Rinke Klein Entink
University of Twente, Enschede

Statistical Models for Responses and Response Times

Samenstelling promotiecommisie

Voorzitter	Prof. Dr. H. W. A. M. Coonen
Promotor	Prof. Dr. W. J. van der Linden
Assistent-promotor	Dr. Ir. G. J. A. Fox
Leden	Prof. Dr. Ir. T. J. H. M. Eggen
	Prof. Dr. C. A. W. Glas
	Prof. Dr. H. L. J. van der Maas
	Prof. Dr. R. R. Meijer
	Prof. Dr. F. Tuerlinckx

Klein Entink, Rinke Hermen
Statistical Models for Responses and Response Times
Proefschrift Universiteit Twente, Enschede. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN: 978-90-365-2755-2
Druk: PrintPartners Ipskamp B.V., Enschede
Cover designed by Annemieke Vliegen

STATISTICAL MODELS FOR RESPONSES AND RESPONSE TIMES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 22 januari 2009 om 15.00 uur

door

Rinke Hermen Klein Entink

geboren op 9 december 1980
te Heiloo

”Ik geloof dat ik voorlopig iets geestelijks ga doen, voor mijn gedachtenleven. Maar wat?”

Heer Bommel - De Viridiaandinges

Preface

After a few years of work and fun, this is it: my thesis about models for responses and response times, on which I worked at the Research Methodology, Measurement and Data Analysis (OMD) group of the University of Twente. I would like to thank everyone at OMD for the good times I had working there. Some people I would like to thank in particular: Anke Weekers for the enjoyable years that we shared a room and for assisting me as paranimf during my defence. Gerrie Croonen for coming over from Vienna to assist me as paranimf. Annemieke Vliegen, who designed the cover of my thesis. Jörg-Tobias Kuhn for the fruitful collaboration on Chapter 3. Wim van der Linden for his advice, pleasant cooperation and helping me to look over the borders of Twente and finally Jean-Paul Fox for the pleasant cooperation and his enthusiasm that greatly encouraged me.

Rinke Klein Entink

Contents

1 Introduction

Inferences about test takers in educational testing have been primarily based on their responses to the items. Information from the time needed to complete an item has been largely ignored because the recording of response times was unpractical with paper-and-pencil tests. Nowadays, computer based testing makes the collection of response times straightforward. However, the question what to do with this additional data source remains.

Of course, we can ignore the response times and base our judgements about test takers or the quality of a test on the response patterns only. Consider Jan and Paul, two students who both want the job of teaching assistant for our 2nd year statistics course. To make sure the best of the two is hired, the teacher gives them a two-item test on regression analysis, both multiple-choice items. As it happens, Jan has both items correct but Paul gives an incorrect answer to item 1. When their responses to these items are all we know about them, we conclude that Jan has the best knowledge about regression analysis. But when looking at the time they took to complete the test, we see that Paul answered the items in 35 and 46 seconds. Jan was quick and answered the items in 5 and 6 seconds, respectively. That, however, appears to be a bit too quick, since reading the item and looking up the appropriate information from the table should take more time, casting doubt on Jan's results. Upon asking, Jan is so honest to admit he just guessed both items. Jan picked the correct answers by sheer luck and it was Paul who answered seriously thus we hire him for the job. Had we looked at the responses only, a valuable source of information from the response times would have been missed.

For sure, the above example is just a simple illustration and a more advanced approach is needed to evaluate the advantages of response time information. This thesis will discuss statistical methods for the joint analysis of responses and response times on test items.

1.1 Measuring Ability and Speed

The field of psychometrics is concerned with the theory and methods of educational and psychological measurement (psycho; relating to the mind, metric; refers to measurement). Think of, for instance, the measurement of reading ability or a personality trait like extraversion. However, someone's knowledge about the fall of the Roman empire is not directly observable, but only trough its manifestations. Estimates of someone's ability level, therefore, are usually obtained by administrating tests or questionnaires.

A body of theory is then required to make inferences about the unobserved abilities of the test takers from their observed responses to the items. Item response theory (IRT) has been developed for exactly that. IRT is a theory that describes the use of mathematical models for measuring abilities or attitudes from test data. IRT models describe the probability of a correct response of a test taker to an item as a function of his/her ability and the characteristics of the item. For example, unidimensional IRT models, like the 2-parameter model, have an ability parameter for every test taker while for each item a difficulty and discrimination parameter are present (Lord & Novick, 1968; Embretson & Reise, 2000). The normal-ogive formulation of the 2-parameter IRT model is given by

$$E(Y) = \Phi(a\theta - b), \tag{1.1}$$

where $E(Y)$ denotes the probability of giving a correct response ($Y = 1$) to the item, given ability level θ, and $\Phi(\cdot)$ denotes the normal cumulative distribution function. The item characteristics are described by the discrimination parameter a and the difficulty level of the item b. The basic idea is that a higher ability leads to a higher probability of giving a correct response.

Following that same basic idea an equivalent can be formulated for a response time model: a higher speed of working leads to a lower expected response time. That is, a person specific parameter is incorporated into a model for response times that accounts for individual differences between test takers. Such a parameter can be found in response time models presented by, for instance, Scheiblechner (1979); Thissen (1983); Maris (1993); Schnipke and Scrams (1997) and van der Linden (2006). Where the test is supposed to measure ability as the underlying construct for the responses, it can be assumed to measure speed as the underlying construct for the response times as well.

However, time differences between items should be included in the model, too. Items in a test usually vary in their difficulty, but it is reasonable that they vary in time intensity as well. Think, for instance, of an item with a text passage where a missing word has to be filled in versus an item where one has to summarize the text passage in 50-70 words. These differences in time intensity between items are therefore modeled by an item parameter, λ. This parameter reflects the time needed to solve an item and can be seen as the analogue of the difficulty parameter b. However, it is not necessarily so that a more time intensive item is also more difficult.

Response times are skewed to the right because they are restricted to be greater than zero. The log-transform has been applied often to account for this skewness (Schnipke & Scrams, 1997; van der Linden, 2006). Assuming that the log-response time T follows a linear model, then $E(T) = -\zeta + \lambda$, where ζ denotes the level of speed of the test taker. This model was suggested earlier by van der Linden (2006). However, a question might show less variability around its mean time intensity λ than predicted by $\boldsymbol{\zeta}$. Such an effect can be considered as the discriminative power of an item and therefore a time discrimination parameter ϕ is introduced. This parameter controls the decrease in expected response time on an item for a one step increase in speed of a test taker. It is the analogue of the discrimination parameter a in Equation 1.1. Subsequently, the log-response time T follows a linear model according to:

$$E(T) = -\phi\zeta + \lambda. \tag{1.2}$$

Comparing (1.2) with (1.1), it can be seen that the minus sign is now in front of the person parameter, reflecting on the one hand that a higher speed of working leads to lower response times and on the other hand that more time intensive items lead to higher response times.

An illustration of the effect on time intensity on the expected response time is given in Figure 1.1. In this figure, both Item Characteristic Curves (ICC) for the IRT model (left) and Response Time Characteristic Curves (RTCC) (right) are plotted against the latent trait. The ICCs illustrate that the probability of a giving a correct response increases with ability. Vice versa, the RTCCs show that the expected response time decreases with 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. The above RTCC curve reflects the most time intensive item ($\lambda = 4$). Given the level of speed, the expected response times are higher for this item than for the item with $\lambda = 3$.

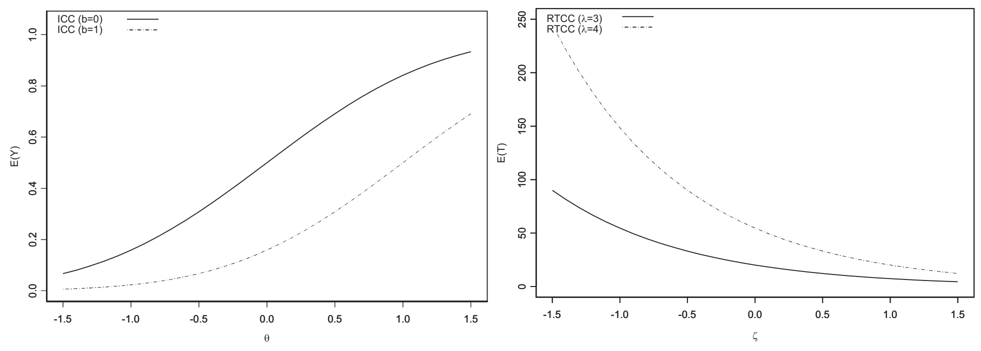


Fig. 1.1. ICC (left) and RTCC (right) curves for two items with different time intensity and difficulty but equal discrimination parameters.

The effect of item discrimination on the ICCs and RTCCs is illustrated in Figure 1.2. 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.

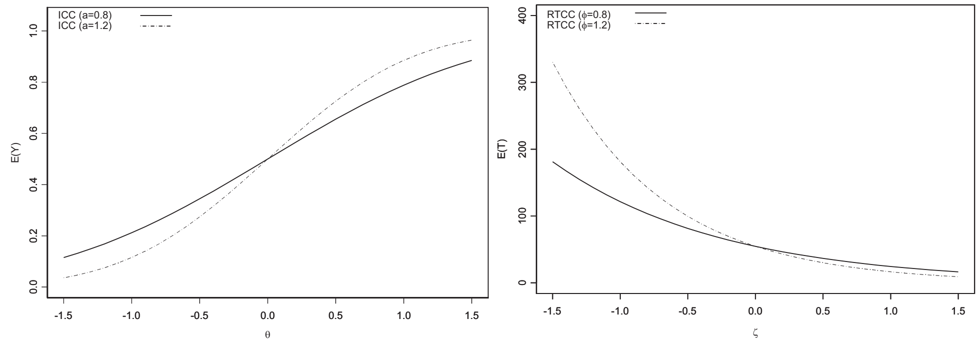


Fig. 1.2. ICC and RTCC curves for two items with different discrimination parameters, where $b = 0$ and $\lambda = 4$

The RTCC thus shows the decrease in expected response time as function of speed. One way to look at this curve is that the RTCC represents the expected response times for a population of test takers with different speed levels. Another viewpoint could be the shift in expected response time if a test taker chooses to work faster. It can be seen that, the faster a test taker works, the lower the difference in expected RTs on the items is. This is the equivalent of a high ability candidate who has almost equal probabilities of a correct response on two low difficulty items.

These measurement models for ability and speed are the core of the methods presented in this thesis. In the next section will be discussed how these univariate models generalize to a multivariate model that forms the starting point for the following chapters.

1.2 Modeling Covariation Between Responses and Response Times

In multivariate statistical analyses the interest is usually focussed on the dependencies between outcome variables. Advantages of joint inferences over univariate inferences are that they are often more informative, leading to a better understanding of the subject, and possible gains in efficiency of the analyses with respect to parameter estimation. Possible interesting questions are what the joint analysis of speed and ability tells us about test taker behavior (remember Jan and Paul) or how response time information can enhance test development. To answer these kind of questions, the possible dependencies between the responses and response times have to be modeled.

A common view in IRT modeling is to see a person as a random draw from a population of test takers (Holland, 1990). By specifying a population distribution for the ability parameters of the test takers, this naturally leads to a hierarchical model, where the responses to the items are seen as nested within subjects. That is, the ability parameter is a random person effect that models the heterogeneity between test takers. Thereby, ability is assumed to explain all associations between the responses on different items: Responses on two items in a test are assumed to be *locally independent* given ability. The local independence assumption has a long tradition in IRT (e.g. Lord, 1980; Holland & Rosenbaum, 1986). Further, the population distribution for ability is usually assumed to be normal.

Analogously, the response times on the items can be seen as nested within test takers, too. Again this leads to a hierarchical model, where the speed parameter is the random subject effect, drawn from a population of test takers. Thereby, these random effects model the heterogeneity in the observed response times between test takers. Therefore, conditional on the random speed parameters, there should be no covariation left between the response times on different items. In other words, conditional independence is assumed in the response time model as well. In both measurement models, the conditional (local) independence assumption between the observations on different items implies that the random effects should be constant over the items. That is, a test taker is assumed to work at a constant level of ability and a constant level of speed during a test.

The ability and speed parameters are, of course, nested within the test takers. Instead of formulating two separate population models (one for ability and one for speed), it is attractive to allow for covariation between the two traits. Assuming a multivariate normal distribution for the population model enables the modeling of dependencies between the responses and the response times that can be attributed to the test takers. The idea of modeling dependencies between multivariate outcomes via the random effects structure has been used earlier. Examples can be found in Snijders and Bosker (1999); Gueorguieva (2001) and Liu and Hedeker (2006) and a recent overview of this topic was given in McCulloch (2008). It was van der Linden (2007) who proposed this generalization to a hierarchical framework to study responses and response times on test items simultaneously.

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 an item these parameters should capture all possible covariation between the responses and response times. That is, the responses and response times on an item are assumed to be conditionally independent given the levels of ability and speed of the test takers.

The other possible source of covariation between the responses and response times results from the items in the test. For instance, this happens when more difficult items tend to be more time intensive. The hierarchical model is readily extended by assuming a population model for the item parameters that is similar to the population model for ability and speed.

Thereby, the hierarchical framework is obtained as proposed by van der Linden (2007) that forms the starting point for this thesis. This framework accounts for two possible sources of covariation between the responses and response times: the test takers and the items. However, this model is rather descriptive than explanatory. In this thesis, structural models will be proposed that address underlying causes of possible dependencies between the responses and response times, pertaining to both the person and the item side. The development of these models also requires the development of the necessary estimation methods and ways to test hypotheses. In the next section will be described how all these topics are divided over the remainder of this thesis.

1.3 Outline

Assessing the differences in ability between test takers or groups of test takers is one of the main goals of testing. In Chapter 2 (published as Klein Entink, Fox, & van der Linden, in press), the focus is on a better understanding of these differences in ability and speed. A multilevel model is developed that allows the incorporation of covariates for explaining differences between individuals and groups of test takers. Bayesian Markov Chain Monte Carlo methods are presented to estimate all model parameters concurrently. Model specific test statistics are derived to evaluate hypotheses about covariates and group differences relating to ability and speed.

Chapter 3 (Klein Entink, Kuhn, Hornke, & Fox, in press) is a study to the relationships between item characteristics and item content for tests with a cognitive, rule-based design. A structural model on the side of the item parameters allows the determination of both time intensity and difficulty of design rules in the test. This allows a better understanding of the relationships between item characteristics and item content. The application of the model is illustrated using a large-scale investigation of figural reasoning ability.

The use of Box-Cox transformations to obtain different distributional shapes for response time models were considered in Chapter 4 (Klein Entink, van der Linden, & Fox, in press). Box-Cox transformations aim to transform skewed data (like response times) to normality, which has great advantages because of its conjugacy with the larger modeling framework. A transformation-invariant implementation of the Deviance Information Criterium (DIC) is developed that allows for comparing model fit between models with different transformation parameters. The performance of a Box-Cox normal model is investigated using simulation studies and a real data example. Showing an enhanced description of the shape of the response time distributions, its application in an educational measurement context is discussed.

Detailed simulation studies are performed in Chapter 5 (van der Linden, Klein Entink, & Fox, in press) to show how additional response time information might affect the estimation of IRT model parameters. It is argued that response times can improve IRT parameter estimates with respect to bias as well as accuracy of the estimates.

Chapter 6 explores the generalized mixed model framework for the joint analysis of responses and response times. It is shown that some analyses can be performed within the class of multivariate generalized linear mixed models, thereby providing fast and easy to use statistical methods for inferences with commercial available software. However, the modeling possibilities are restricted to Rasch-kind models, not allowing for discrimination parameters (or guessing) in the two measurement models.

This thesis concludes with a discussion and some suggestions for further research.

2

A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers

Summary. Response times on test items are easily collected in modern computerized testing. When collecting both (binary) responses and (continuous) response times on test items, it is possible to measure the accuracy and speed of test takers. To study the relationships between these two constructs, the model is extended with a multivariate multilevel regression structure which allows the incorporation of covariates to explain the variance in speed and accuracy between individuals and groups of test takers. A Bayesian approach with Markov chain Monte Carlo (MCMC) computation enables straightforward estimation of all model parameters. Model-specific implementations of a Bayes factor (BF) and deviance information criterium (DIC) for model selection are proposed which are easily calculated as byproducts of the MCMC computation. Both results from simulation studies and real-data examples are given to illustrate several novel analyses possible with this modeling framework.

2.1 Introduction

Response times (RTs) on test items can be a valuable source of information on test takers and test items, for example, when analyzing the speededness of the test, calibrating test items, detecting cheating, and designing a test (e.g., Bridgeman & Cline, 2004; Wise & Kong, 2005; van der Linden & Guo, in press; van der Linden, Breithaupt, Chuah, & Zang, 2007; van der Linden, 2008). With the introduction of computerized testing, their collection has become straightforward.

It is important to make a distinction between the RTs on the test items and the speed at which a test taker operates throughout the test, especially when each person takes a different selection of items, as in adaptive testing. For two different test takers, it is possible to operate at the same speed but produce entirely different RTs because the problems formulated in their items require different amounts of information to be processed, different problem-solving strategies, etc. Models for RTs should therefore have separate parameters for the test takers' speed and the time intensities of the items.

Another potential confounding relationship is that between speed and accuracy. It is well known that, on complex tasks, these two are different constructs (see, for instance Kennedy, 1930; Schnipke & Scrams, 2002). Tate (1948) was one of the first to examine the relationship between speed and accuracy on different tests. He concluded that, for a controlled level of accuracy, each test takers worked at a constant speed. Furthermore, test takers working at a certain speed do not necessarily demonstrate the same accuracy.

Some of these findings can be explained by the well-known speed-accuracy trade-off (e.g., Luce, 1986). The trade-off reflects the fact that speed and accuracy are main determinants of each other. Also, they are negatively related. When a test taker chooses to increase his speed, then his accuracy decreases. But once his speed is fixed, his accuracy remains constant. Observe that this trade-off involves a within-person constraint only; it does not enable us to predict the speed or accuracy of one person from another taking the same test. In order to model the relationship between speed and accuracy adequately, we therefore need a model with different levels. This multilevel perspective has not yet been dominant in the psychometric literature on RT modeling. Instead, attempts have been made to integrate speed parameters or RTs into traditional single-level response models (Verhelst, Verstralen, & Jansen, 1997) or, reversely, response parameters into RT models (Thissen, 1983). However, a hierarchical framework for modeling responses and RTs was introduced in van der Linden (2007). The framework has separate first-level models for the responses and RTs. For the response model, a traditional item-response theory (IRT) model was chosen. For the RTs, a lognormal model with separate person and item parameters was adopted, which has nice statistical properties and fitted actual response time data well (van der Linden, 2006). At the second level, the joint distributions of the person and item parameters in the two first-level models were modeled separately.

Observe that, because the framework in this chapter does not model a speed-accuracy tradeoff, it can be used just as well to analyse responses and RTs to instruments for non-cognitive domains, such as attitudes scales or personality questionnaires.

Because the first-level parameters capture all systematic variation in the RTs, they can be assumed to be conditionally independent given the speed parameter. Likewise, the responses and RTs are assumed to be conditionally independent given the ability and speed parameter. Such assumptions of conditional independence are quite common in hierarchical modeling but may seem counterintuitive in the current context, where the speed-accuracy trade-off is often taken to suggest that the frequency of the correct responses increases if the RTs go up. However, this confusion arises when the earlier distinction between speed and RT is overlooked. The trade-off controls the choice of the levels of speed and accuracy by the individual test taker whereas the conditional independence assumptions address what happens with his response and RT distributions after the levels of speed and accuracy have been fixed.

Besides being a nice implementation of the assumptions of local independence for RTs and responses, this framework allows for the incorporation of explanatory variables to identify factors that explain variation in speed and accuracy between individuals who may be nested within groups. The current chapter addresses this possibility; its goal is to extend the framework with a third level with regression and group effects and to make this result statistically tractable. The result is a multivariate multilevel model for mixed response variables (binary responses and continuous RTs). At the person level, just as in the original framework, it allows us to measure both accuracy and speed. Test takers can therefore be compared to each other with respect to these measures. But at the higher levels the extended framework also allows us to identify covariates and group memberships that explain the measures as well as their relationships. Also, the item parameters are allowed to correlate.

Analysis of the extended model is performed in a fully Bayesian way. The motivation for the Bayesian treatment is its capability of handling complex models with many parameters that take all possible sources of variation into account. A new Gibbs sampling procedure (Geman & Geman, 1984; Gelfand & Smith, 1990) was developed which applies not only to the current framework but to the entire class of nonlinear multivariate multilevel models for mixed responses with balanced and unbalanced designs. All parameters can be estimated simultaneously without the need to fine-tune any parameters to guarantee convergence, for instance, as in a Metropolis-Hastings (MH) algorithm. Proper prior distributions can be specified that can be used both to incorporate a set of identifying restrictions for the model and to reflect the researcher's ideas about the parameter values and uncertainties. The estimation method can also handle incomplete designs with data missing at random.

A model-specific implementation of the Bayes factor (Kass & Raftery, 1995) and the deviance information criterion (DIC) (Spiegelhalter, Best, Carlin, & Linde, 2002) is given, which can be used (i) to test specific assumptions about the distribution of speed and accuracy in a population of test takers and (ii) to iteratively build a structural multivariate multilevel component for the latent person parameters with fixed and random effects. Both statistics can be computed as by-products of the proposed Gibbs sampler. The DIC requires an analytic expression of the deviance associated with the likelihood of interest. Such an expression is offered for the multivariate multilevel model given the complete data, which includes augmented continuous data given the binary responses (Albert, 1992), integrating out both random person parameters and other random regression effects at the level of groups of respondents. The posterior expectation of this complete DIC is taken over the augmented data using the output from the MCMC algorithm. Properties of the DIC, as well as the Bayes factor, were analyzed in a study with simulated data.

In the next sections, we describe the entire model, specify the prior distributions, discuss the Gibbs sampler, and show how to apply the Bayes factor and the DIC to the current model. Then, in a simulation study, the performance of the Gibbs sampler is addressed, whereby our interest is particularly in estimating the parameters in the structural component of the model. In a second simulation study, the relationships between the person parameters and the tests of multivariate hypotheses using the Bayes factor and the DIC are explored. Finally, the results from real-data examples are given and a few suggestions for extensions of the model are presented.

2.2 A Multivariate Multilevel Model

Various sources contribute to the variation between responses and RTs on test items. The total variation can be partitioned into variation due to (i) the sampling of persons and items, (ii) the nesting of responses within persons and items, and (iii) the nesting of persons within groups.

Two measurement models describe the distributions of the binary responses and continuous RTs at level 1 of the framework. At level 2, two correlation structures are posited to allow for the dependencies between the level 1 model parameters. First, the person parameters for ability and speed, denoted as $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, respectively, are modeled to have a multivariate normal regression on covariates $\mathbf{x}$, while group differences between these parameters are explained as a function of group-level covariates $\mathbf{w}$ at a third level. By specifying a higher-level regression structure for these random person parameters, it becomes possible to partition their total variance into within-group and between-group components. As a result, we are able to draw inferences about the person parameters for different groups simultaneously. Second, a correlation structure for the item parameters in the two measurement models is specified.

The model can be used for various analyses. First, the analysis might focus on the item parameters; more specifically, the relationships between the characteristics of the items in the domain covered by the test. For example, we may want to know the correlation between the time intensity and difficulty parameters of the items. Second, the analysis could be about the structural relationships between explanatory information at the individual and/or group levels and the test takers' ability and speed. For example, the variance components of the structural model help us to explore the partitioning of the variance of the speed parameters across the different levels of analysis. Third, the interest might be in the random effects in the model, e.g., to identify atypical individuals or groups with respect to their ability or speed.

Level-1 Measurement Models for the Responses and RTs

The probability of person $i = 1, \ldots, n_j$ in group $j = 1, \ldots, J$ answering item $k = 1, \ldots, K$ correctly ($y_{ijk} = 1$) is assumed to follow the three-parameter normal ogive model:

$$P(y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k, c_k) = c_k + (1 - c_k)\Phi(a_k\theta_{ij} - b_k), \tag{2.1}$$

where $\Phi(.)$ denotes the normal distribution function, θ_{ij} the ability parameter of test taker ij, and a_k, b_k and c_k the discrimination, difficulty and guessing parameters of item k, respectively.

Typically, as the result of a natural lower bound at zero, RT distributions are skewed to the right. A family that describes this characteristic well is the log-normal distribution (van der Linden, 2006; Schnipke & Scrams, 1997). Let t_{ijk} denote the log-response time of person i in group j on item k. We apply a normal model for t_{ijk}, with a mean depending on the speed at which the person works, denoted as ζ_{ij}, and the time intensity of the item, λ_k. A higher λ_k represents an item that is expected to consume more time. On the other hand, a higher ζ_{ij} means that the person works faster and a lower RT is expected. A parameter ϕ_k is introduced, which can be interpreted as a time discrimination parameter.

The response-time model at level 1 is given by:

$$t_{ijk} = -\phi_k \zeta_{ij} + \lambda_k + \epsilon_{\zeta_{ijk}}, \tag{2.2}$$

where $\epsilon_{\zeta_{ijk}} \sim N(0, \tau_k^2)$. Notice that the interpretation of the model parameters in (2.2) results in a different location of the minus sign compared to the IRT model. Also, there is a correspondence of the RT model with IRT models for continuous responses; for the latter, see, for instance, Mellenbergh (1994) and Shi and Lee (1998).

Multivariate Two-Level Model for the Person Parameters

The interest is in the relationships between the person parameters and the effects of potential explanatory variables. For convenience, we use the same set of explanatory variables for both types of person parameters; the generalization to the case of different variables is straightforward. Let $\mathbf{x}_j$ denote a known $n_j \times Q$ covariate matrix (with ones in the first column for the intercept) and $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{2j})$ a $Q \times 2$ matrix of regression coefficients for group $j = 1, ..., J$. The coefficients are treated as random but they can be restricted to be common to all groups, leading to the case of one fixed effect.

The regression of the two sets of person parameters at the individual level is defined by:

$$\theta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}_{1j} + e_{\theta_{ij}} \tag{2.3}$$

$$\zeta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}_{2j} + e_{\zeta_{ij}}. \tag{2.4}$$

The two sets of regression equations are allowed to have correlated error terms; $(e_{\theta_{ij}}, e_{\zeta_{ij}})$ is taken to be bivariate normal with zero means and covariance matrix $\boldsymbol{\Sigma}_P$:

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}. \tag{2.5}$$

It is straightforward to extend the random effects model to explain variance in the $\boldsymbol{\beta}$'s by group level covariates (Snijders & Bosker, 1999). For instance, test takers can be grouped according to their social economic background or because they are nested within different schools. Although different covariates can be included for the Q intercept and slope parameters, for convenience, it will be assumed that the same covariate matrix is used for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The covariates for the Q parameters of group j are contained in a matrix $\mathbf{w}_j$ of dimension $Q \times S$. That is, in total there are S covariates for each group, including the ones for the intercepts. The random effects $\boldsymbol{\beta}_{1j}$ and $\boldsymbol{\beta}_{2j}$ are then modeled as:

$$\boldsymbol{\beta}_{1j} = \mathbf{w}_j \gamma_1 + \mathbf{u}_{1j} \tag{2.6}$$

$$\boldsymbol{\beta}_{2j} = \mathbf{w}_j \gamma_2 + \mathbf{u}_{2j}, \tag{2.7}$$

where $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are the vectors of regression coefficients of length S. The group-level error terms, $(\mathbf{u}_{1j}, \mathbf{u}_{2j})$, are assumed to be multivariate normally distributed with means zero and covariance matrix $\mathbf{V}$. More stable parameter estimates can be obtained by restricting this covariance matrix to be block-diagonal with diagonal matrices $\mathbf{V}_1$ and $\mathbf{V}_2$, each of dimension $Q \times Q$. In this case, the random effects in the regression of $\boldsymbol{\theta}$ on $\mathbf{x}$ are allowed to correlate but they are independent of those in the regression of $\boldsymbol{\zeta}$ on $\mathbf{x}$. This choice will be made throughout this chapter. Note that when $(\mathbf{x}\boldsymbol{\beta}_1, \mathbf{x}\boldsymbol{\beta}_2) = (\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\zeta) = \boldsymbol{\mu}_P$, the model as proposed by van der Linden (2007) is obtained as a special case.

Let $\boldsymbol{\theta}_j$ and $\boldsymbol{\zeta}_j$ denote the vectors of length n_j of the person parameters of group j. The entire structural multivariate multilevel model can now be presented as:

$$\operatorname{vec}\bigl(\boldsymbol{\theta}_j, \boldsymbol{\zeta}_j\bigr) = \bigl(\mathbf{I}_2 \otimes \mathbf{x}_j^t\bigr)\operatorname{vec}\bigl(\boldsymbol{\beta}_j\bigr) + \operatorname{vec}\bigl(\mathbf{e}_{\theta_j}, \mathbf{e}_{\zeta_j}\bigr) \tag{2.8}$$

$$\operatorname{vec}\bigl(\boldsymbol{\beta}_j\bigr) = \bigl(\mathbf{I}_2 \otimes \mathbf{w}_j\bigr)\operatorname{vec}\bigl(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\bigr) + \operatorname{vec}\bigl(\mathbf{u}_{1j}, \mathbf{u}_{2j}\bigr), \tag{2.9}$$

where *vec* denotes the operation of vectorizing a matrix. We refer to these two models as level 2 and 3 models, respectively. Marginalizing over the random regression effects in (2.8) and (2.9), the distribution of $\operatorname{vec}\bigl(\boldsymbol{\theta}_j, \boldsymbol{\zeta}_j\bigr)$ becomes

$$\operatorname{vec}\bigl(\boldsymbol{\theta}_j, \boldsymbol{\zeta}_j\bigr) \sim N\Bigl(\bigl(\mathbf{I}_2 \otimes \mathbf{x}_j \mathbf{w}_j\bigr)\boldsymbol{\gamma}, \bigl(\mathbf{I}_2 \otimes \mathbf{x}_j\bigr)\mathbf{V}\bigl(\mathbf{I}_2 \otimes \mathbf{x}_j\bigr)^t + \Sigma_P \otimes \mathbf{I}_{n_j}\Bigr). \tag{2.10}$$

The structural component of the model allows a simultaneous regression analysis of all person parameters on explanatory variables at the individual and group levels while taking into account the dependencies between the individuals within each group. As a result, among other things, conclusions can be drawn as to the size of the effects of the explanatory variables on the test takers' ability and speed as well as the correlation between these person parameters. Note that hypotheses on these effects can be tested simultaneously.

