latent variables increases. It has been argued that when the number of dimensions exceeds 5, the MML is practically infeasible (see Wood et al., 2002). An alternative is to adopt a Bayesian approach to model fitting; however, this is beyond the scope of the present article. In the next section, we discuss various possibilities for the specification of $f(.)$.

[1] In some models (e.g., the model by van der Linden, 2007), the correlation between τp and θp is a parameter in the joint distribution of τp and θp. In our model, the correlation between τp and θp is a parameter in the cross-relation function $f(.)$. We therefore have to fix the correlation parameter in the joint distribution of τp and θp to 0, which changes the interpretation of τp, but—as we show in this article—the model can still be equivalent to a model with correlated τp and θp (i.e., the van der Linden 2007, model).

SPECIAL CASES OF B-GLIRT: SPECIFYING THE FUNCTION $f(.)$

Our discussion of the special cases of B-GLIRT will focus on linear, curvilinear, and interaction forms for the cross-relation in $f(.)$.For each special case we propose new models and show how existing models from the psychometric literature fit in B-GLIRT. Specifically, we consider the models in Table 1. In the table, specifications of the cross-relations are given for various models. We will derive these specifications below. The models considered here and in Table 1 are chosen because they have been influential in the psychometric literature (see the overview by van der Linden, 2009). As noted before, all models can be fit using standard software. For each model we discuss below, we provide Mplus scripts in the supplementary material. We will refer to these scripts throughout the text.

Linear Form of the Cross-Relation $f(.)$

A linear cross-relation is especially suitable for ability tests as it will generally imply that the higher the underlying ability, the faster the responses will be, which seems appropriate for ability tests (see van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Usually, the relation between ability and response time will thus be negative. However, positive linear relations are conceivable due to better time management of the higher ability subjects or non-speeded testing (see Klein Entink et al., 2009; example 2). Two popular models in the psychometric literature that embrace the idea of a linear

TABLE 1
Examples of Special Instances of B-GLIRT From the Psychometric Literature for Different Forms and Specifications of the Cross-Relation Function f(θ_p; $\boldsymbol{\rho}$)

Form	Specification	Link Functions	Instance	Source
Linear	$f\left(\theta_p;\boldsymbol{\rho}\right) = -\rho_1\varphi_i\theta_p$	$g_X(.)$: logit/ probit $g_T(.)$: identity	Hierarchical Model	van der Linden (2007) Ranger & Ortner (2012) Fox, et al. (2007) Klein Entink, et al. (2009) Glas & van der Linden (2010)
	$f\left(\theta_p;\boldsymbol{\rho}\right) = -\rho_1\alpha_i\theta_p$	$g_X(.)$: logit/ probit $g_T(.)$: identity	Ability Model	Thissen (1983) Furneaux (1961)
Interaction	$f\left(\theta_p;\boldsymbol{\rho}\right) = \rho_1\theta_p + \rho_2\theta_p\tau_p$	$g_X(.)$: logit/ probit $g_T(.)$: identity	Speed-Ability Interplay	Larson & Alderton (1990) Partchev & De Boeck (2012) De Boeck & Partchev (2012)
Curvilinear	$f\left(\theta_p;\boldsymbol{\rho}\right) = \rho_{1i}^{*} + \rho_{2i}^{*}\theta_p =$ $= 2\rho_1\alpha_i\beta_i\theta_p + \rho_1\alpha_i^2\theta_p^2$	$g_X(.)$: logit/ probit $g_T(.)$: identity	Distance-Difficulty	Ferrando & Lorenzo-Seva (2007a)
		$g_X(.)$: identity $g_T(.)$: identity	Distance-Difficulty	Ferrando & Lorenzo-Seva (2007b)
None	—	$g_X(.)$: logit/ probit $g_T(.)$: -	IRT with time	Roskam (1987) Wang and Hanson (2005)
		$g_X(.)$: - $g_T(.)$: c-l-l or cumm. logit	Prop. Hazard / Acc. Failure Time	Ranger & Kuhn (2011)

Note. c-l-l: complementary log–log link. The specifications are derived and elaborated upon in the body text when discussing the corresponding models. The ρ_1 and ρ_2 parameters are the parameters modeling the cross-relation. Parameters α_i, β_i and ϕ_i belong to the measurement models of θ_p and τ_p. See Eq. 5 and Eq. 6.

relation between speed and ability are the hierarchical model of van der Linden (2007) and the ability model of Thissen (1983).

The Model of van der Linden (2007)

Van der Linden's hierarchical model is hierarchical in the sense that it consists of two levels. At the first level, the observed responses are linked to the latent ability variable θ_p. Van der Linden (2007) originally proposed a 3PM to do so. As the 3PM is not part of the generalized linear framework adopted in this article, we follow Fox (2010, p. 227), Fox et al. (2007), Molenaar, Tuerlinckx, & Van der Maas (in press), Ranger & Ortner (2012), and Ranger (2013) and use a 2PM as a measurement model for the responses. For the response times, the observed data, T_{pi}, are linked to the underlying latent speed variable, τ_p, through a lognormal model (Samejima, 1973):

$$\ln T_{pi} = \omega_{pi} + \lambda_i - \varphi_i\tau'_p \tag{7}$$

where ω_{pi} is a normally distributed variable with variance. Note that the speed parameter $\tau_p{}'$ has a reversed scale as compared to the B-GLIRT latent speed variable τ_p in Equation (4) because of the minus sign before φ_i.

The specification in Equation (7) is used by Fox et al. (2007), Klein Entink et al. (2009), and Fox (2010, chapter 8) and implemented in the R package "cirt" (Fox et al., 2007) which was specially developed to fit this model. In this specification, the time discrimination parameter is modeled as a slope parameter for the latent speed variable.[2]

At the second level of the model, both measurement models are connected by means of linear correlations between their item and person parameters. Note that the model is thus a hierarchical crossed random effects model, as the item and the person parameters are considered to be random variables. Here, as already noted, we assume random person effects only, although it is possible to include random item parameters. Note that Ranger and Ortner (2012) also advocated the use of random person effects only. In a simulation study, it was established that the omission of the random item effects does not affect parameter estimates (Molenaar et al., in press). In addition to the above, van der Linden (2007) and Fox et al. (2007) specified prior distributions for the item and person parameters as they estimated the hierarchical model in a Bayesian framework. As we do not consider Bayesian estimation procedures in this article, we do not need to specify priors.

To formulate the hierarchical model within B-GLIRT, it needs to be recognized that taking the logarithm of T_{pi} in Equation (5), makes the model equivalent to a linear factor model on the log-response times. Thus together with a 2PM

[2]In the original introduction of the model by van der Linden (2007) and in Glas and van der Linden (2010), a slightly different specification is used. In this alternative specification the time discrimination parameter, φi, is the precision of the lognormal distribution of the response times. Thus, within Equation (7), the van der Linden (2007) and Glas and van der Linden (2010) notation can be obtained by fixing φi to 1 for all i. In that case, is interpreted as the time discrimination parameter.

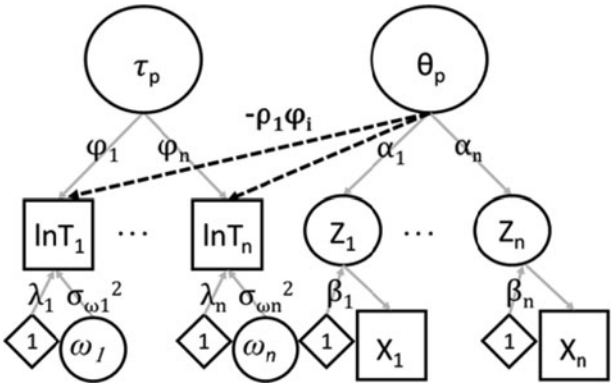

FIGURE 2 A schematic representation of the hierarchical model for responses and response times.

