Psychological Methods
2009, Vol. 14, No. 1, 54–75

© 2009 American Psychological Association
1082-989X/09/$12.00 DOI: 10.1037/a0014877

Evaluating Cognitive Theory: A Joint Modeling Approach Using Responses and Response Times

Rinke H. Klein Entink
University of Twente

Jörg-Tobias Kuhn
University of Münster

Lutz F. Hornke
RWTH Aachen University

Jean-Paul Fox
University of Twente

In current psychological research, the analysis of data from computer-based assessments or experiments is often confined to accuracy scores. Response times, although being an important source of additional information, are either neglected or analyzed separately. In this article, a new model is developed that allows the simultaneous analysis of accuracy scores and response times of cognitive tests with a rule-based design. The model is capable of simultaneously estimating ability and speed on the person side as well as difficulty and time intensity on the task side, thus dissociating information that is often confounded in current analysis procedures. Further, by integrating design matrices on the task side, it becomes possible to assess the effects of design parameters (e.g., cognitive processes) on both task difficulty and time intensity, offering deeper insights into the task structure. A Bayesian approach, using Markov Chain Monte Carlo methods, has been developed to estimate the model. An application of the model in the context of educational assessment is illustrated using a large-scale investigation of figural reasoning ability.

Keywords: rule-based item design, response times, item response theory

An important facet reflecting cognitive processes is captured by response times (RTs). In experimental psychology, RTs have been a central source for inferences about the organization and structure of cognitive processes (Luce, 1986). However, in educational measurement, RT data have been largely ignored until recently, probably because of the fact that recording RTs for single items in paper-and-pencil tests seemed difficult. With the advent of computer-based testing, item RTs have become easily available to test administrators. Taking RTs into account can lead to a better understanding of test and item scores, and it can result in practical improvements of a test, for example, by investigating differential speededness (van der Linden, Scrams, & Schnipke, 1999).

The systematic combination of educational assessment techniques with RT analysis remains a scarcity in the literature. The purpose of the present article is to present a model that allows the integration of RT information into an item response theory (IRT) framework in the context of educational assessment. More specifically, the approach advanced here allows for the simultaneous estimation of ability and speed on the person side, while offering difficulty and time-intensity parameters pertaining to specific cognitive operations on the item side. First, we briefly outline the cognitive theory for test design and IRT models that are capable of integrating cognitive theories into educational assessment. Next, current results from the RT literature with respect to educational measurement are summarized. We then develop a new model within a Bayesian framework that integrates both strands of research, and we demonstrate its application with an empirical example.

Cognitive Theory in Educational Assessment

One of the core interests of psychological research pertains to the analysis of cognitive processes. Research para-

Rinke H. Klein Entink and Jean-Paul Fox, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, the Netherlands; Jörg-Tobias Kuhn, Department of Psychology, University of Münster, Münster, Germany; Lutz F. Hornke, Department of Industrial and Organizational Psychology, RWTH Aachen University, Aachen, Germany.

Software for the application presented in this article is freely available as a package for the R statistical environment on www.kleinentink.eu

Correspondence concerning this article should be addressed to Rinke H. Klein Entink, University of Twente, Department of Research Methodology, Measurement, and Data Analysis, P.O. Box 217, 7500 AE Enschede, the Netherlands. E-mail: r.h.kleinentink@gw.utwente.nl

digms in cognitive psychology often assume that to successfully solve a task or test item, a subject must perform certain associated cognitive processes, either serially or in parallel. Subjects then can be differentiated on the basis of the processing times necessary for specific processes. For example, an important strand of research in cognitive psychology is concerned with analyzing parameters from individual RT distributions on simple tasks, and these parameters can be theoretically connected to psychological processes, such as attention fluctuations or executive control (Schmiedek, Oberauer, Wilhelm, Süss, & Wittmann, 2007; Spieler, Balota, & Faust, 2000). Complex models of reaction times obtained in experimental settings have been developed, for example, focusing on a decomposition of reaction times, comparing competing models or complex cognitive architectures (Dzhafarov & Schweickert, 1995). However, many of the reaction times analyzed in experimental psychology are based on very elementary tasks that are often psychophysical in nature (van Zandt, 2002). The central difference between RT analysis in experimental psychology and educational assessment lies with the cognitive phenomena under investigation and the complexity of the tasks involved. In experimental psychology, research commonly focuses on elementary cognitive processes related to stimulus discrimination, attention, categorization, or memory retrieval (e.g., Ratcliff, 1978; Rouder, Lu, Morey, Sun, & Speckman, 2008; Spieler et al., 2000). In this research tradition, mostly simple choice tasks are utilized that usually do not tap subjects' reasoning or problem-solving abilities. Further, in experimental research on reaction times with mathematical processing models, the focus has often been either on RTs *or* accuracy scores but not on both at the same time, with item parameters sometimes not being modeled at all (e.g., Rouder, Sun, Speckman, Lu, & Zhou, 2003; but see Ratcliff, 1978, for an approach that allows to simultaneously model experimental accuracy and RT data). This is due to the fact that such models often imply a within-subject design with many replications of the same simple items, a procedure not usually followed in educational measurement.

Things look differently in educational assessment. Here, differentiating subjects according to latent variables (e.g., intelligence) as measured by psychological tests is of primary interest. Latent variables represent unobservable entities that are invoked to provide theoretical explanations for observed data patterns (Borsboom, Mellenbergh, & van Heerden, 2003; Edwards & Bagozzi, 2000). Recently, cognitive theories pertaining to test design as well as latent variable models have been merged in the field of educational assessment to provide meaningful results. To improve construct valid item generation, contemporary test development often incorporates findings from theories of cognition (Mislevy, 2006). With respect to construct validation, Embretson (1983, 1998) has proposed a distinction between construct representation, involving the identification of cognitive components affecting task performance, and nomothetic span, which refers to the correlation of test scores with other constructs. Whereas traditional methods of test construction have almost exclusively focused on establishing correlations of test scores with other measures to establish construct validity (nomothetic span), contemporary test development methods focus on integrating parameters reflecting task strategies, processes, or knowledge bases into item design (construct representation). Hence, the cognitive model on which a test is founded lends itself to direct empirical investigation, which is a central aspect of test validity (Borsboom, Mellenbergh, & van Heerden, 2004). After a set of cognitive rules affecting item complexity has been defined on the basis of prior research, these rules can be systematically combined to produce items of varying difficulty. In a final step, the theoretical expectations can then be compared with empirical findings.

The integration of cognitive theory into educational assessment is usually based on an information-processing approach and assumes unobservable mental operations as fundamental to the problem-solving process (Newell & Simon, 1972). The main purpose of educational assessment under an information-processing perspective is to design tasks that allow conclusions pertaining to the degree of mastery of some or all task-specific mental operations that an examinee has acquired. That is, by specifying a set of known manipulations of task structures and contents a priori, psychological tests can be built in a rule-based manner, which in turn allows more fine-grained analyses of cognitive functioning (Irvine, 2002). In the process of test design, it is therefore entirely feasible (and generally desirable) to experimentally manipulate the difficulty of the items across the test by selecting which cognitive operations must be conducted to solve which item correctly. Hence, as is outlined in the next section, some extensions of classical IRT models are capable of modeling the difficulty of cognitive components in a psychometric test. The basic requirement for such a procedure, however, is a strong theory relating specific item properties to the difficulty of the required cognitive operations (Gorin, 2006). Because classical test theory focuses on the true score that a subject obtains on a whole test, that is, on the sum score of correct test items, it is not well-suited to model cognitive processes on specific test items. In contrast, numerous IRT models have been developed that are capable of doing so (cf. Junker & Sijtsma, 2001; Leighton & Gierl, 2007).

In the context of educational assessment, language-free tests lend themselves to rule-based item design, which can be understood as the systematic combination of test-specific rules that are connected to cognitive operations. A large body of research in rule-based test design has focused on figural matrices tests, which allow the assessment of reasoning ability with nonverbal content. In these tests, items consist usually of nine cells organized in 3 × 3 matrices, with each cell except the last one containing one or more geometric elements. The examinee is supposed to detect the rules that meaningfully connect these elements across cells and to correctly apply these rules to find the content of the empty cell. A typical item found in such a test is given in Figure 1.

Several investigations into the structure and design of cognitive rules in figural matrices tests have been conducted. Jacobs and Vandeventer (1972) and Ward and Fitzpatrick (1973) have reported taxonomies of rules utilized in the item design of existing figural matrices tests. In a review of the current literature, Primi (2001) has described four main design factors (*radicals* according to Irvine, 2002) that affect item difficulty: (1) number of elements, (2) number of rules, (3) type of rules, and (4) perceptual organization of elements. In line with recent research, the two first radicals are associated with the amount of information that must be processed during working on an item (Carpenter, Just, & Shell, 1990; Mulholland, Pellegrino, & Glaser, 1980): More information requires more working memory capacity and additionally results in longer RTs (Embretson, 1998). Working memory, which is a construct grounded in cognitive psychology that has repeatedly been shown to correlate highly with intelligence (e.g., Engle, Tuholski, Laughlin, & Conway, 1999), refers to the cognitive system that is capable of simultaneously processing and storing information. Carpenter et al. (1990) assumed that in addition to working memory capacity, abstraction capacity—that is, the ability to represent information in a more conceptual way—plays a role in item solving: Examinees that are capable of processing item features in a more abstract fashion are more capable of discovering the correct solution. The third radical (type of rules) has been studied in several studies (e.g., Bethell-Fox, Lohman, & Snow, 1984; Carpenter et al., 1990; Embretson, 1998; Hornke & Habon, 1986; Primi, 2001). In one study (Carpenter et al., 1990), which analyzed performance on the Advanced Progressive Matrices (Raven, 1962), evidence was presented that easier rules taxing the working memory system less are considered before harder ones. On the basis of this finding, Embretson (1998) proposed that the difficulty of understanding and applying item rules correctly is related to working memory capacity. Finally, the fourth radical, perceptual organization, refers to how the figural elements in an item are grouped. For example, Primi (2001) distinguished between *harmonic* and *disharmonic* items, in which the latter introduce conflicting combinations between visual and conceptual figural elements, whereas the former display more congruent relationships. Primi showed that perceptual organization had a strong effect on item difficulty, even stronger than the number and type of rules (i.e., radicals taxing working memory capacity). In contrast, both Carpenter et al. and Embretson found larger effects of item features relating to working memory.

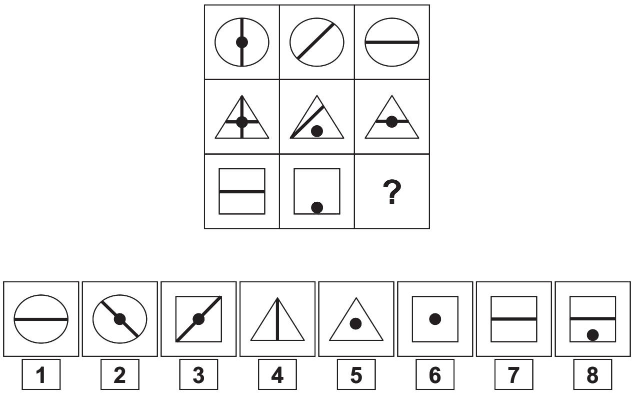

Figure 1. Example of a figural reasoning item.