Multivariate Model for the Item Parameters

An empirical distribution for the item parameters is specified such that for each item the vector $\boldsymbol{\xi}_k = (a_k, b_k, \phi_k, \lambda_k)$ is assumed to follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_\phi, \mu_\lambda)$:

$$\boldsymbol{\xi}_k = \boldsymbol{\mu}_I + \boldsymbol{e}_I, \boldsymbol{e}_I \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I), \tag{2.11}$$

where $\boldsymbol{\Sigma}_I$ specifies the covariance structure.

The assumption introduces a correlation structure between the item parameters. For example, it may be expected that easy items require less time to be solved than more difficult items. If so, the time intensity parameter correlates positively with the item difficulty parameter. The guessing parameter of the response model has no analogous parameter in the RT measurement model (since there is no guessing aspect for the RTs). Therefore, it does not serve a purpose to include it in this multivariate model and an independent prior for this parameter is specified below.

2.3 Exploring the Multivariate Normal Structure

The observed response data are augmented using a procedure that facilitates the statistical inferences. Besides, as will be shown in the next section, these augmentation steps allow for a fully Gibbs sampling approach for estimation of the model.

First, an augmentation step is introduced according to Beguin and Glas (2001). A variable $s_{ijk} = 1$ when a person ij knows the correct answer to question k and is $s_{ijk} = 0$ otherwise. Its conditional probabilities are given by:

$$\begin{aligned}
P(s_{ijk} = 1 | y_{ijk} = 1, \theta_{ij}, a_k, b_k, c_k) &= \frac{\Phi(a_k\theta_{ij} - b_k)}{\Phi(a_k\theta_{ij} - b_k) + c_k(1 - \Phi(a_k\theta_{ij} - b_k))} \\
P(s_{ijk} = 0 | y_{ijk} = 1, \theta_{ij}, a_k, b_k, c_k) &= \frac{c_k(1 - \Phi(a_k\theta_{ij} - b_k))}{\Phi(a_k\theta_{ij} - b_k) + c_k(1 - \Phi(a_k\theta_{ij} - b_k))} \\
P(s_{ijk} = 1 | y_{ijk} = 0, \theta_{ij}, a_k, b_k, c_k) &= 0 \\
P(s_{ijk} = 0 | y_{ijk} = 0, \theta_{ij}, a_k, b_k, c_k) &= 1.
\end{aligned}$$

Second, following Albert (1992), continuous latent responses z_{ijk} are defined:

$$z_{ijk} = a_k\theta_{ij} - b_k + \epsilon_{\theta_{ijk}}, \tag{2.12}$$

where the error terms are standard normally distributed and $\mathbf{s}$ is taken to be a matrix of indicator variables for the events of the components of $\mathbf{z}$ being positive. When the guessing parameters are restricted to be zero, it follows immediately that $s_{ijk} = y_{ijk}$ with probability one and the 2-parameter IRT model is obtained.

Statistical inferences can be made from the complete data due to the following factorization:

$$p\big(\mathbf{y}, \mathbf{z}, \mathbf{s}, \mathbf{t} \mid \mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_P, \mathbf{V}\big) = p\big(\mathbf{y} \mid \mathbf{z}, \mathbf{s}\big)p(\mathbf{s}|\mathbf{c})p\big(\mathbf{z}, \mathbf{t} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_P, \mathbf{V}\big). \tag{2.13}$$

Our interest is in exploring the structural relationships between ability and speed. Therefore, the term on the far right-hand side of (2.13) will be explored in more detail now. This likelihood can be taken to be that of a normal multivariate multilevel model,

$$\begin{aligned}
p\big(\mathbf{z}, \mathbf{t} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \gamma, \boldsymbol{\Sigma}_P, \mathbf{V}\big) = \int\int\int & p\big(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}\big)p\big(\mathbf{t} \mid \boldsymbol{\zeta}, \boldsymbol{\phi}, \boldsymbol{\lambda}\big)p\big(\boldsymbol{\zeta}, \boldsymbol{\theta} \mid \boldsymbol{\beta}, \Sigma_P\big) \\
& \times p\big(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{V}\big) d\boldsymbol{\theta} d\boldsymbol{\zeta} d\boldsymbol{\beta}.
\end{aligned} \tag{2.14}$$

Therefore, all factors in this decomposition are multivariate normal densities. The first two factors occur because of the independence of the responses and response times given the latent person parameters. The last two factors represent levels 2 and 3 of the model.

Inference from this multivariate hierarchical model simplifies when taking advantage of some of the properties of the multivariate normal distribution. For example, let us assume for a moment that the item parameters are fixed and known and define $(\tilde{\mathbf{z}}_{ij}, \tilde{\mathbf{t}}_{ij}) = (\mathbf{z}_{ij} + \mathbf{b}, \mathbf{t}_{ij} - \boldsymbol{\lambda})$. Levels 1 and 2 of the model can then be represented by the following multivariate hierarchical structure:

$$\begin{bmatrix} \theta_{ij} \\ \zeta_{ij} \\ \cdots \\ \tilde{z}_{ij1} \\ \vdots \\ \tilde{z}_{ijK} \\ \cdots \\ \tilde{t}_{ij1} \\ \vdots \\ \tilde{t}_{ijK} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{x}_{ij}^t \boldsymbol{\beta}_{1j} \\ \mathbf{x}_{ij}^t \boldsymbol{\beta}_{2j} \\ \cdots \\ a_1 \theta_{ij} \\ \vdots \\ a_K \theta_{ij} \\ \cdots \\ -\phi_1 \zeta_{ij} \\ \vdots \\ -\phi_K \zeta_{ij} \end{bmatrix}, \left[\begin{array}{cc|c|c} \sigma_\theta^2 & \rho & \sigma_\theta^2 \mathbf{a}^t & -\rho \boldsymbol{\phi}^t \\ \rho & \sigma_\zeta^2 & \rho \mathbf{a}^t & -\sigma_\zeta^2 \boldsymbol{\phi}^t \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{a}\sigma_\theta^2 & \mathbf{a}\rho & \mathbf{a}\sigma_\theta^2 \mathbf{a}^t + \mathbf{I}_K & -\mathbf{a}\rho\boldsymbol{\phi}^t \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ -\boldsymbol{\phi}\rho & -\boldsymbol{\phi}\sigma_\zeta^2 & -\boldsymbol{\phi}\rho\mathbf{a}^t & \boldsymbol{\phi}\sigma_\zeta^2\boldsymbol{\phi}^t + \boldsymbol{\tau}^2 \end{array} \right] \right). \tag{2.15}$$

This representation provides insight in the complex correlational structure hidden in the data and entails several possible inferences. It also helps us to derive some of the conditional posterior distributions for the Gibbs sampling algorithm (e.g., the conditional posterior distributions of the latent person parameters given the augmented data). For a general treatment of the derivation of conditional from multivariate normal distributions, see, for instance, Searle, Casella, and McCulloch (1992).

Parameter ρ, which controls the covariance between the θs and ζs, plays an important role in the model. It can be considered to be the bridge between the separate measurement models for ability and speed. Therefore, its role within the hierarchical structure will be explored in more detail.

The conditional covariance between the latent response variables and RTs on items $k = 1, \ldots, K$ is equal to $cov(a_k\theta_{ij} - b_k + \epsilon_{\theta_{ijk}}, -\phi_k\zeta_{ij} + \lambda_k + \epsilon_{\zeta_{ijk}}) = -a_k\rho\phi_k$, due to independence between the residuals as well as the residuals and the person parameters. Since a_k and ϕ_k are positive, the latent response variables and RTs, and hence the responses and RTs, correlate negatively when ρ is positive. So, in spite of conditional independence between the responses and RTs given the person parameters, their correlation is negative.

The conditional distribution of θ_{ij} given ζ_{ij} is normal:

$$\theta_{ij} \mid \zeta_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2, \sigma_\zeta^2, \rho \sim N\big(\mathbf{x}_{ij}^t \boldsymbol{\beta}_{1j} + \rho\sigma_\zeta^{-2}(\zeta_{ij} - \mathbf{x}_{ij}^t \boldsymbol{\beta}_{2j}), \sigma_\theta^2 - \rho^2\sigma_\zeta^{-2}\big). \tag{2.16}$$

A greater covariance ρ between the person parameters gives a greater reduction of the conditional variance of θ_{ij} given ζ_{ij}. The expression also shows that the amount of information about θ_{ij} in ζ_{ij} depends both on the precision of measuring the speed parameter and its correlation with the ability parameter.

From (2.15), it also follows that the conditional expected value of θ_{ij} given the complete data is equal to

$$\begin{aligned} E\big(\theta_{ij} \mid \boldsymbol{\beta}_j, \zeta_{ij}, \tilde{\mathbf{z}}_{ij}, \mathit{\Sigma}_P, \mathbf{a}, \mathbf{b}\big) &= \mathbf{x}_{ij}^t \boldsymbol{\beta}_{1j} + \rho\sigma_{\zeta}^{-2}(\zeta_{ij} - \mathbf{x}_{ij}^t \boldsymbol{\beta}_{2j}) \\ &\quad + \sigma_\theta^2 \mathbf{a}^t(\mathbf{a}\sigma_\theta^2\mathbf{a}^t + \mathbf{I}_K)^{-1}(\tilde{\mathbf{z}}_{ij} - \mathbf{a}\mathbf{x}_{ij}^t\boldsymbol{\beta}_{1j}) \\ &= \big(\mathbf{a}^t\mathbf{a} + \sigma_{\theta|\zeta}^{-2}\big)^{-1} \qquad (2.17) \\ &\quad \Big(\mathbf{a}^t\tilde{\mathbf{z}}_{ij} + \sigma_{\theta|\zeta}^{-2}\big(\mathbf{x}_{ij}^t\boldsymbol{\beta}_{1j} + \rho\sigma_\zeta^{-2}\big(\zeta_{ij} - \mathbf{x}_{ij}^t\boldsymbol{\beta}_{2j}\big)\big)\Big). \end{aligned}$$

The conditional expected value of θ_{ij} consists of two parts: one part representing the information about θ_{ij} in the (augmented) response data and another the information through the multivariate regression on $\mathbf{x}_{ij}$. For $\rho = 0$, (2.17) reduces to

$$E\big(\theta_{ij} \mid \boldsymbol{\beta}_{1j}, \tilde{\mathbf{z}}_{ij}, \sigma_\theta^2, \mathbf{a}, \mathbf{b}\big) = \big(\mathbf{a}^t\mathbf{a} + \sigma_\theta^{-2}\big)^{-1}\big(\mathbf{a}^t\tilde{\mathbf{z}}_{ij} + \sigma_\theta^{-2}\mathbf{x}_{ij}^t\boldsymbol{\beta}_{1j}\big). \qquad (2.18)$$

This expression can be recognized as the precision-weighted mean of the predictions of θ_{ij} from the (augmented) response data and from the linear regression of θ on $\mathbf{x}$ (see, for instance, Fox & Glas, 2001). Comparing (2.18) with (2.17), it can be seen that when $\rho > 0$, the expected value of θ_{ij} increases for test takers who work at a greater than average speed; that is, a test taker's ability is predicted to be higher when the same response pattern is obtained at a higher speed (i.e., in a shorter expected time on the same set of items).

In (2.15), in addition to the responses and RTs, the random test takers were the only extra source of heterogeneity. But another level of heterogeneity was added in (2.9), where the test takers were assumed to be nested within groups and the regression effects were allowed to vary randomly across them. Also, the item parameters correlate in (2.11). Because of these random effects and correlations, the marginal covariances between the measurements change.

We conclude this discussion with the following comments:

- In (2.15), a special structure (compound symmetry) for the covariance matrix of the residuals at the level of individuals was shown to exist. This structure may lead to more efficient inference. For a general discussion of possible parameterizations and estimation methods for multivariate random effects structures, see, for instance, Harville (1977), Rabe-Hesketh and Skrondal (2001) and Reinsel (1983).
- Linear multivariate three-level structures for continuous responses are discussed, among others, in Goldstein (2003), and Snijders and Bosker (1999). As already indicated, the covariance structure of the level-3 random regression effects is assumed to be block diagonal. This means that the parameters in the regression of $\boldsymbol{\theta}$ on $\mathbf{x}$ are conditionally independent of those in the regression of $\boldsymbol{\zeta}$ on $\mathbf{x}$. It is possible to allow these parameters to correlate but this option is unattractive when the dimension of the covariance matrix becomes large. Typically, the covariance matrix is then poorly estimated (Laird & Ware, 1982).

- For the same reason, the covariance matrix of the fixed effects in (2.9) is assumed to be block diagonal. The Bayesian approach in the next sections allows us to specify different levels of prior information about this matrix.

2.4 Bayesian Estimation using Gibbs Sampling

In Bayesian statistics, inferences are made from the posterior distribution of the model parameters. Markov chain Monte Carlo (MCMC) methods enable us to simulate random draws from this distribution. Summary statistics can then be used to estimate the parameters or functionals of interest. A useful feature of MCMC methods is that they remain straightforward and easy to implement when the complexity of the model increases. Also, they allow for the simultaneous estimation of all model parameters. Since the current model is quite complex and has many parameters, we need these advantages to estimate the model. For general introduction to Gibbs sampling, see Gelman, Carlin, Stern, and Rubin (2004) and Gelfand and Smith (1990). MCMC methods for IRT models are discussed by Albert (1992) and Patz and Junker (1999).

A new Gibbs sampling scheme was developed to deal with the extension of the model. Further, the scheme differs from that in van der Linden (2007) by its increased efficiency; it samples both types of person parameters in one step, taking into account the identifying restrictions, and avoids an MH step in the sampling of the item parameters due to better capitalization on the regression structure of the model. The full conditional distributions of all model parameters for the scheme are given in the appendix.

The remainder of this section discusses the priors and identifying restrictions we use.

2.4.1 Prior Distributions

The parameter c_k is the success probability in the Binomial distribution for the number of correct guesses on item k. A Beta prior with parameters $B(b_1', b_2')$ is chosen, which is the conjugate for the Binomial likelihood and thus leads to a Beta posterior.

For the residual variance τ_k^2 a conjugate inverse Gamma prior is assumed with parameters g_1 and g_2.

A normal inverse-Wishart prior is chosen for the mean vector $\boldsymbol{\mu}_I$ and covariance matrix $\boldsymbol{\Sigma}_I$ of the item parameters. The family of priors is conjugate for the multivariate normal distribution (Gelman et al., 2004). Thus,

$$\boldsymbol{\Sigma}_I \sim Inverse - Wishart(\boldsymbol{\Sigma}_{I_0}^{-1}, \nu_{I_0}) \tag{2.19}$$

$$\boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I \sim N\big(\boldsymbol{\mu}_{I0}, \boldsymbol{\Sigma}_I / \kappa_{I0}\big). \tag{2.20}$$

A vague proper prior follows if ν_{I0} is set equal to the minimum value for the degrees-of-freedom parameter and a diagonal variance matrix with large values is chosen.

Likewise, a normal inverse-Wishart prior is chosen for the fixed parameters $\boldsymbol{\gamma}$ of the multivariate random-effects structure of the person parameters in (2.9),

$$\boldsymbol{\gamma} \mid \boldsymbol{V} \sim N(\boldsymbol{\gamma}_0, \boldsymbol{V}/\kappa_{V_0}). \tag{2.21}$$

The covariance matrix $\mathbf{V}$ of the level-3 random group effects $(\mathbf{u}_{1j}, \mathbf{u}_{2j})$ is assumed to also have an inverse-Wishart prior with scale matrix $\mathbf{V}_0$ and degrees of freedom ν_{V0}.

The prior for the covariance matrix of the person parameters, $\boldsymbol{\Sigma}_P$, is chosen to give special treatment because the model is not yet identified.

Prior for $\boldsymbol{\Sigma}_P$ with Identifying Restrictions

The model can be identified by fixing the scales of the two latent person parameters. A straightforward way of fixing the scale of the ability parameter is to set the mean equal to zero and the variance to one. To avoid a tradeoff between $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$ the time discrimination parameters are restricted to $\prod_{k=1}^{K} \phi_k = 1$. When these are restricted to $\boldsymbol{\phi} = \mathbf{1}$ the lognormal RT model as proposed by van der Linden (2006) is obtained. Then, for the speed parameter, since RTs have a natural unit, we only have to fix the origin of its scale and set it equal to its population mean. Note that a multivariate probit model is identified by fixing the diagonal elements of the covariance matrix (Chib & Greenberg, 1998) but that, because of the special nature of the RTs, in the current case only one element of $\boldsymbol{\Sigma}_P$ has to be fixed.

Generally, two issues arise when restricting a covariance structure. First, defining proper priors for a restricted covariance matrix is rather difficult. For example, for the conjugate inverse-Wishart prior, there is no choice of parameter values that reflects a restriction on the variance of the ability parameter such as that above. For the multinomial probit model, McCulloch and Rossi (1994) tackled this problem by specifying proper diffuse priors for the unidentified parameters and reporting the marginal posterior distributions of the identified parameters. However, it is hard to specify prior beliefs about unidentified parameters. Second, for a Gibbs sampler, sampling from a restricted covariance matrix requires extra attention. Chib and Greenberg (1998) defined individual priors on the free covariance parameters but, as a result, the augmented data had to be sampled from a special truncated region and the values of the free covariance parameter could only be sampled using an MH step. However, such steps involve the specification of an effective proposal density with tuning parameters that can only be fixed through a cumbersome process. A general approach for sampling from a restricted covariance matrix can be found in Browne (2006), but this is also based on an MH algorithm.

Here, a different approach is taken that allows us to specify proper informative priors and facilitate the implementation of the Gibbs sampler. A prior is chosen such that $\sigma_\theta^2 = 1$ with probability one. Hence, covariance matrix $\boldsymbol{\Sigma}_P$ always equals:

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} 1 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}. \tag{2.22}$$

Using (2.8) and (2.22), the conditional distribution of ζ_{ij} given θ_{ij} has density

$$\zeta_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j, \rho, \sigma_\zeta^2 \sim N\big(\mathbf{x}_{ij}^t \boldsymbol{\beta}_{2j} + \rho(\theta_{ij} - \mathbf{x}_{ij}^t \boldsymbol{\beta}_{1j}), \tilde{\sigma}_\zeta^2\big)$$

where $\tilde{\sigma}_\zeta^2 = \sigma_\zeta^2 - \rho^2$. Parameter ρ can be viewed as the slope parameter in a normal regression problem of ζ_{ij} on θ_{ij} with variance $\tilde{\sigma}_\zeta^2$. Specifying a normal and inverse gamma as conjugate priors for these parameters,

$$\rho \sim N(\rho_0, \sigma_\rho^2), \tag{2.23}$$

$$\tilde{\sigma}_\zeta^{-2} \sim Gamma(g_1, g_2), \tag{2.24}$$

their full conditional posterior distributions become

$$\rho \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\beta}, \rho_0, \sigma_\rho^2 \sim N\big(\varDelta\big(\rho_0 \sigma_\rho^{-2} + \big(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}_1\big)^t \tilde{\sigma}_\zeta^{-2}\big(\boldsymbol{\zeta} - \mathbf{x}\boldsymbol{\beta}_2\big)\big), \varDelta\big) \tag{2.25}$$

$$\tilde{\sigma}_\zeta^{-2} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\beta}, \rho \sim Gamma\Big(g_1 + N/2, g_2 + \varXi^t \varXi/2\Big), \tag{2.26}$$

where $\varDelta = \Big(\big(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}_1\big)^t \tilde{\sigma}_\zeta^{-2}\big(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}_1\big) + \sigma_\rho^{-2}\Big)^{-1}$ and $\varXi = \big(\boldsymbol{\zeta} - \mathbf{x}\boldsymbol{\beta}_2\big) - \rho\big(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}_1\big)$.

Since $|\boldsymbol{\Sigma}_P| = \sigma_\zeta^2 - \rho^2 = \tilde{\sigma}_\zeta^2$ and $\tilde{\sigma}_\zeta^2 > 0$, it follows that the determinant $|\boldsymbol{\Sigma}_P| > 0$. The latter is sufficient to guarantee matrix $\boldsymbol{\Sigma}_P$ to be positive semi-definite.

When implementing a Gibbs sampler, the random draws of the elements of covariance matrix $\boldsymbol{\Sigma}_P$ in (2.22) can be constructed from the samples drawn from (2.25)–(2.26). These draws will show greater autocorrelation due to this new parametrization. This implies that more MCMC iterations are needed to cover the support of the posterior distribution adequately, a measure that only involves a (linear) increase in the running time of the sampler. On the other hand, convergence of the algorithm is easily established without having to specify any tuning parameter. Finally, this procedure also enables straightforward implementation of the data augmentation procedure since the **z**s can be drawn from a normal distribution truncated at zero, where **s** indicates when **z** is positive.

The key element of the present approach is the specification of a proper prior distribution for the covariance matrix with one fixed diagonal element and the construction of random draws from the matrix from the corresponding conditional posterior distribution. For the multinomial probit model, the approach was also followed by McCulloch, Polson, and Rossi (2000). For completeness, we also mention an alternative approach. Barnard, McCullogh, and Meng (2000) formulated a prior directly for the identified parameters. In order to do so, they factored the covariance matrix into a diagonal matrix with standard deviations and a correlation matrix, and specified an informative prior for the latter. This prior was then incorporated into a Griddy-Gibbs sampler. However, such algorithms can be slow and require the choices of a grid size and boundaries. Boscardin and Zhang (2004) followed a comparable approach but used a parameter-extended MH algorithm for sampling values from the conditional distribution of the correlation matrix.

2.5 Model Selection Methods

A model comparison method is often based on a measure of fit and some penalty function based on the number of free parameters for the complexity of the model. A bias-variance trade-off exists between these two quantities since a more complex model often leads to less bias but a less complex model involves more accurate estimation. Two well-known criteria of model selection based on a deviance fit measure are the Bayesian information criterion (BIC) (Schwarz, 1978) and Akaike's information criterion (AIC) (Akaike, 1973). These criteria depend on the effective number of parameters in the model as a measure of model complexity. A drawback of these measures is that they are often difficult to calculate for hierarchical models: Although the nominal number of parameters follows directly from the likelihood, the prior distribution imposes additional restrictions on the parameter space and reduces its effective dimension. In a random-effects model, the effective number of parameters depends strongly on the higher-level variance parameters. When the variance of the random effects approaches zero, all random effects are equal and the model reduces to a simple linear model with one mean parameter. But when the variance goes to infinity, the number of free parameters approaches the number of random effects.

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) for model comparison when the number of parameters is not clearly defined. The DIC is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity described below.

An alternative method for model selection that can handle complex hierarchical models is the Bayes factor; for a review, see Kass and Raftery (1995). The Bayes factor is based on a comparison of marginal likelihoods but its implementation is hampered by its critical dependence on the prior densities assigned to the model parameters. It is known that the Bayes factor tends to favor models with reasonably vague proper priors; see, for instance, Berger and Delampady (1987) and Sinharay and Stern (2002). An advantage of the Bayes factor is its clear interpretation as the change in the odds in favor of the model when moving from the prior to the posterior distribution (Lavine & Schervish, 1999).

In one of the empirical examples below, the focus is on the structural multivariate model for the person parameters. It will be shown that a DIC can be formulated for choosing between models that differ in the fixed and/or random part of the structural model. In addition, a Bayes factor for selecting between the IRT measurement model for binary responses and the model extended with the hierarchical structure for responses and RTs is presented.

2.5.1 Model Selection using the DIC

The DIC requires a closed-form likelihood. Our interest is focused on the likelihood of the structural parameters in the model; accordingly, all random effect parameters can be integrated out. Besides, the variances, covariances, and items parameters are considered as nuisance parameters, and their values are assumed to be known. So, a DIC will be derived for the complete-data likelihood with the random effects integrated out. Subsequently, the posterior expectation of the DIC over the augmented data will be taken. The same procedure was proposed for mixture models by DeIorio and Robert (2002).

Let $\mathbf{z}^*_{ij} = vec(\mathbf{z}_{ij} + \mathbf{b}, \mathbf{t}_{ij} - \boldsymbol{\lambda})$ and $\mathbf{H}_P = (\mathbf{a} \oplus -\boldsymbol{\phi})$. From (2.15), Conditional on $\mathbf{s}$, the measurement models for ability and speed can be summarized as

$$\mathbf{z}^*_{ij} = \mathbf{H}_P \boldsymbol{\Omega}_{ij} + \mathbf{e}_{ij}, \tag{2.27}$$

where $\mathbf{e}_{ij} \sim N(0, \boldsymbol{C})$, with $\boldsymbol{C} = (\mathbf{I}_K \oplus \mathbf{I}_K \boldsymbol{\tau}^2)$ a diagonal matrix with in the left upper square $\mathbf{1}$ and in the right lower square $\boldsymbol{\tau}$ on its diagonal, and $\boldsymbol{\Omega}_{ij}$ =vec$\left(\theta_{ij}, \zeta_{ij}\right)$. The focus is on the structure of $\boldsymbol{\Omega}$. Using the factorization in (2.13), the standardized deviance is

$$D(\boldsymbol{\Omega}) = \sum_{ij} \left(\mathbf{z}^*_{ij} - \mathbf{H}_P \boldsymbol{\Omega}_{ij}\right)^t \boldsymbol{C}^{-1} \left(\mathbf{z}^*_{ij} - \mathbf{H}_P \boldsymbol{\Omega}_{ij}\right). \tag{2.28}$$

The DIC is defined as

$$\begin{aligned} DIC &= \int \left[DIC \mid \mathbf{z}\right] p(\mathbf{z}|\mathbf{y}) d\mathbf{z} && (2.29)\\ &= \int \left[D(\bar{\boldsymbol{\Omega}}) + 2p_D\right] p(\mathbf{z} \mid \mathbf{y}) d\mathbf{z} && (2.30)\\ &= E_{\mathbf{z}} \left[D(\bar{\boldsymbol{\Omega}}) + 2p_D \mid \mathbf{y}\right] && (2.31) \end{aligned}$$

where $\bar{\boldsymbol{\Omega}}$ equals the posterior mean and p_D is the effective number of parameters given the augmented data. The latter can be shown to be equal to the mean deviance minus the deviance of the mean. Hence,

$$
\begin{align}
p_D &= \overline{D(\boldsymbol{\Omega})} - D(\bar{\boldsymbol{\Omega}}) \tag{2.32}\\
&= E_{\boldsymbol{\Omega}}\big[D(\boldsymbol{\Omega}) \mid \mathbf{z}^*\big] - D(E(\boldsymbol{\Omega} \mid \mathbf{z}^*)) \notag\\
&= E_{\boldsymbol{\Omega}}\Big[\sum_{ij}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t \boldsymbol{C}^{-1}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\Big] - D(E(\boldsymbol{\Omega} \mid \mathbf{z}^*)) \notag\\
&= tr\Big[\sum_{ij} E_{\boldsymbol{\Omega}}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t \boldsymbol{C}^{-1}\Big] \notag\\
&\quad - tr\Big[\sum_{ij}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)^t \boldsymbol{C}^{-1}\Big] \tag{2.33}\\
&= \sum_{ij} tr\Big[E_{\boldsymbol{\Omega}}\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)^t \boldsymbol{C}^{-1} \notag\\
&\quad - \big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)\big(\mathbf{z}^*_{ij} - \mathbf{H}_P E(\boldsymbol{\Omega} \mid \mathbf{z}^*)\big)^t \boldsymbol{C}^{-1}\Big] \tag{2.34}\\
&= \sum_{ij} tr\big[\boldsymbol{C}^{-1}\ \mathrm{var}(\mathbf{e}_{ij} \mid \mathbf{z}^*_{ij})\big] \tag{2.35}\\
&= \sum_{ij} tr\big[\boldsymbol{C}^{-1}[\mathrm{var}(\mathbf{e}_{ij}) - \mathrm{cov}(\mathbf{e}_{ij}, \mathbf{z}^*_{ij})\mathrm{var}(\mathbf{z}^*_{ij})^{-1}\mathrm{cov}(\mathbf{e}_{ij}, \mathbf{z}^*_{ij})]\big] \tag{2.36}\\
&= \sum_{ij} tr\big[\mathrm{var}(\mathbf{z}^*_{ij})^{-1}\mathrm{var}\big(\mathbf{H}_P\boldsymbol{\Omega}_{ij}\big)\big] \tag{2.37}\\
&= \sum_{ij} tr\Big[\big(\mathbf{H}_P\mathbf{x}_{ij}\mathbf{w}_j\varSigma_\gamma\mathbf{w}^t_j\mathbf{x}^t_{ij}\mathbf{H}^t_P + \mathbf{H}_P\mathbf{x}_{ij}V\mathbf{x}^t_{ij}\mathbf{H}^t_P + \mathbf{H}_P\varSigma_P\mathbf{H}^t_P + \boldsymbol{C}\big)^{-1} \notag\\
&\qquad \big(\mathbf{H}_P\mathbf{x}_{ij}\mathbf{w}_j\varSigma_\gamma\mathbf{w}^t_j\mathbf{x}^t_{ij}\mathbf{H}^t_P + \mathbf{H}_P\mathbf{x}_{ij}V\mathbf{x}^t_{ij}\mathbf{H}^t_P + \mathbf{H}_P\varSigma_P\mathbf{H}^t_P\big)\Big], \tag{2.38}
\end{align}
$$

where $tr(\cdot)$ denotes the trace function, i.e., the sum of the diagonal elements. The expectation is taken with respect to the posterior distribution of $\boldsymbol{\Omega}$. The terms in (2.35) can be recognized as the posterior variances of the residuals whereas those in (2.37) follow from the fact that, because of independence, the variance of $\mathbf{z}^*_{ij}$ equals the sum of the variance of $\mathbf{H}_P\boldsymbol{\Omega}_{ij}$ and $\mathbf{e}_{ij}$.