for the responses, the model is an oblique two-factor model with dichotomous indicators for the ability factor and continuous indicators for the speed factor (see Figure 2; see also Ranger & Ortner, 2012; Molenaar, in press). The model is thus a generalized linear latent variable model: however it is not yet obvious that it is part of B-GLIRT, as the form of the function $f(.)$ is unspecified. We therefore rewrite the model assuming that the latent variables are uncorrelated at first. We transform $\tau_p{}'$ from Equation (7) using $\tau_p{}' = -$. As a result, if the two latent variables are identified by fixing $\sigma_\theta{}^2 = 1$ and $\sigma_\tau{}^2 = 1-\rho_1{}^2$ (so that the variance of $\tau_p{}'$ equals 1), the correlation between $\tau_p{}'$ and θ_p is given by ρ_1 as in the original model of van der Linden (2007). In addition, the minus sign of τ_p in the transformation is due to the reversed scale of $\tau_p{}'$ with respect to τ_p. Incorporating this transformation of $\tau_p{}'$ in the model for the log-response times, we obtain:

$$lnT_{pi} = \lambda_i + \varphi_i\tau_p - \varphi_i\rho_1\theta_p + \omega_{pi}. \tag{8}$$

Thus, the B-GLIRT model for the response times becomes:

$$\begin{aligned} \mathrm{E}\left(lnT_{pi}\right) &= \lambda_i + \varphi_i\tau_p - \varphi_i\rho_1\theta_p \\ \text{with var } \left(\ln T_{pi}\right) &= \sigma^2_{\omega i} \end{aligned} \tag{9}$$

in which we recognize that $f(\theta_\mathrm{p};\, \boldsymbol{\rho}) =$. Note that the alternative identification constraint, $\sigma_\tau{}^2 = 1 - \rho_1{}^2$, is sufficient to identify τ_p. However, this constraint is only necessary to enable interpretation of ρ_1 as a correlation coefficient and to put the parameters in Equation (9) on the scale used in the original model of van der Linden (2007). Relaxing this constraint (by identifying the model by fixing $\sigma_\tau{}^2 = 1$ or by fixing $\varphi_i = 1$ for some i) will not affect model fit but it will result in a different scale for the parameters. Mplus code to fit the model can be found in Appendix A of the supplementary material.

The Model of Thissen (1983)

A model that is closely related to the hierarchical model is the model for ability tests by Thissen (1983). As in the hierarchical model, responses are linked to the latent ability variable by a 2PM and log response times are linked to the underlying latent speed variable by a linear model. However, in Thissen's model, both measurement models are linked by linear regressions of the log response times on the latent ability variable, specifically:

$$\ln T_{pi} = \mu + \lambda_i + \tau_p - \rho_1\left(\alpha_i\theta_p + \beta_i\right) + \omega_{pi} \tag{10}$$

(see Thissen, 1983). The parameters in this model have the same interpretation as in the hierarchical model, except that due to the positive sign for τ_p, the latent speed variable is already correctly scaled in B-GLIRT [Equation (6)]. In addition, μ is a general intercept for the log-response times, and ρ_1 is a general slope parameter in the regression of the log response times on the latent ability variable.

The model by Thissen (1983) lacks a time discrimination parameter φ_i. Here, we include it to allow differences in time discrimination across items. This additional parameter could be fixed to 1 to obtain the original Thissen model. Expanding the parentheses in Thissen's model, and adding the time discrimination parameter, we obtain:

$$\ln T_{pi} = \mu + \lambda_i + \tau_p - \rho_1\alpha_i\theta_p + \rho_1\beta_i + \omega_{pi}. \tag{11}$$

Note that in this equation, μ and the term $\rho_1\beta_i$ can both be absorbed in λ_i without affecting model fit or the scale of the parameters. Then, the B-GLIRT model is given by:

$$\begin{aligned} \mathrm{E}\left(\ln T_{pi}\right) &= \lambda_i + \varphi_i\tau_p - \rho_1\alpha_i\theta_p \\ \text{with var } \left(\ln T_{pi}\right) &= \sigma^2_{\omega i}, \end{aligned} \tag{12}$$

in which we see that the cross-relation function is given by $f(\theta_\mathrm{p};\, \boldsymbol{\rho}) = -\,\rho_1\alpha_i\theta_p$. A schematic representation is given in Figure 3. Mplus code to fit the model can be found in Appendix B of the supplementary material.

An interesting result of the above is that it can be seen that when all discrimination parameters, α_i, are equal (i.e., a 1PM;

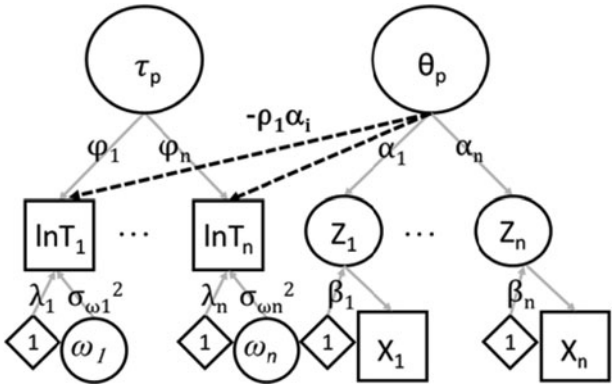

FIGURE 3 A schematic representation of the model by Thissen (1983).

Rasch, 1960), and all time discrimination parameters, φ_i, are equal (i.e., an essentially tau equivalent factor model; Lord & Novick, 1968), the Thissen (1983) model and the hierarchical model of van der Linden (2007) are equivalent. Note that the original models by Thissen and van der Linden did not incorporate a slope parameter φ_i; that is, the measurement model for the response times was already an essential tau equivalent model.

Linear Interactions in the Cross-Relation *f*(.)

When the cross-relation in $f(.)$ is linear, there is no interplay between speed and accuracy (see van der Linden, 2009). This postulation could be questioned, as in the field of mathematical psychology, a close interplay is commonly expected between speed and accuracy (e.g., Luce, 1986). In B-GLIRT, the interplay between speed and accuracy can be modeled by a linear interaction term in the cross-relation function $f(.)$ as follows:

$$\mathrm{E}\left(\ln T_{pi}\right) = \lambda_i + \varphi_i \tau_p + \rho_1 \theta_p + \rho_2 \tau_p \theta_p. \tag{13}$$

Note that to enable modeling of the speed-ability interaction we also needed to include the main effect of θ_{p} on the log-response times (Nelder, 1994). From the above it can be seen that $f\left(\theta_p; \boldsymbol{\rho}\right) = \rho_1 \theta_p + \rho_2 \tau_p \theta_p$. Note that when there is no interaction, i.e., $\rho_2 = 0$, and $\varphi_i = 1$, the model reduces to the hierarchical model of van der Linden (2007). In addition, when $\rho_2 = 0$, $\varphi_i = 1$, and $\alpha_i = 1$, the model is equivalent to the model by Thissen (1983). Figure 4 displays a schematic representation of the model.

The motivation for the form of the speed and accuracy interplay in Equation (13) is given by the "worst performance rule" (Larson & Alderton, 1990). This hypothesis states that fast responses contain less information about ability than slow responses. This hypothesis has been supported empirically (see Coyle, 2003, for a review). In Equation (13) it can be seen that for $\rho_2 > 0$, more variance in the log-response times is due to θ for slower responses (i.e., higher positions of τ_p), which is exactly what is predicted by the worst performance rule. As shown by Van Ravenzwaaij, Brown, & Wagenmakers (2011) the worst performance rule also follows from the popular diffusion model from mathematical psychology (Ratcliff, 1978). In addition, other authors have shown that slow and fast responses contain different information about ability. That is, in the application of the so-called IRTree models by De Boeck & Partchev (2012) and Partchev & De Boeck (2012), a different measurement model was found for the fast responses as compared to slow responses. This is in line with the ideas that are proposed here. However, a difference with the present approach is that the IRTree model includes a within-subjects effect (subjects can switch between the different measurement models during the test), while the present approach does not. It could therefore be seen as a between-subjects version of the IRTree model. Mplus code to fit the model can be found in Appendix C of the supplementary material.

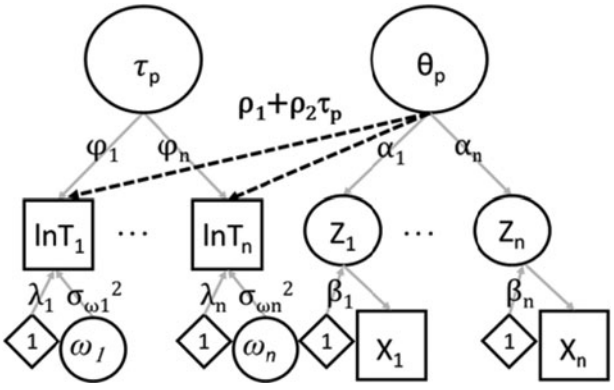

FIGURE 4 A schematic representation of the B-GLIRT model subject to a linear interaction cross relation function.