Psychometric Analysis of Rule-Based Test Items

The analysis of tests and items with a cognitive design is usually cast in an IRT framework (Rupp & Mislevy, 2007). One of the most basic IRT model is the Rasch model (Rasch, 1960). It is the building block for numerous more advanced IRT models. The Rasch model assumes unidimensionality and local item dependence, which can be regarded as equivalent to each other (McDonald, 1981). In the Rasch model, the probability of a correct response of examinee i, $i = 1, 2, \ldots, N$ to a test item k, $k = 1, 2, \ldots, K$ is given by

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{exp(\theta_i - b_k)}{1 + exp(\theta_i - b_k)}, \qquad (1)$$

where θ_i denotes the ability of test taker i, and b_k denotes the difficulty of item k. The Rasch model represents a saturated model with respect to the items, because each item has its own difficulty parameter. Therefore, the model does not allow any statements pertaining to the cognitive operations that are assumed to underlie performance on the items. Another IRT model, the linear-logistic test model (LLTM; Fischer, 1973), which is nested in the Rasch model, allows the decomposition of the item difficulties b_k such that

$$P(Y_{ik} = 1|\theta_i, q_k, \eta) = \frac{exp(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)}{1 + exp(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)}, \qquad (2)$$

where the η_j, $j = 1, \ldots, J$, are so-called "basic parameters" representing the difficulty of a specific design rule or cognitive operation in the items, and the q_{kj} are indicators reflecting the presence or absence of a rule j in item k. The LLTM is therefore capable of determining the difficulty of specific cognitive operations that must be carried out to solve an item.

Both the Rasch model and the LLTM assume that all items discriminate equally well across examinees. This is a rather strict assumption that can be relaxed. The 2 parameter logistic (2PL) model (Lord & Novick, 1968) is defined as

$$P(Y_{ik} = 1|\theta_i, a_k, b_k) = \frac{exp(a_k(\theta_i - b_k))}{1 + exp(a_k(\theta_i - b_k))}, \quad (3)$$

with a_k denoting the item discrimination parameter of item k. The 2PL model therefore is an extension of the Rasch model in that it allows the estimation of item-specific difficulty and discrimination parameters. Conceptually connecting the 2PL model with the LLTM, Embretson (1999) suggested the 2PL-constrained model, which is given by

$$P(Y_{ik} = 1|\theta_i, q_k, \eta, \tau) = \frac{exp\left[\left(\sum_{j=1}^{J} q_{kj}\tau_j\right)\left(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j\right)\right]}{1 + exp\left[\left(\sum_{j=1}^{J} q_{kj}\tau_j\right)\left(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j\right)\right]}, \quad (4)$$

with τ_j reflecting the basic parameters of the J design variables with respect to item discrimination. In addition to decomposing item difficulties, this model can therefore check whether the presence of certain design features in an item enlarge or decrease its discriminatory power. The 2PL-constrained model is nested in the 2PL model and therefore allows a direct comparison of model fit.

Both the LLTM and the 2PL-constrained model make the strong assumption that all item difficulties can be perfectly predicted from the basic parameters, that is, there is no error term in the regression of the item difficulties and/or discrimination parameters on the respective item design features. An implication of this assumption is that all items with the same design structure must have the same item parameters; for example, in the LLTM, all items with the same design vector q_{kj} must have the same item difficulty b_k. It has been shown that there can still be considerable variation in the item difficulties after accounting for item design features (Embretson, 1998). To take this into account, an error term must be introduced into the model. Janssen, Schepers, and Peres (2004) have presented an application of this approach for the LLTM, where item difficulty b_k is decomposed as

$$b_k = \sum_{j=1}^{J} q_{kj}\eta_j + \epsilon_k, \quad (5)$$

with $\epsilon_k \sim N(0, \sigma_\epsilon^2)$. The error term ε_k now captures the residual variance not explained by the design parameters. This approach can be generalized to the 2PL-constrained model as well, that is, the discrimination parameter a_k can be assumed to show variation between structurally equal items. A framework allowing the analysis of such effects has been suggested by Glas and van der Linden (2003) and by De Jong, Steenkamp, and Fox (2007). By allowing random error in these models, the amount of variance that is explained by the cognitive design in the item parameters can be evaluated and, hence, the quality of the proposed cognitive model can be assessed.

RTs in Educational Assessment

Traditional data analysis in educational assessment is founded on accuracy scores. Results obtained in classical, unidimensional IRT models (such as the 2PL model) usually provide information on person and item parameters: For each person, a person parameter reflecting latent ability is estimated, and for each item, a difficulty parameter and a discrimination parameter are obtained (Embretson & Reise, 2000). In such models, RTs are not modeled. However, RTs are easily available in times of computerized testing, and they can contain important information beyond accuracy scores. For example, RTs are helpful in detecting faking behavior on personality questionnaires (Holden & Kroner, 1992), and they can provide information on the speededness of a psychometric test or test items (Schnipke & Scrams, 1997; van der Linden et al., 1999) or aberrant response patterns (van der Linden & van Krimpen-Stoop, 2003). Apart from these issues pertaining to test administration, RTs in psychometric tests with a design potentially contain vital information concerning the underlying cognitive processes. Importantly, RT analysis may allow new insights into cognitive processes that transcend those obtained by IRT modeling. For example, one might be interested in the relationship between the difficulty and time intensity of a cognitive process: Are difficult processes the most time-intensive? What is the relationship between latent ability (e.g., intelligence) and speed? Does the test format affect this relationship? To investigate these questions, a unified treatment of accuracy scores and RTs is required.

Three different strategies have been used in the past to extract RT-related information from psychometric tests. Under the first strategy, RTs are modeled *exclusively*. This strategy is usually applied to speed tests that are based on very simple items administered with a strict time limit for which accuracy data offer only limited information. For example, in his linear exponential model, Scheiblechner (1979) has suggested that the RT T for person i responding to item k is exponentially distributed with density

$$f(t_{ik}) = (\tau_i + \gamma_k)exp[-(\tau_i + \gamma_k)t_{ik}], \quad (6)$$

where τ_i is a person speed parameter, and γ_k is an item speed parameter. Analogous to the LLTM, the item speed parameter (γ_k) can now be decomposed into component processes that are necessary to solve the item:

$$\gamma_k = \sum_{j=1}^{J} a_{kj}\eta_j, \tag{7}$$

where η_j indicates the speed of component process j, and a_{kj} is a weight indicating whether component process j is present in item k. Maris (1993) suggested a similar model, on the basis of the gamma distribution. Note that these models focus on RTs exclusively, whereas accuracy scores are not taken into consideration.

A second strategy chosen by several authors implies a *separate* analysis of RTs and accuracy scores. For example, Gorin (2005) decomposed the difficulty of reading comprehension items using the LLTM, and in a second step regressed the log-transformed RTs on the basic parameters. A similar approach was chosen by Embretson (1998) and Primi (2001) with a figural reasoning task, whereas Mulholland et al. (1980) used analyses of variance to predict RTs by item properties in a figural analogies test separately for correct and wrong answers, respectively (cf. Sternberg, 1977). In contrast, Bejar and Yocom (1991) compared both difficulty parameters and the shape of cumulative RT distributions of item isomorphs, that is, parallel items, in two figural reasoning test forms. Separate analyses provide some information on both accuracy scores and RTs, but the relation between these two variables cannot be modeled, as they are assumed to vary independently. For an analysis that overcomes this difficulty, a model is needed that can simultaneously estimate RT parameters and IRT parameters. This has been done in a third strategy of analyses on the basis of the *joint modeling* of both RTs and accuracy scores. Recently, several models have been proposed for the investigation of RTs in a psychometric test within an IRT framework. One of the first models was introduced by Thissen (1983), which describes the log-transformed RT of person i to item k as

$$log(T_{ik}) = \upsilon + s_i + u_k - bz_{ik} + \epsilon_{ik}, \tag{8}$$

with $\epsilon_{ik} \sim N(0, \sigma^2_\epsilon)$. In this model, υ reflects the overall mean log RT, s_i and u_k are person- and item-related slowness parameters, respectively, $-b$ represents the log-linear relation between RT and ability, z_{ij} is the logit estimated from a 2PL model, and ε is an error term. The new parameter in this model is $-b$, which reflects the relationship of ability and item difficulty with the RT. The model suggested by Thissen (1983) is rather descriptive than explanatory in nature in that it does not provide a decomposition of item parameters reflecting cognitive operations (but see Ferrando & Lorenzo-Seva, 2007, for a recent extension of Thissen's model to questionnaire data).

The model proposed by Roskam (1997), which conceptually is very similar to the model by Verhelst, Verstralen, and Jansen (1997), specifies the probability of a correct response of person i to item k as

$$P(Y_{ik} = 1|T_{ik}) = \frac{\theta_i T_{ik}}{\theta_i T_{ik} + \varepsilon_k} = \frac{exp(\xi_i + \tau_{ik} - \sigma_k)}{1 + exp(\xi_i + \tau_{ik} - \sigma_k)}, \tag{9}$$

where θ_i represents the person ability, ε_k is item difficulty, T_{ik} is RT, and ξ_i, τ_{ik}, and σ_k represent the natural logarithms of θi, T_{ik}, and ε_k, respectively. In this model, RT is parameterized as a predictor for the solution probability of item k by person i. As can be seen, if T_{ik} goes to infinity, the probability of a correct solution approaches 1 irrespective of item difficulty. The model, therefore, is more suitable for speed tests than for power tests, because items in a speed test usually have very low item difficulties under conditions without a time limit. This is not the case for items in a power test, even with a moderate time limit.

A model more suitable for power tests under time-limit conditions was proposed by Wang and Hanson (2005), who extended the traditional three-parameter logistic model by including RTs as well as parameters reflecting item slowness and person slowness, respectively:

$$P(Y_{ik} = 1|\theta_i, \rho_i, a_k, b_k, c_k, d_k, T_{ik}) = c_k + \frac{1 - c_k}{1 + e^{(-1.7a_k[\theta_i - (\rho_i d_k/T_{ik}) - b_k])}}, \tag{10}$$

where a_k, b_k, and c_k are the discrimination, difficulty, and guessing parameter of item k; and θ_i is a parameter for person i. d_k is an item slowness parameter; ρ_i is a person slowness parameter; and T_{ik} is the RT of subject i on item k. In this model, RTs are treated as an additional predictor, but in contrast to the model by Roskam (1997), as RT goes to infinity, a classical three-parameter logistic model is obtained. A similar model for speeded reasoning tests, with presentation time as an experimentally manipulated variable, was developed by Wright and Dennis (1999) in a Bayesian framework. The model allows the dissociation of time parameters with respect to persons and items, thereby avoiding the problematic assumptions as above. However, a major problem here pertains to the RTs, which are modeled as fixed parameters. It is a common assumption across the literature that RT is a random variable (Luce, 1986). By treating a variable assumed to be random as fixed, systematic bias in parameter estimation can occur. Further, the joint distribution of item responses and RTs cannot be analyzed. The model by Wang and Hanson (2005), therefore, can only be regarded as a partial model, as stated by the authors.

A different approach was chosen by van Breukelen (2005). He used a bivariate mixed logistic regression model, predicting log-normalized RTs as well as the log-odds of correct responses simultaneously. For the log-odds, the model assumed the log-normalized RTs and item-related design parameters with random effects (Rijmen & De Boeck, 2002) as predictors. Similarly, the RTs were predicted by item-related design parameters as well as accuracy scores. However, this approach can be problematic. van Breukelen (2005), for example, took the log-normalized RTs into account but did not specify parameters reflecting the test taker's speed or the time intensity of the items. If RTs are both regarded as a person-related predictor and as being implicitly equal to processing speed, as was done in the model by van Breukelen, the assumption is made that the time intensity of the items is equal, although their difficulties are not. This assumption can be avoided by including explicit time parameters in the model, reflecting the time intensity of the items and the speed of the test takers, respectively.