DICs of nested models are computed by restricting one or more variance parameters in (2.38) to zero. Also, (2.38) can be estimated as a by-product of the MCMC algorithm; that is, the output of the algorithm can be used to estimate the posterior means of the model parameters in the second term of (2.32) and to integrate the DIC over the item parameters to obtain the first term. (In the current application, the item parameters are the nuisance parameters.)

Usually the variance parameters are unknown. Then the DIC has to be integrated over their marginal distribution, too. In fact, the correct Bayesian approach would be to integrate the joint posterior over the nuisance parameters to obtain the marginal posterior of interest. However, this approach is not possible since no closed-form expression of the DIC can be obtained for this marginal posterior. Thus, our proposal does not account for the unknown variances. (2.38) reflects the effective number of parameters of the proposed model without the additional variability in the posterior because of the unknown covariance parameters. The more general case with unknown covariance parameters is complex, and no simple correction seems available. But Vaida and Blanchard (2005) showed that, for a mixed-effects model, the correction for unknown covariance parameters is negligible asymptotically. So, it seems safe to assume that their effect on the estimate of (2.32) becomes only apparent when the covariance parameters are estimated less precisely.

2.5.2 Model Selection using the Bayes factor

The question we address is if the use of the RTs in the hierarchical model proves to be beneficial for making inferences about the ability parameter. As no benefits can be obtained when the correlation $r(\theta, \zeta) = \varrho = 0$ (i.e., independence between θ and ζ), a Bayes factor is defined to test whether the data support fitting a model M_1 between θ and ζ or the null model $M_0 \subset M_1$ with independence. For an introduction to Bayes factors, see Berger and Delampady (1987); Kass and Raftery (1995).

Both models are given equal prior weight. Therefore, the Bayes factor can be presented as

$$BF = \frac{p(\mathbf{y}, \mathbf{t} \mid M_0)}{p(\mathbf{y}, \mathbf{t} \mid M_1)} \tag{2.39}$$

$$= \frac{\int p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}, \mathbf{t} \mid M_0) d\mathbf{z}}{\int p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}, \mathbf{t} \mid M_1) d\mathbf{z}} \tag{2.40}$$

$$= \frac{\int p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}, \mathbf{t} \mid \varrho = 0) d\mathbf{z}}{\int \int p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}, \mathbf{t} \mid \varrho) \pi(\varrho) d\varrho d\mathbf{z}}. \tag{2.41}$$

A popular family of conjugate priors for the correlation coefficient has the form $(1 - \varrho^2)^\nu$ on its support, $0 \leq \varrho \leq 1$ (P. M. Lee, 2004). For $\nu = 0$, a uniform distribution is obtained. For $\nu = 5$, a half-normal distribution is approximated. For $\nu \to \infty$, the prior assigns probability 1 to $\varrho = 0$, which yields model M_0. To assess the sensitivity of the Bayes factor to the specification of the prior density, a variety of members from the family can be chosen.

2.6 Simulation Study

In the first study, different data sets were simulated and the parameters were re-estimated to check the performance of the Gibbs sampler. In the second study, the properties of the proposed Bayes factor in (2.41) were investigated for data sets generated for different values of ϱ and different choices of prior distributions. We also checked the rejection region for the null hypothesis. In the third study, the characteristics of the proposed DIC test were analyzed.

2.6.1 Parameter Recovery

Datasets were simulated for the following structural component of the model:

$$\begin{pmatrix}\theta_{ij}\\ \zeta_{ij}\end{pmatrix} = \begin{pmatrix}\gamma_{00}+u_{0j}^{(\theta)}\\ \gamma_{10}+u_{1j}^{(\zeta)}\end{pmatrix} + \begin{pmatrix}x_{ij}\left(w_j\gamma_{01}+u_{1j}^{(\theta)}\right)\\ x_{ij}\left(w_j\gamma_{11}+u_{2j}^{(\zeta)}\right)\end{pmatrix} + \begin{pmatrix}e_{1ij}\\ e_{2ij}\end{pmatrix},$$

where $\mathbf{e}_{ij} \sim N(0, \varSigma_P)$, $\mathbf{u}^{(\theta)} \sim N(0, \mathbf{V}_1)$ and $\mathbf{u}^{(\zeta)} \sim N(0, \mathbf{V}_2)$. The model had the same set of explanatory variables in the regression of each latent parameter and had random intercepts and slopes. The intercepts and slopes were taken to be independent of the residuals and across the person parameters. The true values of the structural parameters used in the study are given in Table 2.1. The values of the explanatory variables $\mathbf{x}$ and $\mathbf{w}$ were drawn from a standard normal distribution. For the responses, the 2PL model was assumed and the item parameters were drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}_I = (1, 0, 1, 0)$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_I$ with all variances equal to .5. Negative values of $\boldsymbol{\phi}$ and $\mathbf{a}$ were simply ignored. Responses and RTs were simulated for $N = 1,000$ persons nested in 50 groups each taking 20 items.

In the estimation procedure, the following hyperparameters were used: Scale matrices $\varSigma_{I0}$ and $\varSigma_{\gamma 0}$ were chosen to be diagonal with elements .01 to indicate vague proper prior information, and we set $\boldsymbol{\mu}_{I0} = (1, 0, 1, 0)$ and $\boldsymbol{\gamma}_0 = \mathbf{0}$. Besides, a vague normal prior with parameters $\mu_\rho = 0$ and $\sigma^2_\rho = 10$ was specified for ρ.

The MCMC procedure was iterated $50,000$ times and the first $5,000$ iterations were discarded when the means and posterior standard deviations of the parameters were estimated.

The accuracy of the parameter estimates was investigated by comparing them to their true values. The results for the parameters in the structural model are given in Table 1. Both the estimates of the fixed parameters and the variance components are in close agreement with the true values. (Note that γ_{00} and γ_{10} are zero due to the identifying restrictions.) Although not shown here, the same close agreement was observed for the item parameter estimates.

2.6.2 Sensitivity of the Bayes Factor

Usually, we will have little prior information about the correlation of the person parameters. Therefore, it is important to know how the Bayes factor behaves for a relatively vague prior distribution of the correlation $\varrho = \rho^2/\sqrt{\sigma^2_\theta \sigma^2_\zeta}$. In total, 500 data sets were simulated for different values of $\varrho \in [0, 1]$ and an empty structural model for the person parameters. All other specifications were identical to those in the preceding study. The Bayes factor in (2.41) was calculated using an importance sampling method (Newton & Raftery, 1994). For each data set, the calculations were repeated for different priors for the correlation parameter.

Following Lee (2004), a reference prior for ϱ was used, which led to

$$BF(\nu) = \frac{\int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}, \mathbf{t} \mid \varrho = 0)d\mathbf{z}}{\int\int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}, \mathbf{t} \mid \varrho)\mathcal{C}\left(1-\varrho^2\right)^{\nu} d\varrho d\mathbf{z}}, \tag{2.42}$$

Table 2.1. Simulated and estimated values of the structural parameters.

Fixed parameters		True value	EAP	SD
	γ_{00}	0.00	0.00	-
	γ_{01}	4.00	3.77	0.23
	γ_{10}	0.00	0.00	-
	γ_{11}	3.00	2.99	0.12
Variance components		True value	EAP	SD
$\boldsymbol{\Sigma}_P$				
	Σ_{11}	1.00	1.00	-
	Σ_{12}	0.50	0.55	0.04
	Σ_{22}	1.00	1.07	0.06
$\mathbf{V}_1$				
	$\mathbf{V}_{11}$	1.00	1.00	0.25
	$\mathbf{V}_{12}$	0.50	0.48	0.22
	$\mathbf{V}_{22}$	1.00	1.13	0.35
$\mathbf{V}_2$				
	$\mathbf{V}_{11}$	1.00	1.07	0.23
	$\mathbf{V}_{12}$	0.50	0.47	0.17
	$\mathbf{V}_{22}$	1.00	0.86	0.19

with $\mathcal{C}$ the normalizing constant. According to Jefreys' scheme (Kass & Raftery, 1995), $1/BF(\nu) > 3$ implies evidence against the null hypothesis of $\varrho = 0$ given the value of ν.

The results are shown in Figure 2.1, where the dotted line indicates $log(BF) = 0$. For true values of ϱ close to zero or larger than .35, the Bayes factor yielded the same conclusion for all chosen priors. More specifically, it favored the null model for all values of ϱ below .1 but the alternative model for all values larger than .35. It can also be seen that the estimated Bayes factors are higher (and thus favor the null model more frequently) for lower values of ν, which correspond to the less informative priors. For $\varrho \in [.20, .35]$, the prior distribution of ϱ was the major determinant of the Bayes factor favoring the null or the alternative model.

It can be concluded that the Bayes factor is sensitive to the prior choice for ϱ. Figure 2.1 gives a clear idea about the variation of the Bayes factor for a class of prior distributions. This information can be used in real-world applications when a prior for ϱ needs to be selected but the information about this parameter is poor.

2.6.3 Iterative Model Building using the DIC

In this study, it was investigated whether the DIC can be used to choose between models with different fixed and/or random terms in the structural component of

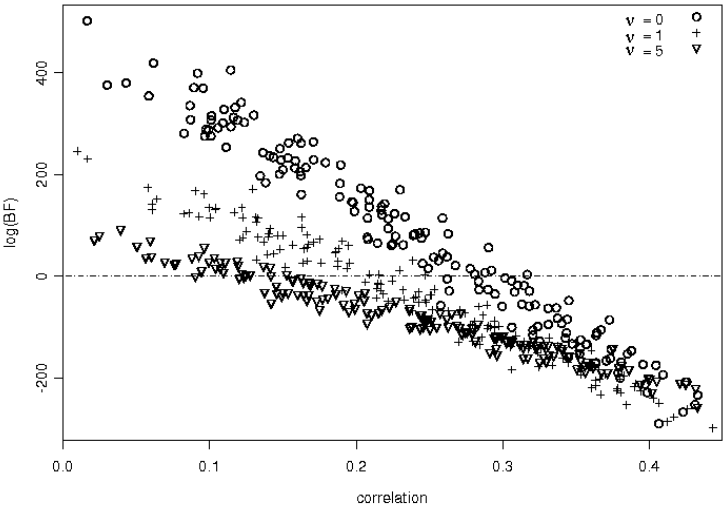


Fig. 2.1. Log(BF) as a function of the correlation between accuracy and speed for 3 different priors for ϱ.

the model for the person parameters. Data were simulated for 1,000 persons nested in 20 groups each taking 20 items using a model that is explained below. The setup was the same as in the earlier parameter recovery study; the only difference was that w_j was set equal to one for all j.

Table 2.2 summarizes the calculations of the DIC for four different models. $\bar{D}$ is the estimated posterior mean deviance; $D(\hat{\boldsymbol{\Omega}})$ is the deviance for the posterior mean of the parameter values.

Table 2.2. Deviance summaries for the four models in the simulation study.

Model	$\bar{D}$	$D(\hat{\boldsymbol{\Omega}})$	p_D	DIC
1,Two-Level, fixed parameters	40168	38184	1984	42152
2,Empty Two-level	40161	38194	1967	42129
3,Two-level + $\boldsymbol{\Omega} \mid \mathbf{x}$	40172	38206	1966	42139
4,Three-level + $\boldsymbol{\Omega} \mid \mathbf{x}$	40165	38290	1875	42039

Model 1 was an empty two-level model with fixed parameters for $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, which was obtained by setting ρ to 0 and the variances of each person parameters equal to $1,000$. Model 2 was an empty two-level model that ignored any group structure for the test takers. In Model 3, the two-level structure was extended with a covariate of $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$ but no group structure was assumed. Model 4 was the true model under which the data were simulated; this model did have both the covariate and the group structure for the test takers. Identification of the models was obtained via the restriction $\prod_{k=1}^{K} a_k = 1$ on the item parameters. This choice is motivated as follows: In Model 4, the variability of the person parameters is explained by $\boldsymbol{V}$ and the covariates. When estimating Model 1, 2 or 3 from the data simulated under Model 4, this variability should be captured by $\boldsymbol{\Sigma}_P$ as extra residual variation. Therefore, $\boldsymbol{\Sigma}_P$ was left unrestricted; otherwise, the variance would have been traded with the estimated $\boldsymbol{a}$ parameters, which might have led to misinterpretation of the results.

As expected, the DIC values we found suggested that Model 4 was best performing. Particularly, it can be seen that when the grouping of the test takers was ignored, this led to an increase in the effective number of parameters. Note that the optimal model choice is not necessarily the best fitting model, but a tradeoff between model fit and the number of parameters used.

2.7 Empirical Examples

In this section, two empirical examples illustrate the use of several developments that were presented in this chapter.

2.7.1 First Example

A data set of 286 persons who had taken a computerized version of a 22-item personality questionnaire was analyzed. The respondents were Psychology and Social Sciences undergraduates from a university in Spain. The majority of the students was between 18 and 30 years old (age variable), and this group consisted of 215 girls and 71 boys (gender variable). The questionnaire consisted of two scales of 11 dichotomous items measuring neuroticism and extraversion. The neuroticism dimension assesses whether a person is prone to experience unpleasant emotions and is emotionally unstable and the extraversion dimension measures sociability, enthusiasm and arousal of pleasure. According to the five factor model, these two dimensions summarize part of the covariation among personality traits.

As already indicated, because it does not assume anything about the relationship between the speed at which the individual test takers work and the latent trait represented in the response model, the modeling framework can also be applied to personality questionnaires. For this domain, it is also interesting to study the responses and RTs simultaneously. In addition to the statistical advantages of multivariate modeling of the data over separate univariate modeling, such a study would allow us to infer, for instance, how differences in speed levels between subgroups of test takers correlate with differences in their personality traits.

From earlier results it was known that there is a moderate negative dependency between neuroticism and extraversion (Becker, 1999; McCrea & Costa, 1997). Here, interest was focused on the differences between students with respect to these two personality dimensions given age and gender. Additionally, variation in the respondent's speed-levels with respect to neuroticism and extraversion was explored. This part of the study involved the estimation of the covariances between the personality traits and the latent speed-levels.

Since this test consisted of yes - no personality questions, the 2PL model was chosen as the measurement model for the responses. For the RTs, the measurement part was specified by (2.2). The following structural part was specified to explore the variation in speed and the traits as a function of both background variables:

$$\begin{pmatrix}\theta_i\\ \zeta_i\end{pmatrix} = \begin{pmatrix}\gamma_{00}\\ \gamma_{10}\end{pmatrix} + \begin{pmatrix}\gamma_{01}\text{Male}_i + \gamma_{02}\text{Age}_i\\ \gamma_{11}\text{Male}_i + \gamma_{12}\text{Age}_i\end{pmatrix} + \begin{pmatrix}e_{1i}\\ e_{2i}\end{pmatrix}, \tag{2.43}$$

where $\boldsymbol{e} \sim N(0, \boldsymbol{\Sigma}_P)$. Further, γ_{00} and γ_{10} denote the intercepts, γ_{01} and γ_{11} represent the effects of being male, and γ_{02} and γ_{12} represent the effects of age. The age vector contained the age of the test takers on a continuous scale.

Four models were fitted to the data: (1) null model without covariates, (2) and (3) structural multivariate model with age and gender as a covariate, respectively, and (4) full structural multivariate model with both age and gender as covariates. For estimation, proper noninformative priors were specified, with all prior variance components set at 100 and the covariances at 0. The MCMC algorithm was iterated 50,000 times; the first 10,000 iterations were discarded as the burn-in period.

Posterior predictive checks were used to evaluate several assumptions of the model. An important assumption of the model is that of local independence. Therefore, an odds ratio statistic was used to test for possible dependencies between response patterns of items. For an impression of overall fit of the response model, an observed score statistic was estimated to assess if the model was able to replicate the observed score patterns. For a detailed description of these two statistics, see Sinharay (2005) and Sinharay, Johnson, and Stern (2006). To assess the fit of the RT model van der Linden and Guo (in press) proposed a Bayesian residual analysis. That is, 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 $p \approx \sum_{m=1}^{M} \Phi\left(t_{ik}|\zeta_i^{(m)}, \phi_k^{(m)}, \lambda_k^{(m)}\right)/M$, from M iterations from the MCMC chain. According to the probability integral transform theorem, under a good fitting model, these probabilities should be distributed $U(0,1)$. Model fit can then be checked graphically by plotting the posterior p-values against their expected values under the $U(0,1)$ distribution.

The posterior checks of the model were based on 1000 replicated data sets from the posterior distribution. The fitted IRT model replicated the responses well, as the observed sum score statistic did not point at any significant flaws for neither of the two scales. The odds ratio statistic indicated that, for two item combinations on the neuroticism scale and for four item combinations on the extraversion scale, a violation of local independence might exist. However, given all possible item combinations, these possible violations of local independence did not give any reason to doubt the unidimensionality of the scales. As indicated by minor deviations in the lower tail and in the middle of the $U(0,1)$ distributions, the RT model tended to slightly underpredict the quicker responses (see Figure 2.3 in the Appendix). Overall, however, model fit was satisfactory.

Table 2.3. DIC values for 4 models fitted to the neuroticism and extraversion scales.

	Neuroticism			Extraversion		
Model	p_D	$D(\hat{\Omega})$	DIC	p_D	$D(\hat{\Omega})$	DIC
1 (null)	572	3839	4983	572	3767	4911
2 (gender)	572	3846	4990	572	3774	4918
3 (age)	572	3941	5085	572	3838	4982
4 (full)	572	3943	5087	572	3827	4971

Table 2.3 gives the calculated DIC values for the four models and the two scales. Comparing the results for model (1) and model (3), the DIC criterium yielded no significant difference in performance between boys and girls on both scales. That is, for the neuroticism as well as the extraversion scale, no mean significant difference between boys and girls was found, neither in the latent personality traits, nor in the speed of working on the test. Neither did the age of the test takers explain any significant amount of variation in the personality traits and speed levels.

Next, the respondents were clustered with respect to their estimated extraversion scores. The clustering was such that the intervals of respondents' scores in each cluster had equal probability mass under a normal model for the population distribution. The sample size of 286 respondents allowed a grouping of respondents in eight different clusters of extraversion levels. Note that the clusters were obtained from an estimated population model, and that they jointly represented the entire score range.

It was investigated whether the grouping of respondents with respect to the extraversion scores explained any variation in respondents' neuroticism scores. Further, the influence of the background variables was explored. The following multivariate random effects structural model was specified:

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} = \begin{pmatrix} \gamma_{00} + u_{0j}^{(\theta)} \\ \gamma_{10} + u_{1j}^{(\zeta)} \end{pmatrix} + \begin{pmatrix} \text{Male}_{ij}\left(\gamma_{01} + u_{01j}^{(\theta)}\right) + \text{Age}_{ij}\left(\gamma_{02} + u_{02j}^{(\theta)}\right) \\ \text{Male}_{ij}\left(\gamma_{11} + u_{11j}^{(\zeta)}\right) + \text{Age}_{ij}\left(\gamma_{12} + u_{12j}^{(\zeta)}\right) \end{pmatrix} + \begin{pmatrix} e_{1ij} \\ e_{2ij} \end{pmatrix}, \tag{2.44}$$

where $\mathbf{e}_{ij} \sim N(0, \Sigma_P)$, $\mathbf{u}^{(\theta)} \sim N(0, \mathbf{V}_1)$ and $\mathbf{u}^{(\zeta)} \sim N(0, \mathbf{V}_2)$. In (2.44), the intercepts and slope coefficients for the regression on the neuroticism scores and the speed levels were treated as random across clusters of extraversion levels. These random effects were allowed to correlate both within the regression on the neuroticism scores and within the regression on the speed levels. Also, the error terms at the individual level were allowed to correlate since the speed levels and neuroticism scores were clustered within individuals.

Five models were fitted to the neuroticism scale by restricting one or more parameters to zero: (1) the null model with fixed intercepts by restricting $\mathbf{V}_1$ and $\mathbf{V}_2$ to be zero; (2) the empty multivariate random effects model (without covariates) with free covariance parameters; (3) - (4) a multivariate random effects model including a random regression effect for gender and age, respectively and (5) the full model as specified in (2.44).

Using the proper noninformative priors described earlier, the models were estimated using 50,000 iterations of the Gibbs sampler, where 10,000 iterations were discarded because of the burn-in. The DIC value for each of the five models was estimated using (2.31) since our interest was focused primarily on the structural model on the speed levels and neuroticism scores. The estimated DIC values are given in Table 2.4.

Table 2.4. DIC values for five models fitted to the neuroticism scale, accounting for a grouping of respondents in extraversion levels.

Model	p_D	$D(\hat{\boldsymbol{\Omega}})$	DIC
1 (null)	572	3839	4983
2 (empty)	521	3858	4899
3 (gender)	536	3866	4938
4 (age)	572	3860	5004
5 (full)	572	3873	5017

It can be seen that the empty multivariate random-effects model had a smaller effective number of model parameters relative to the null model and was to be preferred given the DIC values of both models. The estimated deviance increased slightly for Models 3 - 5, which can be attributed to additional sampling variance introduced by the covariates. Note that in the empty multivariate random-effects model, the individual random-effect parameters were modeled as group-specific random effects at the level of the clusters of extraversion scores (a third level in the model) and that this led to a serious reduction in the effective number of model parameters. It can be concluded that the grouping of respondents according to their extraversion levels explained a substantial amount of variation in the speed levels as well as the neuroticism scores. The estimated correlation between the neuroticism scores and the speed levels was .30 (with a standard deviation of .07), which justified the multivariate modeling approach. Intraclass correlation coefficients were calculated to asses the amount of variability in the individual neuroticism scores and the speed levels due to the grouping of respondents in clusters of extraversion levels. The intraclass correlation estimates for neuroticism and the speed trait were based on the MCMC output for the empty multivariate random effects model. The estimates were

$$
\begin{aligned}
\mathrm{ICC}_\theta &\approx \frac{1}{M}\sum_{m=1}^{M}\frac{V_{11}^{(m)}}{V_{11}^{(m)}+\sigma_\theta^{2(m)}} = .12 \\
\mathrm{ICC}_\zeta &\approx \frac{1}{M}\sum_{m=1}^{M}\frac{V_{22}^{(m)}}{V_{22}^{(m)}+\sigma_\zeta^{2(m)}} = .07,
\end{aligned}
$$

where $m = 1, \ldots, M$ denotes the number of iterations after burn-in. It follows that 12% of the variability in the neuroticism trait could be explained by the grouping of the respondents by their extraversion levels. It is surprising that 7% of the variability in speed levels was located at the group level. This means that the clustering of the respondents via the estimated extraversion levels explained a significant amount of variation in the individual speed levels corresponding to the neuroticism test. The explanation is supported by the estimated correlation between both speed parameters, which was .76. Note that this relatively high correlation between the individual speed levels on the two tests also supports the assumption of stationary speed during testing. Finally, the DIC values show that the covariates did not explain any variation in the trait or speed levels. It can be seen that the introduction of random regression parameters for the background variables did not lead to any reduction in the effective number of parameters since the covariates did not explain any variation within the grouped neuroticism scores. Neither did they for the entire sample of neuroticism scores.

2.7.2 Second Example

In this example, the data set studied earlier by Wise, Kong, and Pastor (2007) was analyzed. This data set included 388 test takers who each answered 65 items of a computer-based version of the *Natural World Assessment Test* (NAW-8). This test is used to assess the quantitative and scientific reasoning proficiencies of college students. It was part of a required education assessment for mid-year sophomores by a medium-sized university. Covariates for the test takers such as their SAT scores, gender (GE), a self-report measure of citizenship (CS) and a self-report measure of test effort (TE) were available. Citizenship was a measure of a test taker's willingness to help the university collecting its assessment data, whereas test effort reflected the importance of the test to the test taker. The number of response options for the items varied between 2 and 6.

The 3PL model was chosen as the measurement model for the responses. In the estimation procedure, the same hyperparameter values as in the simulation study above were used to specify vague proper prior knowledge. The model was estimated with 20,000 iterations of the Gibbs sampler, and the first 10,000 iterations were discarded as the burn-in. The odds ratio statistic indicated that for less than 4 % of the possible item combinations there was a significant dependency between two items. The replicated response patterns under the posterior distribution matched the observed data quite well, as shown by the observed sum score statistic. From the posterior residual check it followed that the RT model described the data well. The estimated time discrimination parameters varied over $[.25, 1.65]$, indicating that the items discriminated substantially between test takers of different speed. This result was verified by testing the RT model with $\boldsymbol{\phi} = 1$ against the RT model where $\boldsymbol{\phi} \neq 1$ using the DIC. The estimated DIC's were 85780 and 84831 for the restricted and for the unrestricted RT model, respectively.

Table 2.5 gives the estimated covariance components and correlations between the level 1 parameters. The correlation between the person parameters was esti-

Table 2.5. Estimated covariance components and correlations.

Variance components		EAP	SD	Cor
$\boldsymbol{\Sigma}_P$				
	Σ_{11}	1.00	-	1.00
	Σ_{12}	$-.38$	0.02	-.76
	Σ_{22}	0.25	0.02	1.00
$\boldsymbol{\Sigma}_I$				
	Σ_{11}	0.15	0.04	1.00
	Σ_{12}	$-.11$	0.04	$-.53$
	Σ_{13}	0.05	0.02	0.41
	Σ_{14}	0.02	0.04	0.09
	Σ_{22}	0.33	0.07	1.00
	Σ_{23}	0.06	0.03	0.34
	Σ_{24}	0.07	0.05	0.21
	Σ_{33}	0.10	0.02	1.00
	Σ_{34}	0.10	0.03	0.51
	Σ_{44}	0.35	0.06	1.00

mated to be $-.76$. The Bayes factor for testing the null hypothesis of this correlation being zero, clearly favored the alternative for the range of possible priors given in the simulation study above. Therefore, for this data set, fitting the hierarchical model has to be favored over the alternative of independence between the two constructs. An explanation for this strong negative dependency might be that higher-ability candidates have more insight in their test behavior and, therefore, are better at time management. A negative correlation between speed and ability also often suggests a non-speeded test, because it implies that higher ability test takers who take their time do not run out of time towards the end of the test.

As shown earlier by van der Linden et al. (2007), response times can be a valuable tool for diagnosing differential speededness. Thereby, checks on the assumption of stationary speed during the test are particularly useful. For each test taker, the standardized residuals $e_{ijk} = (t_{ijk} - (\lambda_k - \phi_k \zeta_{ij}))/\tau_k$ were calculated. When the stationary speed assumption holds, a test taker's residual pattern shows randomly varying residuals which almost all will lie between $[-2, 2]$. However, a test taker running out of time will show a deviation of this assumption towards the end of the test. In such a case, this result is misfit of the RT model, because of larger residuals for the test taker on these last items. In Figure 2.2, residual patterns of the RT model for 16 test takers are shown. An aberrancy can be seen in the last figure, where for some items the test taker responded unusually fast. However, a graphical check of the residual patterns of all the test takers did not reveal any structural aberrancies. Therefore, there were no indications of speededness for this test.

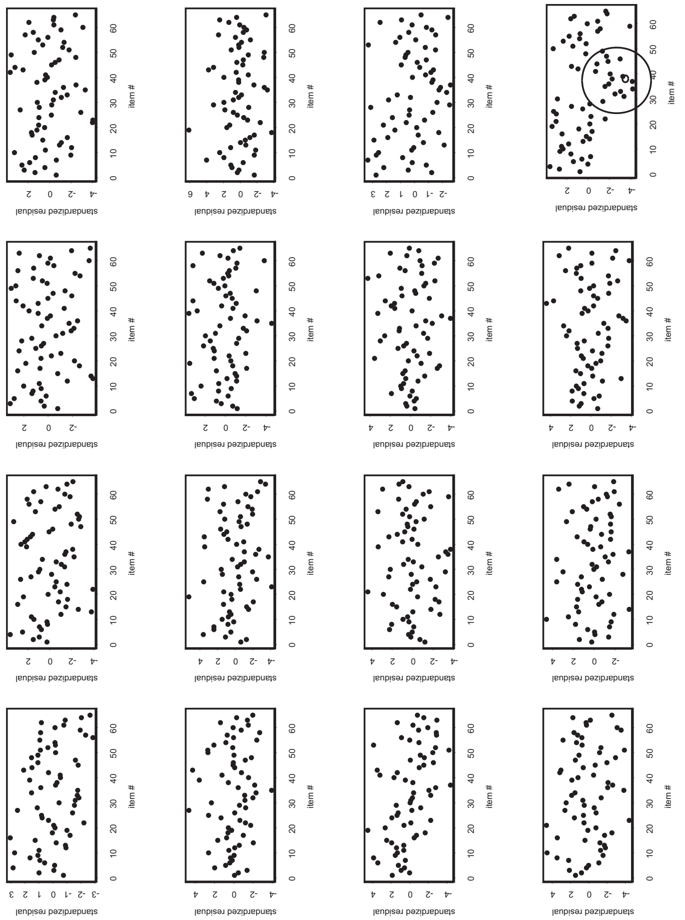


Fig. 2.2. Standardized residual patterns for the RT model for 16 selected test takers.