Curvilinear Form of the Cross-Relation *f*(.)

In the models above, the cross-relation function was motivated by the hypothesis that in ability tests, the relation between speed and ability is generally linear with a possible interaction. This is unlikely to hold for personality tests. Specifically, within response time modeling of personality test data, the distance–difficulty hypothesis postulates that the closer a subject's ability, θ_p, is located to the difficulty, β_i, of a given item, the more time it takes for that subject to answer the item (Ferrando & Lorenzo-Seva 2007a; 2007b; see also Kuiper, 1981). This effect is supported in various empirical studies (e.g., Kuncel, 1973; Holden, Fekken, & Cotton, 1991). Ferrando and Lorenzo-Seva (2007a) formulated a general model based on this hypothesis to analyze response and response times on personality questionnaires for binary items. Ferrando and Lorenzo-Seva (2007a) propose a 2PM for the responses, and for the log-response times they propose:

$$\ln T_{pi} = \mu + \lambda_i + \tau_p + \rho_1 \delta_{pi} + \omega_{pi} \tag{14}$$

with $\delta_{pi} = |\alpha_i \theta_p + \beta_i|$, which means that the log-response times are regressed on the absolute (weighted) distance between θ_p and β_i.[3] All other parameters are similar to those in the models discussed previously. Note that when δ_{pi} is taken to be $\alpha_i \theta_p + \beta_i$ (without the absolute value) the model is equivalent to Thissen's model.

As the function for δ_{pi} contains absolute signs, the model is not a B-GLIRT or generalized linear latent variable model. A reparameterization is possible but cumbersome: If we assume that θ_p has a normal distribution with mean 0 and variance 1, then δ_{pi} has a folded normal distribution with parameters,

[3]Ferrando & Lorenzo-Seva (2007a) use $\alpha_i(\theta_p - \beta_i)$ in the 2PM and in the definition of δ_{pi}. We chose to use $\alpha_i \theta_p + \beta_i$ as it connects better to the generalized linear framework. This is only a reparametrization and will not affect modeling results.

$\mu = \beta_i$ and $\sigma = \alpha_i$. The log-response times can then be linearly regressed on this folded normally distributed latent variable, δ_{pi}, resulting in regression parameter ρ_1. However, this is computationally intensive, as we need an extra latent variable for each item. In addition, the use of folded normal variables is not common practice in generalized linear latent variable models. Ranger (2013) proposes the use of a quadratic function for δ_{pi}:

$$\delta_{pi} = \left(\alpha_i\theta_p + \beta_i\right)^2. \tag{15}$$

Note that this is conceptually the same; that is, still reflects the absolute distance between θ_p and β_i although this difference is now squared. Using this new definition for δ_{pi}, by expanding the brackets we arrive at:

$$\ln T_{pi} = \lambda_i + \varphi_i\tau_p + \rho_1\alpha_i^2\theta_p^2 + 2\rho_1\alpha_i\beta_i\theta_p + \rho_1\beta_i^2 + \omega_{pi}, \tag{16}$$

again introducing a time discrimination parameter, φ_i. In this equation, the term $\rho_1\beta_i^2$ can be omitted as it will be absorbed in λ_i. Specifying $\rho_{1i}{}^* = 2\rho_1\alpha_i\beta_i$ and $\rho_{2i}{}^* = \rho_1\alpha_i^2$, we obtain:

$$\mathrm{E}\left(\ln T_{pi}\right) = \lambda_i + \varphi_i\tau_p + \rho_{1i}^*\theta_p + \rho_{2i}^*\theta_p^2$$
$$\text{with var}\left(\ln T_{pi}\right) = \sigma_{\omega i}^2. \tag{17}$$

That is, $= \rho_{1i}{}^*\theta_p + \rho_{2i}{}^*\theta_p{}^2 = 2\rho_1\alpha_i\beta_i\theta_p + \rho_1\alpha_i{}^2\theta_p{}^2$. See Figure 5 for a schematic representation of the model. Note that $\rho_{1i}{}^*$ and $\rho_{2i}{}^*$ are not free parameters, as they are a function of the underlying parameter ρ_1. The restrictions in $\rho_{1i}{}^*$ and $\rho_{2i}{}^*$ cannot be relaxed without affecting the likelihood function. Mplus code to fit the model can be found in Appendix D of the supplementary material.

Generalization for Polytomous Responses

The model discussed above is appropriate for dichotomous personality items. However, in many cases, personality questionnaires consist of Likert scale items. Ferrando and Lorenzo-Seva (2007b) proposed a model related to the model above for Likert scale items. In this model an LFM is imposed on the categorical responses, which might be suboptimal when modeling Likert scales with few categories (i.e., less than 7 answer categories; see Dolan, 1994). We therefore show how the model above can be easily extended to incorporate a graded response model (GRM; Samejima, 1969) for the responses, which is more appropriate for ordinal response scores in the case of few answer categories (i.e., 3 to 6; Dolan, 1994).

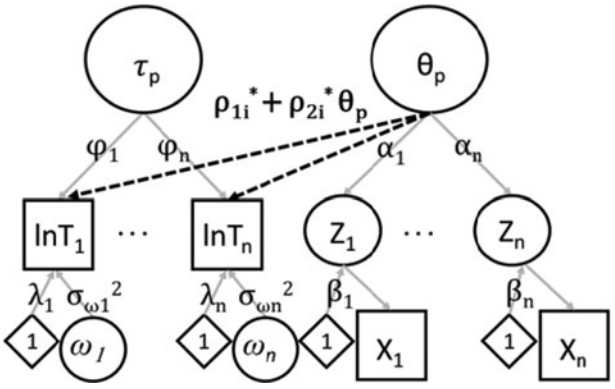

FIGURE 5 Schematic representation of the generalization of the distance–difficulty model within B-GLIRT.

To extend the distance–difficulty model above, we replace the two-parameter model for the responses with the GRM. In the case of items with C answer categories for each item, the GRM contains c-1 category difficulty parameters, β_{ic}, and one item discrimination parameter, α_i. As with the 2PM, this model can be seen as a generalized linear model. As each item now has multiple difficulty parameters, Equation (15) is not applicable because it assumes only one difficulty parameter for each item. We therefore propose:

$$f\left(\theta_p\right) = \left(\alpha_i\theta_p + o_i\right)^2 \tag{18}$$

where o_i is the middle difficulty parameter. For instance, in the case of four ordinal answer categories, we have three thresholds: β_{i1}, β_{i2}, and β_{i3}, thus $o_i = \beta_{i2}$. In case of an uneven number of answer categories, o_i is taken as the point in between the two surrounding difficulties, for example, in the case of five categories, we have β_{i1}, β_{i2}, β_{i3}, and β_{i4}, and $o_i = (\beta_{i2} + \beta_{i3})/2$. Doing so, o_i represents the middle of the answer scale; that is, the point on θ_p at which subjects have maximum uncertainty about whether they should answer in an upper or lower category. Mplus code to fit the model can be found in Appendix E of the supplementary material.

To end, we stress that this is only one possible way to incorporate the distance–difficulty hypothesis with Likert item data. Ranger (2013) and Ranger and Ortner (2011) considered alternative functions for Equation (18) besides a quadratic function. However, these functions are not generalized linear and are therefore not considered here.

Relation to the Hierarchical Model

In the hierarchical model for responses and response times discussed above (van der Linden, 2007), the relation between the latent ability and speed variables is modeled via a linear relation. Thus, for the analysis of personality tests, we need an extension of the hierarchical model that is able to capture possible nonlinear effects predicted by the distance–difficulty hypothesis. In the hierarchical model, we regress the latent speed variable on the latent ability variable using a curvilinear function to enable testing whether the relation between speed and ability departs from a linear function, that is:

$$\tau_p = \rho_1\theta_p + \rho_2\theta_p^2 + \tau_p'. \tag{19}$$

In Equation (13), ρ_1 reflects the linear effect of θ_p on τ_p, ρ_2 reflects the quadratic effect, and τ_p' is the residual term. Note that we do not have an intercept, as it is not identified in a single group application. In addition, by fixing ρ_2 to 0, the

model is made equivalent to the original hierarchical model presented earlier.