To conclude, several IRT models have been developed recently that are capable of incorporating RTs, but these suffer from some conceptual or statistical drawbacks for the application to time-limited tests. Further, they cannot relate the design structure of the utilized items to the RTs and accuracy scores simultaneously. A model that can overcome these difficulties, on the basis of the model developed by van der Linden (2007), is described below.

A Model for Response Accuracies and RTs

With responses and RTs, we have two sources of information on a test. The first provides us with information on the response accuracy of test takers on a set of items. The RTs result from the required processing time to solve the items. Naturally, test takers differ in their speed of working, and different items require different amounts of cognitive processing to solve them. This leads us to consider RTs as resulting from person effects and item effects, a separation similar to that made in IRT. A framework will be developed that deploys separate models for the responses and the RTs as measurement models for ability and speed, respectively. At a higher level, a population model for the person parameters (ability and speed) is deployed to take account of the possible dependencies between the person parameters. This hierarchical modeling approach was recently introduced by van der Linden (2007). The focus of this article, however, is on the item parameter side. A novel model is presented where the item parameters of both measurement models can be modeled as a function of underlying design factors.

Level 1 Measurement Model for Accuracy

The probability that person $i = 1, \ldots, N$ answers item $k = 1, \ldots, K$ correctly ($Y_{ik} = 1$), is assumed to follow the 2-parameter normal ogive model (Lord & Novick, 1968):

$$P(Y_{ik} = 1 | \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k), \tag{11}$$

where θ_i denotes the ability parameter of test taker i, and a_k and b_k denote the discrimination and difficulty parameters of item k, respectively. $\phi(.)$ denotes the cumulative normal distribution function. This normal ogive form of the 2-parameter IRT model is adopted for computational convenience, as was shown by Albert (1992). Its latent variable form lends itself perfectly for Bayesian estimation and is given by

$$Z_{ik} = a_k\theta_i - b_k + \varepsilon_{\theta ik}, \tag{12}$$

where $Z_{ik} \geq 0$ when $Y_{ik} = 1$ and $Z_{ik} < 0$ otherwise and with $\varepsilon_{\theta ik} \sim N(0, 1)$. With this data augmentation approach (Albert, 1992; Lanza, Collins, Schafer, & Flaherty, 2005), it is possible to change from dichotomous response variables to continuous latent responses. Also, as is shown below, after a suitable transformation of the RTs to normality, the simultaneous distribution of the responses and RTs turns out to be a bivariate normal one. This allows us to view the entire structure as a multivariate normal model, thereby simplifying the statistical inferences as well as the estimation procedure.

Level 1 Measurement Model for Speed

As a result of a natural lower bound at zero, the distribution of RTs is skewed to the right. Various types of distributions are able to describe such data. For instance, the Poisson, Gamma, Weibull, inverse normal, exponential, and lognormal distributions have been employed to describe RT distributions in psychometric applications. The reader is referred to Maris (1993), Roskam (1997), Rouder et al. (2003), Thissen (1983), van Breukelen (1995), Schnipke and Scrams (1997), and van der Linden (2006) for examples. However, in this application, the log-normal model is chosen to model the RT distributions for specific reasons. First of all, Schnipke and Scrams and van der Linden have shown that the lognormal model is well suited to describe such distributions, and it generally performs well with respect to model fit, as we experienced during the analyses of several data sets. Second, the lognormal model fits well within the larger framework for responses and RTs. It is assumed that the log-transformed RTs are normally distributed. Thereby, as mentioned above, the simultaneous distribution of the latent responses and log-transformed RTs can be viewed as a bivariate normal one. This is a strong advantage over other possible RT distributions, because its generalization to a hierarchical model becomes straightforward. Also, the properties of multivariate normal distributions are well known (Anderson, 1984), which simplifies the statistical inferences.

By analogous reasoning, an RT model will be developed that is similar in structure to the 2-parameter IRT model. Test takers tend to differ in their speed of working on a test; therefore, a person speed parameter ζ_i is introduced. Like ability in IRT, speed is assumed to be the underlying construct for the RTs. Also, it is assumed that test takers work with a constant speed during a test and that, given speed, the RTs on a set of items are conditionally independent. That is, the speed parameter captures all the systematic variation within the population of test takers. These assumptions are similar to the assumptions of constant ability and conditional independence in the IRT model.

However, test takers do not divide their time uniformly over the test, because items have different time intensities. The expected RT on an item is modeled by a time intensity parameter λ_k. Basically, an item that requires more steps to obtain its solution can be expected to be more time intensive, which is then reflected in a higher time intensity. It can be seen that λ_k is the analogue of the difficulty parameter b_k, reflecting the time needed to solve the item. As an example, running 100 m will be less time consuming than running 200 m. Clearly, the latter item takes more steps to be solved and will have a higher time intensity. An illustration of the effect on time intensity on the expected RTs is given in Figure 2. In this figure, item characteristic curves (ICCs) for the IRT model (left figure) and response time characteristic curves (RTCCs) (right figure) are plotted against the latent trait. The RTCCs show the decrease in expected RT as function of speed. For both measurement models, two curves are plotted that show the shift in probability/time as a result of a shift in difficulty/time intensity. In this example, the above curve would reflect running 200 m, whereas the lower curve reflects the expected RTs on the 100-m distance. Note, however, that it is not necessarily so that running 200 m is more difficult than 100 m.

Now, for the expectation of the log-RT T_{ik} of person i on item k we have obtained that $E(T_{ik}) = -\zeta_i + \lambda_k$. However, a straightforward yes–no question might show less variability around its mean λ_k than predicted by ζ_i. Such an effect can be considered as the discriminative power of an item, and therefore a time discrimination parameter ϕ_k is introduced. This parameter controls the decrease in expected RT on an item for a one-step increase in speed of a test taker. It is the analogue of the discrimination parameter a_k in Equation 12. The effect of item discrimination on the ICCs and RTCCs are illustrated in Figure 3. It can be seen that the difference in expected RTs between test takers working at different speed levels is less for the lower discriminating item.

Finally, the log-RT T_{ik} of person i on item k follows a normal model according to

$$T_{ik} = -\phi_k\zeta_i + \lambda_k + \epsilon_{\zeta ik}, \tag{13}$$

where $\epsilon_{\zeta ik} \sim N(0, \sigma_k^2)$ models the residual variance.

Level 2 Model for the Person Parameters

In IRT, it is common to view observations as nested within persons. Local independence between observations is assumed conditional on the ability of a test taker. That is, a test taker is seen as a sample randomly drawn from a population distribution of test takers. Usually, a normal population model is adopted, so

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2). \tag{14}$$

The local independence assumption has a long tradition in IRT (e.g., Lord, 1980; Holland & Rosenbaum, 1986). Similarly, the speed parameter models the heterogeneity in the observed RTs between test takers. Therefore, conditional on the (random) speed parameters, there should be no covariation left between the RTs on different items. In other words, conditional independence is assumed in the RT model as well. Now a third assumption of conditional independence follows from the previous two. If test takers work with constant speed and constant ability during a test, then within

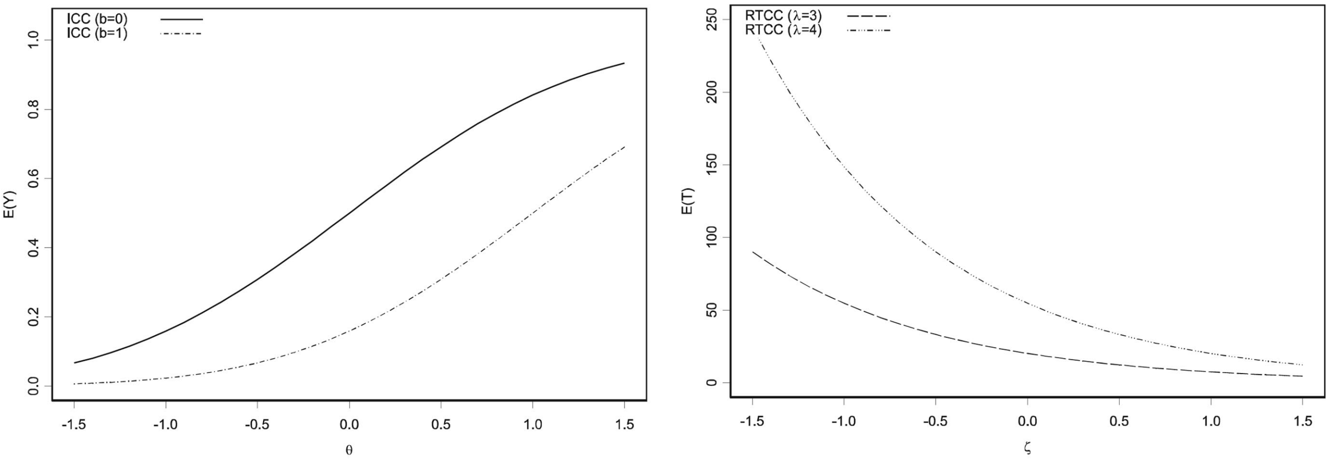

Figure 2. Item characteristic curves (ICCs) and response time characteristic curves (RTCCs) for two items with differing time intensity and difficulty, where $a = \phi = 1$.

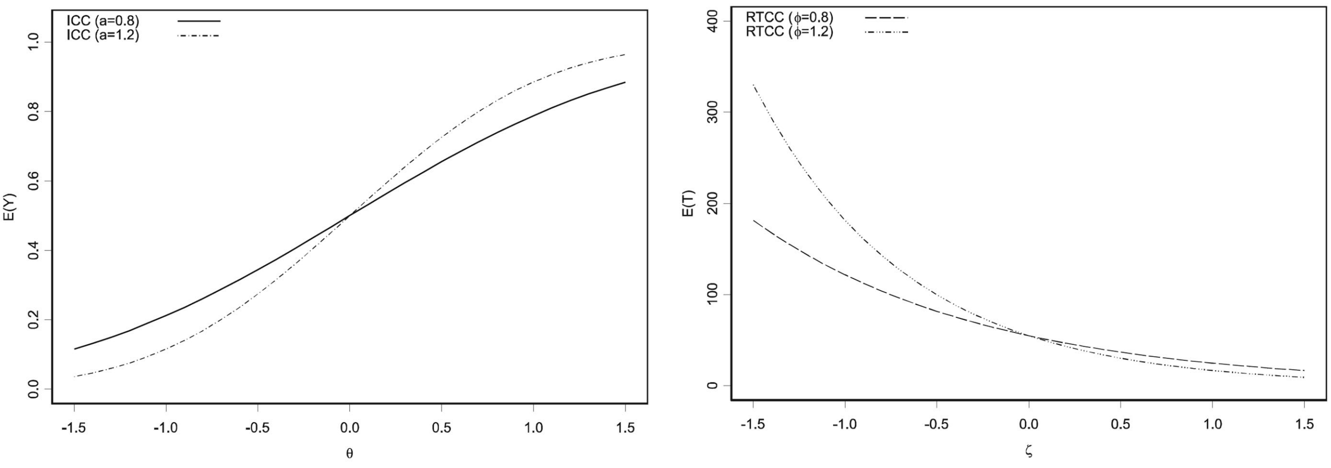

Figure 3. Item characteristic curves (ICCs) and response time characteristic curves (RTCCs) for two items with differing discrimination, where $b = 0$, and $\lambda = 4$.

an item these parameters should capture all possible covariation between the responses and RTs. That is, the responses and RTs on an item are assumed to be conditionally independent given the levels of ability and speed of the test takers. At the second level of modeling, these person parameters are assumed to follow a bivariate normal distribution:

$$(\theta_i, \zeta_i) = \mu_P + e_{P,}\ e_P \sim N\left(0, \sum\nolimits_P\right), \qquad (15)$$

where $\mu_p = (\mu_\theta, \mu_\zeta)$, and the covariance structure is specified by

$$\sum\nolimits_P = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\theta\zeta} & \sigma_\zeta^2 \end{bmatrix}. \qquad (16)$$

The parameter $\sigma_{\theta\zeta}$ in the model for the person parameters reflects possible dependencies between speed and accuracy of the test takers. For instance, when $\sigma_{\theta\zeta}$ is negative, this means that persons who work faster than average on the test are expected to have below-average abilities. When $\sigma_{\theta\zeta} = 0$, there is independence between ability and speed. However, this is not necessarily equivalent to independence between the responses and RTs, because such a dependency can occur via the item side of the model as well, as is discussed below.