Subsequently, the following structural model on the person parameters was specified to identify possible relationships of ability and speed with the covariates:

$$\begin{pmatrix}\theta_i\\ \zeta_i\end{pmatrix} = \begin{pmatrix}\gamma_{00}\\ \gamma_{10}\end{pmatrix} + \begin{pmatrix}\mathrm{SAT}_i\gamma_{01} + \mathrm{TE}_i\gamma_{02} + \mathrm{GE}_i\gamma_{03} + \mathrm{CS}_i\gamma_{04}\\ \mathrm{SAT}_i\gamma_{11} + \mathrm{TE}_i\gamma_{12} + \mathrm{GE}_i\gamma_{13} + \mathrm{CS}_i\gamma_{14}\end{pmatrix} + \begin{pmatrix}e_{1i}\\ e_{2i}\end{pmatrix}, \quad (2.45)$$

where $\mathbf{e}_i \sim N(0, \boldsymbol{\Sigma}_P)$. Several hypotheses about this model were tested. First, the composite null hypothesis H_{01} of both γ_{01} and γ_{11} equal zero was tested. Second, the null hypotheses H_{02} that γ_{02} and γ_{12} equal zero were evaluated and, similarly, the hypotheses H_{03} (γ_{03} and γ_{13} equal zero) and H_{04} (γ_{04} and γ_{14} equal zero) were evaluated. Finally, by iterative model building, the composite hypothesis H_{05} of the effects $\gamma_{03}, \gamma_{04}, \gamma_{12}, \gamma_{13}$ and γ_{14} equal to zero was tested. Testing these hypotheses corresponds to comparing models that differ only in their fixed part. This can be easily done via a Bayes factor because, by using a result known as the Savage-Dickey density ratio (Dickey, 1971; Verdinelli & Wasserman, 1995), these Bayes factors are easy to obtain in reduced computation time.

The hypotheses that gender and citizenship had no effect on ability and speed were confirmed. Also, the estimated .95 HPD regions of their effects (and their .90 HPD regions too) included 0, which was another indication that these covariates did not have any explanatory power in speed and ability. However, the SAT scores and test effort explained a significant amount of variation between the person parameters. This result implies the following reduced model:

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix} + \begin{pmatrix} \mathrm{SAT}_i \gamma_{01} + \mathrm{TE}_i \gamma_{02} \\ \mathrm{TE}_i \gamma_{12} \end{pmatrix} + \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}, \tag{2.46}$$

where $\mathbf{e}_i \sim N(0, \boldsymbol{\Sigma}_P)$. The Bayes Factors for the several nested models and the final estimates of the regression parameters are given in Table 2.6.

Table 2.6. Estimated Bayes factors and regression parameters for the structural models

Hypothesis	log(BF)
H_{01}	−5.0
H_{02}	−21.3
H_{03}	16.4
H_{04}	16.3
H_{05}	16.2

Fixed parameters	EAP	SD
γ_{01}	.25	.03
γ_{02}	.25	.03
γ_{12}	-.22	.02

Intuitively, a positive relationship of TE with ability should have been expected. That is, test takers scoring higher on the TE-scale should have been expected to differ from test takers who care less about their results. Also, when the test is relatively more important to the candidate, he/she can be expected to try harder and spend more time on each item to get better results. The negative relationship of TE with speed is also in agreement with this hypothesis since a lower speed results in higher expected RTs. As expected, the SAT score shows a positive relationship with ability. However, there was no significant effect for SAT with respect to the speed of working of the test takers.

2.8 Discussion

A framework for a multivariate multilevel modeling approach was given in which the latent response parameters are measured using conjoint IRT models for the response and response time data. The IRT models for speed and ability are based on the assumption of conditional independence. This means that the ability parameter (speed parameter) is the assumed underlying construct for the response data (response time data). As a result, for each individual, at the level of measurements the responses and response times are local independent given the latent person parameters. The correlation structure between the person parameters is specified at a higher level. The correlation between speed and accuracy in the population of respondents can be tested via a Bayes factor. As the empirical examples showed, the correlation between ability and speed is not necessarily positive. The sign of this relationship will probably depend on the type of test and the test conditions. That is, sometimes hard work will pay off (e.g. a test with strict time limit) while for another setting "take your time" might be the best advice. RTs can give insight about the best strategy of test taking, which is useful information for both test takers and test developers. Other model selection issues related to the structural model on the person parameters can be handled via the proposed DIC which can be computed as a by product of the MCMC algorithm. It was shown that the MCMC algorithm performed well and enabled simultaneous estimation.

The class of multivariate mixted effect models has not received much attention in the literature. Schafer and Yucel (2002) developed an MCMC implementation for the linear multivariate mixed effect model with incomplete data that does converge rapidly for a small number of large groups but it is limited to two levels of nesting. Shah, Laird, and Schoenfeld (1997) extended the EM-algorithm of Laird and Ware (1982) to deal with linear bivariate mixed models. Also, for some applications, it may be possible to stack the columns of the response matrix and apply standard software for univariate mixed models (e.g., *SAS Proc Mixture*; *S-Plus Nlme*). However, this approach quickly becomes impossible when the number of individuals per group and/or the number of variables grows. The MCMC algorithm developed in this project, which is available in *R* from the authors upon request, may help researchers to analyze nonlinear multivariate multilevel mixed response data. This implementation is not limited to small numbers of variables or responses and can handle multiple random effects.

The model in this chapter can easily be extended, for example, to deal with polytomous response data. MCMC algorithms for polytomous IRT models can be found in Fox (2005); Patz and Junker (1999), and Johnson and Albert (1999), among others. The necessary adjustment of the MCMC algorithm consists of replacing the random draws from the parameters in the three-parameter normal-ogive IRT model with those in a polytomous model. Although several studies have shown the log-normal model to yield satisfactory fit to RTs on test items, the hierarchical framework can be used with other measurement models for RTs, for example, to deal with RT distributions with a different skewed or that require heavier tails to be more robust against outliers.

If subpopulations of test takers follow different strategies to solve the items, differences in the joint distribution of accuracy and speed can be expected. To model them, a mixture modeling approach with different latent classes for different strategies can be used (see, for instance, Rost, 1990). This procedure can also be used to relate the popularity of different strategies to covariates.

Finally, the relationship between accuracy and speed may differ across groups of items, for instance, when they are organized in families of cloned items (Glas & van der Linden, 2003) or are presented with a testlet structure (Bradlow, Wainer, & Wang, 1999). In order to deal with such cases, the hierarchical framework has to be extended with a group structure for items. The consequences of this extension and the extensions above for the MCMC method of estimation and hypothesis testing still have to be explored.

2.9 Appendix: Gibbs Sampling Scheme

The Gibbs sampler iteratively samples from the full conditional distributions of all parameters. The full conditional distributions are specified below.

Sampling of Structural Model Parameters.

The sampling of the augmented data is described in (2.12).

As for the person parameters, observe that (2.15) can be written as

$$\begin{bmatrix} \boldsymbol{\Omega}_{ij} \\ \mathbf{z}^*_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{x}^t_{ij}\boldsymbol{\beta}_j \\ \mathbf{H}_P\boldsymbol{\Omega}^t_{ij} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_P & \mathbf{\Sigma}_P\mathbf{H}^t_P \\ \mathbf{H}_P\mathbf{\Sigma}_P & \mathbf{H}_P\mathbf{\Sigma}_P\mathbf{H}^t_P + \mathbf{C} \end{bmatrix} \right), \tag{2.47}$$

where the matrix notation is the same as in (2.27), with $\boldsymbol{z}^*_{ij} = vec(z_{ijk}+b_k, t_{ijk}-\lambda_k)$ and $\mathbf{H}_P = (\mathbf{a} \oplus -\boldsymbol{\phi})$. From the fact that (2.47) is multivariate normal, it follows for the full conditional distribution of the person parameters that

$$\boldsymbol{\Omega}_{ij} \mid \mathbf{z}^*_{ij}, \boldsymbol{\Sigma}_P, \boldsymbol{\beta} \sim N\big(E(\boldsymbol{\Omega}_{ij} \mid \boldsymbol{z}^*_{ij}), var(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}^*_{ij})\big), \tag{2.48}$$

with

$$E\big(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}^*_{ij}, \boldsymbol{\Sigma}_P, \boldsymbol{\beta}\big) = \mathbf{x}^t_{ij}\boldsymbol{\beta}_j + \mathbf{H}_P\boldsymbol{\Sigma}_P(\mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}^t_P + \mathbf{C})^{-1}(\mathbf{z}^*_{ij} - \mathbf{H}_P(\mathbf{x}^t_{ij}\boldsymbol{\beta}_j)^t), \tag{2.49}$$

and

$$var\big(\boldsymbol{\Omega}_{ij} \mid \mathbf{z}^*_{ij}, \boldsymbol{\Sigma}_P, \boldsymbol{\beta}\big) = \boldsymbol{\Sigma}_P - \boldsymbol{\Sigma}_P\mathbf{H}^t_P\big(\mathbf{H}_P\boldsymbol{\Sigma}_P\mathbf{H}^t_P + \mathbf{C}\big)^{-1}\mathbf{H}_P\boldsymbol{\Sigma}_P \tag{2.50}$$

This result involves an efficient sampling scheme since the values of both person parameters are obtained in just one step.

The derivation of the full conditional distribution of regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is analogous. From (2.47) and (2.9), it follows that the $\boldsymbol{\beta}_j$s are multivariate normal with mean

$$E\big(\boldsymbol{\beta}_j \mid \boldsymbol{\Omega}_j, \boldsymbol{\Sigma}_P, \mathbf{V}, \boldsymbol{\gamma}\big) = \mathbf{w}_j\boldsymbol{\gamma} + \mathbf{x}_j\mathbf{V}\big(\mathbf{x}_j\mathbf{V}\mathbf{x}^t_j + \boldsymbol{\Sigma}_P\big)^{-1}(\boldsymbol{\Omega}_j - \mathbf{x}_j\mathbf{w}_j\boldsymbol{\gamma}), \tag{2.51}$$

and variance,

$$var\big(\boldsymbol{\beta}_j \mid \boldsymbol{\Omega}_j, \boldsymbol{\Sigma}_P, V\big) = \mathbf{V} - \mathbf{V}\mathbf{x}_j^t\big(\mathbf{x}_j\mathbf{V}\mathbf{x}_j^t + \boldsymbol{\Sigma}_P\big)^{-1}\mathbf{x}_j\mathbf{V}. \tag{2.52}$$

Likewise, from (2.9) and (2.21), the (fixed) coefficients $\boldsymbol{\gamma}$ are multivariate normal distributed with mean

$$E\big(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{V}, \mathbf{V}_0, \kappa_{V_0}, \boldsymbol{\gamma}_0\big) = \boldsymbol{\gamma}_0 + \mathbf{w}\mathbf{V}^*\big(\mathbf{w}\mathbf{V}^*\mathbf{w}^t + \mathbf{V}\big)^{-1}\big(\boldsymbol{\beta} - \mathbf{w}\boldsymbol{\gamma}_0\big), \tag{2.53}$$

and variance

$$var\big(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{V}, \mathbf{V}_0, \kappa_{V_0}\big) = \mathbf{V}^* - \mathbf{V}^*\mathbf{w}^t\big(\mathbf{w}\mathbf{V}^*\mathbf{w}^t + \mathbf{V}\big)^{-1}\mathbf{w}\mathbf{V}^*, \tag{2.54}$$

where $\mathbf{V}^* = \mathbf{V}/\kappa_{V_0}$.

The full conditional distribution of covariance matrix $\boldsymbol{\Sigma}_P$ was already introduced in the section on the identifying prior structure.

Sampling of the Remaining Parameters.

As for the item parameters, a regression structure analogous to that of the person parameters in (2.15) can be found. Let $\boldsymbol{\Lambda}_k = (a_k, b_k, \phi_k, \lambda_k)^t$ and $\mathbf{H}_I = (\boldsymbol{\theta}, -\mathbf{1}_N) \oplus (-\boldsymbol{\zeta}, \mathbf{1}_N)$. The item parameters are the coefficients of the regression of $\boldsymbol{z}_k^*$ on $\mathbf{H}_I$. Combined with the prior in (2.11), this observation leads to a multivariate normal posterior distribution of the item parameters with mean,

$$E\big(\boldsymbol{\Lambda}_k \mid \mathbf{z}_k^*, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_I\big) = \boldsymbol{\mu}_I + \mathbf{H}_I\boldsymbol{\Sigma}_I\big(\mathbf{H}_I\boldsymbol{\Sigma}_I\mathbf{H}_I^t + \boldsymbol{C}_{2K}\big)^{-1}\big(\boldsymbol{z}_k^* - \mathbf{H}_I\boldsymbol{\mu}_I\big) \tag{2.55}$$

and variance,

$$var(\boldsymbol{\Lambda}_k \mid \mathbf{z}_k^*, \boldsymbol{\Omega}, \boldsymbol{\Sigma}_I) = \boldsymbol{\Sigma}_I - \boldsymbol{\Sigma}_I\mathbf{H}_I^t\big(\mathbf{H}_I\boldsymbol{\Sigma}_I\mathbf{H}_I^t + \boldsymbol{C}_{2K}\big)^{-1}\mathbf{H}_I\boldsymbol{\Sigma}_I, \tag{2.56}$$

where $\boldsymbol{C}_{2K} = \mathbf{I}_K \oplus \mathbf{I}_K\boldsymbol{\tau}^2$.

The parameters of the distribution of the item parameters follow a multivariate normal distribution; see (2.11). The normal inverse-Wishart prior in (2.19) and (2.20) is conjugate for the multivariate normal distribution (Gelman et al., 2004). The resulting posterior distribution also belongs to the normal inverse-Wishart family:

$$p(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \mid \boldsymbol{\Lambda}, \boldsymbol{\mu}_{I_0}, \boldsymbol{\Sigma}_{I_0}, \kappa_{I_0}, \nu_{I_0}) \sim N - Inv - Wishart\big(\boldsymbol{\mu}_*, \boldsymbol{\Lambda}^*, \kappa, \nu\big) \tag{2.57}$$

with parameters:

$$\begin{aligned}
\boldsymbol{\mu}_* &= \frac{\kappa_{I_0}}{\kappa_{I_0} + K}\boldsymbol{\mu}_{I_0} + \frac{K}{\kappa_{I_0} + K}\bar{\boldsymbol{\Lambda}},\\
\kappa &= \kappa_{I_0} + K,\\
\nu &= \nu_{I_0} + K,\\
\boldsymbol{\Lambda}^* &= \boldsymbol{\Sigma}_{I_0} + \mathbf{S} + \frac{\kappa_{I_0}K}{\kappa_{I_0} + K}\big(\bar{\boldsymbol{\Lambda}} - \boldsymbol{\mu}_{I_0}\big)\big(\bar{\boldsymbol{\Lambda}} - \boldsymbol{\mu}_{I_0}\big)^t,
\end{aligned}$$

where $\mathbf{S} = \sum_{k=1}^{K} \left(\boldsymbol{\Lambda}_k - \bar{\boldsymbol{\Lambda}}\right)\left(\boldsymbol{\Lambda}_k - \bar{\boldsymbol{\Lambda}}\right)^t$.

Likewise, for the fixed parameters $\boldsymbol{\gamma}$, an Inverse-Wishart prior was specified. So the posterior is an Inverse-Wishart

$$V \mid \boldsymbol{\gamma}_0, \boldsymbol{\beta}, \nu_{V_0}, \kappa_{V_0}, \mathbf{V}_0 \sim Inv - Wishart_{\nu_V}\left(\boldsymbol{\Psi}^{-1}\right) \tag{2.58}$$

with parameters

$$\begin{aligned}
\nu_V &= \nu_{V_0} + J \\
\boldsymbol{\Psi} &= \mathbf{V}_0 + \boldsymbol{S} + \frac{\kappa_{V_0} J}{\kappa_{V_0} + J}\left(\mathbf{w}\boldsymbol{\gamma} - \mathbf{w}\boldsymbol{\gamma}_0\right)\left(\mathbf{w}\boldsymbol{\gamma} - \mathbf{w}\boldsymbol{\gamma}_0\right)^t \\
\boldsymbol{S} &= \sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{w}_j\gamma\right)\left(\boldsymbol{\beta}_j - \mathbf{w}_j\gamma\right)^t.
\end{aligned}$$

For the 3PL model, an additional augmentation step is introduced according to Beguin and Glas (2001). A variable $s_{ijk} = 1$ when a person ij knows the correct answer to question k and is $s_{ijk} = 0$ otherwise. Its conditional probabilities are given by (2.12). Subsequently, $z_{ijk} \sim N(a_k\theta_{ij} - b_k, 1)$, truncated at the left of 0 when $s_{ijk} = 0$ and truncated at the right when $s_{ijk} = 1$.

It was already noted that the posterior of the guessing parameters is a Beta distribution:

$$c_k \sim Beta(b_1' + s_k, b_2' + n_k - s_k), \tag{2.59}$$

where n_k is the number of people who do not know the answer and s_k is the number of people who guessed the answer correctly.

For the residual variance of the RT model τ_k^2, with an Inverse-Gamma prior, the posterior is again an Inverse-Gamma distribution with parameter $g_1 + N/2$ and scale parameter $g2 + (\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))^t(\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))/2$.

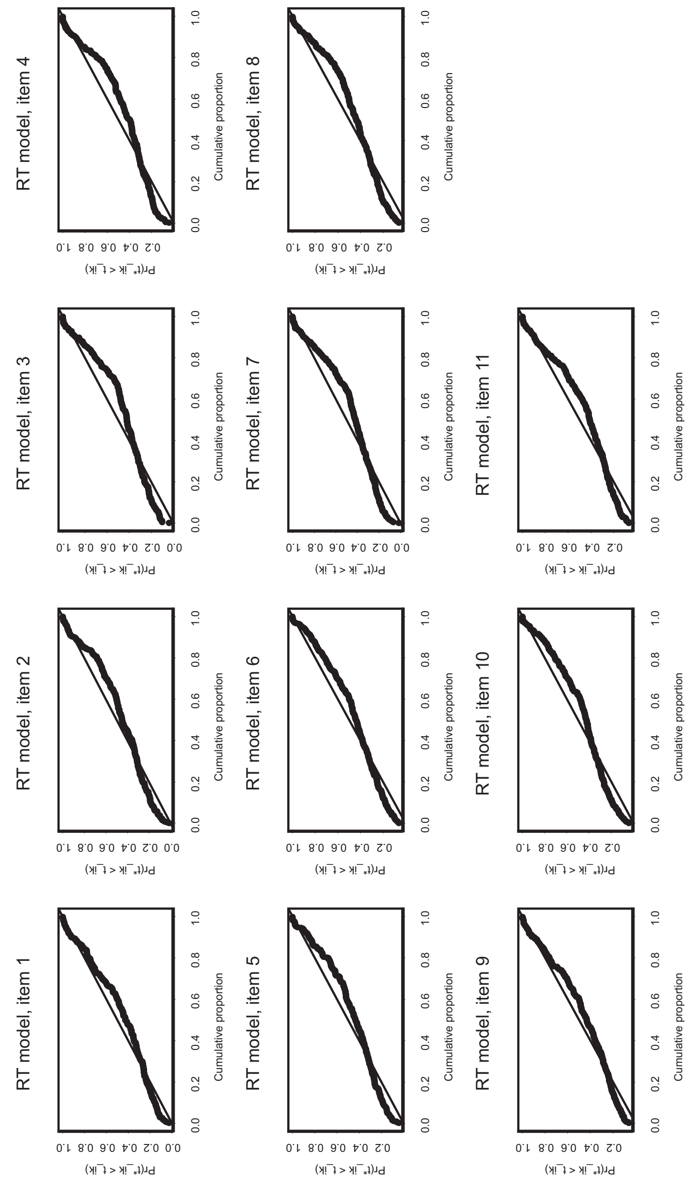


Fig. 2.3. Probabilities of $P(t^*_{ik} < t_{ik}|\mathbf{y}, \mathbf{t})$ against their expected values under the $U(0, 1)$ distribution for the 11 items of the neuroticism scale, Example 2.

3

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

Summary. The analysis of performance in computer-based educational assessment is often confined to accuracy scores. Response times, although being an additional source of information, are either neglected or analyzed separately. In this chapter, a model is developed that allows the simultaneous analysis of accuracy scores and response times of cognitive tests with a rule-based design. Further, the model is capable of determining both time intensity and difficulty of design rules in the test, thus dissociating information that is often confounded in current assessment procedures. This allows a better understanding of the relationships between item characteristics and item content. The application of the model is illustrated using a large-scale investigation of figural reasoning ability.

3.1 Introduction

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

The systematic combination of educational assessment techniques with response time analysis remains a scarcity in the literature. The purpose of the present article is to present a model which allows the integration of response time 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 will 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 response time 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 demonstrate its application with an empirical example.

3.1.1 Cognitive Theory in Educational Assessment

One of the core interests of psychological research pertains to the analysis of cognitive processes. Research paradigms in cognitive psychology often assume that in order 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 based on the processing times necessary for specific processes. For example, an important strand of research in cognitive psychology is concerned with analyzing parameters from individual response time distributions on simple tasks, and these parameters can be theoretically connected to psychological processes like 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, e.g., 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 response time 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, in press; 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 response times *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, and Zhou, 2003; but see Ratcliff, 1978 for an approach that allows to simultaneously model experimental accuracy and response time 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 in order 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 in order to provide meaningful results. In order 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 in order 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). Once a set of cognitive rules affecting item complexity has been defined based on 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 will be 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, i.e. 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. Leighton & Gierl, 2007; Junker & Sijtsma, 2001).

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 9 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 which meaningfully connect these elements across cells, and to correctly apply these rules in order to find the content of the empty cell. A typical item found in such a test is given in Figure 3.1.

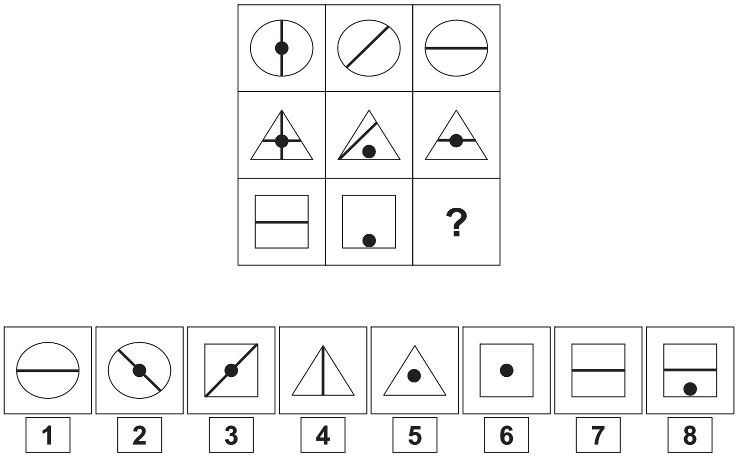


Fig. 3.1. Example of a figural reasoning item.

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) report taxonomies of rules utilized in the item design of existing figural matrices tests. In a review of the current literature, Primi (2001) describes four main design factors (”radicals” according to Irvine, 2002) which 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 response times (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) assume that in addition to working memory capacity, abstraction capacity, i.e. 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. Based on 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, where the latter introduce conflicting combinations between visual and conceptual figural elements, whereas the former display more congruent relationships. Primi (2001) 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. (1990) and Embretson (1998) found larger effects of item features relating to working memory.

3.1.2 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 and Mislevy, in press). One of the most basic IRT model is the Rasch model (Rasch, 1980). 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

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{exp(\theta_i - b_k)}{1 + exp(\theta_i - b_k)}, \tag{3.1}$$

where θ_i denotes the ability of test taker i and b_k 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 underly 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

$$P(Y_{ik} = 1|\theta_i, q_k, \boldsymbol{\eta}) = \frac{exp(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)}{1 + exp(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)}, \tag{3.2}$$

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 in order 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 model (2PL model; Lord & Novick, 1968) is defined as

$$P(Y_{ik} = 1|\theta_i, a_k, b_k) = \frac{exp(a_k(\theta_i - b_k))}{1 + exp(a_k(\theta_i - b_k))}, \tag{3.3}$$

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

$$P(Y_{ik} = 1|\theta_i, q_k, \boldsymbol{\eta}, \boldsymbol{\tau}) = \frac{exp[\sum_{j=1}^{J} q_{kj}\tau_j(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)]}{1 + exp[\sum_{j=1}^{J} q_{kj}\tau_j(\theta_i - \sum_{j=1}^{J} q_{kj}\eta_j)]}, \tag{3.4}$$

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, i.e. 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). In order to take this into account, an error term must be introduced into the model. Janssen, Schepers, and Peres (2004) present an application of this approach for the LLTM, where item difficulty b_k is decomposed as

$$b_k = \sum_{j=1}^{J} q_{kj}\eta_j + \epsilon_k, \tag{3.5}$$

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, i.e. 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 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.

3.1.3 Response Times in Educational Assessment

Traditional data analysis in educational assessment is founded on accuracy scores. Results obtained in classical, unidimensional IRT models like 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 and a discrimination parameter are obtained (Embretson & Reise, 2000). In such models, response times are not modeled. However, response times are easily available in times of computerized testing, and they can contain important information beyond accuracy scores. For example, response times 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, response times in psychometric tests with a design potentially contain vital information concerning the underlying cognitive processes. Importantly, response time 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? In order to investigate these questions, a unified treatment of accuracy scores and response times is required.

Three different strategies have been used in the past to extract response time-related information from psychometric tests. Under the first strategy, response times are modeled *exclusively*. This strategy is usually applied to speed tests which 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) suggests that the response time T for person i responding to item k is exponentially distributed with density

$$f(t_{ik}) = (\tau_i + \gamma_k)exp[-(\tau_i + \gamma_k)t_{ik}], \tag{3.6}$$

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:

$$\gamma_k = \sum_{j=1}^{J} a_{kj}\eta_j, \tag{3.7}$$

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, based on the gamma distribution. Note that these models focus on response times exclusively, whereas accuracy scores are not taken into consideration.

A second strategy chosen by several authors implies a *separate* analysis of response times 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 response times 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 ANOVAs to predict response times 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 response time distributions of item isomorphs, i.e. parallel items, in two figural reasoning test forms. Separate analyses provide some information on both accuracy scores and response times, 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 response time parameters and IRT parameters. This has been done in a third strategy of analyses based on the *joint modeling* of both response times and accuracy scores. Recently, several models have been proposed for the investigation of response times in a psychometric test within an IRT framework. One of the first models was introduced by Thissen (1983), which describes the log-transformed response time of person i to item k as

$$log(T_{ik}) = \upsilon + s_i + u_k - bz_{ik} + \epsilon_{ik}, \tag{3.8}$$

with $\epsilon_{ik} \sim N(0, \sigma_\epsilon^2)$. In this model, υ reflects the overall mean log response time, s_i and u_k are person- and item-related slowness parameters, respectively, $-b$ represents the log-linear relation between response time 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 response time. 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.

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

$$P(Y_{ik} = 1|T_{ik}) = \frac{\theta_i T_{ik}}{\theta_i T_{ik} + \varepsilon_k} = \frac{exp(\xi_i + \tau_{ik} - \sigma_k)}{1 + exp(\xi_i + \tau_{ik} - \sigma_k)}, \tag{3.9}$$

where θ_i represents the person ability, ε_k is item difficulty, T_{ik} is response time, and ξ_i, τ_{ik} and σ_k represent the natural logarithms of θ_i, T_{ik} and ε_k, respectively. In this model, response time is parametrized 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 (3PL) model by including response times as well as parameters reflecting item slowness and person slowness, respectively:

$$P(Y_{ik} = 1|\theta_i, \rho_i, a_k, b_k, c_k, d_k, T_{ik}) = c_k + \frac{1 - c_k}{1 + e^{(-1.7a_k[\theta_i - (\rho_i d_k / t_{ik}) - b_k])}}, \tag{3.10}$$

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 response time of subject i on item k. In this model, response times are treated as an additional predictor, but in contrast to the model by Roskam (1997), as response time goes to infinity, a classical 3PL 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 response times, which are modeled as fixed parameters. It is a common assumption across the literature that response time 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 response times 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 lognormalized response times as well as the log-odds of correct responses simultaneously. For the log-odds, the model assumed the lognormalized response times and item-related design parameters with random effects (Rijmen & DeBoeck, 2002) as predictors. Similarly, the response times 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 response times into account, but did not specify parameters reflecting the test-taker's speed or the time intensity of the items. If response times 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 (2005), 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 response times, 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 response times and accuracy scores simultaneously. A model that can overcome these difficulties, based on the model developed by van der Linden (2007), will be described below.