Within B-GLIRT this model can be written as:

$$\begin{aligned} \mathrm{E}\left(\ln T_{pi}\right) &= \lambda_{\mathrm{i}} + \varphi_{\mathrm{i}}\left(\rho_1\,\theta_p + \rho_2\theta_p^2 + \tau_p^{'}\right) \\ &= \lambda_{\mathrm{i}} + \phi_{\mathrm{i}}\tau_p^{'} + \rho_1\varphi_i\theta_p + \rho_2\varphi_i\theta_p^2 \text{ with var} \\ &\quad \left(\ln T_{pi}\right) = \sigma_{\omega\mathrm{i}}^2. \end{aligned} \tag{20}$$

Note the model has the same form as the distance–difficulty model by Ferrando and Lorenzo-Seva (2007a)—see Equation (17)—and only differs with respect to the slope coefficients for θ_p and ${\theta_p}^2$.

Other Models

Some existing models from the literature do fall in B-GLIRT but lack either a measurement model for the responses or the response times. For instance, Ranger and Kuhn (2012) proposed a model for response times only. The model is flexible in that it leaves open the exact distribution of response times. In the model, the response times are categorized. Specifically, in the case of dichotomized response times, they propose:

$$\begin{aligned} g_{\mathrm{T}}\left[E\left(T_{pi}\right)\right] &= log\left(\frac{\left[1 - P\left(T_{pi} = 1\middle|\,\tau_p\right)\right]^{-c_i} - 1}{c_i}\right) \\ &= \varphi_i\tau_p + \lambda_i \text{ with } \mathrm{c_i} > 0 \end{aligned} \tag{21}$$

where $\mathrm{c_i}$ is an item-specific shape parameter that is estimated from the data. When $\mathrm{c_i} \to 0$ the link function $g_{\mathrm{T}}(.)$ converges to the complementary log–log function, and when $c_i = 1$, $g_{\mathrm{T}}(.)$ is equal to a logit function. For all values in between 0 and 1, the link function has a form in between these two functions. Thus, although the model as a whole is not generalized linear, these two special cases ($\mathrm{c_i} \to 0$ and $\mathrm{c_i} = 1$) are. Within B-GLIRT these instances of the model by Ranger and Kuhn can thus be used in the case of categorical response times. The researcher only needs to choose an appropriate measurement model for the responses (e.g., 2PM), and a function for $f(.)$ possibly—but not necessarily—from Table 1.

Another example is the model proposed by Roskam (1987). In this model, a 1PM is adopted for the responses and the log response times are linearly regressed on the residuals, that is:

$$\mathrm{E}(Z_{pi}) = g_{\mathrm{X}}[\mathrm{E}(X_{pi})] = \theta_p + \beta_i + \ln T_{pi}. \tag{22}$$

See Figure 6. This approach thus lacks a measurement model for the response times. In addition, the cross-relation between the responses and the response times is specified in the ability measurement model. A related model is the model by Wang and Hanson (2005). The difference with Roskam's model is that $1/T_{pi}$ is used instead of $\ln T_{pi}$, and a random slope is introduced in the regression of $\mathrm{E}(\mathrm{Z_{pi}})$ and $1/T_{pi}$.

Gaviria (2005) proposed a 2PM for the responses and:

$$\ln\left(\frac{T_{pi} - T_0}{A}\right) = \alpha_i\left(\theta_p + \beta_i\right) + \omega_{pi} \tag{23}$$

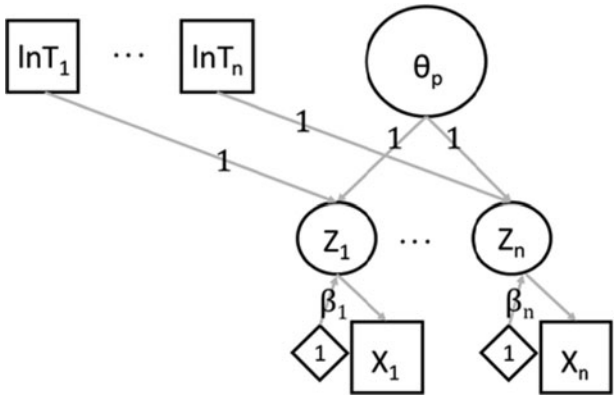

FIGURE 6 The Roskam (1987) model.

for the response times. As $\mathrm{T_0}$ is a known constant, the model is similar to the ability model by Thissen (1983) discussed above, without a latent speed variable (i.e., $\varphi_i = 0$) and $\lambda_i = \ln(A)$ for all i. Thus conceptually, it is a one-factor model on the responses and response time data.

APPLICATIONS

Application 1: Predictive Validity in a Chess Ability Dataset

We applied the models discussed in this article to a dataset on chess ability. The data consisted of the scores of 259 subjects on the "choose a move A" scale of the Amsterdam Chess Test (ACT; van der Maas & Wagenmakers, 2005). This scale consists of 40 chess puzzles divided over three subscales: tactical skill (20 items), positional skill (10 items), and endgame skill (10 items). Items consisted of pictures showing a configuration of chess pieces on the chess board. Respondents were asked to select the best possible move. Responses were coded correct (1) or incorrect (0), and response times were recorded. We conducted the analysis for each subscale separately.

An appealing feature of the dataset is that two interesting covariates are available, the subject's age (mean 30.86, sd 14.92, min 11, max 70) and the subject's "Elo rating" (mean 1882, sd 301, min 1169, max 2629). First, "age" could be an interesting covariate to relate to the speed factor score estimates of the B-GLIRT models as speed of responding is assumed to increase with age (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004). Second, the "Elo rating" is a strong external criterion measure of chess ability based on the number of wins and losses of a given chess player in all official chess games he or she ever played. This variable could thus be regarded as a "gold standard" for chess ability. It would therefore be interesting to see how much variance estimates of θ_{p} share with this variable to see which operationalization of ability has more predictive validity. Note that we

use factor score estimates to calculate $R^2_{\theta,\text{elo}}$ and $R^2_{\text{age},\tau}$ as these are used in practice to make inferences about individual subjects. In the present application we estimated the factor scores using maximum a posteriori estimation.

To be able to meaningfully compare the factor score correlations, we also determined the shared variance between Elo and the sum score (as a proxy for θ_p) and between age and the average response time (as a proxy for τ_p). For $R^2_{elo,\theta}$ we found 0.46, 0.43, and 0.44 for the tactical, positional, and end-game subscales respectively. For $R^2_{age,\theta}$ we found 0.32, 0.09, and 0.19 respectively.

All models in all applications (1–4) were fitted in Mplus (Muthén & Muthén, 2007). Generally, if inferences about a specific parameter were made, we used a 0.05 level of significance. In the present application we use MML estimation (as WLS could not be used for the models that include nonlinear effects in the cross-relation function). As we fit various models, we consider the AIC, BIC, and the sBIC to establish comparative fit. As indicated before, for all these indices a lower value indicates a better fitting model. Note that the use of fit indices likes AIC and BIC is associated with difficulties when comparing models with different numbers of random effects (see, e.g., Greven & Kneib, 2010; Vaida & Blanchard, 2005). However, in the present approach we have two random effects, θ_p and τ_p, across all models. Possible quadratic and interaction terms like ${\theta_p}^2$ and $\theta_p\tau_p$ do not constitute additional random effects as they are deterministic functions of the random effects already in the model. That is, the likelihood function of the curvilinear models does not contain additional integrals over the quadratic terms. This ensures that the AIC, BIC and sBIC are comparable across models.

We fit a 2PM to the responses and a LFM to the log response times. We specified various cross-relations in $f(.)$ as discussed in this article, including a linear form [both the hierarchical form in Equation (9) and the Thissen form in Equation (12)], a form with a linear interaction between θ_p and τ_p, and a curvilinear form [both the distance difficulty instance of Equation (17) and the form in (20)]. In addition, we also fitted a model in which the two measurement models were unconnected (i.e., $f(\theta_p; \boldsymbol{\rho}) = 0$ in Figure 1).

Results

See Tables 2, 3, and 4 for the results concerning the tactical, positional, and end-game items, respectively. It can be seen that the introduction of the cross-relations among the measurement models almost always improves the goodness of fit of the model in terms of AIC, BIC, and sBIC. In all subscales but the end-game scale, the model with a linear form of $f(.)$ based on the Thissen (1983) model, fitted best. In case of the end-game scale, the model according to the distance–difficulty hypothesis fit best (Table 4) suggesting a curvilinear relation between ability and speed. As the end-game items were administered last, this unexpected relation between speed and ability might be due to low ability subjects who started guessing due to a loss of motivation or tiredness while the higher ability subjects were still motivated to give accurate responses.