This hierarchical approach, which was first presented by van der Linden (2007), models a connection between the two Level 1 measurement models. Note that Equation 15 is a population model and is therefore entirely different from what is known as the speed–accuracy trade-off (Luce, 1986). The latter is a within-person phenomenon, reflecting the trade-off between accuracy and speed of working for a specific test taker, and it is often assumed to be negative. That is, it assumes that a test taker chooses a certain speed level of working and, given that speed, attains a certain ability. If he or she chooses to work faster, the trade-off then predicts that this test taker will make more errors and, as a result, will attain a lower ability. On the contrary, the model given in Equation 15 describes the relationship between ability and speed at the population level. It is perfectly reasonable that, within a population, the dependency between ability and speed is positive, reflecting that faster working test takers are also the higher ability candidates. In the analysis of real test data, we have found positive as well as negative dependencies between ability and speed (Klein Entink, Fox, & van der Linden, in press).

So far, the model is equivalent to that presented by van der Linden (2007) and as described in Fox, Klein Entink, and van der Linden (2007). Another possible bridge between the two Level 1 models can be built on the item side. That one is developed now and presents a novel extension of the model that allows us to describe item parameters as a function of underlying cognitive structures, which is the focus of this article.

Level 2 Model for the Item Parameters

The hierarchical approach is easily extended to the item side of the model. As discussed in the overview in the introduction section, several approaches have been developed to model underlying item design structures in IRT. However, some of these approaches made rather strict assumptions by incorporating the design model into the IRT model. We present an approach in which this is avoided by introducing possible underlying design features at the second level of modeling.

Interest goes out to explaining differences between items resulting from the item design structure. Because the characteristics of the items are represented by their item parameters, it seems straightforward to study the differences in the estimated item parameters as a function of the design features. Moreover, it should be possible to assess to what extend the differences in these parameters can be explained by the design features. To do so, the hierarchical modeling approach is extended to the item side of the model first. Similarly to Equation 15, the vector $\xi_k = (a_k, b_k, \phi_k, \lambda_k)$ is assumed to follow a multivariate normal distribution,

$$\xi_k \sim N\left(\mu_I, \sum\nolimits_I\right), \tag{17}$$

where Σ_I specifies the covariance structure of the item parameters:

$$\sum\nolimits_I = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{a\phi} & \sigma_{b\phi} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{a\lambda} & \sigma_{b\lambda} & \sigma_{\phi\lambda} & \sigma_\lambda^2 \end{bmatrix}. \tag{18}$$

Σ_I is the second bridge between the Level 1 models. It allows us to study dependencies between the item parameters. For instance, if there is a dependency between item difficulty and time intensity, this would be reflected by the covariance component between these parameters. For instance, a positive estimate for $\sigma_{b\lambda}$ indicates that more difficult items also tend to be more time consuming.

Now suppose we have a test in which items are formulated using either squares, circles, or triangles and we are interested whether such items differ in their difficulty. This leads us to consider the following model, in which we develop an analysis of variance approach to model the effects of each rule. That is, the means of the item parameters are decomposed into a general mean and deviations from that mean as a result of the underlying item construction rules used to formulate the items. To reflect the three symbols used to formulate the items, two dummy variables are constructed. The first variable, denoted by A_1 of length K, contains a 1 for circles, a 0 for triangles, and a -1 for squares. The second variable A_2 contains a 0 for circles, a 1 for triangles, and also a -1 for squares. Now, following Equation 5, the difficulty of item k can be modeled as

$$b_k = \gamma_0^{(b)} + A_{1k}\gamma_1^{(b)} + A_{2k}\gamma_2^{(b)} + e_k^{(b)}. \tag{19}$$

This indicator variable approach models the difficulty of item k as a deviation from the base level γ_0 as a result of the figure used to construct the item. That is, if item k is constructed using circles, its difficulty is predicted by $\gamma_0 + \gamma_1$. If there are triangles used, its difficulty is given by $\gamma_0 + \gamma_2$. In the case that squares are used, its difficulty is modeled as $\gamma_0 - \gamma_1 - \gamma_2$. Note that when γ_3 denotes the effect for squares, it must equal $-\gamma_1 - \gamma_2$ because otherwise the model is over parameterized. Let $A = (1, A_1, A_2)$ and $\gamma^{(b)} = (\gamma_0, \gamma_1, \gamma_2)^t$, then the model can be represented as

$$b_k = A_k\gamma^{(b)} + e_k^{(b)}. \tag{20}$$

In the previous example, the interest was only in dissociating the heterogeneity in the item difficulty parameters into three possible groups of items. However, if we are interested in validating a cognitive model that underlies the item design, it makes sense to extend the model to the other item parameters as well. The full multivariate model for the item parameters can be generalized to

$$a_k = A_k\gamma^{(a)} + e_k^{(a)}, \tag{21}$$

$$b_k = A_k\gamma^{(b)} + e_k^{(b)}, \tag{22}$$

$$\phi_k = A_k\gamma^{(\phi)} + e_k^{(\phi)}, \tag{23}$$

$$\lambda_k = A_k\gamma^{(\lambda)} + e_k^{(\lambda)}, \tag{24}$$

where the error terms are assumed to follow a multivariate normal distribution, that is $e \sim N(0, \Sigma_I)$. This is a generalization of Equation 5, not only by allowing for residual variance in other item parameters than b but by modeling covariance components between the item parameters as well. Further, A is a design matrix containing zeros and ones denoting which construction rules are used for each item; $\gamma^{(a)}$, $\gamma^{(b)}$, $\gamma^{(\phi)}$, and $\gamma^{(\lambda)}$ are the vectors of effects of the construction rules on discrimination, difficulty, time discrimination, and item time intensity, respectively.

The complete model structure is represented in Figure 4. The ovals denote the measurement models for accuracy (left) and speed (right). The circles at Level 2 denote the covariance structures that connect the Level 1 model parameters. The structural model is denoted by the square box. The square containing A_I denotes the design matrix containing item specific information that allows for explaining variance between the item parameters. This approach is not limited to rule based test construction but can just as well be used to test hypotheses of, for instance, differences in cognitive processing when data are presented in a table versus presented in a figure.

By the conditional independence assumption and by taking the possible dependencies to a second level of modeling, this framework becomes very flexible. It allows for the incorporation of any measurement model for either accuracy or speed. For example, the measurement model for the dichotomous responses could be replaced by a model for polytomous items. When needed, independence between the two Level 1 models can be obtained by restricting Σ_I and Σ_P to be diagonal matrices. However, the strength of the framework comes from the simultaneous modeling of two data sources on test items. The two likelihoods at Level 1, linked via the covariance structures at Level 2, allow us to use the RTs as collateral information in the estimation of the response parameters and vice versa.

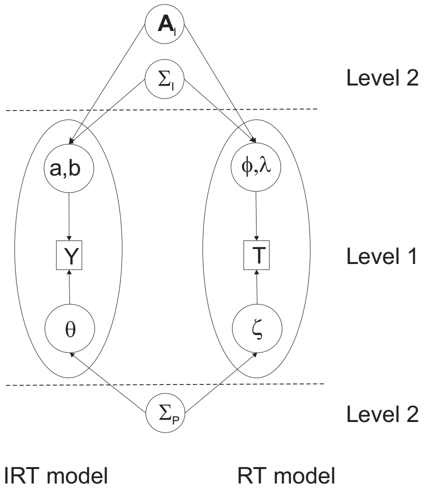

Figure 4. Schematic representation of the modeling structure. IRT = item response theory; RT = response time.

Bayesian Inference and Estimation

This section deals with the statistical treatment of the model. The model is estimated in a fully Bayesian framework. Before discussing the estimation procedures, however, first the basic principles of the Bayesian approach are introduced. For a general introduction to the Bayesian approach and its estimation methods, see Gelman, Carlin, Stern, and Rubin (2004). Bayesian estimation of IRT models is discussed in, for instance, Albert (1992), Patz and Junker (1999), and Fox and Glas (2001).

Bayesian Approach

In the classical approach to statistics, a parameter μ is assumed to be an unknown, but fixed, quantity. A random sample from a population indexed by μ is obtained. On the basis of the observed sample, the value of μ can be estimated. Instead, in the Bayesian approach, μ is assumed to be random. That is, there is uncertainty about its value, which is reflected by specifying a probability distribution for μ. This is called the *prior distribution* and reflects the subjective belief of the researcher before the data are seen. Subsequently, a sample is obtained from the distribution indexed by μ, and the prior distribution is then updated. The updated distribution is called the *posterior* and is obtained via Bayes' rule. Let $p(\mu)$ denote the prior and $f(\mathrm{x}|\mu)$ denote the sampling distribution, then the posterior density of $\mu|\mathrm{x}$ is

$$p(\mu|\mathrm{x}) = f(\mathrm{x}|\mu)p(\mu)/m(\mathrm{x}), \quad (25)$$

where $m(\mathrm{x})$ denotes the marginal distribution of x (Casella & Berger, 2002, p. 324).

Markov Chain Monte Carlo (MCMC) Methods

The posterior distributions of the model parameters are the objects of interest in Bayesian inference. For simple models, obtaining these estimates can be done analytically. However, for complex models as presented above, it is impossible to do so. Sampling based estimation procedures, known as MCMC methods, however, solve these problems easily. A strong feature of these methods is that their application remains straightforward, whereas model complexity may increase.

The MCMC algorithm applied in this article is known as the Gibbs sampler (Geman & Geman, 1984). To obtain samples from the posterior distributions of all model parameters, a Gibbs sampling algorithm requires that all the conditional distributions of the parameters can be specified. Basically, a complex multivariate distribution from which it is hard to sample is broken down into smaller univariate distributions, conditional on the other model parameters, from which it is easy to draw samples. After giving the algorithm some arbitrary starting values for all parameters, it alternates between the conditional distributions for M iterations. Thereby, every step depends only on the last draws of the other model parameters. Hence, (under some broad conditions) a Markov Chain is obtained that converges toward a target distribution. It has been shown that if the number of iterations goes to infinity, the target distribution can be approximated with any accuracy (Robert & Casella, 2004).