3.2 A Model for Response Accuracies and Response Times

With responses and response times (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 item response theory. A framework will be developed that deploys separate models for the responses and the response times 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 (see Figure 4). This hierarchical modeling approach was recently introduced by van der Linden (2007). The focus of this chapter, 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 two-parameter normal ogive model (Lord & Novick, 1968):

$$P(Y_{ik} = 1 | \theta_i, a_k, b_k) = \Phi(a_k \theta_i - b_k), \tag{3.11}$$

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. $\Phi(\cdot)$ denotes the cumulative normal distribution function. The 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:

$$Z_{ik} = a_k \theta_i - b_k + \epsilon_{\theta_{ik}}, \tag{3.12}$$

where $Z_{ik} \geq 0$ when $Y_{ik} = 1$ and $Z_{ik} < 0$ otherwise and with $\epsilon_{\theta_{ik}} \sim N(0, 1)$. 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 will be 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 response times 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 (1997) and van der Linden (2006) 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 response times 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, since 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 a 100 meters will be less time consuming than running 200 meters. 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 3.2. In this figure, Item Characteristic Curves (ICC) for the IRT model (left figure) and Response Time Characteristic Curves (RTCC) (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 the 200 meters, while the lower curve reflects the expected RTs on the 100 meters distance. Note, however, that it is not necessarily so that running 200 meters is more difficult than the 100 meters.

Now for the expectation of the log-response time 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 3.12. The effect of item discrimination on the ICCs and RTCCs are illustrated in Figure 3.3. 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-response time T_{ik} of person i on item k follows a normal model according to:

$$T_{ik} = -\phi_k \zeta_i + \lambda_k + \epsilon_{\zeta_{ik}}, \tag{3.13}$$

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

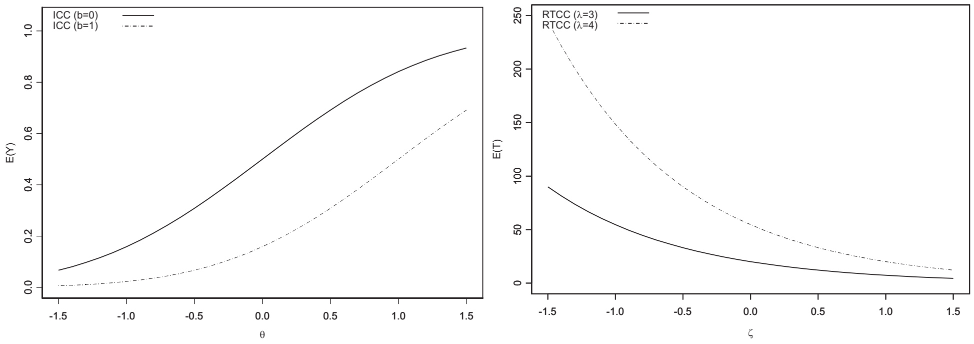


Fig. 3.2. ICC (left) and RTCC (right) curves for two items with different time intensity and difficulty but equal discrimination parameters.

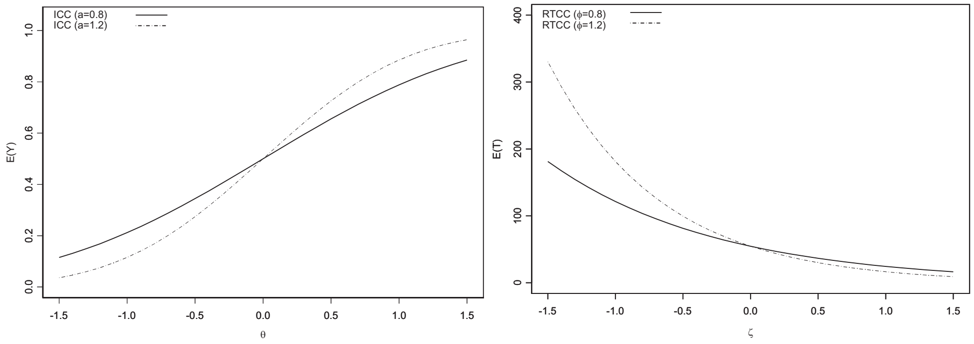


Fig. 3.3. ICC and RTCC curves for two items with differing discrimination, where $b = 0$ and $\lambda = 4$

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

$$\theta_i \sim N(\mu_\theta, \sigma_\theta^2) \tag{3.14}$$

However, together with the RT model there are now two traits that describe each test taker, ability and speed. At the second level of modeling, these person parameters are assumed to follow a bivariate normal distribution:

$$(\theta_i, \zeta_i) = \boldsymbol{\mu}_P + \boldsymbol{e}_P, \boldsymbol{e}_P \sim N(\mathbf{0}, \boldsymbol{\Sigma}_P), \tag{3.15}$$

where $\boldsymbol{\mu}_P = (\mu_\theta, \mu_\zeta)$ and the covariance structure is specified by:

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}. \tag{3.16}$$

The parameter ρ in the model for the person parameters reflects possible dependencies between speed and accuracy of the test takers. For instance, when ρ is negative, this means that persons who work faster than average on the test are expected to have below-average abilities. When $\rho = 0$, there is independence between ability and speed. However, this is not necessarily equivalent to independence between the responses and RTs, since such a dependency can occur via the item side of the model as well, as will be 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 3.15 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 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 3.15 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 will be developed now and will present 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 chapter.

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, 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 will present an approach where 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. Since 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 3.15, the vector $\boldsymbol{\xi}_k = (a_k, b_k, \phi_k, \lambda_k)$ is assumed to follow a multivariate normal distribution,

$$\boldsymbol{\xi}_k \sim N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) \tag{3.17}$$

where $\boldsymbol{\Sigma}_I$ specifies the covariance structure of the item parameters:

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{a\phi} & \sigma_{b\phi} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{a\lambda} & \sigma_{b\lambda} & \sigma_{\phi\lambda} & \sigma_\lambda^2 \end{bmatrix}. \tag{3.18}$$

$\boldsymbol{\Sigma}_I$ 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 where items are formulated using either squares, circles or triangles and we are interested if as such items differ in their difficulty. This leads us to consider the following model, where we develop an ANOVA 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 $\mathbf{A}_1$ of length K, contains a 1 for circles, a 0 for triangles and -1 for squares. The second variable $\mathbf{A}_2$ contains a 0 for circles, a 1 for triangles and also a -1 for squares. Now, following Equtaion 3.5, the difficulty of item k can be modeled as

$$b_k = \gamma_0^{(b)} + A_{1k}\gamma_1^{(b)} + A_{2k}\gamma_2^{(b)} + e_k^{(b)}. \tag{3.19}$$

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 $\gamma_0 + \gamma_1$. Are there triangles used, its difficulty is given by $\gamma_0 + \gamma_2$. In the case the 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 $-\gamma_1 - \gamma_2$ since otherwise the model is over parameterized. Let $\mathbf{A} = (\mathbf{1}, \mathbf{A}_1, \mathbf{A}_2)$ and $\boldsymbol{\gamma}^{(b)} = (\gamma_0, \gamma_1, \gamma_2)^t$, then the model can be represented as

$$b_k = \mathbf{A}_k \boldsymbol{\gamma}^{(b)} + e_k^{(b)} \tag{3.20}$$

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:

$$a_k = \mathbf{A}_k \boldsymbol{\gamma}^{(a)} + e_k^{(a)} \tag{3.21}$$

$$b_k = \mathbf{A}_k \boldsymbol{\gamma}^{(b)} + e_k^{(b)} \tag{3.22}$$

$$\phi_k = \mathbf{A}_k \boldsymbol{\gamma}^{(\phi)} + e_k^{(\phi)} \tag{3.23}$$

$$\lambda_k = \mathbf{A}_k \boldsymbol{\gamma}^{(\lambda)} + e_k^{(\lambda)}, \tag{3.24}$$

where the error terms are assumed to follow a MVN distribution, that is $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I)$. This is a generalization of Equation 3.5, not only by allowing for residual variance in other item parameters than $\boldsymbol{b}$, but by modeling covariance components between the item parameters as well. Further, $\mathbf{A}$ is a design matrix containing zeros and ones denoting which construction rules are used for each item, $\boldsymbol{\gamma}^{(a)}, \boldsymbol{\gamma}^{(b)}, \boldsymbol{\gamma}^{(\phi)}, \boldsymbol{\gamma}^{(\lambda)}$ 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 3.4 below. 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 $\mathbf{A}_I$ 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 $\boldsymbol{\Sigma}_I$ and $\boldsymbol{\Sigma}_P$ 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.

3.3 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,

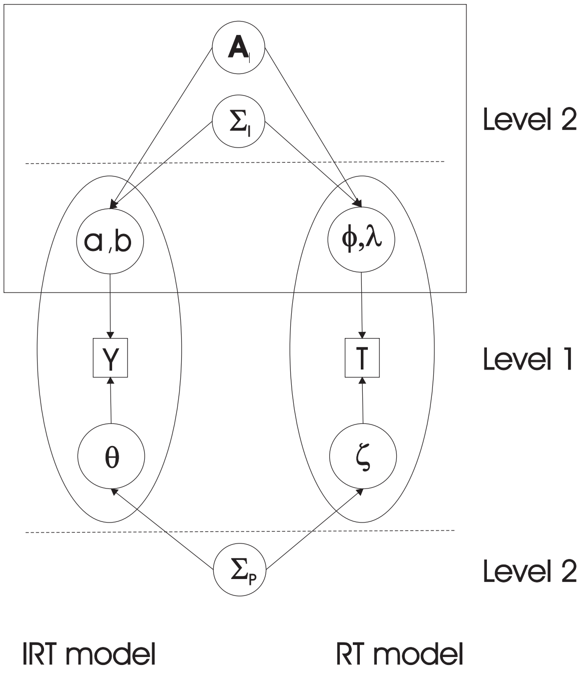


Fig. 3.4. Schematic representation of the modeling structure.

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 et al. (2004). Bayesian estimation of IRT models is discussed in, for instance, Albert (1992), Patz and Junker (1999) and Fox and Glas (2001).

3.3.1 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. Based on 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 $f(\boldsymbol{x}|\mu)$ denote the sampling distribution, then the posterior density of $\mu|\mathbf{x}$ is

$$p(\mu|\mathbf{x}) = f(\mathbf{x}|\mu)p(\mu)/m(\mathbf{x}), \tag{3.25}$$

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

3.3.2 Markov Chain Monte Carlo 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 Markov Chain Monte Carlo (MCMC) methods, however, solve these problems easily. A strong feature of these methods is that their application remains straightforward, while model complexity may increase.

The MCMC algorithm applied in this paper 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 towards 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, 1999).

To illustrate the approach, consider estimation of the RT model given by Equation 3.13. For simplicity of this example, we assume that $\boldsymbol{\phi} = \mathbf{1}$ and independence from the response model. First, (independent) prior distributions for $\boldsymbol{\zeta}, \boldsymbol{\lambda}$ and $\boldsymbol{\sigma}^2$ are specified. Now, since it does not depend on $m(\mathbf{t})$ up to some constant, the posterior distribution is proportional to $p(\boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2|\mathbf{t}) \propto f(\mathbf{t}|\boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2)p(\boldsymbol{\zeta})p(\boldsymbol{\lambda})p(\boldsymbol{\sigma}^2)$. After providing the algorithm with starting values $\boldsymbol{\zeta}^{(0)}, \boldsymbol{\lambda}^{(0)}$ and $\boldsymbol{\sigma}^{2^{(0)}}$, the algorithm proceeds as follows:

- At iteration m, draw the person parameters $\boldsymbol{\zeta}$ from $p(\boldsymbol{\zeta}|\boldsymbol{\lambda}^{(m-1)}, \boldsymbol{\sigma}^{2^{(m-1)}}, \mathbf{t})$.
- Using the new values $\boldsymbol{\zeta}^{(m)}$, draw $\boldsymbol{\lambda}$ from $p(\boldsymbol{\lambda}|\boldsymbol{\zeta}^{(m)}, \boldsymbol{\sigma}^{2^{(m-1)}}, \mathbf{t})$.
- Using the new values $\boldsymbol{\lambda}^{(m)}$, draw $\boldsymbol{\sigma}^2$ from $p(\boldsymbol{\sigma}^2|\boldsymbol{\zeta}^{(m)}, \boldsymbol{\lambda}^{(m)}, \mathbf{t})$.
- 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 like 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 if 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 if 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). Since 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 chapter.

Model Checking and Evaluation

In a Bayesian framework, goodness of fit tests can be performed using posterior predictive checks (Gelman, Meng, & Stern, 1996; Gelman et al., 2004). Model fit can be evaluated by comparing replications of the data $\mathbf{x}^{rep}$, 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 $T(\mathbf{x}|\boldsymbol{\mu})$ (for example, mean squared error), where $\mathbf{x}$ denotes the data and $\boldsymbol{\mu}$ 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:

$$p^* = P(T(\mathbf{x}^{rep}, \boldsymbol{\mu}) \geq T(\mathbf{x}, \boldsymbol{\mu})|\mathbf{x}), \tag{3.26}$$

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 if 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 et al. (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

$$u_{ik} \approx \sum_{m=0}^{M} \Phi\big(t_{ik} | \zeta_i^{(m)}, \phi_k^{(m)}, \lambda_k^{(m)}\big)/M, \tag{3.27}$$

from M iterations from the MCMC chain. According to the probability integral transform theorem (Casella & Berger, 2002, pp. 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 criterium 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:

$$BF = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}. \tag{3.28}$$

The marginal likelihood is the average of the density of the data taken over all parameter values admissible by the prior. That is: $p(\mathbf{y}|M) = \int p(\mathbf{y}|\boldsymbol{\gamma}, M)p(\boldsymbol{\gamma}|M)d\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the vector of model parameters. Since 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, while 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: $p(\mathbf{y}|M_1, \boldsymbol{\gamma} = \mathbf{0}) = p(\mathbf{y}|M_0)$. When this special case holds, computation of the Bayes Factor for testing M_1 versus M_0 simplifies to evaluating the marginal posterior density $p(\boldsymbol{\gamma}|\mathbf{y}, M_1)$ at $\boldsymbol{\gamma} = \mathbf{0}$. This result is known as the Savage-Dickey density ratio (Dickey, 1971; Verdinelli & Wasserman, 1995):

$$BF = \frac{p(\boldsymbol{\gamma} = \mathbf{0}|\mathbf{y}, M_1)}{p(\boldsymbol{\gamma} = 0|M_1)}, \tag{3.29}$$

where $p(\boldsymbol{\gamma} = 0|M_1)$ is the evaluation of the restriction under the prior density of model M_1. Using this result greatly reduces the computational burden, since 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 3.23

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

denoting the regression at level 2 of the model parameters $\boldsymbol{b}$ on design matrix $\mathbf{A}$, where we dropped the superscript (b) in the error term for the moment. Then, the proportion explained variance in the $\boldsymbol{b}$-parameters is given by

$$R^2 = 1 - \frac{E\big(\frac{1}{K-1}\sum_{i=1}^{K} e_k^2\big)}{E\big(\frac{1}{K-1}\sum_{i=1}^{K}(b_k - \bar{b})^2\big)}, \tag{3.30}$$

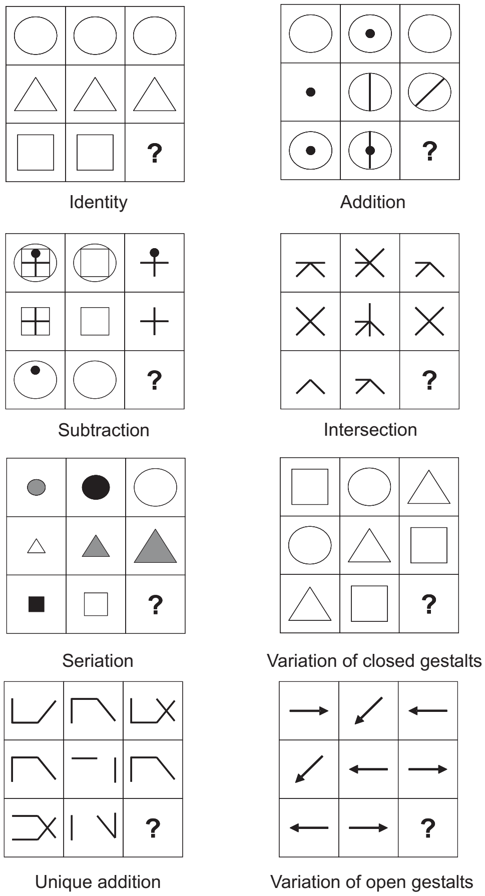
where E denotes the posterior mean. Using the MCMC algorithm, these expectations can be obtained by averaging over the draws from the posterior distribution. When the model explains almost all variability in $\boldsymbol{b}$, the R^2 statistic will be close to 1. If the R^2 statistic is close to 0, then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

3.4 Empirical Example

The application of the model is illustrated using a large-scale investigation of figural reasoning ability, based on 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, not only will we try to validate the cognitive model on the item difficulties, but also by examining the time intensities of the items.

3.4.1 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 distinguish 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, substraction, intersection, seriation, variation of closed gestalts, unique addition and variation of open gestalts (see Figure 3.5). 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 subtraction requires that elements occuring 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



Identity

Addition

Subtraction

Intersection

Seriation

Variation of closed gestalts

Unique addition

Variation of open gestalts

Fig. 3.5. Set of operations used in item construction.

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 3.5). 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 3.5: The rule refers to shape, but not to texture). In solving items with embedded components, an examinee must conduct an additional mental search operation in order to discover to which part of a figural element a rule relates (see unique addition in Figure 3.5). Finally, Hornke and Habon (1986) allowed relations between figural elements to occur across rows, columns or both (e.g., see addition in Figure 3.5 for columnwise direction and subtraction for row- and columnwise direction, respectively). As can be seen from Figure 1, which represents a typical item from the test, 8 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).

3.4.2 Data Set

Data from 30,000 examinees and 456 items were available. 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 in order to have all design rules sufficiently present. Since the links between the groups had to be maintained, the first 6422 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, since 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, based on these estimates. However, for this example it is preferred to analyze the accuracies and response times jointly, since 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).

3.4.3 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:

$$b_k = \gamma_0^{(b)} + \gamma_{rule1}^{(b)} + \gamma_{rule2}^{(b)} + \gamma_{p.o.}^{(b)} + \gamma_{rc}^{(b)} + e_k^{(b)} \tag{3.31}$$

$$\lambda_k = \gamma_0^{(\lambda)} + \gamma_{rule1}^{(\lambda)} + \gamma_{rule2}^{(\lambda)} + \gamma_{p.o.}^{(\lambda)} + \gamma_{rc}^{(\lambda)} + e_k^{(\lambda)}, \tag{3.32}$$

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), $\gamma_{p.o.}$ the effect of the perceptual organization, γ_{rc} the effect of the row-wise or colum-wise operation and e_k models the unexplained variability.

We will restrict ourselves to the difficulty and time intensity parameters for practical reasons, since 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, since the design matrix $\mathbf{A}$ involves many parameters. Necessarily, also interaction effects like 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 3.31 and 3.32 it is possible to formulate more restricted models to test some hypotheses. The following four models will be considered:

- Let M_0 denote the restricted model where $(b_k, \lambda_k) = (\gamma_0^{(b)}, \gamma_0^{(\lambda)}) + (e_k^{(b)}, e_k^{(\lambda)})$. It assumes that there is no explanatory effect for difficulty or time intensity resulting from the cognitive design.
- 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.
- Model M_2 extends model M_1 by including also the design rules (two per item) that were used to formulate each item.
- 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 if there results any effect on difficulty or time intensity from the row wise, column wise, or row and column wise organization of the rules.

3.4.4 Design Matrix

To construct the design matrix $\mathbf{A}$ 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 if 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.

3.5 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 will be evaluated.

Estimation

Identification of the model is obtained by setting $\boldsymbol{\mu}_P = (\mu_\theta, \mu_\zeta) = (0, 0)$, 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, in order 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 asses convergence of the MCMC chains. Graphical checks like 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 if the fitted response model was able to describe the observed response patterns. To check if the introduction of an item discrimination parameter was necessary, the fit of the Rasch model was

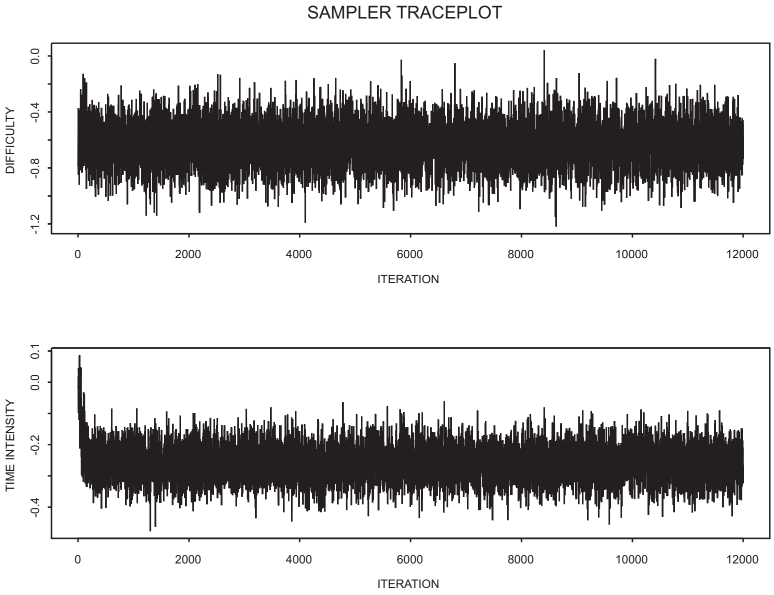


Fig. 3.6. Trace plots for the effect of the identity rule on difficulty (above) and time intensity (below) under model M_2.

compared with that of the 2-paramater normal ogive IRT model. 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 $0, \ldots, K$ items correct. Figure 3.7 shows this statistic for the 2-parameter model, where the line denotes the observed number of test takers with k correct items and the dots with 95% HPD intervals denote the summary of the 1,000 replications under the model. It appeared that the Rasch model was unable to capture the observed data patterns, but from Figure 7 can be seen that the 2-paramater normal ogive model performed quite well. The odds ratio statistic pointed at some item combinations where 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 correct scores. Only for a low percentage of all the possible item combinations (2.6 %) a significant p-value was found.

To test if 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 3.27. 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, since no systematic aberrancies were found, it was concluded that model fit was satisfactory for this data set.

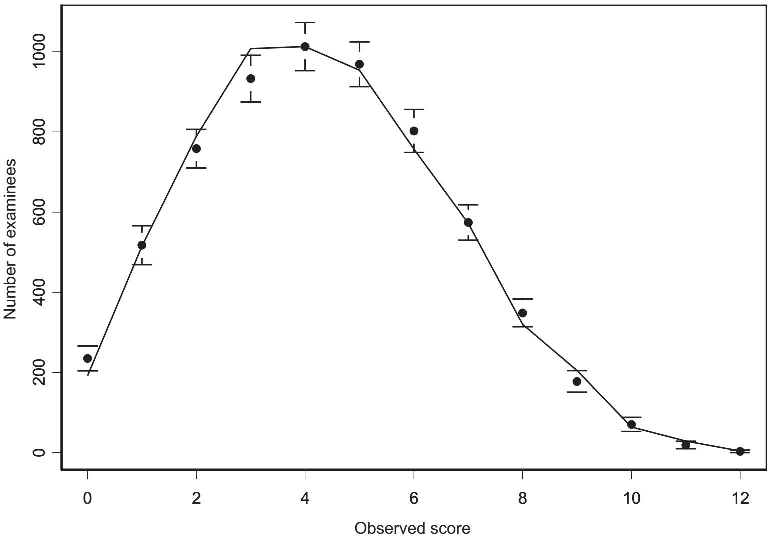


Fig. 3.7. Observed sum scores (line) and model predicted sum scores (dots with .95 HPD regions).

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 3.1 gives the posterior means (EAP) and posterior standard deviations (SD) for the variance and covariance parameters of $\boldsymbol{\Sigma}_P$ and $\boldsymbol{\Sigma}_I$. 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 3.15) 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 indentification restrictions. The estimated correlation between ability and speed was strongly negative (-.61). Interestingly, this tells us that in general it were the higher ability candidates who took more time to solve the items.

Similarly, the estimates for $\boldsymbol{\Sigma}_I$ contain information about the items in the test. From Table 3.1 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 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.

Table 3.1. Estimated covariance components and correlations, obtained for model M_0.

Variance components	EAP	SD	cor
$\boldsymbol{\Sigma}_P$			
σ^2_θ	1.00	-	1.00
ρ	$-.30$	0.01	-.61
σ^2_ζ	0.24	0.01	1.00
$\boldsymbol{\Sigma}_I$			
σ^2_a	0.10	0.01	1.00
σ_{ab}	$-.17$	0.03	$-.61$
$\sigma_{a\phi}$	0.01	0.01	$-.02$
$\sigma_{a\lambda}$	$-.05$	0.01	$-.54$
σ^2_b	0.71	0.08	1.00
$\sigma_{b\phi}$	$-.01$	0.02	$-.05$
$\sigma_{b\lambda}$	0.19	0.04	.68
σ^2_ϕ	0.05	.01	1.00
$\sigma_{\phi\lambda}$	0.01	.01	.15
σ_λ	0.10	0.01	1.00

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 will be evaluated against each other. The estimated Bayes Factors and R^2 statistics for the four models are given in Table 3.2. The Bayes factor was estimated for models $M_0 - M_2$ against model M_3. That is, for model M_2 the density ratio $BF_{23} = \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_3)}$ 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, based on this data set we have to reject the hypothesis that the row wise versus colum 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 it, 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 items. Since row wise versus column wise operations concerns an easy rotation (left-right versus 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 will proceed assuming that model M_2 is the appropriate choice.

Table 3.2. Estimated proportions of explained variance and Bayes factors for the models M_0 - M_3.

Model	$R^2(\boldsymbol{b})$	$R^2(\boldsymbol{\lambda})$	Bayes factor
M_0	0.00	0.00	$\exp(-7)$
M_1	0.13	0.08	$\exp(-5)$
M_2	0.34	0.37	$\exp(46)$
M_3	0.37	0.39	$\exp(0)$

Estimated Effects

The estimated effects for model M_2 can be found in Table 3.3. From the Highest Posterior Density (HPD) (see Box & Tiao, 1973, pp. 123) regions of the estimated parameters 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 ’non-significant rules’ should not be ignored, but interpreted as a set of rules leading to approximately equal response times (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 3.8 below, 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 to items with integrated or separated components. So item construction rules with a positive effect on difficulty also lead to higher expected response times. These findings are in line with finding that more information requires more working memory capacity and additionally results in longer response times (Embretson, 1998).

It is interesting to evaluate the estimated effects on time intensity on the time scale since this gives us results that are more intuitive to interpret. For model

Table 3.3. 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	$[-.15, 0.59]$	4.01	$[3.88, 4.15]$
γ_1 (separated)	-.14	$[-.29, 0.02]$	-.02	$[-.07, 0.02]$
γ_2 (integrated)	-.18	$[-.31, -.04]$	-.09	$[-.14, 0.05]$
γ_3 (embedded)	0.33	$[0.17, 0.47]$	0.11	$[0.06, 0.17]$
γ_4 (identity)	-.63	$[-.93, -.34]$	-.26	$[-.36, -.14]$
γ_5 (addition)	0.22	$[-.05, 0.51]$	0.00	$[-.10, 0.11]$
γ_6 (substraction)	0.17	$[-.12, 0.45]$	0.02	$[-.08, 0.12]$
γ_7 (intersection)	0.47	$[0.16, 0.77]$	0.08	$[-.04, 0.18]$
γ_8 (unique addition)	0.42	$[0.13, 0.70]$	0.17	$[0.06, 0.27]$
γ_9 (seriation)	-.06	$[-.35, 0.22]$	-.10	$[-.21, 0.00]$
γ_{10} (variation of closed gestalts)	-.05	$[-.31, 0.21]$	-.11	$[-.21, -.02]$
γ_{11} (variation of open gestalts)	0.05	$[-.25, 0.33]$	0.01	$[-.11, 0.10]$

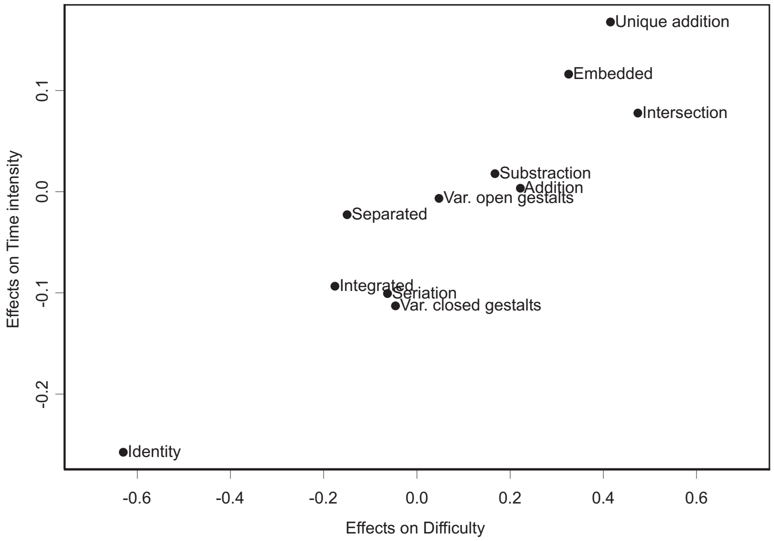


Fig. 3.8. EAP's of the effects of the item construction rules on item difficulty against their effects on time intensity for model M_2.