If we consider the shared variance between Elo rating and θ_p we see that the response times do contain information about θ_p and Elo, as indicated by the increase of $R^2_{elo,\theta}$ in the models with $f(.)$ specified as compared to the model without $f(.)$. The results follow the same pattern as the fit indices. That is, the model with a linear cross-relation according to the Thissen model resulted in the highest $R^2_{elo,\theta}$ for the tactical and positional scales, while the curvilinear cross-relation according to the distance–difficulty hypothesis resulted in the highest $R^2_{elo,\theta}$ in case of the end-game scale.

Concerning $R^2_{age,\theta}$, it is interesting to see that this coefficient is generally the smallest for the linear form of $f(.)$ according to the Thissen model. This coefficient is 0.07 at best while some other models even exceed 0.30. This illustrates that the operationalization of the speed variable is highly different across these models. In the former, the speed variable is a method factor while in the latter model, τ_p can be interpreted as a more substantive variable.

A final conclusion is that in most cases the unrestricted version of the model, in which the time discrimination parameter φ_i is allowed to differ across items, fits better than the restricted version in which φ_i is fixed to be equal across items (as in most original sources, e.g., van der Linden, 2007; Thissen, 1983; Ferrando & Lorenzo-Seva, 2007a; 2007b). This is judged by the fit indices; however, in the unrestricted models ${R_{\text{Elo},\theta}}^2$ is generally not larger than in the restricted models.

Application 2: Measurement Invariance

Measurement invariance (MI; Meredith, 1993; Mellenbergh, 1989) refers to the notion that the item parameters from a measurement model are invariant across populations. In the absence of MI, these populations cannot be meaningfully compared on the latent variable using the scores on the items from that measurement model. When response times come into play, the only work we are aware of is that of Glas and van der Linden (2011), who outlined a procedure to test the equality of the time intensity and item difficulty parameters. In B-GLIRT, the time discrimination and item discrimination parameters can also be tested on invariance. In addition, procedures for testing MI with respect to responses and response times in other models (e.g., in Thissen, 1983; or Ferrando & Lorenzo-Seva; 2007a) have not yet been developed. Within B-GLIRT, it is straightforward to test for MI in such instances. We illustrate this in the present section.

We investigate MI on the responses and response times of the items from the tactical subscale of the ACT with respect to computer experience. As the ACT was administered by computer, it could be argued that more experienced computer users might be quicker in responding irrespective of their chess ability. Part of the ACT data collection was a

TABLE 2
Model Fit Indices, Shared Variance of θ_p and elo ($R_{elo,\theta}^2$)o, Shared Variance of Age and τ_p Correlation ($R_{age,\tau}^2$), and Estimates for the Cross Parameters in f(.) for the Tactical Chess Items

Model f(.)		npar	AIC	BIC	sBIC	$R_{elo,\theta}^2$	$R_{age,\tau}^2$	ρ_1	ρ_2
no f(.)	*restr*	81	11077	11356	11100	0.47	0.31	—	—
	unrestr	100	10837	11183	10866	0.47	0.37	—	—
Linear (Thissen)	*restr*	82	10568	**10851**	10591	0.51	0.07	0.22 (0.02)	—
	unrestr	101	**10558**	10907	**10587**	0.51	0.03	0.24 (0.02)	—
Linear (hierarchical)	*restr*	82	11011	11295	11035	0.48	0.32	–0.59 (0.06)	—
	unrestr	101	10714	11063	10743	0.49	0.35	–0.75 (0.05)	—
Interaction	*restr*	83	11013	11300	11037	0.49	0.23	–0.16 (0.02)	0.06 (0.12)
	unrestr	102	10805	11157	10834	0.47	0.32	0.14 (0.05)	0.25 (0.16)
Curvilinear (distance–difficulty)	*restr*	82	10640	10923	10663	0.50	0.34	–0.06 (0.01)	—
	unrestr	107	11615	11985	11646	0.51	0.37	–0.05 (0.01)	—
Curvilinear (hierarchical)	*restr*	83	11013	11299	11036	0.49	0.32	−0.75 (0.12)	−0.06 (0.07)
	unrestr	103	10716	11069	10745	0.49	0.35	1.14 (0.17)	0.02 (0.09)

Note. restr denotes that all φ_i were constrained to be equal. *unrestr* denotes that the φ_i parameters were free to vary over items. The model with no f(.) denotes a model with a 2PM on the responses and a linear factor model on the log response times without any cross-relations. In the case of the hierarchical model, we identified the model such that ρ_1 can be interpreted as the correlation between τ_p and θ_p. See the section on the hierarchical model for explanation. For the AIC, BIC, and sBIC, the optimal values are in boldface.

TABLE 3
Model Fit Indices, Shared Variance of θ_p and elo ($R_{elo,\theta}^2$), Shared Variance of Age and τ_p Correlation ($R_{age,\tau}^2$), and Estimates for the Cross Parameters in f(.) for the Positional Chess Items

Model f(.)		npar	AIC	BIC	sBIC	$R_{elo,\theta}^2$	$R_{age,\tau}^2$	ρ_1	ρ_2
no f(.)	*restr*	41	4895	5036	4907	0.45	0.08	—	—
	unrestr	50	4829	5002	4843	0.45	0.09	—	—
Linear (Thissen)	*restr*	42	4771	**4917**	4783	0.50	0.04	0.16 (0.03)	—
	unrestr	51	**4762**	4938	**4776**	0.48	0.05	0.13 (0.04)	—
Linear (hierarchical)	*restr*	42	4859	5004	4871	0.46	0.09	–0.51 (0.08)	—
	unrestr	51	4782	4958	4796	0.48	0.10	–0.57 (0.08)	—
Interaction	*restr*	43	4844	4993	4857	0.45	0.04	–0.18 (0.03)	0.09 (0.02)
	unrestr	52	4808	4988	4823	0.45	0.08	–0.11 (0.05)	0.08 (0.02)
Curvilinear (distance–difficulty)	*restr*	42	4857	5002	4869	0.47	0.09	–0.01 (0.00)	—
	unrestr	51	4805	4981	4819	0.46	0.10	–0.01 (0.01)	—
Curvilinear (hierarchical)	*restr*	43	4850	4998	4862	0.42	0.09	0.53 (0.14)	0.37 (0.17)
	unrestr	53	4776	4955	4791	0.48	0.10	0.65 (0.16)	0.23 (0.14)

Note. See Table 2 for explanation.

TABLE 4
Model Fit Indices, Shared Variance of θ_p and elo ($R_{elo,\theta}^2$), Shared Variance of Age and τ_p Correlation ($R_{age,\tau}^2$), and Estimates for the Cross Parameters in f(.) for the End-Game Chess Items

Model f(.)		npar	AIC	BIC	sBIC	$R_{elo,\theta}^2$	$R_{age,\tau}^2$	ρ_1	ρ_2
no f(.)	*restr*	41	5504	5646	5516	0.44	0.21	—	—
	unrestr	50	5484	5657	5498	0.44	0.24	—	—
Linear (Thissen)	*restr*	42	5269	5414	5281	0.49	0.06	0.19 (0.03)	—
	unrestr	51	5245	5422	5260	0.50	0.04	0.22 (0.03)	—
Linear (hierarchical)	*restr*	42	5485	5630	5497	0.46	0.22	–0.41 (0.08)	—
	unrestr	51	5438	5614	5453	0.48	0.26	–0.60 (0.09)	—
Interaction	*restr*	43	5486	5635	5499	0.46	0.15	–0.13 (0.03)	−0.01 (0.02)
	unrestr	52	5487	5666	5501	0.43	0.26	0.09 (0.18)	−0.02 (0.04)
Curvilinear (distance–difficulty)	*restr*	42	**5234**	**5379**	**5246**	0.50	0.20	–0.05 (0.01)	—
	unrestr	51	5235	5411	5249	0.50	0.20	–0.05 (0.01)	—
Curvilinear (hierarchical)	*restr*	43	5486	5635	5499	0.46	0.22	–0.45 (0.11)	0.01 (0.09)
	unrestr	53	5439	5619	5454	0.47	0.26	–0.78 (0.21)	0.08 (0.12)

Note. See Table 2 for explanation.