To illustrate the approach, consider estimation of the RT model given by Equation 13. For simplicity of this example, we assume that $\phi = 1$ and independence from the response model. First, (independent) prior distributions for ζ, λ, and σ^2 are specified. Now, because it does not depend on $m(\mathrm{t})$ up to some constant, the posterior distribution is proportional to $p(\zeta, \lambda, \sigma^2|\mathrm{t}) \propto f(\mathrm{t}|\zeta, \lambda, \sigma^2)p(\zeta)p(\lambda)p(\sigma^2)$. After providing the algorithm with starting values $\zeta^{(0)}$, $\lambda^{(0)}$, and $\sigma^{2(0)}$, the algorithm proceeds as follows:

1. At iteration m, draw the person parameters ζ from $p(\zeta|\lambda^{(m-1)}, \sigma^{2(m-1)}, \mathrm{t})$.

2. Using the new values $\zeta^{(m)}$, draw λ from $p(\lambda|\zeta^{(m)}, \sigma^{2(m-1)}, \mathrm{t})$.

3. Using the new values $\lambda^{(m)}$, draw σ^2 from $p(\sigma^2|\zeta^{(m)}, \lambda^{(m)}, \mathrm{t})$.

4. Increment m with 1 and repeat the above steps for M iterations.

Now M values for both parameters have been obtained. Before descriptive statistics as the posterior mean and posterior variance can be obtained, issues such as autocorrelation of the samples and convergence of the Markov chain must be checked. Most statistical software packages provide means to obtain autocorrelations. Convergence can be checked by making trace plots, that is, plotting the drawn samples against their iteration number. This allows for a visual inspection to determine whether stationarity has been reached. Dividing the MCMC chain into two or more subsets of equal sample size and comparing the posterior mean and standard deviation also provides information on convergence. Another approach is rerunning the algorithm using different starting values. This is also helpful to determine whether the chain has really converged to a global optimum. Other (numerical) methods to assess convergence issues are discussed in Gelman et al. (2004; Section 11.6). Because the first samples are influenced by the starting values, a "burn-in" period is used, which means that the first samples of the chain are discarded. The posterior means and variances of the parameters are then obtained from the remaining Q samples. This is usually done by checking convergence of the chain and when this seems to be reached, running the algorithm for another few thousand iterations on which the inferences can be based. The BOA software for use in the SPLUS or R statistical environment provides several of these diagnostic tools (numerical and graphical) to assess convergence of the MCMC chains (Smith, 2007).

The model presented above lends itself for a fully Gibbs sampling approach. This is a feature of the multivariate normality of the responses and RTs after the data augmentation step. The derivation of the conditional distributions for the Gibbs sampling algorithm is discussed in the Appendix of this article. The algorithm has been implemented in Visual FORTRAN Pro 8.0.

Model Checking and Evaluation

In a Bayesian framework, goodness of fit tests can be performed using posterior predictive checks (Gelman et al., 2004; Gelman, Meng, & Stern, 1996). Model fit can be evaluated by comparing replications of the data x^{rep}, drawn from the posterior predictive distribution of the model, with the observed data. A discrepancy between model and data is measured by a test quantity $T(\mathrm{x}, \mu)$ (e.g., mean squared error), where x denotes the data, and μ denotes the vector of model parameters. A Bayesian p value, p^*, can be estimated as the probability that the replicated data under the model are more extreme than the observed data:

$$p* = P(T(\mathrm{x}^{rep}, \mu) \geq T(\mathrm{x}, \mu)|\mathrm{x}), \tag{26}$$

whereby p values close to 0 or 1 indicate extreme observations under the model. Using the drawn samples each iteration of the Gibbs sampler, these estimates of the p values are easily obtained as a by product from the MCMC chain. For more details, see Gelman et al. (1996).

Next, appropriate test quantities have to be chosen. An important assumption of the model is that of local independence. Therefore, an odds ratio statistic was used to test for possible violations of local independence between response patterns on items. For an impression of the overall fit of the response model, an observed score statistic was estimated to assess whether the model was able to replicate the observed response patterns of the test takers. For a detailed description of these two statistics, see Sinharay (2005) and Sinharay, Johnson, and Stern (2006).

Residual analysis is another useful means to examine the appropriateness of a statistical model. The basic idea is that the observed residuals, that is, the difference between the observed values and the expected values under the model, should reflect the assumed properties of the error term. To assess the fit of the RT model, van der Linden and Guo (in press) proposed a Bayesian residual analysis. More specifically, by evaluating the actual observation t_{ik} under the posterior density, the probability of observing a value smaller than t_{ik} can be approximated by

$$u_{ik} \approx \sum_{m=0}^{M} \Phi(t_{ik}|\zeta_i^{(m)}, \phi_k^{(m)}, \lambda_k^{(m)})/M, \tag{27}$$

from M iterations from the MCMC chain. According to the probability integral transform theorem (Casella & Berger, 2002, p. 54), under a good fitting model, these probabilities should be distributed $U_{ik} \sim U\ (0, 1)$. Model fit can then be checked graphically by plotting the posterior p values against their expected values under the $U\ (0, 1)$ distribution. When the model fits well, these plots should approximate the identity line.

Model Selection

Research hypotheses are usually reformulated so that two competing statistical models are obtained that explain the observed data. An appropriate test criterion then has to be selected that evaluates these two models with respect to their explanatory power for the data. The Bayes factor (Kass & Raftery, 1995; Klugkist, Laudy, & Hoijtink, 2005) can be used to test a model M_1 against another model M_0 for the data at hand. The Bayes factor is defined as the ratio of the marginal likelihoods of these models:

$$BF = \frac{p(y|M_0)}{p(y|M_1)}. \tag{28}$$

The marginal likelihood is the average of the density of the data taken over all parameter values admissible by the prior—that is, $p(\mathrm{y}|M) = \int p(\mathrm{y}|\gamma, M)p(\gamma|M)d\gamma$, where γ is the vector of model parameters. Because the Bayes factor weighs the two models against each other, a value near one means that both models are equally likely. A value of 3 or greater is considered to be strong evidence in favor of the null model, whereas on the contrary a value near zero favors the larger model as the best explanation for the data (Kass & Raftery, 1995).

In the special case that model M_0 is nested in model M_1, that is, $M_0 \subset M_1$, we can express model M_0 as a restriction of M_1: $p(y|M_1, \gamma = 0) = p(y|M_0)$. When this special case holds, computation of the Bayes factor for testing M_1 versus M_0 simplifies to evaluating the marginal posterior density $p(\gamma|y, M_1)$ at $\gamma = 0$. This result is known as the Savage–Dickey density ratio (Dickey, 1971; Verdinelli & Wasserman, 1995):

$$BF = \frac{p(\gamma = 0|y, M_1)}{p(\gamma = 0|M_1)}, \tag{29}$$

where $p(\gamma = 0|M_1)$ is the evaluation of the restriction under the prior density of model M_1. Using this result greatly reduces the computational burden because it allows to evaluate different models from the estimated marginal density of the effects under the largest model.

Explained Variance

A Bayesian R^2 statistic is proposed to assess the proportion of explained variance in the item parameters by the design rules. Gelman and Pardoe (2006) presented a R^2 statistic in Bayesian multilevel framework. For the difficulty parameters, we had from Equation 23

$$b_k = \mathrm{A}\gamma^{(b)} + e_k,$$

denoting the regression at Level 2 of the model parameters b on design matrix A, where we dropped the superscript (b) in the error term for the moment. Then, the proportion explained variance in the b parameters is given by

$$R^2 = 1 - \frac{E\left(\frac{1}{K-1}\sum_{i=1}^{K} e_k^2\right)}{E\left(\frac{1}{K-1}\sum_{i=1}^{K}(b_k - \bar{b})^2\right)}, \tag{30}$$

where E denotes the posterior mean. Using the MCMC algorithm, these expectations can be obtained by averaging over the draws from the posterior distribution. When the model explains almost all variability in b, the R^2 statistic will be close to 1. If the R^2 statistic is close to 0, then the variability in b almost equals the average variance of the errors.

Empirical Example

The application of the model is illustrated using a large-scale investigation of figural reasoning ability, on the basis of an earlier study by Hornke and Habon (1986). In their study, the rule based test design was evaluated on the difficulty parameters using the LLTM modeling approach. Although the data set they used in that study is different form the one used here, the underlying item design is the same. In our study, we not only try to validate the cognitive model on the item difficulties but we also try to validate the cognitive model by examining the time intensities of the items.

Principles of the Test

The current empirical example is based on the rule framework proposed by Hornke and Habon (1986) and Hornke (2002). These authors have distinguished between three types of radicals that largely correspond to those mentioned by Primi (2001): type of rules, number of rules, and perceptual organization of elements. Eight different rules were used for item design: identity, addition, subtraction, intersection, seriation, variation of closed gestalts, unique addition, and variation of open gestalts (see Figure 5). Identity implies that the same figural element occurs three times. For addition, a subject needs to mentally superimpose figural elements in the first two cells of a row or column, whereas

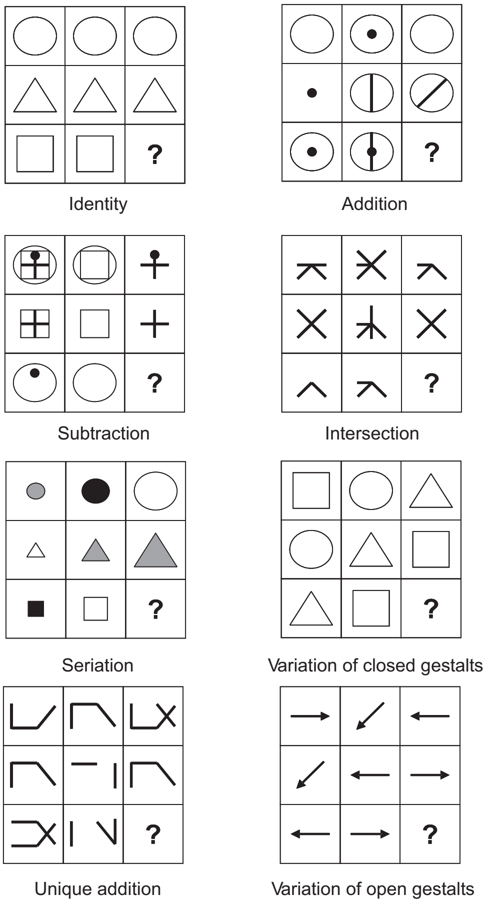

Figure 5. Set of operations used in item construction.

subtraction requires that elements occurring both in the first and second cell are omitted in the third. Intersection implies that only elements that occur in the first two cells can be present in the third cell of a row or column. Seriation means that a transformation of a figural element between the first and second cell is repeated from the second to the third cell (e.g., size). Further, variation of gestalts (both of closed and open ones) means that the sequence of presentation of figural elements is varied. Finally, unique addition means that only figural elements that occur once in the first two cells are also present in the third cell of a row or column. The number of rules per item varied between one and two.

Further, Hornke and Habon (1986) introduced a radical similar to perceptual organization: The figural components of an item could be either separated, integrated, or embedded. Separated components were designed to be easily distinguishable (e.g., see identity and variation of closed gestalts in Figure 5). Integrated components demand that design rules are related to different facets of the same figural element and that the figural element that relates to the rule operating in the item is identified (e.g., see seriation in Figure 5; the rule refers to shape but not to texture). In solving items with embedded components, an examinee must conduct an additional mental search operation to discover to which part of a figural element a rule relates (see unique addition in Figure 5). Finally, Hornke and Habon allowed relations between figural elements to occur across rows, columns, or both (e.g., see addition in Figure 5 for column-wise direction and subtraction for row- and column-wise direction, respectively). As can be seen from Figure 1, which represents a typical item from the test, eight solution alternatives were available to subjects. Some of the alternatives were partially correct (see Alternatives 3, 6, 7, and 8 for correct application of identity only; see Alternative 5 for correct application of unique addition only).