M_2 the estimates for the mean were $\gamma_0^{(\lambda)} = 4.01$ and for the variance $\hat{\sigma}^2_\lambda = .07$. The inverse transformation to the time scale is then given by $\exp(\hat{\lambda} + \hat{\sigma}^2_\lambda/2) = \exp(4.01 + .07/2) = 57$ seconds. 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 response times between an ”identity item” and a ”mean item” and would be:

$\exp(\hat{\gamma}_0+\hat{\gamma}_1+\hat{\sigma}^2/2)-\exp(\hat{\gamma}_0+\hat{\sigma}^2/2) = \exp(4.01-0.25+.04)-\exp(4.01+.04) = -12$ 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 response time of -20 seconds. It is also interesting to evaluate the easiest and most difficult item in the subset. For the easiest item, $\hat{\lambda} = 2.99$, while for the most difficult item we found that $\hat{\lambda} = 3.52$. On the time scale, this leads to a difference in expected response times of 14 seconds. 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.

3.6 Discussion

The aim of this chapter was to show that an IRT-based approach to response times (RTs) can contribute to the evaluation and testing 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. 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 pre-programmed (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 towards 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).

Other practical implications of the proposed method relate to item and test construction. A test is developed in order 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 describes 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 in order to measure the ability of test takers up to a certain accuracy. RT information and, more specifically, the time intensity of an item now provides 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 based on 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 chapter, 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.

3.7 Appendix: Estimation

The model can be identified by setting the means of the person parameters $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ to zero, so that $\boldsymbol{\mu}_P = 0$ and restricting the variance of $\boldsymbol{\theta}$ to 1, so $\sigma^2_\theta = 1$. Fox et al. (2007) provided a Gibbs sampling solution where these identifying restrictions are directly included into the prior distributions. The same authors describe a straightforward and efficient Gibbs sampling scheme for the hierarchical model, except for a step for sampling the design effects $\boldsymbol{\gamma}$ of the item parameters. Therefore, the sampling steps will be described below, but only the sampling step for the design effects is given explicitly here.

- Step 1 Sample augmented response data according to Equation 3.12.
- Step 2 Draw (θ_i, ζ_i) simultaneously from $\theta_i, \zeta_i | \mathbf{z}_i, \mathbf{t}_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P$.
- Step 3 Draw $(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ from $\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P | \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\Sigma}_{P0}, \boldsymbol{\mu}_{P0}$, where $\boldsymbol{\Sigma}_{P0}, \boldsymbol{\mu}_{P0}$ denote the hyperprior parameters
- Step 4 Draw $(a_k, b_k, \phi_k, \lambda_k)$ from $a_k, b_k, \phi_k, \lambda_k | \mathbf{z}_k, \mathbf{t}_k, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_I$
- Step 5 Draw $\boldsymbol{\Sigma}_I$ from $\boldsymbol{\Sigma}_I | \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_{I0}$
- Step 6 Draw $\boldsymbol{\gamma} | \mathbf{A}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_{\gamma_0}$ This step is specified below.

Let $\boldsymbol{\gamma} = \text{vec}\left(\boldsymbol{\gamma}^{(a)}, \boldsymbol{\gamma}^{(b)}, \boldsymbol{\gamma}^{(\phi)}, \boldsymbol{\gamma}^{(\lambda)}\right)$ and $\boldsymbol{\Omega}_I = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\phi}, \boldsymbol{\lambda})$, where vec denotes the operation of vectorizing a matrix. Furthermore, let $\mathbf{A}_I = (\mathbf{I}_4 \otimes \mathbf{A})$. This enables rewriting Equations 3.21 - 3.24 to:

$$\text{vec}(\boldsymbol{\Omega}_I) = \mathbf{A}_I \boldsymbol{\gamma} + \text{vec}(\boldsymbol{e}), \tag{3.33}$$

where $\text{vec}(\boldsymbol{e}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I \otimes \mathbf{I}_K)$. Next, a conjugate normal prior is chosen for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} \sim N(\boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_{\gamma_0}) \tag{3.34}$$

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

$$\boldsymbol{\gamma} | \mathbf{A}_I, \boldsymbol{\Omega}_I, \boldsymbol{\Sigma}_I, \boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_{\gamma_0} \sim N\left(\frac{\hat{\boldsymbol{\Sigma}}_\gamma^{-1}\hat{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\gamma_0}^{-1}\boldsymbol{\gamma}_0}{\hat{\boldsymbol{\Sigma}}_\gamma^{-1} + \boldsymbol{\Sigma}_{\gamma_0}^{-1}}, \left(\hat{\boldsymbol{\Sigma}}_\gamma^{-1} + \boldsymbol{\Sigma}_{\gamma_0}^{-1}\right)^{-1}\right), \tag{3.35}$$

where $\hat{\boldsymbol{\Sigma}}_\gamma$ and $\hat{\boldsymbol{\gamma}}$ are the common least squares estimates, which can be derived from Equation 3.33. For Gibbs sampling of the other model parameters, see Fox et al. (2007).

4

A Box-Cox Normal Model for Response Times

Summary. The log-transform has been a convenient choice in response time modeling on test items. However, motivated by a dataset of the Medical College Admission Test where the lognormal model violated the normality assumption, the possibilities of the broader class of Box-Cox transformations for response time modeling are investigated. After an introduction and an outline of a broader framework for analyzing responses and response times simultaneously, the performance of a Box-Cox normal model for describing response times is investigated using simulation studies and a real data example. A transformation-invariant implementation of the Deviance Information Criterium (DIC) is developed that allows for comparing model fit between models with different transformation parameters. Showing an enhanced description of the shape of the response time distributions, its application in an educational measurement context is discussed extensively.

4.1 Introduction

Recording response times (RTs) on test items is common practice nowadays. Thereby, besides the response patterns, an additional source of information is available to test developers and testing agencies. For instance, RTs can be helpful to improve the design of a test or study the response behavior of test takers. However, an appropriate statistical treatment of the RTs is required before making any inferences.

Response time experiments have been a major source of inferences about cognitive processes in experimental psychology (Luce, 1986). To illustrate the type of experiments and the kind of data that arise from them, we give the following three examples. Schmiedek et al. (2007) performed experiments using simple speed tasks to study attention fluctuation and working memory. One of the experiments reported by these authors was a verbal classification task where participants had to classify single words into categories of animals or plants. Ratcliff and Rouder (1998) performed experiments to study stimulus discrimination, where participants had to classify the intensity of an array of pixels on a monitor as high or low. An example of a time pressure study of the well known speed-accuracy tradeoff can be found in van der Lubbe, Jaśkowski, Wauschkuhn, and Verleger (2001). There, participants had to respond before the space between an inner circle and an outer circle was filled. Typically, experiments like these consist of many repetitions of the same simple task. The data that arise from such experiments are the RTs (usually in the order of milliseconds) and accuracy measures (correct/incorrect). For the joint analysis of RT and accuracy data traditional ANOVA methods have been used up till recently (van der Lubbe et al., 2001), with inferences based on the mean RTs and the mean proportion correct (PC) scores. An approach that provides more detail into the analysis and relates RTs and accuracy explicitly is the diffusion model presented by Ratcliff (1978). For more recent references and approaches that are related to the diffusion model see, for instance, Ratcliff and Tuerlinckx (2002); Wagenmakers, van der Maas, and Grasman (2007) and Brown and Heathcote (2008).

In educational assessment, measurement had to be based on response accuracy only for a long time. This limitation was overcome with the introduction of computerized test adminstration, which made the accurate collection of RTs feasible. Thereby, an additional source of information on test items and test takers has become available. For instance, when students are not motivated for a test, this might lead to lower RTs as a result of guessing behavior, something that cannot be easily seen from accuracy data alone. Therefore, there is a need to incorporate RTs into the analysis of test data and study responses and RTs simultaneously. However, there are some important differences in the data collection process compared to the procedures in experimental psychology. First, in experimental psychology the RTs are linked directly to theoretical cognitive phenomena which are evaluated, for instance, using elementary two-choice tasks, whereas in educational measurement the tasks (items) are of a much higher cognitive complexity. As a result, the observed RTs are in the range of seconds up to some minutes. Where experiments measured in milliseconds need to take account of a lower bound on the RTs, this can safely be ignored in educational measurement due to the size of the measurements. Also, in educational assessment multiple items are administered that are answered only once, contrary to the within-subject replications found in experimental psychology. These differences lead to a somewhat different approach to the joint modeling of RTs and accuracy data on test items than that mentioned above.

In educational testing, item response theory (IRT) models have served as measurement models for a latent construct, ability, which is assumed to underly the accuracy data. Very different from the diffusion model, RTs are not included in IRT models. Instead, it will be assumed that individual differences between test takers in their observed RTs result from differences in speed. That is, speed will be assumed to be the latent construct underlying the RTs and a separate measurement model is required for measuring it. At a higher (second) level the relationships between the two measurement models are modeled to account for possible dependencies between the RTs and the accuracy data. This leads to a framework of modeling that allows for the simultaneous analysis of RTs and accuracy data on test items. IRT models have been well developed, but models for RTs have had much less attention in the psychometric literature. In this paper, motivated by an empirical problem, we focus on models for RTs that are flexible in their distributional shapes and fit well into the framework for the simultaneous analysis of responses and RTs.

Typically, RTs are non-negative and, as a result, their distribution is positively skewed. Various types of distributions are able to describe such data and have been extensively studied, for instance, in the field of lifetime modeling. Examples are the Poisson, Gamma, Weibull, inverse normal, exponential and lognormal distributions. For discussions on the use of these distributions for modeling RTs 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, 2002); van der Linden (2006). In practice, it is difficult to determine which distribution would fit the RT data best. The lognormal model has been a convenient choice, with good results regarding model fit (Thissen, 1983; Schnipke & Scrams, 1997; van der Linden et al., 1999; van der Linden, 2006). Besides, it permits the use of the nice statistical properties of a normal model for the log-transformed RTs. A normal model easily allows for a decomposition of the mean into item and person effects. van der Linden (2006) introduced a lognormal model for describing response times.

Nevertheless, the analysis of a computerized version of the Medical College Admission Test (MCAT) revealed that the log-transformed RTs do not always satisfy the normality assumption (Section 4.6.2). A Bayesian residual analysis indicated that the skewness of the RT distributions was not always captured well for the MCAT data. In such cases, it would be desirable to evaluate the fit of the model against other distributions. For instance, a Gamma model might be more appropriate for describing the structure of the skewness Maris (1993). But, fitting and evaluating different RT models can be laborious and is not desirable from a practical perspective. A more general approach for describing any RT distribution would be preferred.

The Box-Cox transformation has been widely used to model skewed distributions. For instance, it finds application in life time / failure time models in industry or in the empirical determination of functional relationships in the field of economics. Nonetheless, as far as known by the authors, it has not found application in the psychometric literature of response time modeling. Therefore, the class of Box-Cox transformations is considered in this study. Using a whole class of transformations gives the researcher more freedom in analyzing response time data. It allows one to choose an appropriate transformation in order to obtain normally distributed data. Box and Cox (1964) proposed a power transformation as a function of an unknown parameter ν, which contains the log-transform as a special case:

$$T^{(\nu)} = \begin{cases} \frac{T^{\nu}-1}{\nu} & (\nu \neq 0), \\ \log T_{ik} & (\nu = 0), \end{cases} \tag{4.1}$$

where T denotes the original time and $T^{(\nu)}$ denotes the Box-Cox transformed time. Note that the log transform for $\nu = 0$ is defined in order to obtain a family of transformations over a continuous range of ν.

To illustrate the flexibility in shape of the Box-Cox density, consider response times T that follow a Box-Cox normal density with parameters (ν, λ, τ^2), where $\nu \neq 0$, given by:

$$f(t) = t^{(\nu-1)} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\left(\frac{t^{(\nu)}-\lambda}{\tau}\right)^2\right) \tag{4.2}$$

An impression of the different shapes this density can take is shown in Figure 4.1. For $(\lambda, \tau) = (6, 1)$, the density $f(t)$ is plotted for some values of $\nu \in (0, 1]$ over a range of $T \in [0, 80]$.

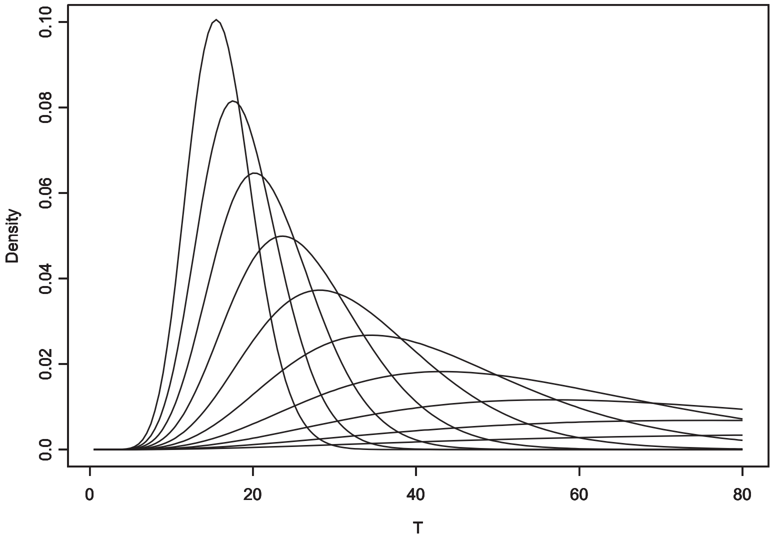


Fig. 4.1. Box-Cox normal densities as a function of the transformation parameter $\nu\epsilon(0, 1]$ and with fixed mean and variance $\lambda = 6$, $\tau^2 = 1$. The density with the leftmost and highest peak is $\nu = .05$, the flattest density is the one with $\nu = 1$

Besides flexibility, it is interesting to evaluate its benefits with respect to model fit as well as the interpretation of the parameters and its behavior in a larger framework for the simultaneous analysis of responses and RTs. Below, the modeling framework is introduced first, followed by the method of estimation. Thereafter, the problem of how to obtain the moments of the Box-Cox distribution is discussed. These moments are helpful to characterize, for instance, the skewness of the RT distributions. The presentation of a few tools for evaluating the fit of the model is then followed by empirical examples as well as simulation studies that address the research questions above. A discussion of the advantages and disadvantages of the Box-Cox approach for modeling RTs concludes this paper.

4.2 A Framework for the Simultaneous Analysis of Responses and Response Times

The interest of researchers is often focussed on studying responses or response times alone. However, since both data sources contain information on the same items and test takers, it can be advantageous to study them simultaneously. For instance, the interest may be in the relationship between speed and accuracy of test takers or the testing of the common assumption that more difficult items also are more time intensive. Therefore, a framework that allows for modeling dependencies between responses and RTs is outlined here.

Measurement models at level 1 separate the variability in the observed responses and response times into item and person effects. Just as ability, the speed of the test takers is assumed to be the underlying construct for the RTs. Further, it is assumed that speed and ability of the test takers are fixed during the test. This assumption leads to conditional independence of the responses and response times of a test taker given the latent traits, which is a key feature of this model. At level 2, a correlation structure models the dependencies between the level 1 model parameters.

Not only can the Box-Cox normal model improve the description of the skewness of the data but, due to the transformation to normality, it also fits this hierarchical framework nicely. Contrary to a Weibull or Gamma RT model, the BC model allows the use of easy to implement conjugate normal models for the item and person parameters at level 2, which enables a straightforward Gibbs sampling approach for estimation of the model parameters as well.

4.2.1 Response Model

In IRT, it is assumed that the variability in observed response patterns on test items can be separated into item and person effects. Within an item, the variability between the responses of different test takers is the results from differences in their ability, denoted by θ. The higher one's ability, the higher the probability of giving a correct response. Within a test, there are differences between items regarding their difficulty. The probability that a test taker answers an item correctly depends on the difference between the difficulty b of the item and his or her ability. The way the item distinguishes between test takers of different ability is described by the discrimination parameter a.

Assuming that the probability that person $i = 1, \ldots, N$ answers item $k = 1, \ldots, K$ correctly ($Y_{ik} = 1$) follows the two-parameter normal ogive model,

$$P(Y_{ik} = 1 | a_k, \theta_i, b_k) = \Phi(a_k \theta_i - b_k), \tag{4.3}$$

or, in its latent response formulation,

$$P(Y_{ik} = 1 | a_k, \theta_i, b_k) = \int_0^\infty P(z_{ik}; a_k \theta_i - b_k) dz, \tag{4.4}$$

where $Z_{ik} \geq 0$ when $Y_{ik} = 1$ and $Z_{ik} < 0$ otherwise. The model is given in its latent variable form for computational convenience, as introduced by Albert (1992).

4.2.2 Response-Time Model

Analogously, it is assumed that the variability in observed response time patterns on test items can be separated into item and person effects. For instance, it never happens that a group of test takers finish the test in the same time. Some persons are working faster than others. This leads to the assumption that, within an item, the variability in response times results from differences in speed of working of the test takers. Therefore, a personality trait for speed is introduced, denoted by ζ. That is, the speed parameter is assumed to be the underlying construct for the RTs, just as the ability parameter is for the responses. It is assumed that during a test, a person works at a fixed speed. In general, within a test, test takers do not spent equal time on the items. Some items require more time to be solved (it is often assumed that this concerns the more difficult items). As an example, solving $2 + 5 =?$ involves less steps than solving $2 + 5 + 7 =?$ and, therefore, it can be expected that the latter is more time intensive. For representing these differences in time intensity of items, an item parameter λ is introduced. This parameter can be seen as the time-analogue of the difficulty parameter. The parameter ϕ reflects the way the item distinguishes between test takers of different speed levels.

The generalization to a Box-Cox normal model then leads to a linear model for the transformed RTs:

$$T^{(\nu)} = \begin{cases} \frac{T_{ik}^{\nu}-1}{\nu} \sim N(-\phi_k\zeta_i + \lambda_k, \tau_k^2) & (\nu \neq 0), \\ \log T_{ik} \sim N(-\phi_k\zeta_i + \lambda_k, \tau_k^2) & (\nu = 0). \end{cases} \tag{4.5}$$

For notational convenience, the superscript will be dropped and T_{ik} will denote the Box-Cox transformed time from now on.

4.2.3 Second-Level Models

At the second level of modeling, the person parameters are assumed to follow a multivariate normal distribution. Let $\boldsymbol{\xi}_i = (\theta_i, \zeta_i)$, then:

$$\boldsymbol{\xi}_i = \boldsymbol{\mu}_P + \boldsymbol{e}_P, \boldsymbol{e}_P \sim N(\mathbf{0}, \boldsymbol{\Sigma}_P), \tag{4.6}$$

where $\boldsymbol{\mu}_P = (\mu_\theta, \mu_\zeta)$ and the covariance structure is specified by:

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}. \tag{4.7}$$

Here, ρ denotes the covariance between the two person parameters. A positive estimate for ρ indicates a positive dependence between ability and speed, meaning that a person who works faster than average also tends to have an above-average ability.

Similarly, it can be assumed that the item parameters follow a multivariate normal distribution. Let $\boldsymbol{\Omega}_k = (a_k, b_k, \phi_k, \lambda_k)$, then:

$$\boldsymbol{\Omega}_k = \boldsymbol{\mu}_I + \boldsymbol{e}_I, \boldsymbol{e}_I \sim N(\mathbf{0}, \boldsymbol{\Sigma}_I), \tag{4.8}$$

where $\boldsymbol{\mu}_I = (\mu_a, \mu_b, \mu_\phi, \mu_\lambda)$ and the covariance structure is specified by:

$$\boldsymbol{\Sigma}_I = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{\phi a} & \sigma_{\phi b} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{\lambda a} & \sigma_{\lambda b} & \sigma_{\lambda\phi} & \sigma_\lambda^2 \end{bmatrix} . \tag{4.9}$$

This covariance structure allows for the investigation of dependencies between the item parameters. For instance, one can test the common assumption that more difficult items also take more time to be solved. Doing so would amount to evaluating the null hypothesis $H_0 : \sigma_{b\lambda} = 0$ against the alternative $H_a : \sigma_{b\lambda} > 0$.

4.3 Bayesian Estimation Using MCMC Methods

The model is estimated using a fully Bayesian approach using straightforward Markov Chain Monte Carlo (MCMC) methods. In a Bayesian approach, inferences are made from the posterior distribution $p(\theta|x)$. Using Bayes' rule, the posterior is obtained from the observed data x, a realization of $X \sim f(x|\theta)$, combined with available prior information, specified as $p(\theta)$. An introduction to Bayesian inference can be found, for instance, in Box and Tiao (1973).

Estimation of the posterior distributions of the model parameters requires evaluating integrals, which, analytically, for complex models, can be an impossible task. A solution to this problem is to use simulations to approximate the densities. Markov Chain Monte Carlo methods, such as the Gibbs sampler (Geman & Geman, 1984) and the Metropolis-Hastings algorithm (Chib & Greenberg, 1998), are useful for drawing samples from the posterior distributions of the model parameters. Although computationally intensive, these methods remain straightforward when model complexity increases. Gelman et al. (2004) provide an introduction to MCMC methods; a more advanced text is Robert and Casella (1999).

Since our interest is in the Box-Cox normal RT model, the sampling steps for the transformation parameter and the RT model parameters are given explicitly below. Sampling of the other model parameters is outlined in the Appendix.

Identification

The response time model can be identified by setting $E(\zeta) = 0$, which fixes the mean. By specifying $\prod_{k=1}^{K} \phi_k = 1$, a tradeoff between σ_ζ^2 and ϕ_k is avoided. Identification of the hierarchical model can be obtained by fixing the location of the latent traits by $\boldsymbol{\mu}_P = \mathbf{0}$. Further, the scale of the ability trait can be fixed in two ways: either by setting $\sigma_\theta^2 = 1$ or by setting $\prod_{k=1}^{K} a_k = 1$.

Sampling the Box-Cox Parameter

A normal model for the transformed response times is assumed. The likelihood with respect to the original response times is given by:

$$p(\mathbf{t}|\nu)\boldsymbol{J}(\nu,\mathbf{t}^*), \tag{4.10}$$

where $\mathbf{t}^*$ denotes the original response times and $\boldsymbol{J}(\nu,\mathbf{t}^*)$ the Jacobian of the transformation. For $\nu = 0$, the Jacobian equals t^{*-1}; when $\nu \neq 0$, it equals $t^{*(\nu-1)}$. Different priors for ν were studied by Box and Cox (1964), However, these were outcome dependent; that is, they were was dependent on the observations. Pericchi (1981) did propose non-informative priors for the transformation parameter that where not outcome dependent. However, these priors were derived in order to obtain analytic results on the value of ν.

A main problem is that there does not seem to exist a conjugate prior for ν (as far as known by the authors), so a Gibbs sampling step for the parameter is not feasible. However, for a sampling based approach, the choice of a family of priors is less critical. For that reason a Metropolis-Hastings (MH) step is proposed, the advantage being that any chosen prior for ν is easily implemented in the MH-step. At iteration m, a new value ν^*, sampled from a proposal density $\varphi(\nu^*|\nu)$, is accepted with probability:

$$\min\left\{1, \frac{p(\nu^*|\mathbf{t})}{p(\nu^{m-1}|\mathbf{t})} \times \frac{\varphi(\nu^{m-1}|\nu^*)}{\varphi(\nu^*|\nu^{m-1})}\right\}, \tag{4.11}$$

otherwise $\nu^m = \nu^{m-1}$.

When the optimal transformation is the logartihm, the distribution should converge to $E(\nu) = 0$. However, although a posterior mean of approximately 0 can be obtained, a value of $\nu^{(m)} = 0$ will practically never be sampled since it has probability 0. To accommodate for the log-transform, consider a critical value C such that when $|\nu^{(m)}| < C$, then $\nu^{(m)} = 0$ with probability .5. Tuning of the value of C is required, whereby (based on our experience) a value of .05 can be considered a good starting point.

Sampling the Item and Person Parameters

Below the conditional posterior distributions of the person and items parameters of the RT model are presented. Together with the sampling step for the transformation parameter, these steps constitute the MCMC algorithm for the RT model.

- The person speed parameters $\boldsymbol{\zeta}$ are the parameters of the linear regression of $-\mathbf{T}_i + \boldsymbol{\lambda}$ on $\boldsymbol{\phi}$. Assuming a normal prior $\zeta_i \sim N(\mu_\zeta, \sigma^2_\zeta)$, the resulting posterior is again normal with

$$\zeta_i|\mathbf{t}_i, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \nu \sim N\left(\frac{\sigma_\zeta^{-2}\mu_\zeta + \sum_{k=1}^K \tau_k^{-2}\phi_k(\lambda_k - t_{ik})}{\sigma_\zeta^{-2} + \sum_{k=1}^K \phi_k^2\tau_k^{-2}}, (\sigma_\zeta^{-2} + \sum_{k=1}^K \phi_k^2\tau_k^{-2})^{-1}\right) \tag{4.12}$$

- The item parameters $(\boldsymbol{\phi}, \boldsymbol{\lambda})$ are the coefficients of the regression of $\mathbf{T}_k$ on $\mathbf{X} = (-\boldsymbol{\zeta}, \mathbf{1})$. Assuming a normal prior, $\phi_k, \lambda_k \sim N(\boldsymbol{\mu}_{\phi,\lambda}, \boldsymbol{\Sigma}_{\phi,\lambda})$, the posterior distribution is given by:

$$\phi_k, \lambda_k|\mathbf{t}_k, \tau_k^2, \boldsymbol{\zeta}, \nu \sim N\left(\frac{\boldsymbol{\Sigma}_{\phi,\lambda}^{-1}\boldsymbol{\mu}_{\phi,\lambda} + \tau_k^{-2}\mathbf{X}^t\mathbf{t}_k}{\boldsymbol{\Sigma}_{\phi,\lambda}^{-1} + \mathbf{X}^t\mathbf{X}\tau_k^{-2}}, (\boldsymbol{\Sigma}_{\phi,\lambda}^{-1} + \mathbf{X}^t\mathbf{X}\tau_k^{-2})^{-1}\right) \tag{4.13}$$

- For the residual variance τ_k^2, a conjugate inverse Gamma prior with parameters $Inv{-}Gamma(g_1, g_2)$ is assumed. The posterior is then again an inverse Gamma distribution with parameter $g_1 + N/2$ and scale parameter $g2 + (\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))^t(\mathbf{t}_k - (-\phi_k\boldsymbol{\zeta} + \lambda_k))/2$.

4.4 Moments of the Response-Time Distributions

We will use the first three moments about zero of the distributions to assess the differences between the lognormal and Box-Cox normal models. More specifically, it is expected that these models will differ in their third moment, which characterizes the skewness of the distribution. Therefore, only the estimation of the first three moments of the distributions is considered in this study.

How to obtain the moments of the lognormal distribution is well known. However, the moments of the Box-Cox normal distribution are not so straightforward to estimate, except for some specific transformations, such as $\nu = 2$ or $\nu = .5$. Freeman and Modarres (2006) studied the properties of the inverse Box-Cox transformation. Let $Y = (X^\nu - 1)/\nu$, $Z = (Y - \mu)/\sigma$ and $Y \sim N(\mu, \sigma^2)$. Then X is power-normal distributed, or $X \sim PN(\nu, \mu, \sigma^2)$. The authors derived the rth moment of X as

$$E(X^r) = (\nu\mu + 1)^{r/\nu} + \sum_{i=1}^{\infty} \frac{1}{i!}(\nu\mu + 1)^{r/\nu - i}\sigma^i E(Z^i) \prod_{j=0}^{i-1}(r - j\nu), \tag{4.14}$$

for $\nu \neq 0$. Moreover, they showed that these moments can be approximated by $E(X^r) \approx (\nu\mu + 1)^{r/\nu} + \sum_{\text{Even } i>0} \frac{\sigma^i}{2^{i/2}(i/2)!}(\nu\mu + 1)^{r/\nu - i}E(Z^i)\prod_{j=0}^{i-1}(r - j\nu)$, where $i > 0$ and even. When $\nu = 0$ the moments of the lognormal distribution can be approximated by $E(X^r) = \exp(r\mu + \frac{r^2\sigma^2}{2})$. These results will be used to approximate the moments of the distributions.

From these raw moments, the second central moment, which corresponds to the variance, and the third standardized moment, which is a measure for the skewness of the distribution, are obtained.

4.5 Evaluating Model Fit

Model fit will be evaluated using two methods: (i) Baysian residual analysis by evaluating the posterior probabilities under the model, and (ii) a Deviance Information Criterion (DIC).

The transformed values t_{ij} are evaluated under their predictive density under the RT model. Subsequently, the probability $P(T_{ik} < t_{ik}|\mathbf{y}, \mathbf{t})$ can be approximated by

$$P(T_{ik} < t_{ik}|\mathbf{y},\mathbf{t}) \approx \sum_{m=0}^{M} \Phi\big(t_{ik}|\zeta_i^{(m)}, \phi_k^{(m)}, \lambda_k^{(m)}, \nu_k^{(m)}\big)/M \tag{4.15}$$

from the M iterations of the MCMC chain. Now, the *probability integral transformation theorem* (e.g., Casella & Berger, 2002) implies that under the true model these probabilities are distributed as $U(0,1)$. This feature allows evaluation of the model fit. To do so, the calculated probabilities of the items are plotted against their expected values under the $U(0,1)$ distribution. If the underlying distribution really is $U(0,1)$, these plots should be approximately linear.