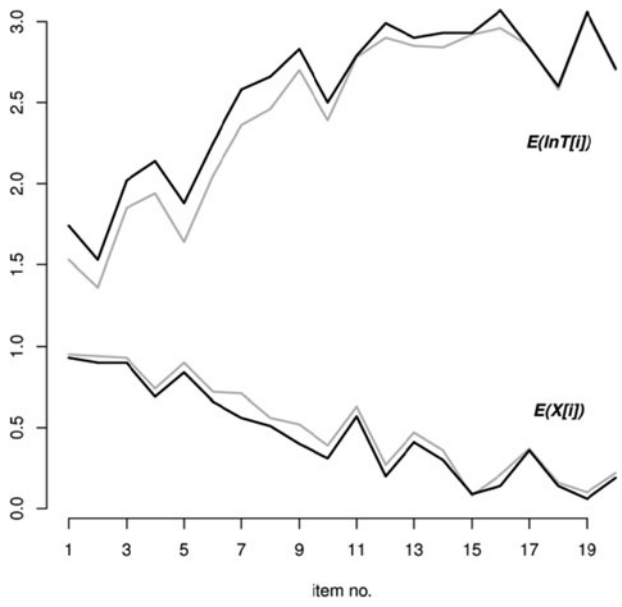

FIGURE 7 Proportion correct $E(X_i)$, and mean log response time $E(\ln T_i)$ for the 20 items of the tactical subscale of the Amsterdam Chess Test in the computer experienced group (grey line) and the computer inexperienced group (black line).

background question in which subjects had to indicate whether they had experience using computers or not. The mean and standard deviation of the response times and the proportion correct of the computer-experienced and computer-inexperienced groups are plotted in Figure 7. As can be seen, the proportion of correct responses in the computer-experienced group are uniformly higher on all items except item 15. In addition, average log response times are uniformly lower for all items except item 17 in the experienced group. The question arises whether these systematic differences are due to underlying differences on the latent variables τ_p and θ_p, or whether these differences reflect bias with respect to group membership (e.g., differences in the time intensity parameters).

We test for MI in the responses and response times of these items of the ACT. As in the analyses above, a linear cross-relation for $f(.)$ based on Thissen (1983) fit these data best. We therefore focus on testing MI within this model. However, other instances of B-GLIRT are equally amenable to this procedure. In the data of the ACT (259 subjects), 70 subjects indicated that they were inexperienced with computers and 162 subjects indicated that they were experienced with computers. For the remaining 27 subjects, responses to this variable were missing. These subjects are excluded from the analysis.

We tested for MI by following the stepwise procedure of Millsap and Yun-Tein (2004). These authors propose fitting a series of increasingly restrictive models to the item responses within each group to test the equality of measurement model parameters across groups. If all restrictions hold, a meaningful comparison of the groups can be made on the latent

TABLE 5
Fit Statistics for the Models in the Measurement Invariance Application

Model	Restricted pars	Npar	AIC	BIC	sBIC
1	—	203	10822	11521	10878
2	α_i	184	10797	11432	10848
3	α_i, φ_i	165	10819	11387	10864
4	α_i, ρ_1	183	10801	11432	10852
5	α_i, β_i	165	10771	11340	10817
6	α_i, β_i, λ_i	146	10748	11251	10788
7	α_i, β_i, λ_i, ${\sigma_{\omega i}}^2$	126	**10740**	**11174**	**10775**

Note. For the AIC, BIC, and sBIC, best values are in boldface.

variables. Specifically, they propose to test sequentially for (1) equality of the slope parameters, (2) equality of the intercept parameters, and (3) equality of the residual variances. As we have parameters from two measurement models (for the responses and the response times), this implies the following steps in the present analysis: (1) testing for invariance of α_i, φ_i, and ρ_1; (2) testing for invariance of β_i and λ_i; and (3) testing for invariance of ${\sigma_{\omega i}}^2$. Including a baseline model in which all parameters are free to vary across groups, this approach results in seven steps, each step introducing an additional equality constraint.

Results

Results of the model fitting are displayed in Table 5 . We started with a baseline model (Model 1) in which all parameters are free to vary across groups. In both groups, we identified the model by fixing the mean and variance of θ_p and τ_p to equal 0 and 1, respectively. In Model 2 we imposed an equality restriction on α_i, freeing the variance of θ_p in the inexperienced group, which was estimated to be 2.04 (0.68). All fit indices indicated that this model fit better than the previous model: That is, the equality restriction α_i was judged to be tenable. In Model 3, the time discrimination parameters, φ_i, were constrained to be equal across groups, and the variance of τ_p was free. This parameter was estimated to be 1.23 (0.57). As can be seen from Table 5, the AIC and sBIC (but not BIC) indicated that this model fit worse than Model 2. Thus, based on the estimated parameters and the AIC and sBIC, we concluded that the φ_is differ across groups.

The parameter estimates (not displayed) indicated that time discrimination parameters where estimated to be higher in the inexperienced group for all but two items. We therefore proceeded by leaving the φ_i parameters free and restricting ρ_1 to be equal across groups (Model 4). Compared to our previously accepted Model 2, this model fit worse as both the AIC and sBIC were larger. In the baseline model ρ_1 was estimated to be 0.26 (0.03) and 0.19 (0.03) in the computer experienced and inexperienced group, respectively. We therefore proceeded by leaving ρ_1 unconstrained and fixing β_i to be equal across groups, freeing the mean of θ_p in the

inexperienced group (Model 5). As the mean of θ_p in the experienced group is fixed to 0 for identification purposes, the mean in the inexperienced group reflects the mean difference on θ_p between the groups. As can be seen in Table 5, all model fit indices improved as compared to our previously accepted Model 2. Thus, we can consider β_i to be invariant across groups. The mean in the computer inexperienced group was estimated to be –0.46 (0.20), indicating that this group has less chess ability on average.

Next, we fitted Model 6 in which we retained the restrictions on α_{i} and β_{i} and additionally introduced equality restrictions on λ_i across groups, freeing the mean of τ_{p} in the inexperienced group. This model fit better than Model 5, which we previously accepted based on all fit indices. The mean of τ_p in the inexperienced group was estimated to be 0.03 (0.19). Finally, we fit Model 7 to the data in which the residual response time variance, ${\sigma_{\omega \mathrm{i}}}^2$, was constrained to be equal across groups. This restriction resulted in an improvement of the model fit as compared to Model 6 in terms of AIC, BIC, and sBIC. In this final model, the means of θ_p and τ_{p} in the experienced group were estimated to be –0.55 (0.20) and 0.04 (0.19), respectively.

It can be concluded from the above that MI is present in the responses but not in the response times. Because φ_i and ρ_1 were shown to be non-invariant across groups, the response times are a biased indicator for both the speed factor, τ_p (as judged by the non-invariance of φ_i), and the ability factor, θ_p (as judged by the non-invariance of ρ_1). Because the time discrimination parameters φ_i were estimated to be higher for the inexperienced group for most i, it could be concluded that individual differences in response times in this group are more due to speediness than is true for the experienced group. In addition, the estimate of cross-relation parameter ρ_I was larger in the experienced group, suggesting that individual differences in response times are more due to ability in this group. However, as the item parameters in the measurement model of θ_p, (α_{i} and β_i) displayed invariance across groups, the responses are not biased with respect to computer experience. That is, θ_p could be meaningfully compared across groups using the responses only.

If both response times and responses are used to assess mean differences in proficiency across groups, such as when using a score rule combining both (Maris & van der Maas, 2012), the above illustrates the importance of testing for MI across groups in both the response measurement model and the response time measurement model. If MI is absent in one of the measurement models, the group comparison will be biased.

Application 3: Investigating Local Independence and Modeling Multidimensionality

As noted by van der Linden and Glas (2010) and Glas and van der Linden (2010), models like the B-GLIRT in Equations (3) and (4) assume three types of local independence: (1) responses are independent conditional on θ_p, (2) response times are independent conditional on τ_p, and (3) responses and response times are independent conditional on θ_p and τ_p. Van der Linden and Glas (2010) and Glas and van der Linden (2010) proposed specific tests to investigate these three types of local independence in the hierarchical model of van der Linden (2007). Presently, however, it is relatively unclear what to do when a violation is found. Using the methodology outlined in this article, it is straightforward to investigate all three types of local independence and, most importantly, when a violation is detected for some items, the assumption can be relaxed for these items (see also Ranger & Ortner, 2012). We illustrate this below.