Data Set

Data from 30,000 examinees and 456 items were available (Hornke & Wilding, 1997). The test takers were divided over 30 groups who each took 12 items. Each group had six overlapping items with the previous group (and thus also six with the next group), which established the links between the groups. Because analyzing the complete data is computationally unfeasible, a subset of the data was analyzed. The subset chosen was large enough to have all design rules sufficiently present. Because the links between the groups had to be maintained, the first 6,422 test takers were selected, who answered a total of 186 items. From this set, 14 items were removed that were constructed from only one component, because perceptual organization is not involved in these items. (Another approach could be to estimate the IRT and RT models separately on a larger data set and analyze the rule-based design in a second step, on the basis of these estimates. However, for this example, it is preferred to analyze the accuracies and RTs jointly because this allows the estimation of the covariances between the Level 1 model parameters.) In the subset, the least occurring construction rule was the identity rule, which was used in 26 items. The variation of closed gestalts occurred most frequently, in 51 items. So, all construction rules were sufficiently present in the subset to obtain reasonable estimates of their effects. The row-wise and column-wise operations and their combination occurred almost equally in this sample (56, 56, and 60 times, respectively).

Goal of the Study

Basically, the test taker has to decipher and trace back the steps made by the test developer. That is, discovering the separate item components (perceptual organization), determining the row and column directions, and applying the appropriate item construction rule. The analysis aims at testing the assumption that each of these steps contributes to the amount of cognitive processing required to come to the solution of the item. If so, we expect that different combinations of perceptual organization, construction rules, and row-/column-wise organization not only lead to different difficulties of the items but also reflect the amount of time required to solve the item. Therefore, it is expected that different combinations of design features lead to heterogeneity in the difficulties and time intensities of the items. This leads us to consider the following model:

$$b_k = \gamma_0^{(b)} + \gamma_{rule\,1}^{(b)} + \gamma_{rule\,2}^{(b)} + \gamma_{p.o.}^{(b)} + \gamma_{rc}^{(b)} + e_k^{(b)}, \tag{31}$$

$$\lambda_k = \gamma_0^{(\lambda)} + \gamma_{rule\,1}^{(\lambda)} + \gamma_{rule\,2}^{(\lambda)} + \gamma_{p.o.}^{(\lambda)} + \gamma_{rc}^{(\lambda)} + e_k^{(\lambda)}, \tag{32}$$

where γ_0 denotes the baseline for difficulty/time intensity, which allows us to view the other effects as deviations from this base level. Further, γ_{rule} denotes the effect for the rule used (one of the eight), $\gamma_{p.o.}$ denotes the effect of the perceptual organization, γ_{rc} denotes the effect of the row-wise or column-wise operation, and e_k models the unexplained variability.

We restrict ourselves to the difficulty and time intensity parameters for practical reasons because only a low amount of 12 items per person was available. These parameters can be estimated with more precision than the discrimination parameters. Therefore, they lend themselves better for this study because the design matrix A involves many parameters. Necessarily, also interaction effects, such as Rule × Search operation, had to be ignored. Their incorporation would lead to $3 \times 8 = 24$ additional effects to be estimated, which is unfeasible unfortunately. Therefore, although a strict assumption, the design effects are assumed to be independent from each other.

From Equations 31 and 32 it is possible to formulate more restricted models to test some hypotheses. The following four models are considered:

1. Let M_0 denote the restricted model in which $(b_k, \lambda_k) = (\gamma_0^{(b)}, \gamma_0^{(\lambda)}) + (e_k^{(b)}, e_k^{(\lambda)})$. It assumes that there is no explanatory effect for difficulty or time intensity resulting from the cognitive design.

2. Model M_1 includes the effects for perceptual organization. That is, we want to test whether there is a difference in difficulty and/or time intensity between items where the two components are either embedded, integrated, or separated.

3. Model M_2 extends model M_1 by including also the design rules (two per item) that were used to formulate each item.

4. Model M_3 is the full model that includes all effects (unless, of course, the testing of its more restricted versions reveals otherwise). It allows to test whether there results any effect on difficulty or time intensity from the row-wise, column-wise, or row- and column-wise organization of the rules.

Design Matrix

To construct the design matrix A for this study, the indicator variable approach described earlier is used. Thereby, we used the information available from the item writing process. The first variable is a 1 for all items, to reflect the incorporation of the general mean for either difficulty or time intensity. The next two indicator variables differ per item according to its perceptual organization. The second indicator variable took the value of 1 for separated components, the value of 0 for integrated components, and the value of -1 for embedded components. Similarly, the third variable took the value of 0 for separated components, the value of 1 for integrated components, and the value of -1 for embedded components. As a result, the deviation from the base level γ_0 for embedded components equals $-\gamma_1 - \gamma_2$. For the eight design rules, an indicator variable was used that reflected whether the design rule was used to construct that item (denoted by a 1) or not (denoted by 0). For the row-wise, column-wise, and both column-wise and row-wise operations, two indicator variables were constructed similar to the way the perceptual organization of the items was modeled. Finally, a design matrix of dimension $K \times 13$ was obtained in this way, reflecting the full model M_3. To fit model M_2, the design matrix can simply be restricted by specifying all row-/column-wise indicators to be 0. The design matrix for M_1 and M_0 can be obtained similarly.

Analysis

In this section, the results of the analysis are discussed. First, estimation issues and model fit are discussed, followed by the interpretation of the parameter estimates obtained. Second, the hypotheses formulated above are evaluated.

Estimation

Identification of the model is obtained by setting $\mu_p = (\mu_\theta, \mu_\zeta) = (0, 0)$, specifying $\prod_{k=1}^{K}\phi_k = 1$ *and* $\sigma_\theta^2 = 1$ (see the Gibbs sampling algorithm in the Appendix). For estimation, vague proper priors were specified for the covariance components. The priors for the means and regression coefficients were chosen to be 0, except for the variance components $\mu_0^{(a)} = \mu_0^{(\phi)} = 1$. First, model M_0 was fitted to the data, to assess model fit and evaluate the estimated correlation structures. Subsequently, we fitted the other three models to the data. We used 12,000 iterations of the algorithm to estimate the model.

The BOA package was used to assess convergence of the MCMC chains. Graphical checks—such as autocorrelation, trace plots, and estimated densities of the parameters—are easily assessable with this software. For illustration, Figure 6 shows the two trace plots for the effect of the identity rule on difficulty and time intensity, respectively. Several statistics to assess convergence that were provided by this package were evaluated. Additionally, we used three runs with different random starting points for model M_0. The estimates from these three chains all converged to approximately the same marginal densities, indicating that convergence was reached. Finally, the estimated convergence statistics suggested to discard the first 1,200 iterations as burn-in. The final estimates of the posterior means and variances of all model parameters were therefore based on the last 10,000 iterations.

Model Fit

The fit of response model was assessed using posterior predictive checks. More specifically, the observed score statistic was used to see whether the fitted response model was able to describe the observed response patterns. To check whether the introduction of an item discrimination parameter was necessary, the fit of the Rasch model was compared with that of the 2-parameter normal ogive IRT model. A total of 1,000 replicated data sets under the posterior density were used to assess the fit of the model. The observed score statistic evaluates the number of test takers with 0, . . ., K items correct. Figure 7 shows this statistic for the 2-parameter model, where the lines denote the observed number of test takers with k correct items, and the dots with 95% highest posterior density (HPD) intervals denote the summary of the 1,000 replications under the model. It appeared that the Rasch model was unable to

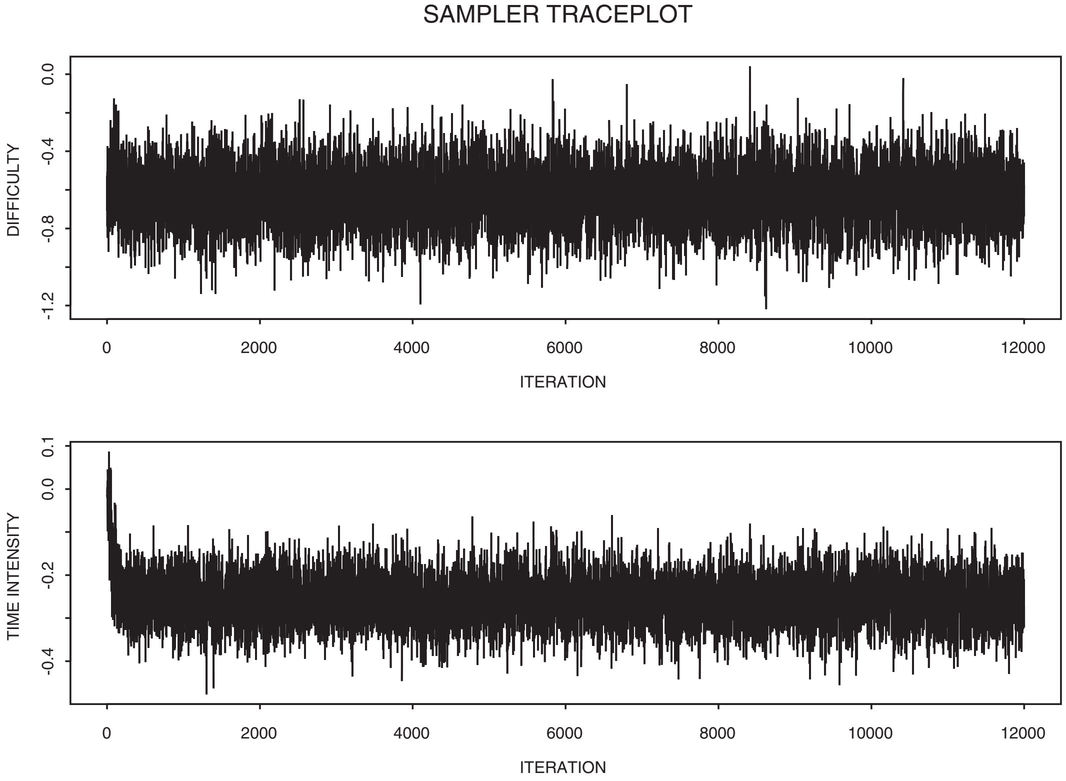

Figure 6. Trace plots for the effect of the identity rule on difficulty (above) and time intensity (below) under model M_2.

capture the observed data patterns, but from Figure 7, it can be seen that the 2-paramater normal ogive model performed quite well. The odds ratio statistic pointed at some item combinations in which a possible dependency might exist (a *p* value < 2.5 or > 97.5). However, upon inspection of the data, these appeared to be very hard items with a relative low proportion of correct scores. A significant *p* value was found only for a low percentage of all the possible item combinations (2.6%).

To test whether there were any systematic patterns in the RT data that were not captured by the model, the Bayesian residual check was used. Again, 1,000 iterations of the Gibbs sampler were used to estimate the posterior probabilities as given by Equation 27. Besides a graphical check of overall model fit, model fit was examined for each item as well. From the figures, it could be concluded that the underlying distribution was very likely to be U (0, 1) distributed. No serious aberrant patterns could be detected from these graphs. Therefore, because no systematic aberrancies were found, it was concluded that model fit was satisfactory for this data set.

Estimated Population Models

Before discussing the structural model on the item parameters, first the estimated covariance structures under the null model M_0 are discussed. This is of interest because the covariance components between the various parameters reveal the dependencies between the response model and the RT model. Table 1 gives the posterior means and posterior standard deviations for the variance and covariance parameters of Σ_P and Σ_I. From the covariance components, the correlation between the two parameters was estimated as well and is given in the last column.