Graphical model checking can be very helpful to understand in what way a fitted model departs from the data. However, graphical comparison of two competing models can be difficult when they are close. Also, the proposed graphical check does not penalize for model complexity. Therefore, the DIC (Spiegelhalter et al., 2002), which does account for model complexity, should be estimated as well. Besides being a useful test statistic for model comparison, it has the advantage that it is easily obtained as a by-product of the MCMC chain.

The deviance $D(\mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta})$ is given by:

$$\begin{aligned} D(\mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta}) &= -2\log p(\mathbf{t}|\boldsymbol{\phi}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta}) & (4.16)\\ &= N\sum_{k=1}^{K}\log(2\pi\tau_k^2) + \sum_{k=1}^{K}\sum_{i=1}^{N}(t_{ik} - (-\phi_k\zeta_i + \lambda_k))^2/\tau_k^2 \\ &\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\log J_{ik}, & (4.17)\end{aligned}$$

where J_{ik} denotes the Jacobian of the transformation, which is $(t_{ik}^*)^{-1}$ when $\nu = 0$ and $(t_{ik}^*)^{(\nu_k - 1)}$ when $\nu_k \neq 0$, with t_{ik}^* the original observation. The DIC is equal to the deviance plus a penalty term for model complexity, and is given by:

$$DIC = \bar{D} + (\bar{D} - \hat{D}), \tag{4.18}$$

with $\bar{D} \approx \frac{1}{M}\sum_{m=1}^{M} D(\mathbf{t}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\lambda}^{(m)}, \nu^{(m)}, \boldsymbol{\zeta}^{(m)})$, $m = 1, \ldots, M$ denoting the number of iterations of the algorithm, and $\hat{D} \approx E(D(\mathbf{t}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\lambda}, \nu, \boldsymbol{\zeta})|\mathbf{t}^*)$. Spiegelhalter et al. (2002) report that when using the posterior median instead of the posterior mean to estimate the DIC, the term for model complexity is invariant to transformations. However, the DIC is constructed from a likelihood-based term plus the correction for model complexity. Different transformations lead to different scales of the data and thereby affect the likelihood. The Jacobian of the transformation is to guarantee that all DIC values correspond to one common scale (=the original time scale). As a result, this DIC allows the comparison of model fit for different transformations.

4.6 Flexibility of the Box-Cox Normal Model

To illustrate the possibilities of the Box-Cox approach for modeling response times on test items, two examples are given here. In the first example, it will be shown that the Box-Cox normal model can approximate data resulting from Weibull, Gamma and Exponential models. The second example analyzes an empirical data set and shows that model fit can be improved when the lognormal model is generalized to a Box-Cox normal model.

4.6.1 Approximation of Weibull, Gamma and Exponential Data

Aim of this example is to show that if the true underlying distribution of the RTs is Gamma, Weibull or Exponential, the Box-Cox normal distribution can be a good approximation to the RTs.

For our example, we used the empirical mean and variance of the RTs of three items: The first was obtained from a Raven test taken by 300 German army recruits for which (mean, var) = $(64, 1766)$. The second was from a computerized version of the MCAT for which (mean, var) = $(190, 4904)$ seconds. Rouder et al. (2003) used a Weibull distribution to model reaction times and report a typical estimate for the shape parameter of 2. This value was used for the third item. The parameters for the Gamma, Weibull and Exponential distribution were chosen such that they corresponded closely with the estimated means and variances of the selected items.

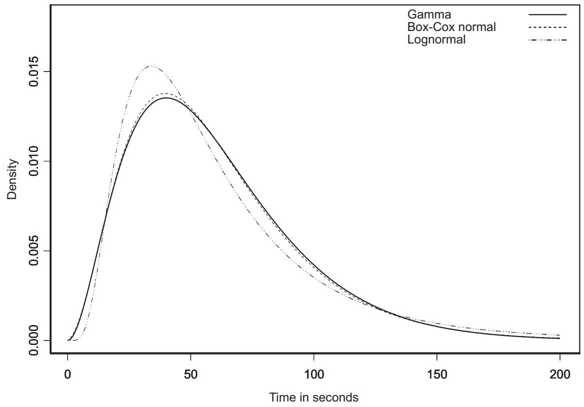


Fig. 4.2. Density of the $Gamma(3, .05)$ function, its lognormal, and its Box-Cox normal approximation.

Subsequently, 10,000 data points were simulated under these models, and the Box-Cox normal model was fitted to the data. From the obtained parameter estimates of the Box-Cox normal model, the density function was plotted together with the density function of the true underlying distribution. In the same figures, the lognormal density was plotted; see Figures 4.2 - 4.4. Furthermore, estimates of the DIC criterion as well as the moments of these distributions were obtained. Table 4.1 summarizes the results.

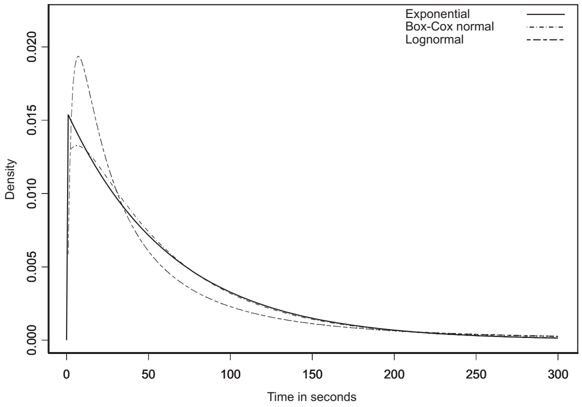


Fig. 4.3. Density of the *Exponential*(1/64) function, its lognormal, and its Box-Cox normal approximation.

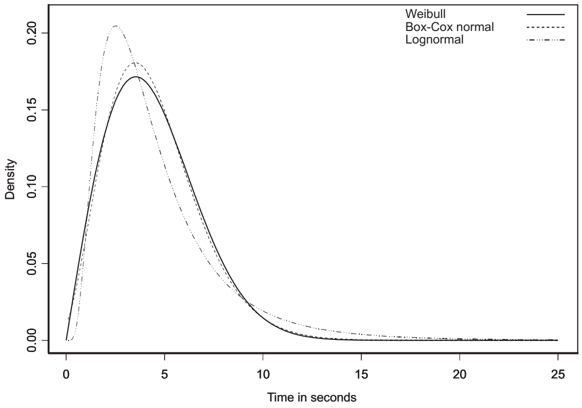


Fig. 4.4. Density of the $Weibull(2, 5)$ function, its lognormal, and its Box-Cox normal approximation.

As can be seen from the figures, the Box-Cox normal model approximated the three chosen distributions quite well. Both the regions of highest density as well as the tails of the distributions are captured. Only for the Exponential model, the lognormal and Box-Cox normal models did have problems to describe the density near 0. From Table 4.1 it can be seen that the means of the lognormal and the Box-Cox densities were quite close, using (4.14). However, especially the skewness of the distributions, the third standardized moment of a distribution, differed substantially. In all cases, the lognormal distribution was more skewed to the right than the Box-Cox normal distribution. For each distribution, one example is given in Figures 4.2 - 4.4. The lognormal model distribution was more peaked. According to the DIC criterion, the best descriptions of the data were obtained with the Box-

Table 4.1. Parameter estimates, estimated moments (mean, variance and third standardized central moment, the skewness) and DIC for the approximation of Gamma, Exponential and Weibull data.

Distribution		Parameters			Moments			Model Fit
Simulated	Approximated	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$E(X)$	$Var(X)$	Skewness	DIC
$Gamma(3,.05)$								
	LN	3.92	0.64	0	61.0	1757	2.39	97231
	BC	7.89	2.04	.31	59.8	1164	1.60	96591
$Gamma(7.4,.04)$								
	LN	5.1	0.39	0	176.9	5217	1.29	111411
	BC	12.1	1.78	.30	175.6	4266	0.89	111221
$Expon(1/64)$								
	LN	3.59	1.27	0	81.2	26403	10.8	105128
	BC	6.27	3.06	.27	59.4	3743	2.40	103373
$Expon(1/200)$								
	LN	4.22	1.28	0	154.2	97762	11.0	117721
	BC	8.34	3.58	.26	126.9	17620	2.43	116022
$Weibull(2,5)$								
	LN	1.32	0.63	0	4.59	10.44	2.46	45744
	BC	2.12	1.17	.53	4.43	5.43	0.77	44066
$Weibull(5,6)$								
	LN	1.65	0.31	0	5.48	3.11	0.99	38246
	BC	4.66	1.62	1.04	5.45	2.30	-0.03	36779

Cox normal model. Of course, this does not prove that the Box-Cox model is well suited to approximate all possible Gamma or Weibull models. However, the aim of this example was to show that, for a typical range of response times, the Box-Cox model does provide a good approximation to such data.

4.6.2 Empirical Example

For this example, the data of 405 test takers on 214 items from a computerized version of the MCAT were analyzed. Six items were omitted from the data set because the algorithm showed convergence problems for them. For the remaining items, only a few observations were missing (less than one percent). These were assumed to be missing at random and were ignored in the estimation procedure. Preliminary analysis showed that the time discrimination parameter did not vary across items using the DIC criterion. The analysis reported below were therefore conducted under the restriction $\boldsymbol{\phi} = 1$.

Step 1

Two models were fitted to the data: model M_1, with the restriction $\nu = 0$ (=LN model) and the more general BC model M_2 with $\nu \neq 0$. The prior for the person parameters was fixed at a mean of $\mu_\zeta = 0$ (for identification) and had a (lowinformative) variance of $\sigma_\zeta^2 = 10$. For the item parameters, (lowinformative) priors $(\mu_\lambda, \sigma_\lambda) = (0, 10)$ were chosen. Since the values for ν are usually within the range of $[-1, 1]$, a slightly informative uniform $U(-4, 4)$ density was specified as prior. The models were estimated using 100,000 iterations of the MCMC algorithm, from which every 10th sample was stored. The reason for doing so is to reduce the auto-correlation between the draws of the transformation and item parameters. The draws of the transformation parameter affects the mean and variance of the distribution on the transformed time scale and therefore influences λ_k and τ_k. It appeared sufficient to discard the first 1,000 stored samples and base the estimates of the model parameters and the model fit criteria on the remaining 9,000 samples. Rerunning the algorithm with different starting values confirmed convergence of the chains.

Table 4.2 gives the estimated DIC for each model. It can be seen that M_2 should be favored over the more restricted model M_1. At the item level, the graphical model check suggested an improvement of model fit for the majority of the items for the BC model. The .95 HPD region of the transformation parameter was estimated as $(.19, .21)$. Since the DIC was calculated by summing the deviance terms over the items, it was straightforward to obtain the estimates of the DIC at the item level as well. From these results, it followed that model M_2 was selected by the DIC over M_1 for 174 of the 214 items.

The graphical posterior check suggested that the LN model assigned somewhat more weight to the middle region and somewhat less to the tails of the distributions. Plotting the estimated densities of the three models for some items confirmed this impression; the plots showed that the density of the BC model was less peaked near its highest density region and has somewhat wider tails than the LN model. On average, these difference resulted in an improved description of this data set.

Table 4.2. Estimated DIC values for the models of the MCAT analysis

Model		DIC
M_1	Lognormal	830132
M_2	Box-Cox	822847
M_3	Box-Cox (item-specific)	820963

Item-Specific Transformation Parameters

The flexibility of the RT model may be improved further by making the transformation parameter item specific. We explored this possibility mainly for theoretical reasons but observe that item-specific transformations also lead to item-specific time scales. As discussed below, we therefore expect the applicability to be low.

Using the DIC, the introduction of the item-specific transformation (model M_3) resulted in improved model fit for 150 items. Except for 17 items, the estimates of the DIC criterion suggested that the improvement in model M_3 relative to model M_1 was significant. The .95 HPD regions of the transformation parameters $\boldsymbol{\nu}$ were consistent with these results as well (zero not being contained within these regions). Overall, the DIC for the complete data set decreased to 820963 for M_3. Although an improvement, the decrease was smaller than for the transition from M_1 to M_2 (see Table 4.2).

In order to illustrate the effect of the Box-Cox transformation for this real-data example, three cases are given in the Figures 4.5, 4.6 and 4.7 in the appendix. These cases where chosen because they reflected a range of parameter values for ν_k. It can be seen that, for Item 86, there was no noticeable difference between the two competing models even though the DIC criterion suggested a slight loss of model fit for the BC model. On the other hand, the BC model showed substantial improvement for Item 4. For Item 15, the result was between the two other items and pointed at a slight improvement in our description of the data.

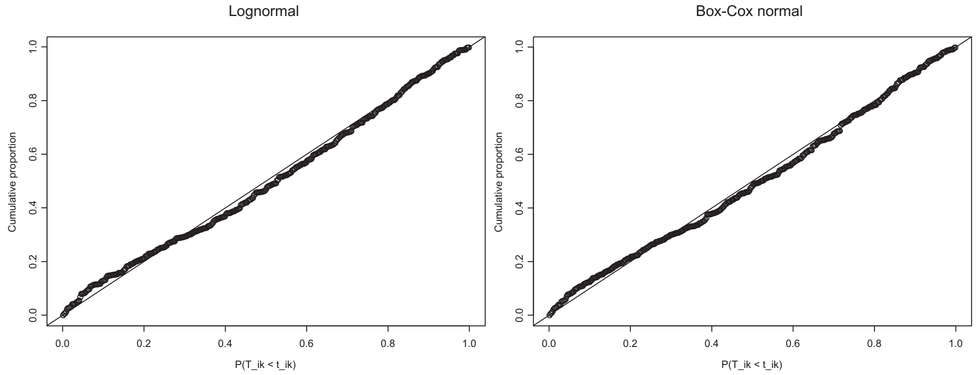


Fig. 4.5. Cumulative probability plots of the posterior probabilities of item 86. Left: lognormal model, $DIC = 4751$, $\nu = 0$. Right: Box-Cox normal model, $DIC = 4761$, $EAP(\nu) = .05$, $.95HPD(\nu) = [.00, .10]$.

4.7 Model Interpretation and Selection

It is interesting to determine the effects of the different transformations on the interpretation of the RT model parameters. They should help us to guide our choice of model for different types of analyses.

Upon transformation, the RTs are assumed to follow a normal model. In educational testing, it is natural to assume variability across persons as well as items.

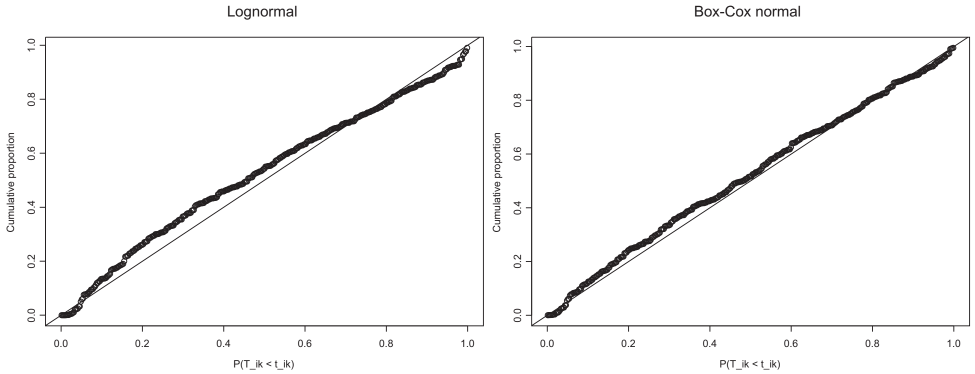


Fig. 4.6. Cumulative probability plots of the posterior probabilities of item 15. Left: lognormal model, $DIC = 4189$, $\nu = 0$. Right: Box-Cox normal model, $DIC = 4142$, $EAP(\nu) = .19$, $.95HPD(\nu) = [.12, .24]$.

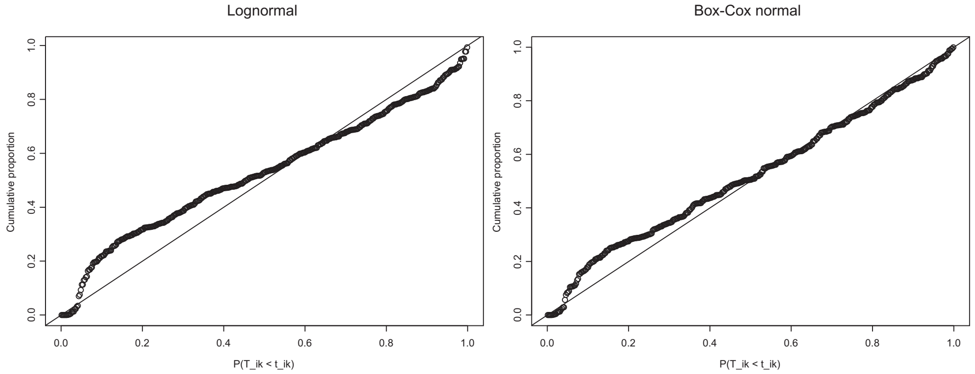


Fig. 4.7. Cumulative probability plots of the posterior probabilities of item 4. Left: lognormal model, $DIC = 3918$, $\nu = 0$. Right: Box-Cox normal model, $DIC = 3810$, $EAP(\nu) = .26$, $.95HPD(\nu) = [.21, .31]$.

Differences in the times required to solve the items are reflected by the time intensities λ_k of the items. That is, if Item 1 is more time intensive than Item 2, this will be reflected as $\lambda_1 > \lambda_2$. From (4.5), it can be seen that then the expected RTs on Item 1 will be higher than those on Item 2 as well: $E(T_{\lambda_1}) > E(T_{\lambda_2})$, which holds for every ζ. For the speed parameters, the relationship with the expected RTs is negative. That is, if Person 1 has a speed of working ζ_1 greater than Person 2 with speed ζ_2, then it holds for every item that for their expected RTs that $E(T_{\zeta_1}) < E(T_{\zeta_2})$. Discrimination parameter ϕ does not affect these relationships. It only controls the rate of decrease in expected RT on an item for one step of increase in speed of a test taker.

For $\nu_k = \nu, k = 1, \ldots, K$ (that is, one common transformation parameter for all items in the test), the interpretation above holds. All parameters are on the same transformed time scale, so it does not make any difference whether $\nu = 0$ or $\nu \neq 0$. Also, the sign of the relationships between $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ or $(\boldsymbol{a}, \boldsymbol{\phi})$, modeled at the second level, remain the same. Note, however, that the interpretations of time intensity and speed of the parameters do not hold for the original time scale. For instance, two items on the log-time scale both with $\lambda = 3$, but with $\tau_1^2 = 2$ and $\tau_2^2 = 4$, have a mean on the time scale of $\exp(3 + 2/2)$ and $\exp(3 + 4/2)$, respectively.

Things are different for item-specific transformations, i.e., when $\nu_k \neq \nu, k = 1, \ldots, K$. These transformations result in item specific scales. As a result, it is impossible to interpret differences between the item parameter estimates directly as differences between item characteristics. To do so, an extra scaling step would be required. Observe that for RTs on multiple tests, each with their own transformation, the same problem occurs and a scaling step would be required as well. Although the scale of the item varies under transformations, it can be seen that for two persons with $\zeta_1 > \zeta_2$ their expected response times (on any item) are still ordered by $E(T_{\zeta_1}) < E(T_{\zeta_2})$, regardless of the transformation. So, by definition, the ranking of the speed parameters is invariant under these transformations. Thus, item-specific transformations do affect the scale of the population distribution, σ_ζ^2, as well as the covariance between ability and speed. But they do not lead to interpretative difficulties for the speed parameters or the dependency between ability and speed.

In practice, however, difficulties might arise in the case of missing data. Even when the missing data are ignorable, the analysis may still result in different scales for different test takers: as the scale is item specific, a test taker who misses an item immediately works on a different speed scale.

In conclusion, the following practical guidelines can be given:

- The case of a common transformation parameter for all items in a test maintains the interpretation of the RT model parameters. It gives the researcher the freedom to fit different distributional shapes to the RT data and admits comparisons between the person and the item parameter estimates for the test.
- When the interest is in parameter estimates for multiple tests, the transformation parameter should be restricted to be common to all tests. Then all parameters are on the same scale, and no additional equating of scale is necessary to make comparisons.
- More general item-specific transformations are mainly of main interest when the focus is on inferences with respect to the ranking of the person parameters. The item parameters are not directly comparable and would require rescaling to a common scale first. An example where the item specific transformation might be of interest is the study of possible aberrant behavior of test takers, for which van der Linden and Guo (in press) presented an approach based on residual analysis. Then, the focus is on the individual person-item combinations and model fit is important to avoid misleading conclusions.

4.8 Discussion

Transformations to normality have obvious and much exploited advantages for the statistical modeling of non-normal data. For modeling response times in a psychometric application, the log-transform has proven to be useful. However, this study was motivated by a data set for which the lognormal model was not able to capture certain aspects of the data. Therefore, the class of Box-Cox transformations was considered, which allows for more flexibility in the description of the data. The examples illustrated how the Box-Cox transformation parameter affects the shape of RT distributions and, as a result, improves the description of the data.

In Section 2, the full modeling framework for responses and RTs on test items was developed to place the RT model in a broader context. A strong feature of the Box-Cox model is that its transforming of the data leaves the standard modeling framework intact.

In educational testing, it makes sense to decompose observed RTs into item effects (time intensity) and person effects (speed). Therefore, the parameters of the RT model presented in (4.5) have a clear interpretation in an educational context. Also, its conjugacy with the MVN level-2 models for the person and item parameters allows for straightforward modeling of the dependencies between the parameters in the level-1 models (van der Linden, 2007; Fox et al., 2007).

Transforming the data instead of the model parameters provides the flexibility of using different distributional shapes for the RTs, while the MCMC algorithm is easily extended with an additional sampling step. On the other hand, for instance, the use of a more flexible 3-parameters Weibull distribution instead of the Box-Cox transformation would require the replacement of the MCMC steps for the current normal RT model by much more complicated procedures since the conjugacy between the level-1 and level-2 models is then lost.

4.9 Appendix: Estimation of the Hierarchical Framework

This section briefly outlines the MCMC algorithm for the full hierarchical framework. A simulation study to illustrate the parameter recovery of this algorithm is given below. The model can be identified by setting the means of the person parameters $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ to zero $(\boldsymbol{\mu}_P = 0)$ and by specifying $\prod_{k=1}^{K} \phi_k = 1$ and $\prod_{k=1}^{K} a_k = 1$ which fixes the scale of the latent variables. Fox et al. (2007) provides a Gibbs sampler that uses identification for the ability scale by restricting $\sigma_\theta^2 = 1$, where the identifying restrictions are directly incorporated into the prior distributions.

4.9.1 Algorithm

Linear Measurement Models for Augmented Data

The vector of augmented data $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ minus the vector of difficulty parameters, $\mathbf{b}^T$, and the similar vector of response times $\log \mathbf{T}_i = (\log T_{i1}, \ldots, \log T_{iK})$

minus the vector of time intensity parameters, $\boldsymbol{\lambda}^t$, are stacked in a vector $\mathbf{Z}_i^*$. Then, both measurement models can be presented as a linear regression structure,

$$\mathbf{Z}_i^* = \big(\mathbf{a} \oplus -\boldsymbol{\phi})(\theta_i, \zeta_i)^t + \mathbf{e}_i \tag{4.19}$$

$$= \mathbf{x}_I \boldsymbol{\xi}_i + \mathbf{e}_i \tag{4.20}$$

where $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{C}_{2K})$ and $\boldsymbol{C}_{2K} = \mathbf{I}_K \oplus \mathbf{I}_K \boldsymbol{\tau}^2$.

Similarly, let $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{nk})^t$ and the vector of log response times, $\log \mathbf{T}_i = (\log T_{1k}, \ldots, \log T_{nk})^t$, to item k be stacked in a vector $\mathbf{Z}_k^*$. Define covariate matrices $\mathbf{H}_\theta$ and $\mathbf{H}_\zeta$ as $\big(\boldsymbol{\theta}, -\mathbf{1}_n\big)$ and as $\big(-\boldsymbol{\zeta}, \mathbf{1}_n\big)$, respectively. A regression structure for the item parameters can be presented as

$$\mathbf{Z}_k^* = \big(\mathbf{H}_\theta \oplus \mathbf{H}_\zeta)(a_k, b_k, \phi_k, \lambda_k)^t + \mathbf{e}_k \tag{4.21}$$

$$= \mathbf{x}_P \boldsymbol{\Omega}_k + \mathbf{e}_k, \tag{4.22}$$

where $\mathbf{e}_k \sim N(\mathbf{0}, \boldsymbol{C}_{2N})$ and $\boldsymbol{C}_{2N} = \mathbf{I}_N \oplus \mathbf{I}_N \tau_k^2$.

Hyperpriors

As a hyperprior for $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, a normal-inverse-Wishart distribution is chosen. That is,

$$\boldsymbol{\Sigma}_I \sim Inv - Wishart_{\upsilon_I}\big(V_I^{-1}\big) \tag{4.23}$$

$$\boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I \sim \mathcal{N}\big(\boldsymbol{\mu}_{I0}, \boldsymbol{\Sigma}_I/\kappa\big), \tag{4.24}$$

where υ_I and V_I are the degrees of freedom and scale matrix of the inverse Wishart distribution, $\boldsymbol{\mu}_{I0}$ is the prior mean and κ the number of prior measurements.

Similarly, as a hyperprior for $\boldsymbol{\Sigma}_P$, an inverse-Wishart distribution is chosen. That is,

$$\boldsymbol{\Sigma}_P \sim Inv - Wishart_{\upsilon_P}\big(V_P^{-1}\big) \tag{4.25}$$

where υ_P and V_P are the degrees of freedom and scale matrix of the inverse Wishart distribution. The mean $\boldsymbol{\mu}_P$ is fixed at $\mathbf{0}$ with probability 1 because of the identification restrictions.

MCMC Algorithm

Estimation of all model parameters for the full hierarchical framework proceeds as follows:

Step 1. Sample augmented response data according to (4.3), given the values for the item and ability parameters.

Step 2. Sample values for the item parameter from $p(\boldsymbol{\Omega}_k \mid \mathbf{Z}_k^*, \boldsymbol{\xi}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ for $(k = 1, \ldots, K)$. A product of a normal likelihood and a normal prior again leads to a normal posterior distribution. So, from (4.22) and (4.8), it follows that

$$\boldsymbol{\Omega}_k \sim MVN(\boldsymbol{\mu}_{\Omega_k}, \boldsymbol{\Sigma}_{\Omega_k}) \tag{4.26}$$

where $\boldsymbol{\Sigma}_{\Omega_k}^{-1} = \mathbf{x}_P^t \boldsymbol{C}_{2N}^{-1} \mathbf{x}_P + \boldsymbol{\Sigma}_I^{-1}$ and $\boldsymbol{\mu}_{\Omega_k} = \boldsymbol{\Sigma}_{\Omega_k}(\mathbf{x}_P^t \boldsymbol{C}_{2N}^{-1} \mathbf{Z}_k^* + \boldsymbol{\Sigma}_I^{-1} \boldsymbol{\mu}_I)$.

Step 3. Sample values for the ability speed parameters from a multivariate normal distribution. Analogous to Step 2, the full conditional posterior distribution is constructed from a multivariate normal likelihood, (4.20) and a multivariate normal prior distribution as

$$\boldsymbol{\xi}_i \sim MVN(\boldsymbol{\mu}_{\xi_i}, \boldsymbol{\Sigma}_{\xi_i}) \tag{4.27}$$

where $\boldsymbol{\Sigma}_{\xi_i}^{-1} = \mathbf{x}_I^t \boldsymbol{C}_{2K}^{-1} \mathbf{x}_I + \boldsymbol{\Sigma}_P^{-1}$ and $\boldsymbol{\mu}_{\xi_i} = \boldsymbol{\Sigma}_{\xi_i}(\mathbf{x}_I^t \boldsymbol{C}_{2K}^{-1} \mathbf{Z}_i^* + \boldsymbol{\Sigma}_P^{-1} \boldsymbol{\mu}_P)$.

Step 4. For the residual variance τ_k^2, a conjugate inverse Gamma prior with parameters $Inv - Gamma(g_1, g_2)$ is assumed. The posterior is then again an inverse Gamma distribution with parameter $g_1 + N/2$ and scale parameter $g2 + (\mathbf{t}_k - (-\phi_k \boldsymbol{\zeta} + \lambda_k))^t (\mathbf{t}_k - (-\phi_k \boldsymbol{\zeta} + \lambda_k))/2$.

Step 5. Draw a new value for ν from a proposal density $\varphi(\nu^* \mid \nu)$ and accept the draw with the probability specified in (4.11).

Step 6. The hyperprior parameters are related to a multivariate normal model for the person parameters, $\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P$, or a multivariate model for the item parameters, $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$.