We applied B-GLIRT to the positional and end-game subscales of the ACT simultaneously. As the 20 items (10 items from each subscale) are likely to violate local independence due to multidimensionality, we think that these data will provide a good illustration of the benefits of the B-GLIRT in terms of detecting violations of local independence and modeling multidimensionality. With respect to the latter, it is particularly interesting as all traditional models from Table 1 are proposed as unidimensional models (meaning one latent speed and one latent ability variable), while in the B-GLIRT multidimensional extensions are straightforward.

We fit the hierarchical model from Equation (9) to the log response times with a 2PM for the responses. As the cross-relation function is linear (see Table 1), we can estimate this model using WLS. As discussed above, this estimation procedure has the advantage of the availability of absolute fit measures. Here we consulted the RMSEA, CFI, and TLI. We follow the guidelines by Schermelleh-Engel et al. (2003) and took a value of the RMSEA smaller than 0.08 as an indication of acceptable model fit and a value smaller than 0.05 as an indication of good model fit. For the CFI and TLI, values larger than 0.95 were taken as indicators of acceptable model fit, and a value larger than 0.97 were taken as indicators of good model fit. In addition, we consult the modification indices that are part of the output of most latent variable software programs. Modification indices are available for constrained parameters and denote the approximate decrease in the log likelihood of the model that would occur if that parameter were freed. Modification are useful tools in investigating local independence, as they can be used to check whether large residual correlations are present within or between the measurement models.

Results

The B-GLIRT model fit the data poorly ($\chi^2\,(91) = 301.367$, RMSEA $= 0.099$). We investigated the source of the misfit by first fitting the measurement models (2PL and a LFM) to the responses and log response times separately (see Table 6 for the results). The 2PL fit relatively well, as both the χ^2 and the RMSEA indicated good model fit but CFI and TLI are still poor. We checked the modification indices to identify

TABLE 6
Model Fit Results for Illustration 3

Data	Model	χ^2	df	χ^2/df	RMSEA	CFI	TLI
Responses only	2PL	61.139	40	1.528	0.048	0.932	0.935
RT only	One-factor model	558.781	170	3.287	0.099	0.791	0.766
	Testlet model	212.347	149	1.425	0.043	0.966	0.957
RT and Responses	Full model	147.748	87	1.698	0.055	0.878	0.931

any residual dependencies in the data. However, modification indices for the residual correlations were all reasonably low. The largest modification index was 4.07 for the residual correlation between items 10 and 11, but freeing this parameter did not result in a significant parameter estimate. We therefore concluded that the local independence of the responses conditional on θ was tenable.

We continued by fitting the one-factor model to the log response time data. This model fit poorly according to all fit measures (see Table 6). Modification indices suggested that this misfit was mainly due to a violation of local independence. Specifically, as expected, the log-response time residuals of items 1–10 seemed to covary, and the log-response time residuals of items 11–20 seemed to covary with modifications roughly between 1 and 40. These residual covariances were thus due to the nature of the item content, as items 1–10 are from the positional subscale while items 11–20 are from the end-game subscale. As both subscales are purported to measure chess ability, we see the residual covariation as method variance. To account for these violations of local independence, we introduced two additional latent variables that accounted for the method variance in the response times of items 1–10 and items 11–20, respectively. The resulting model can be seen as a mono-trait multi-method model (Campbell & Fiske, 1959) or a testlet model (Bradlow, Wainer, & Wang, 1999) with a general speed factor and two correlated method factors (see Figure 8). Below we refer to this model as the testlet model.

As can be seen from the fit statistics in Table 6, the extensions improved the model fit for the response times substantially according to the RSMEA, but the CFI and TLI are still poor. Next, we fit the full model with the 2PL for the responses and the testlet model for the log response times. In this full model, the ability factor was allowed to correlate with the general speed factor and the two-method factors. As can be seen in Table 6, fit was acceptable to good based on the RMSEA and χ^2 but CFI and TLI still indicate that there was potential for improvement. At this stage in model fitting, it is possible to check the third type of local independence: That is, the responses and log response times should be independent conditional on θ_p and τ_p. A violation would be evident when the residuals of the responses tend to correlate with the residuals of the log response times. This would cause large modification indices of the corresponding parameters. However, in this case, all modification indices were small (with the largest indices equal to 4.27). Thus, the unsatisfactory values of the CFI and TLI were not due to misfit. These poor values might have been due to the relatively low inter-item and inter-response time correlations (as the CFI

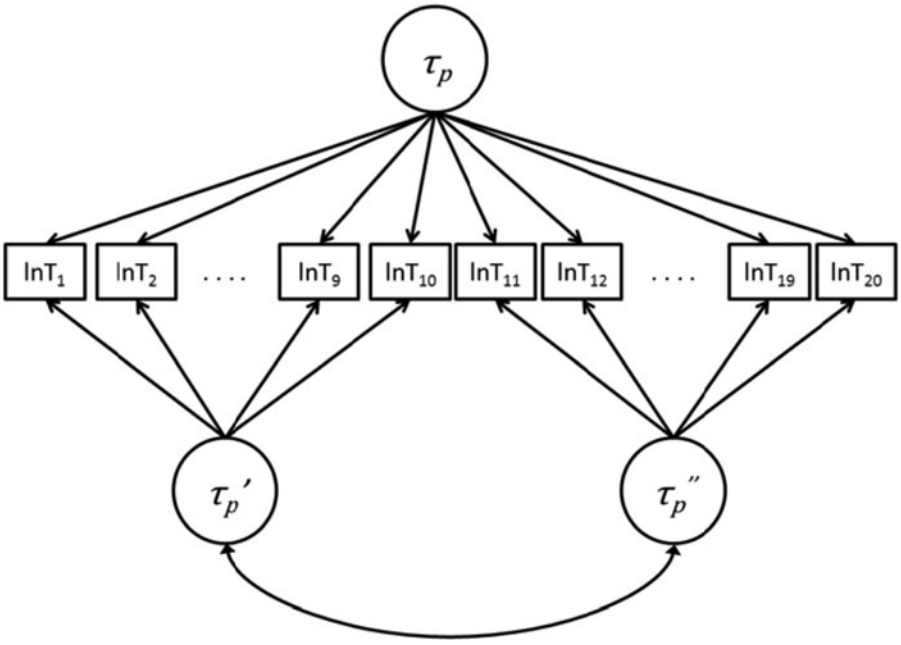

FIGURE 8 The testlet model to account for the multidimensionality in the data in Application 4.

and TLI are incremental model fit indices, they are sensitive to small correlations). Thus, given the acceptable value for RMSEA and the absence of severe model misfit as indicated by the relatively small modification indices, we accept this as the final model. In the final model, ability and speed are correlated 0.91 (SE 0.05).

Application 4: Using the Graded Response Model with Response Times

As the applications above only considered dichotomous data, we now illustrate the flexibility of present approach by applying it to Likert scale data. We analyze the 10 items of th "negative fear of failure" subscale of the motivation test of the ACT (see Application 1 above). Items concern questions about the negative influence of the subjects' fear of losing a chess game. An example of such a question is: "In a hopeless position I have difficulty concentrating." Questions are administered using a five-point Likert scale. To the data we fit a graded response model (Samjima, 1969) to the responses: that is, we used the cumulative logit link in Equation (3) and, as above, a LFM on the log-response times. Generalization to Likert data is straightforward for all models discussed in this article. For the Thissen (1983) model we needed a decomposition similar to that of Equation (18). Specifically, instead of $(\alpha_i\theta_p + \beta_i)$ we use $(\alpha_i\theta_p + o_i)$, where $o_i = (\beta_{i2}+\beta_{i3})/2$, see above.

Results

Results are in Table 7. As can be seen, the fit indices favor the curvilinear form $f(.)$ based on to the hierarchical model. Judged on the fit indices, it is relatively unclear which of the restricted or unrestricted versions fit best as the AIC favors the unrestricted model, the BIC favors the restricted model, and the sBIC is undecided. Note in all models fit, the linear component is never significant. Notwithstanding, it can be conclude that for these personality data, a curvilinear model outperformed the models based on linear cross-relations, which is in line with the distance–difficulty hypothesis (see also Ranger, 2013).

LIMITATIONS AND FUTURE DIRECTIONS

The goal of the present study was to present a generalized linear latent variable modeling approach to the analysis of responses and response times. Adopting a generalized linear framework was shown to be advantageous because of its flexibility and well-developed statistical underpinning. Specifically, in this article we illustrated how existing models could be extended to incorporate quadratic and interaction effects, multiple dimensions, multiple groups, and alternative measurement models. In addition, we illustrated the use of modification indices and model fit statistics that are available within the generalized linear latent variable framework.