The population model of the latent traits θ, ζ (given by Equation 15) provides us with information about the relationship between ability and speed of test takers. Note that the variance component for ability was fixed because of the identification restrictions. The estimated correlation between ability and speed was strongly negative (−.61). Interestingly, this tells us that in general the higher ability candidates took more time to solve the items.

Similarly, the estimates for Σ_I contain information about the items in the test. From Table 1, it can be seen that the most significant correlations between the item parameters are between the discrimination parameters a and item difficulty b (−.61), between discrimination parameter a and time intensity λ (−.54), and between difficulty b and time intensity λ (.68). So the more difficult items tend to be more time consuming as well, which is in line with the common assumption that the more complex cognitive reasoning

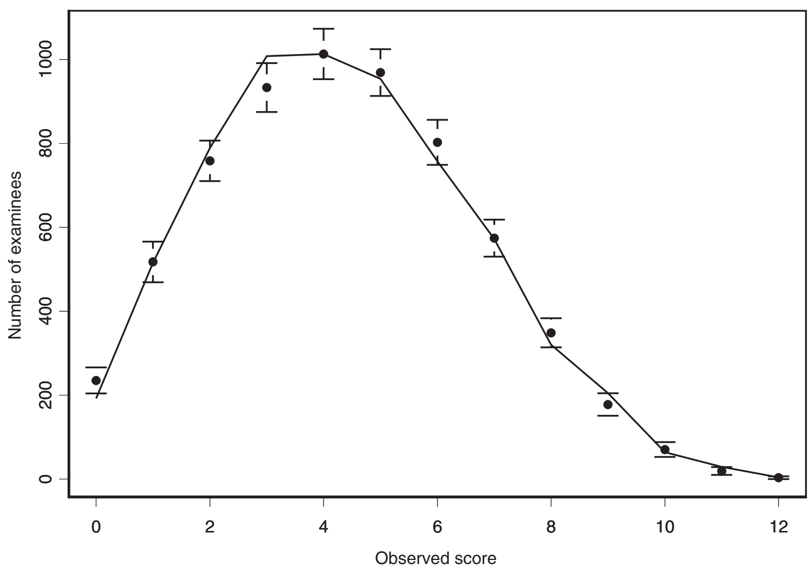

Figure 7. Observed sum scores (lines) and model predicted sum scores (dots with .95 highest posterior density regions).

items require more processing steps by the test taker. From the correlation with the *a* parameters, it follows that the more difficult and more time intensive items tend to discriminate less between test takers with different abilities.

Testing Hypotheses

To assess to which extent the perceptual organization, the specific construction rules, and the organization along rows and columns contribute to the difficulty and time intensity of the items, the four models are evaluated against each other. The estimated Bayes factors and R^2 statistics for the four models are given in Table 2. The Bayes factor was estimated for models $M_0 - M_2$ against model M_3. That is, for model M_2 the density ratio $BF_{23} = \dfrac{p(y|M_2)}{p(y|M_3)}$ was estimated to be $BF_{23} \approx \exp(46)$. The Bayes factor thus strongly favors model M_2 over the larger model M_3. On the other hand, the Bayes factor clearly rejects the two other hypotheses. From these results can be concluded that perceptual organization of the items as well as the item construction rules used provide us with information about the difficulty and time intensity of the items. However, on the basis of this data set, we have to reject the hypothesis that the row-wise versus column-wise organization of the design rules affects item difficulty or time intensity. Moreover, the estimated R^2 statistics appear to be in line with the estimated Bayes factors. That is, the proportion explained variance increases from M_0 to M_3. However, the R^2 statistic improves only slightly from model M_2 to model M_3, which also gives us an indication that the row-wise and column-wise operations do not contribute much to difficulty and time intensity of the

Table 1
Estimated Covariance Components and Correlations Obtained for Model M_0

Variance component	EAP	*SD*	Correlation
Σ_P			
σ_θ^2	1.00		1.00
$\sigma_{\theta\zeta}$	−0.30	0.01	−0.61
σ_ζ^2	0.24	0.01	1.00
Σ_I			
σ_a^2	0.10	0.01	1.00
σ_{ab}	−0.17	0.03	−0.61
$\sigma_{a\phi}$	0.01	0.01	−0.02
$\sigma_{a\lambda}$	−0.05	0.01	−0.54
σ_b^2	0.71	0.08	1.00
$\sigma_{b\phi}$	−0.01	0.02	−0.05
$\sigma_{b\lambda}$	0.19	0.04	0.68
σ_ϕ^2	0.05	0.01	1.00
$\sigma_{\phi\lambda}$	0.01	0.01	0.15
σ_λ	0.10	0.01	1.00

Note. EAP = expected a posteriori estimate.

Table 2
Estimated Proportions of Explained Variance and Bayes Factors for the Models M_0–M_3

Model	R^2 (*b*)	R^2 (λ)	Bayes factor
M_0	.00	.00	exp(−7)
M_1	.13	.08	exp(−5)
M_2	.34	.37	exp(46)
M_3	.37	.39	exp(0)

items. Because row-wise versus column-wise operations concern an easy rotation (left–right vs. up–down) and not, for instance, the rotation of a complex 3D-object, it seems plausible to assume that such items are both equally difficult and time consuming. Regarding these results, we proceed assuming that model M_2 is the appropriate choice.

Estimated Effects

The estimated effects for model M_2 can be found in Table 3. From the HPD (see Box & Tiao, 1973, p. 123) regions of the estimated parameters, it can be seen that for item difficulty, the effects of integrated and embedded components and the effects for the identity, intersection, and unique addition rules significantly deviate from 0. Regarding item time intensity, for the effect for separated components and the effects of identity, unique addition, seriation, and variation of closed gestalts, 0 is not contained in their .95 HPD region. These results imply that applying these specific rules results in a deviation from the overall mean of item difficulty and/or time intensity. For example, when looking at the use of the identity rule to construct a new item, it can be expected that this item is less difficult and also less time consuming than the "mean item" in the test. Note that estimated effects of the other "nonsignificant rules" should not be ignored but interpreted as a set of rules leading to approximately equal RTs (or equal item difficulties, respectively).

Regarding the relative high correlation between item difficulty and item time intensity, it would be expected that the estimated regression effects on both parameters also show dependencies. Indeed, in Figure 8, a plot of the effects for item time intensity against the estimated effects for item difficulty shows a positive trend as well. As expected, items with embedded components appeared to be the most difficult and time intensive, compared with items with integrated or separated components. So, item construction rules with a positive effect on difficulty also lead to higher expected RTs. These findings are in line with the finding that more information requires more working memory capacity and additionally results in longer RTs (Embretson, 1998).

It is interesting to evaluate the estimated effects on time intensity on the time scale because this gives us results that are more intuitive to interpret. For model M_2, the estimates for the mean were $\gamma_0^{(\lambda)} = 4.01$ and for the variance $\hat{\sigma}_\lambda^2 = .07$. The inverse transformation to the time scale is then given by $\exp(\hat{\lambda} + \hat{\sigma}_\lambda^2/2) = \exp(4.01 + .07/2) = 57$ s. Now, take the effect of the identity rule, which is the rule with the strongest effect on the time intensity of an item. On the time scale, the difference in expected RTs between an "identity item" and a "mean item" would be

$$\exp(\hat{\gamma}_0 + \hat{\gamma}_1 + \hat{\sigma}^2/2) - \exp(\hat{\gamma}_0 + \hat{\sigma}^2/2) = \exp(4.01 - 0.25 + .04) - \exp(4.01 + .04) = -12 \text{ seconds.}$$

Similarly, consider the most extreme deviation from the mean time intensity, which can be obtained by constructing a hypothetical item using identity, seriation, and variation of closed gestalts. This leads to a difference in expected RT of −20 s. It is also interesting to evaluate the easiest and most difficult item in the subset. For the easiest item, $\hat{\lambda} = 2.99$, whereas for the most difficult item we found that $\hat{\lambda} = 3.52$. On the time scale, this leads to a difference in expected RTs of 14 s. Although this result might appear small when focusing on a single item, the effects can become large when longer tests (e.g., 20 or more items) are investigated.

Discussion

The aim of this article was to show that an IRT-based approach to RTs can contribute to the evaluation and testing

Table 3
Estimated Effects and .95 HPD Regions for Model M_2

	Difficulty (b)		Time intensity (λ)	
Effect	EAP	.95 HPD	EAP	.95 HPD
μ (intercept)	0.33	[−0.15, 0.59]	4.01	[3.88, 4.15]
γ_1 (separated)	−0.14	[−0.29, 0.02]	−0.02	[−0.07, 0.02]
γ_2 (integrated)	−0.18	[−0.31, −0.04]	−0.09	[−0.14, 0.05]
γ_3 (embedded)	0.33	[0.17, 0.47]	0.11	[0.06, 0.17]
γ_4 (identity)	−0.63	[−0.93, −0.34]	−0.26	[−0.36, −0.14]
γ_5 (addition)	0.22	[−0.05, 0.51]	0.00	[−0.10, 0.11]
γ_6 (substraction)	0.17	[−0.12, 0.45]	0.02	[−0.08, 0.12]
γ_7 (intersection)	0.47	[0.16, 0.77]	0.08	[−0.04, 0.18]
γ_8 (unique addition)	0.42	[0.13, 0.70]	0.17	[0.06, 0.27]
γ_9 (seriation)	−0.06	[−0.35, 0.22]	−0.10	[−0.21, 0.00]
γ_{10} (variation of closed gestalts)	−0.05	[−0.31, 0.21]	−0.11	[−0.21, −0.02]
γ_{11} (variation of open gestalts)	0.05	[−0.25, 0.33]	0.01	[−0.11, 0.10]

Note. EAP = expected a posteriori estimate; HPD = highest posterior density.

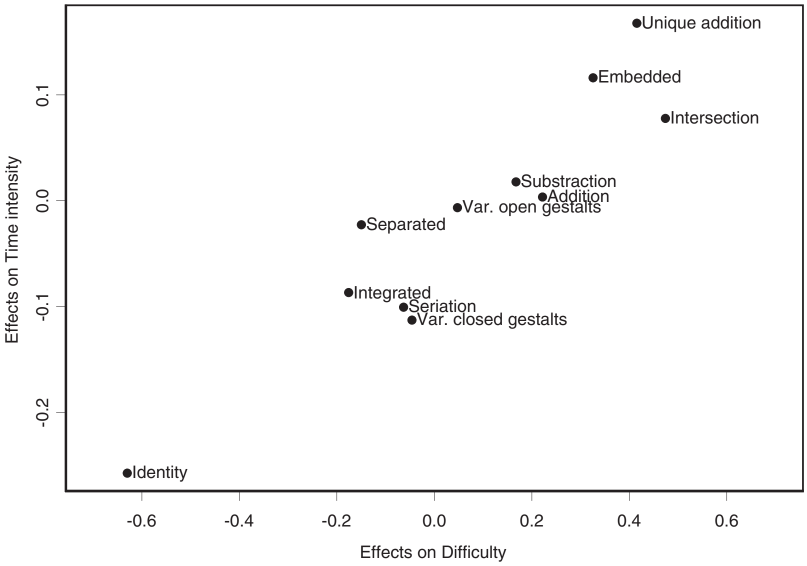

Figure 8. Expected a posteriori estimates (EAPs) of the effects of the item construction rules on item difficulty against their effects on time intensity for model M_2. Var. = variation.

of the cognitive theory underlying tests with a rule-based design. In the context of educational testing, it was argued that RTs may be described by both item and person parameters. By correcting for the speed of individual test takers, it is possible to reveal systematic differences between the items in a test, which were modeled by item discrimination and time intensity parameters, respectively.