Of course, such a framework also has limitations as not all psychometric measurement models can be formulated in generalized linear form. Such cases include the 3PM model by Birnbaum (1968), the model by Thissen & Steinberg (1986) intended to account for guessing, and the nonparametric model of Mokken (1970). In addition the Poisson models proposed by Rasch (1960) fall outside of present framework as these models are not formulated specifically for item responses and response times, but for the number correct and time to complete a task. In addition, the model by Verhelst, Verstralen, and Jansen (1997) cannot be formulated in a generalized linear form as it contains an exponential item parameter.

The approach as presented here is a parametric approach in the sense that it requires an assumption about the distribution of the response times. This approach is somewhat flexible due to the possibility of using different transformations

TABLE 7
Fit Indices and Estimates for the Cross Parameters in $f(.)$ for the Negative Fear Items

Model f(.)		npar	AIC	BIC	sBIC	Cross Parameter(s)
no f(.)	*restr*	71	7358	7604	7379	—
	unrestr	80	7357	7634	7380	—
Linear (Thissen)	*restr*	72	7360	7608	7380	$\rho_1 = 0.01$ (0.03)
	unrestr	81	7358	7638	7381	$\rho_1 = 0.02$ (0.02)
Linear (hierarchical)	*restr*	72	7360	7609	7381	$\rho_1 = 0.01$ (0.03)
	unrestr	81	7359	7639	7382	$\rho_1 = 0.03$ (0.20)
Interaction	*restr*	73	7294	7547	7315	$\rho_1 = -0.02$ (0.02) $\rho_2 = -0.07$(0.03)
	unrestr	82	7295	7578	7318	$\rho_1 = -0.02$ (0.02) $\rho_2 = -0.09$(0.03)
Curvilinear (distance–difficulty)	*restr*	72	7339	7588	7360	$\rho_1 = -0.02$ (0.01)
	unrestr	81	7340	7619	7363	$\rho_1 = -0.02$ (0.01)
Curvilinear (hierarchical)	*restr*	73	7290	**7542**	**7311**	$\rho_1 = -0.13$ (0.13) $\rho_2 = -0.41$ (0.18)
	unrestr	82	**7287**	7571	**7311**	$\rho_1 = -0.12$ (0.13) $\rho_2 = -0.41$ (0.19)

Note. restr denotes that all φ_i were constrained to be equal; *unrestr* denotes that the φ_i parameters were free to vary over items. The model with no $f(.)$ denotes a model with a graded response model on the responses and a linear factor model on the log response times without any cross-relations. For the AIC, BIC, and sBIC, the optimal values are in boldface.

of the raw response times (e.g., log, square root, reciprocal), categorizing of the response times as discussed in Ranger & Kuhn (2011), and using mixtures models. However, if none of these options are appropriate for a given dataset, one might consider a semi-parametric response time model (Loeys, Legrand, Schettino, & Pourtois, 2014; Ranger & Kuhn, 2013; Wang, Chang, Douglas, 2013; Wang, Fan, Chang, & Douglas, 2013). These models are more flexible than the models presented here, as they require at most only a mild distributional assumption for the response times.

As stated earlier, the present modeling approach does allow for random item effects, and due to the advances in software development (e.g., Mplus 7 enables the possibility of factor analysis with random item effects; see Asparouhov & Muthén, 2012), incorporation of these effects is straightforward (see Glas and van der Linden, 2010). De Boeck (2008) discusses different research designs that call for random item effects in the psychometric model. First, random item effects may be necessary to account for the sampling variance that arises when items are sampled from a larger item bank (see Raaijmakers, Schrijnemakers, & Gremmen, 1999). Second, some research designs involve item families. Items within the same family (e.g., addition items) are considered to be more similar as compared to items from a different family (e.g., multiplication items). As the items from a given family are relatively homogenous, one can often focus on the item family parameters instead of the individual item parameters, which call for random item effects (see Janssen, Tuerlinckx, Meulders, & De Boeck, 2000). As a final example of research settings in which random item effects might be necessary, Be Boeck (2008) discusses longitudinal designs. In these designs researchers sometimes rely on randomly drawn items from a larger item pool to prevent the same items being administered multiple times (see e.g., Albers, Does, Ombos, & Janssen, 1989).

In this article, we presented some existing models and introduced some new models in a novel framework that included a so-called cross-relation function. We stress that possibilities are not limited to the specific models presented here. For instance, the cross-relation function could be approximated using polynomials when theory about the exact form of this relationship is lacking. Models could also be extended to investigate within-subjects response time curves or to analyze data from large-scale assessments. In addition, models could be subjected to multivariate mixtures to investigate possible (latent) subgroups of subjects. Finally, marketing researchers are interest in modeling response times and nominal responses (e.g., Haaijer, Kamakura, & Wedel, 2000). This is straightforward in the present approach as Bock's nominal response model (1972) can also be formulated as a generalized linear model (see Mellenbergh, 1994). Thus, we think that the present approach opens a wide variety of modeling possibilities and has many possible applications in psychology and other areas of research.

ACKNOWLEDGMENTS

We thank Leanne Stanley, Mijke Rhemtulla, and Conor Dolan for their comments on previous drafts of this article.

SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed on the publisher's website.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Albers, W., Does, R. J. M. M., Ombos, T., & Janssen, M. P. E. (1989). A stochastic growth model applied to tests of academic knowledge. *Psychometrika*, *54*, 451–466

Arbuckle, J. L. (1997). *Amos (version 3.61) [Computer software]*. Chicago, IL: Small Waters.

Asparouhov, T., & Muthén, B. (2012, May). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Paper presented at the 3rd UK Mplus Users Meeting, London.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. UK: Wiley.

Bentler, P. M. (2006). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (Chap. 17–20). Reading, MA: Addison Wesley.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Boker, S., Neale. M. C., Maes, H. H., Wilde, M., Spiegel, M., Brick, T. et al. (2010) OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, *56*, 81–105.

Coyle, T. R. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *31*, 567–587.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Dolan, C. V., van der Maas, H. L. J., & Molenaar, P. C. M. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods*, *34*, 304–323.

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

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

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*, 1–14.

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology*. London: Pitman.

Gaviria, J.-L. (2005). Increase in precision when estimating parameters in computer assisted testing using response times. *Quality & Quantity*, *39*, 45–69.

Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603–626.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 773–789.

Haaijer, R., Kamakura, W., & Wedel, M. (2000). Response latencies in the analysis of conjoint choice experiments. *Journal of Marketing Research*, *37*, 376–382.

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

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, *4*(2), 170.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modelling with the SIMPLIS command language*. Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL user's guide*. Chicago: Scientific Software International.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54.

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

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

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*(3), 309–325.

Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semiparametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, *67*, 304–327.

Lord, F. M. (1952). *A theory of test scores*. New York: Psychometric Society.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika*, *27*, 392–415.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.

Mokken, R. J. (1970). *A theory and procedure of scale analysis*. 's-Gravenhage: Mouton.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (in press). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391–411.

Muthén, B. O., du Toit, S. H. C., & Spisic D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide (5*th *ed.)* Los Angeles: Muthén & Muthén.

Muthén, B., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*, 489–503.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2006). *Mx: Statistical modeling* (7th ed.). Richmond, VA: VCU, Department of Psychiatry.

Nelder, J. A. (1994). The statistics of linear models: Back to basics. *Statistics and Computing 4*, 221–234.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32.

Raaijmakers, J., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect-fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, *55*, 361–382

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*, 31–47.

Ranger, J., & Kuhn, J. T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics*, *38*(1), 61–80.

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

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

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*, 278–289.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North-Holland.

Rummel, R. J. (1970). *Applied factor analysis*. Northwestern University Press.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17*.

Samejima, F. (1973). Homogeneous case of the continuous response level. *Psychometrika*, *38*, 203–219.

SAS Institute Inc. (2011). *SAS/STAT software: Release 9.3*. Cary, NC: SAS Institute, Inc.

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*, 23–74.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Chapman and Hall/CRC.

Spearman, C. E. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

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

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: *University of Chicago Press*.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*, 351–370.

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

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130.

van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177.

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

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

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*, 381–393.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York: Springer.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144–168.

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2002). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International, Inc.