The hierarchical modeling allowed us to study observed correlations between responses and RTs. Dependencies might arise because of a relationship between ability and speed of test takers, which was modeled at the second level by a population model for the test takers. A population model for the item parameters modeled similar possible dependencies but between the item characteristics of the two Level 1 models. The extension of the population model for the item parameters with a structural component enabled us to relate content specific information about items to the observed differences in their estimated difficulty and time intensity parameters.

The approach worked well for dissociating item difficulty and time intensity as a function of the underlying rule based design of the test. The current sample size, about 300 test takers per item and 12 items per test taker, appeared sufficient for the current analysis. However, this application could benefit from administrating more items per person because this would strengthen the linked design between the items, reduce the number of test takers required, and subsequently reduce the computational burden. That only a proportion of variance in difficulty and time intensity was explained can be attributed to the limited amount of information on the items that was included in the model. The analysis was restricted to the inclusion of effects for perceptual organization, the design rules, and row-wise and column-wise operations. Moreover, these effects were assumed to be independent and additive. As a result, the model lacked a description of possible interactions between rules and perceptual organization. Information regarding the complexity of figures was not included either. That is, items can be similar in structure (the rules used) but different in their symbols and figures. It is reasonable to suspect that difficulty and time intensity also depend on such item features.

In the example, both time intensity as well as item difficulty parameters were decomposed using the same design matrix. However, this condition is not mandatory. A strong cognitive theory might well propose different design matrices for item difficulty and time intensity parameters, respectively. Further, the model can be applied to existing tests with a reconstructed design matrix as well. Reconstruction of the design matrix can occur, for example, by consulting experts for the test who carefully inspect the items.

The model combines discrete and continuous data sources from the same test. This enhances the possibilities for the researcher but brings along computational difficulties as well. However, the Bayesian treatment of the model, using MCMC methods, is able to deal with these issues. Although computationally intensive, the MCMC approach has several advantages that reside in its flexibility. For instance, the user is not limited to preprogrammed (model fit) statistics but can easily compute his/her own statistics of interest from the samples of the MCMC chain. With the developed software, it is just as well possible to include continuous variables related to item content. Think of, for instance, regressing the item parameters on the number of words used to formulate them.

Extensions of the model can be implemented as an additional sampling step, without having to develop an entirely new algorithm. For example, we assumed unidimensionality in both ability and speed of the test takers. To relax this assumption, the model could be extended toward multidimensional IRT models (Adams, Wilson, & Wang, 1997; Embretson, 1997). Accordingly, one might assume that the latent speed of a person is multidimensional as well. Furthermore, it is possible that subgroups of test takers follow different solution strategies. For example, in a spatial rotation test, test takers might use a mental rotation strategy or an analytical strategy for detecting feature matches that do not require mental rotation (Mislevy & Verhelst, 1990). Mixture modeling approaches that deal with such cases can be found in Mislevy and Verhelst (1990) and Rost (1990).

As could have been expected, the results obtained here with respect to the decomposition of item difficulties were very similar to those reported by Hornke and Habon (1986). However, although a strong relationship between difficulty and time intensity was found for the cognitive operations involved, some interesting differences were observed. For example, whereas the rule intersection raised item difficulty, it did not affect time intensity to a significant degree. In contrast, variation of closed gestalts reduced the time intensity of an item, but it had no effect on its difficulty. This suggests that item difficulty and time intensity can be manipulated independently of each other: Cognitive operations that raise item difficulty need not affect time intensity and vice versa. From a theoretical point of view, a subject needs a higher latent ability to solve an item including intersection, that is, the solution probability is affected both by person and item characteristics. In contrast, the RT to solve such an item will exclusively hinge on the subject's speed and not on the presence of intersection. These results suggest that different operationalizations of cognitive processes (accuracy scores, RTs) that are described as capacity-demanding do not necessarily lead to identical results, highlighting the importance of jointly analyzing accuracy and RT data.

Other practical implications of the proposed method relate to item and test construction. A test is developed to measure a specific construct—for instance, mathematical ability. For construction of a test, item selection is primarily based on the information function of the items. The information functions describe how well an item measures the ability of interest and how well it covers the ability range. Using the item information functions allows the optimal design of item subsets or tests to measure the ability of test takers up to a certain accuracy. RT information and, more specifically, the time intensity of an item now provide a second item selection criterion. It is possible to select a group of items so as to minimize the time intensity of the complete set. This would not only minimize the number of items needed to measure ability but minimize test length with respect to time as well. Such applications could be interesting for computerized adaptive testing (CAT). In CAT, an adaptive algorithm is used that selects a new item on the basis of the ability of a test taker estimated from the previously presented items. These CAT algorithms minimize the number of items needed to measure ability up to a specified accuracy. A second optimization of the algorithm would now be possible with respect to total test time.

However, if item writing can be based on cognitive theory, test construction can be done in a more structured way. With the methods proposed in this article, a thorough assessment of how cognitive operations affect task difficulty as well as time intensity becomes feasible. Revealing the differences in the time intensities of tasks provides more detailed insight in their cognitive demands. Thereby, RTs provide tools to evaluate a cognitive theory more thoroughly. This can give a better understanding of the relationships between item characteristics and item content.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15,* 129–137.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8,* 205–238.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review, 97,* 404–431.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260–278.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypothesis on normal location parameters. *The Annals of Mathematical Statistics, 42,* 204–223.

Dzhafarov, E. N., & Schweickert, R. (1995). Decompositions of response times: An almost general theory. *Journal of Mathematical Psychology, 39,* 285–314.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128,* 309–331.

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

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269–286.

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.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics, 48,* 241–251.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 23,* 249–263.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension items: The feasibility of verbal item generation. *Journal of Educational Measurement, 42,* 351–373.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25,* 21–35.

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,* 170–173.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523–1543.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 159–178). Mahwah, NJ: Erlbaum.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369–380.

Hornke, L. F., & Wilding, U. (1997). *Konstanz von Itemparametern bei parallelen Itembanken* [Constancy of item parameters in parallel item banks] (Tech. Rep.). RWTH Aachen University.

Irvine, S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–34). Mahwah, NJ: Erlbaum.

Jacobs, P. I., & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement, 32,* 235–248.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. DeBoeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (in press). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika.* (DOI: 10.1007/S11336–008–9075-Y).

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10,* 477–493.

Lanza, S. T., Collins, L. M., Schafer, J. L., & Flaherty, B. P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods, 10,* 84–100.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications.* New York: Cambridge University Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing items.* Hillsdale, NJ: Erlbaum.

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

Luce, D. R. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58,* 445–469.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12,* 252–284.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30,* 41–70.

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

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

Raven, J. C. (1962). *Advanced progressive matrices, Set II.* London: H. K. Lewis.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26,* 271–285.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods.* New York: Springer.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General, 137,* 370–389.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68,* 589–606.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response models. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theories and applications* (pp. 205–240). Cambridge, England: Cambridge University Press.

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

Schmiedek, F., Oberauer, K., Wilhelm, O., Süss, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136,* 414–429.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method for measuring speededness. *Journal of Educational Measurement, 34,* 213–232.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42,* 375–394.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321.

Smith, B. J. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software, 21,* 1–37.

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 506–526.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84,* 353–378.

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

van Breukelen, G. J. P. (1995). Psychometric and information processing properties of selected response time models. *Psychometrika, 60,* 95–113.

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70,* 359–376.

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

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

van der Linden, W. J., & Guo, F. (in press). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika.*

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23,* 195–210.

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

van Zandt, T. (2002). Analysis of response time distributions. In H. Pashler & J. Wixted (Eds.), *Steven's handbook of experimental* *psychology, Vol. 4: Methodology in experimental psychology* (3rd ed., pp. 461–516). New York: Wiley.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association, 90,* 614–618.

Verhelst, N., Verstralen, H., & Jansen, M. (1997). Models 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, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29,* 323–339.

Ward, J., & Fitzpatrick, F. (1973). Characteristics of matrices items. *Perceptual and Motor Skills, 36,* 987–993.

Wright, D. E., & Dennis, I. (1999). Exploiting the speed-accuracy trade-off. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 231–248). Washington, DC: American Psychological Association.

Appendix

Estimation

The model can be identified by setting the means of the person parameters (θ, ζ) to zero, so that $\mu_P = 0$ and restricting the variance of θ to 1, so $\sigma_\theta^2 = 1$. Fox, Klein Entink, and van der Linden (2007) provided a Gibbs sampling solution in which these identifying restrictions are directly included into the prior distributions. The same authors described a straightforward and efficient Gibbs sampling scheme for the hierarchical model, except for a step for sampling the design effects γ of the item parameters. Therefore, the sampling steps are described below, but only the sampling step for the design effects is given explicitly here.

Step 1. Sample augmented response data according to Equation 12.

Step 2. Draw (θ_i, ζ_i) simultaneously from $\theta_i, \zeta_i | z_i, t_i, a, b, \phi, \lambda, \mu_P, \Sigma_P$.

Step 3. Draw (μ_P, Σ_P) from $\mu_P, \Sigma_P | \theta, \zeta, \Sigma_{P0}, \mu_{P0}$, where Σ_{P0}, μ_{P0} denote the hyperprior parameters.

Step 4. Draw (a_k, b_k, ϕ_k, λ_k) from $a_k, b_k, \phi_k, \lambda_k | z_k, t_k, \theta, \zeta, \gamma, \Sigma_I$.

Step 5. Draw Σ_I from $\Sigma_I | a, b, \phi, \lambda, \gamma, \Sigma_{I0}$.

Step 6. Draw $\gamma | A, a, b, \phi, \lambda, \gamma_0, \Sigma_{\gamma 0}$. This step is specified below.

Let $\gamma = \text{vec}\,(\gamma^{(a)}, \gamma^{(b)}, \gamma^{(\phi)}, \gamma^{(\lambda)})$ and $\Omega_I = (a, b, \phi, \lambda)$, where vec denotes the operation of vectorizing a matrix. Furthermore, let $A_I = (I_4 \otimes A)$. This enables rewriting Equations 21–24 to

$$\text{vec}(\Omega_I) = A_I\gamma + \text{vec}(e), \tag{A1}$$

where $\text{vec}(e) \sim N(0, \Sigma_I \otimes I_K)$. Next, a conjugate normal prior is chosen for γ:

$$\gamma \sim N(\gamma_0, \textstyle\sum_{\gamma 0}). \tag{A2}$$

Subsequently, it follows that the posterior distribution is again normal:

$$\gamma | A_I, \Omega_I, \textstyle\sum_I, \gamma_0, \sum_{\gamma 0} \sim N\left(\dfrac{\hat{\sum}_\gamma^{-1}\hat{\gamma} + \sum_{\gamma 0}^{-1}\gamma_0}{\hat{\sum}_\gamma^{-1} + \sum_{\gamma 0}^{-1}}, \left(\hat{\textstyle\sum}_\gamma^{-1} + \textstyle\sum_{\gamma 0}^{-1}\right)^{-1}\right), \tag{A3}$$

where $\hat{\sum}_\gamma$ and $\hat{\gamma}$ are the common least squares estimates, which can be derived from Equation A1. For Gibbs sampling of the other model parameters, see Fox et al. (2007).

Received August 17, 2007
Revision received November 7, 2008
Accepted November 13, 2008 ■