- The full conditional posterior distribution of $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ has a normal-inverse-Wishart distribution (e.g., Gelman et al., 2004). It follows that

$$p\big(\boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I, \boldsymbol{\mu}_0, \boldsymbol{\Omega}, V_I\big) = N\big(\big(\kappa \boldsymbol{\mu}_0 + K \bar{\boldsymbol{\Omega}}\big)/(\kappa + K), \boldsymbol{\Sigma}_I/(K + \kappa)\big), \tag{4.28}$$

  where $\bar{\boldsymbol{\Omega}} = \sum_k \boldsymbol{\Omega}_k / K$. Subsequently, the full conditional of $\boldsymbol{\Sigma}_I$ is an inverse-Wishart with parameters $K + \upsilon_I$ and scale parameter $V_I + \sum_k \big(\boldsymbol{\Omega}_k - \bar{\boldsymbol{\Omega}}\big)\big(\boldsymbol{\Omega}_k - \bar{\boldsymbol{\Omega}}\big)^t + \frac{\kappa K}{\kappa + K}\big(\bar{\boldsymbol{\Omega}} - \boldsymbol{\mu}_0\big)\big(\bar{\boldsymbol{\Omega}} - \boldsymbol{\mu}_0\big)^t$.
- Similarly, the full conditional of $\boldsymbol{\Sigma}_P$ is an inverse-Wishart with parameters $N + \upsilon_P$ and scale parameter $V_P + \sum_n \big(\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}}\big)\big(\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}}\big)^t + \frac{\kappa N}{\kappa + N}\big(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_{P0}\big)\big(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_{P0}\big)^t$.

4.9.2 Simulation Study for Parameter Recovery

To illustrate the parameter recovery for the algorithm for the full hierarchical framework, a simulation study was performed. We simulated responses under the 2PL model and RTs under the RT-model with $\nu = 0.3$ for 1000 test takers answering 20 items. The ability and speed parameters were randomly drawn from $\theta_i \sim N(0, 1)$, $\zeta_i \mid \theta_i \sim N(0, 1)$ with $\rho = 0.5$ (see Equation (4.7)). The item parameters were randomly drawn according to: $a_k \sim N(1, .1)$, $b_k \sim N(0, 1)$, $\lambda_k \sim N(10, 2)$ and the time discrimination parameters were generated from $\phi_k \sim N(2, .3)$ and subsequently standardized to assure that $\prod_{k=1}^{K} \phi_k = 1$.

The model was identified by setting $\boldsymbol{\mu}_P = \mathbf{0}$, $\sigma_\theta^2 = 1$ and $\prod_{k=1}^{K} \phi_k = 1$. The prior variance σ_ζ^2 was chosen to be noninformative and was set to 100, the prior covariance between ability and speed was chosen to be $\rho_0 = 0$. Priors for the item parameters were noninformative as well and were chosen to be $\boldsymbol{\mu}_{I0} = (1, 0, 1, 0)$ and a diagonal matrix with 10 on its diagonal for the prior covariance matrix. That is, we assumed prior independence between the response and RT model parameters.

The algorithm was run for 100,000 iterations of which every tenth was stored to account for autocorrelation induced by the Box-Cox transformation, since the transformation affects the mean and variance of the RT distribution. From the stored samples, the first 1,000 were discarded as burn-in. The simulated (true) values and the re-estimated values (Expected A Posteriori, EAP) plus standard deviations of the model parameters are given in Table 4.3 below. Graphical inspection of the re-estimated ability and speed parameters showed that their values were in good agreement with their true values. The EAP estimate of the transformation parameter was $E(\nu) = 0.309$ and its .95 highest posterior density region was estimated to be $(0.293, 0.323)$, which includes the true value of $\nu = 0.3$. It can be seen that for this example the parameter recovery of the algorithm was good, even with a moderate number of items.

Table 4.3. Simulated and re-estimated model parameters.

a			b			α			β		
True	EAP	SD	True	EAP	SD	True	EAP	SD	True	EAP	SD
0.92	1.04	0.08	-1.66	-1.77	0.10	0.79	0.85	0.03	7.99	8.14	0.16
0.96	0.96	0.06	0.12	0.14	0.05	0.77	0.78	0.03	10.27	10.49	0.23
1.00	1.03	0.08	1.44	1.53	0.08	1.20	1.20	0.03	8.30	8.49	0.17
1.06	1.07	0.07	-0.96	-0.94	0.06	1.22	1.18	0.03	7.24	7.33	0.14
0.91	0.87	0.06	0.33	0.33	0.05	1.27	1.26	0.03	6.41	6.50	0.11
1.04	0.96	0.06	-0.49	-0.56	0.05	0.85	0.88	0.03	5.50	5.55	0.09
0.84	0.74	0.08	1.79	1.76	0.09	0.82	0.84	0.03	7.63	7.80	0.15
0.93	0.98	0.07	-0.88	-0.85	0.06	1.20	1.23	0.03	6.75	6.92	0.12
0.98	1.01	0.06	0.04	0.08	0.05	0.73	0.73	0.03	9.04	9.26	0.19
0.95	0.86	0.08	1.59	1.47	0.08	1.06	1.05	0.04	4.00	3.99	0.06
0.90	0.94	0.12	-2.60	-2.69	0.18	0.86	0.80	0.03	7.41	7.53	0.14
0.88	0.96	0.09	-1.69	-1.85	0.10	1.04	1.03	0.03	6.85	6.94	0.12
1.08	1.06	0.07	0.69	0.69	0.05	1.06	1.08	0.03	11.03	11.34	0.26
1.00	1.06	0.07	0.89	0.91	0.06	1.15	1.14	0.03	8.31	8.48	0.17
1.11	1.16	0.07	0.05	0.08	0.05	1.01	1.05	0.03	8.72	8.89	0.18
0.68	0.71	0.08	1.73	1.81	0.09	0.95	0.93	0.03	4.87	4.89	0.07
1.02	1.02	0.07	-0.99	-0.97	0.06	1.25	1.21	0.03	8.21	8.32	0.16
1.10	1.23	0.08	-1.26	-1.37	0.08	1.01	1.02	0.03	9.65	9.82	0.21
1.12	1.09	0.06	-0.03	-0.01	0.05	0.88	0.88	0.03	10.22	10.47	0.23
1.11	1.06	0.08	1.67	1.69	0.09	1.14	1.12	0.03	8.77	8.98	0.18

5

IRT Parameter Estimation with Response Times as Collateral Information

Summary. Hierarchical modeling of responses and response times on test items facilitates the use of response times as collateral information in the estimation of the response parameters. Two sources of collateral information are identified: (i) the joint information in the responses and the response times summarized in the estimates of the hyperparameters and (ii) the information in the posterior predictive distribution of the response parameters given the response times. The latter is shown to be a natural empirical prior distribution for the estimation of the response parameters. Unlike traditional hierarchical IRT modeling, where the gain in estimation accuracy is typically paid for by an increase in bias, use of this posterior predictive distribution improves both the accuracy and the bias of IRT parameter estimates. In an empirical study, the improvements are demonstrated for the estimation of the person and item parameters in the 3-parameter response model.

Item response theory (IRT) models belong to the class of models that explain the data for each unit of observation by different parameters. One of the main advantages of a hierarchical approach to this class of models is the possibility of borrowing information on one parameter from the data collected for the units associated with the other parameters. This borrowing is realized through the assumption of a common distribution of the parameters as a second level in the statistical model for the data. The estimator of the parameter then typically compromises between this distribution and the likelihood associated with the data. In doing so, it tends to strike a profitable balance between ignoring the data on the other parameters (separate estimates) and the more reckless assumption that all parameters are identical (pooled estimates). The profit typically occurs in the form of a more favorable tradeoff between a higher efficiency of the inference at the cost of a less serious increase in bias. The profit is reflected in lower mean-squared error of the estimates.

One of the first examples in test theory demonstrating this principle is the classical true-score estimate based on Kelley's regression function,

$$E(T \mid X = x) = \rho_{XX'}x + (1 - \rho_{XX'})\mu_T, \tag{5.1}$$

where X is the observed score of the test taker, T the true score, μ_T the mean true score in the population of test takers, and $\rho_{XX'}$ the reliability of the test (Lord & Novick, 1968, sect. 3.7). An estimate of the test takers true score, $\widehat{\tau}$, is obtained by substituting estimates of μ_T and $\rho_{XX'}$ derived from the marginal distribution of the observed scores into (5.1). The estimate compromises between $X = x$ as a direct estimate of ζ and the estimate of the population parameter μ_T. In the representation in (5.1), the weights are $\rho_{XX'}$ and $1 - \rho_{XX'}$. But, using a well-known variance partition in classical test theory, the estimate can be shown to be also equivalent to the precision-weighted average of x and $\widehat{\mu}_T$ (Novick & Jackson, 1974, Eq. 9.5.11).

As discussed extensively in Novick and Jackson (1974, sect. 9.5), the Kelley estimate is an instance of the more formal problem of estimating many means simultaneously. Later examples of the applications of the same principle of borrowing information to this problem are the estimation of multiple regressions in m groups from normal data in Novick, Jackson, Thayer, and Cole (1972) and the estimation of proportions in m groups from binomial data in Novick, Lewis, and Jackson (1973). An instructive empirical application of the estimation of multiple regressions is the often-cited study of the effects of coaching schools on SAT scores in Rubin (1981).

We have not yet specified the nature of the "distribution of the parameters." In a frequentist approach, the units of observation are typically assumed to be sampled from a population and hence their parameters are taken to be random. The interpretation of Kelley's estimate in classical test theory belongs to this approach. From a Bayesian perspective, the assumption of random sampling from a population is not necessary. When the units of analysis can be assumed to be exchangeable, any density that approximates the distribution of the parameters for the data set becomes a profitable common prior for the inference of their posterior distributions. The difference between a hierarchical model with a population distribution and this empirical Bayes approach resides only in their motivation and interpretation; the more formal aspects of both approaches involve the same two-level structure. For an introduction to the empirical Bayes approach, see, (e.g. Carlin & Louis, 2000, chap. 3).

In order to emphasize that, in this tradition of hierarchical or empirical Bayes modeling, the information that is borrowed is collected simultaneously with the direct information on the parameters, Novick and Jackson (1974, sect. 9.5) introduced the notion of *collateral information.* This term avoids the more temporal connotation in the Bayesian terminology of *prior information*, which seems to suggest that the information should always be present before any data on the parameters is collected. Mislevy and Sheehan (1989) used collateral information about examinees for the calibration of item parameters.

It should be noticed that the use of the term "information" differs from that elsewhere in scientific endeavors, where it is typically taken to imply that observations can be predicted from other variables. However, collateral information in the hierarchical sense does not require the presence of any predicting variables but is already available if the units of observation can be assumed to follow a common distribution. If the assumption holds, as soon as we collect data for the parameters of some of the units, we get information on all of them; for example, about their typical range of values.

In this paper, we combine first-level models for the responses and the response times (RTs) by the test takers on the items with second-level models for the joint empirical distributions of their item and person parameters. As a result, we are not only able to borrow information on the response parameter for one item from the responses collected for the other items but also from the RTs collected for the *same* item. The same borrowing is possible for the response parameter for each person. Because responses and RTs are always recorded simultaneously, the additional information in the RTs is literally collateral. Surprisingly, as will be shown later in this paper, the fact that the collateral information is specific for the individual items and persons leads to improvement of *both the accuracy and the bias of the estimates.* In doing so, the information thus breaks the bias-accuracy tradeoff typical of more traditional hierarchical modeling.

The research in this paper was motivated by the fact that now that computer-based testing has become more dominant, and RTs in this mode of testing are automatically recorded, it would be imprudent to ignore such information. Before showing how joint hierarchical modeling of responses and RTs helps to exploit such information, we first take a closer look at the role of collateral information in the more traditional problem of parameter estimation in a separate IRT model.

5.1 Collateral Information in IRT

The example considered is the estimation of ability parameter θ_i for a test taker i in a response model (e.g., the well-known three-parameter logistic model) of which the item parameters are already known. The case is met, for instance, when a test from a calibrated item pool is used to measure the abilities of several test takers.

Suppose the test takers are from a population with a normal distribution of ability $N(\mu_\theta, \sigma^2_\theta)$, of which the mean μ_θ and variance σ^2_θ have already been estimated with enough precision to treat them as known. Estimates of θ_i that capitalize on this information should be based on the posterior distribution

$$f(\theta_i \mid \mathbf{y}_i, \widehat{\mu}_\theta, \widehat{\sigma}^2_\theta) \propto f(\theta_i; \mathbf{y}_i) f(\theta_i \mid \widehat{\mu}_\theta, \widehat{\sigma}^2_\theta), \tag{5.2}$$

where $\mathbf{y}_i = (y_{i1,\dots,}y_{iK})$ are the responses by i on the K items in the test, $f(\theta_i; \mathbf{y}_i)$ is the likelihood associated with the response vector of i, and $\widehat{\mu}_\theta$ and $\widehat{\sigma}^2_\theta$ are the available estimates of the population parameters.

The mean of this distribution, which is often used as a point estimate of θ_i, is generally known to have a smaller mean square error than an estimate based on the likelihood $f(\theta_i; \mathbf{y}_i)$ only. The decrease is due to the information in the population density $f(\theta_i \mid \widehat{\mu}_\theta, \widehat{\sigma}^2_\theta)$ in the right-hand side of (5.2), which tells us, for instance, where the ability parameters in the population are concentrated and how much they are dispersed. The decrease is paid for by an increase in the bias of the ability estimate toward the mean of the population of test takers or the domain of items.

This combination of effects is an example of the well-known bias-accuracy tradeoff in statistics. However, in hierarchical modeling, the tradeoff is exploited to work to our advantage; the improved accuracy is generally realized at a less serious increase in bias.

The same principle can be shown to hold for the estimation of the item parameters. The only change necessary is the replacement of the population density $f(\theta_i \mid \widehat{\mu}_\theta, \widehat{\sigma}^2_\theta)$ in (5.2) by that for the (joint) distribution of the item parameters.

In (5.2), we assumed that the population parameters μ_θ and σ^2_θ had already been estimated. This was only because we wanted to emphasize that the estimate of θ_i was actually based on *empirical* information about the population distribution. Also, we assumed that the item parameters were already known. Both assumptions are not necessary, though. The same borrowing of information occurs when we fit a hierarchical model with unknown parameters for the items and persons and hyperparameters for their distributions and estimate all unknowns simultaneously. The second-level distribution that is fitted then constrains the estimates of the parameters in the response model in precisely the same way as when it had been fitted previously. (Actually, due to the fact that some of the parameters in the response model are not identifiable, the notion of collateral information becomes more subtle. For example, for the 3PL model, a common choice is to use $\mu_\theta = 0$ and $\sigma^2_\theta = 1$ as identifiability constraints. The choice leads to the replacement of the population density in (5.2) by $f(\theta_i \mid 0, 1)$. The density does not involve any parameter estimation but remains a source of *empirical* information; knowledge of its shape allows us to derive information on one parameter from information on the others.)

5.2 Hierarchical Model

In order to profit fully from the information on the IRT parameters in the RTs, we have to adopt a model for the RTs and to model the common distribution of all person and item parameters. The result is a hierarchical framework with the IRT and RT models as first-level components and population and domain models for the IRT and RT parameters as second-level components. The RT model in (5.4) below was proposed in van der Linden (2006) whereas the extension with the second-level models in (5.5)–(5.9) was introduced in van der Linden (2007). These models are borrowed to demonstrate the benefits of using RTs as collateral information when estimating IRT parameters; other models substituted in the same type of hierarchical framework would do the same job.

5.2.1 IRT and RT Models

As the first-level model for the responses of test takers $i = 1, ..., N$ on items $k = 1, ..., K$, we use the three-parameter normal-ogive (3PNO) model, which gives the probability of a correct response on item k by person i as

$$P(Y_{ik}=1;\theta_i,a_k,b_k,c_k)=c_k+(1-c_k)\Phi(a_k\theta_i-b_k), \tag{5.3}$$

where $\Phi(\cdot)$ denotes the normal distribution function and a_k, b_k, and c_k are the discrimination, difficulty, and guessing parameters for item k, respectively.

Response-time distributions are often approximated well by lognormal distributions (for a review of alternatives, see Schnipke & Scrams, 2002. Therefore, analogous to the IRT model in (5.3), the RTs are modeled with a speed parameter ζ_i for test taker i and time intensity and discrimination parameters λ_k and ϕ_k for item k, respectively. Let T_{ik} denote the RT of test taker i on item k. The lognormal model posits that

$$T_{ik}=-\phi_k\zeta_i+\lambda_k+\epsilon_{ik},\epsilon_{ik}\sim N(0,\tau_k^2). \tag{5.4}$$

Notice that, except for the difference in sign, which is due to the negative relationship between time and speed, the two parameters for speed and time intensity in (5.4) play an identical role as those for the ability and item difficulty in (5.3). However, unlike (5.3), RT distributions have a natural zero and do not involve the estimation of any lower asymptote.

5.2.2 Population and Domain Models

The population model specifies the joint distribution of the person parameters θ and ζ. We assume that the distribution is bivariate normal,

$$(\theta,\zeta)\sim MVN(\mu_{\mathcal{P}},\mathbf{\Sigma}_{\mathcal{P}}), \tag{5.5}$$

where

$$\mu_{\mathcal{P}}=(\mu_\theta,\mu_\zeta) \tag{5.6}$$

and covariance matrix

$$\mathbf{\Sigma}_{\mathcal{P}}=\begin{pmatrix}\sigma_\theta^2 & \rho\\ \rho & \sigma_\zeta^2\end{pmatrix}. \tag{5.7}$$

Likewise, item parameters a_k, b_k, ϕ_k, and λ_k in the response and RT models are assumed to have a multivariate normal distribution,

$$(a,b,\phi,\lambda)\sim MVN(\mu_{\mathcal{I}},\mathbf{\Sigma}_{\mathcal{I}}), \tag{5.8}$$

where

$$\mu_{\mathcal{I}}=(\mu_a,\mu_b,\mu_\phi,\mu_\lambda) \tag{5.9}$$

and matrix $\mathbf{\Sigma}_{\mathcal{I}}$ has all variances and covariances between the item parameters as elements.

This hierarchical model is not yet fully identifiable. In addition to the usual lack of identification for a hierarchical IRT model, the parameters λ_k and ζ_i in the RT model are not identified; addition to a constant to λ_k can be compensated by addition of the same constant to ζ_i. Identifiability is obtained if we set $\mu_{\mathcal{P}}=\mathbf{0}$, $\prod_{k=1}^{K}\phi_k=1$ and $\sigma_\theta^2=1$.

5.2.3 Bayesian Estimation

In the empirical examples later in this paper, the model parameters were estimated using Bayesian estimation with data augmentation and a Gibbs sampler. The method uses normal inverse-Wishart priors for the mean vectors and covariance matrices for the multivariate models in (5.5) and (5.8), which have the convenient property of conjugacy (Gelman et al., 2004, sect. 3.6). For the data augmentation, see Albert (1992) or Johnson and Albert (1999). For a discussion of Gibbs sampling, see Gelman et al. (2004, chap. 11) or Gelfand and Smith (1990). For technical details of the estimation method, see Klein Entink, Fox, and van der Linden (in press) and van der Linden (2007). A package of procedures in the statistical language *R* that implement the method is described in Fox et al. (2007).

5.3 Different Sources of Information

We demonstrate the same principle as in (5.2) but this time for a test taker j with response vector $\mathbf{y}_i = (y_{i1,\ldots,}y_{iK})$ and RT vector $\mathbf{t}_i = (t_{i1,\ldots,}t_{iK})$. Again, without loss of generality, we assume that the second-level means, $\mu_{\mathcal{P}}$ and $\mu_{\mathcal{I}}$, and covariance matrices $\mathbf{\Sigma}_{\mathcal{P}}$ and $\mathbf{\Sigma}_{\mathcal{I}}$, have already been estimated during item calibration. Consequently, θ_i and ζ_i are the only unknown parameters.

The complication we are now faced with is an estimation problem with two separate likelihoods–a primary likelihood that is response based and one associated with the RTs under the lognormal model in (5.4). In order to assess the improvement in the estimation of θ_i relative to (5.2), we try to derive the posterior distribution of θ_i as a product of the primary likelihood and whatever other factors necessary. The comparison of these other factors with the prior distribution of θ_i in (5.2) should allow us to assess the improvement in the estimation of θ_i relative to (5.2).

The posterior distribution of θ_i follows from the joint distribution of θ_i and ζ_i given all known quantities

$$f(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) = \int f(\theta_i, \zeta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) d\zeta_i. \tag{5.10}$$

For the integral, it holds that

$$\int f(\theta_i, \zeta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) d\zeta_i \propto \int f(\mathbf{y}_i, \mathbf{t}_i; \theta_i, \zeta_i) f(\theta_i, \zeta_i \mid \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) d\zeta_i. \tag{5.11}$$

Hence, because of local independence,

$$f(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) \propto \int f(\mathbf{y}_i; \theta_i) f(\mathbf{t}_i; \zeta_i) f(\theta_i, \zeta_i \mid \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}}) d\zeta_i. \tag{5.12}$$

Factorizing $f(\theta_i, \zeta_i \mid \widehat{\mu}_{\mathcal{P}}, \widehat{\mathbf{\Sigma}}_{\mathcal{P}})$, we obtain

$$
\begin{aligned}
f(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) &\propto f(\mathbf{y}_i; \theta_i) \int f(\theta_i \mid \zeta_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) f(\mathbf{t}_i; \zeta_i) f(\zeta_i \mid \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) d\zeta_i \\
&\propto f(\mathbf{y}_i; \theta_i) \int f(\theta_i \mid \zeta_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) f(\zeta_i \mid \mathbf{t}_i) d\zeta_i, \qquad (5.13)
\end{aligned}
$$

where the second step follows from the definition of the posterior distribution of ζ_i as

$$
f(\zeta_i \mid \mathbf{t}_i) \propto f(\mathbf{t}_i \mid \zeta_i) f(\zeta_i \mid \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}). \qquad (5.14)
$$

For the integral in the second line of (5.13), it holds that

$$
\int f(\theta_i \mid \zeta_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) f(\zeta_i \mid \mathbf{t}_i) d\zeta_i \propto f(\theta_i \mid \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}). \qquad (5.15)
$$

Notice that the right-hand side is the posterior predictive density of θ_i given $\mathbf{t}_i$; that is, the probability of the test taker's ability θ_i given his/her speed ζ_i integrated over the posterior distribution of ζ_i given the response times $\mathbf{t}_i$.

Thus, we are able conclude that

$$
f(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}) \propto f(\mathbf{y}_i; \theta_i) f(\theta_i \mid \mathbf{t}_i, \widehat{\mu}_{\mathcal{P}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}). \qquad (5.16)
$$

The result has a simple form that is entirely analogous to (5.2). It shows that, when the RTs are used as collateral information, θ_i is estimated from the same likelihood $f(\mathbf{y}_i; \theta_i)$ associated with the response vector $\mathbf{y}_i$ as when the RTs are ignored but with the original prior distribution of θ in (5.2) replaced by the posterior predictive distribution of θ_i given the RTs $\mathbf{t}_i$ for the test taker.

More generally, the result also answers our earlier question of how to deal with the presence of two different likelihoods in the statistical inference for one kind of parameters in an hierarchical framework as in (5.3)–(5.9). The solution is to keep the likelihood of the primary parameters in tact but absorb the second likelihood in the posterior predictive density of the primary parameters given the information collected for the other parameters. (Our use of the term "posterior predictive density" is motivated by its formal definition as a model density—here: that of θ given ζ— integrated over the posterior distribution of its parameters; see, (e.g. Gelman et al., 2004, sect. 1.3). It is only used for prediction in a thought experiment in which we pretend not to know anything about θ but estimate it from the RTs. Its standard use in Bayesian statistics is for a prediction of a new observation under the same density as for the observations that led to the posterior distribution.)

The result in (5.16) enables us to identify *three* different sources of information on θ_i:

1. The information directly available in $\mathbf{y}_i$ in the first factor of (5.16); that is, the regular likelihood $f(\mathbf{y}_i; \theta_i)$ associated with the response vector.
2. The information summarized in the estimates $\widehat{\mu}_{\mathcal{P}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{P}}$ in the second factor. This information is derived from the vectors of responses and RTs in the entire sample of test takers. These estimates generalize the role of $\widehat{\mu}_\theta$ and $\widehat{\sigma}^2_\theta$ in (5.2).

3. The information in the shape of the posterior predictive distribution of the response parameters given the response times. Unlike the preceding source of information, the information in this vector is unique for each individual test taker.

Notice how the use of the RTs for test taker i as collateral information on θ_i turns the common posterior distribution for all test takers in (5.2) into an *individual* distribution for i. This information highlights an important role played by the RTs. They lead not only to a further decrease of the variance of the posterior distribution of θ_i relative to (5.2) but also to move its location closer to the true ability level of the test taker.

Analogous effects of the RTs can be shown to hold for the estimation of the item parameters. To demonstrate this point, the same argument can be followed with $\widehat{\mu}_{\mathcal{P}}$, $\widehat{\mathbf{\Sigma}}_{\mathcal{P}}$, θ_i, and $\mathbf{t}_i$ replaced by $\widehat{\mu}_{\mathcal{I}}$, $\widehat{\mathbf{\Sigma}}_{\mathcal{I}}$, item-parameter vector $\xi_k = (a_k, b_k, c_k)$ and RT vector $\mathbf{t}_k$, respectively. The result is

$$f(\xi_k \mid \mathbf{y}_k, \mathbf{t}_k, \widehat{\mu}_{\mathcal{I}}, \widehat{\mathbf{\Sigma}}_{\mathcal{I}}) \propto f(\mathbf{y}_k; \xi_k) f(\xi_k \mid \mathbf{t}_k, \widehat{\mu}_{\mathcal{I}}, \widehat{\mathbf{\Sigma}}_{\mathcal{I}}). \tag{5.17}$$

Again, it is not necessary to estimate the second-level parameters prior to estimating θ_i and ξ_k. The collateral information in the RTs is retained if the full hierarchical model is fitted to all response data and RTs simultaneously. The estimates of θ_i and ξ_k are then still constrained by the conditional posterior distributions $f(\theta_i \mid \mathbf{t}_i, \mu_{\mathcal{P}}, \mathbf{\Sigma}_{\mathcal{P}})$ in (5.16) and $f(\xi_k \mid \mathbf{y}_k, \mathbf{t}_k, \mu_{\mathcal{I}}, \mathbf{\Sigma}_{\mathcal{I}})$ in (5.17) while the procedure looks for estimates of $(\mu_{\mathcal{P}}, \mathbf{\Sigma}_{\mathcal{P}})$ and $(\mu_{\mathcal{I}}, \mathbf{\Sigma}_{\mathcal{I}})$ that fit the entire data set best.

5.3.1 Bias-Accuracy Tradeoff

It is important to notice the consequences of the replacement of the common prior distribution of θ for all test takers in (5.2) by the *individual* distribution for θ_i in (5.16). The new prior distribution does not only have a smaller variance but, because of its conditioning on the RTs for each individual test taker, also tends to have a location closer to his/her true ability level. The impact of the former is a further increase of the posterior precision of θ_i; the impact of the latter a decrease in the bias of the posterior mean. The size of both effects is dependent on the (absolute) size of the correlation between the response and RT parameters.

As already argued, parameter estimation in traditional IRT hierarchical modeling is subject to the well-known bias-accuracy tradeoff in statistics. For this type of modeling, the addition of a population model for the distribution of the θ parameters leads both to an increase in the bias and accuracy of their estimates. However, as long as a sensible distribution is chosen, the net result—summarized in the mean square errors of the parameters—is positive.

On the other hand, the posterior predictive density of θ_i in (5.16) is conditional on the actual RTs by each test taker i, and therefore serves as a prior with an individual location for each of them. Unlike the use of a population distribution as a common prior, whose location necessarily compromises between the true values of the individual θs, and hence produces a bias in their estimates, these individual priors avoid the necessity of such a compromise. In addition, they have smaller variances and are thus more informative.

5.4 Empirical Example

Simulation studies were conducted to demonstrate the effect of the use of the collateral information in the RTs on the estimation of IRT model parameters. Obviously, the improvements are dependent on the correlation between the parameters in the second level of the hierarchical framework in (5.3)–(5.9). In this study, we focused on the correlation between the speed and ability parameters, ρ. (The effects for the other parameters are analogous.) The evaluations were therefore conducted for a range of alternative sizes of ρ. In order to assess the improvements of the estimation of the item and person parameters in the response model separately, two different studies with a different setup had to be conduced.

5.4.1 Study 1: Ability Estimation

Responses and RTs were simulated for $N = 1,000$ test takers on a 30-item test for five levels of correlation: $\rho = 0, .25, .50, .75, 1$. Only positive correlations were used; the results generalize immediately to negative correlations.

The parameters a_k were randomly drawn from $U(0.8, 1.2)$. In order to guarantee a test with uniform distribution of the item difficulties across θ, we did not sample the difficulty parameters b_i but used equally spaced values on $(-\frac{31}{15}, \frac{31}{15})$ with steps of $\frac{2}{15}$. The guessing parameter was assumed to be .25 for all items, which represents the case of four-choice items. (This parameter was further ignored in the evaluations.) The time intensity parameters for the RT model were randomly drawn from $\lambda \sim N(5, 1)$ and the discrimination parameters ϕ were fixed to 1.

The ability parameters for the test takers were taken to be the $N = 1,000$ quantiles of the $N(0, 1)$ distribution. This decision guaranteed coverage of the whole θ range and allowed us to assess the statistical quality of the estimators of θ with uniform precision over the range. The speed parameters for the test takers were drawn from the conditional distributions of $\zeta|\theta$ for the assumed correlation ρ, where the marginal distribution of ζ was always $N(0, 1)$. In order to estimate the bias and mean square error of the θ estimates, the entire setup was replicated 10 times for each of the five sizes of the correlation between θ and ζ.

The ability parameters of the test takers were estimated for two different cases: First, all item parameters were assumed to be known and the ability parameters were the only parameters estimated (case of measurement using previously calibrated items). Second, both the item and person parameters were treated as unknown and estimated simultaneously (case of ability estimation in a calibration study). Because the results for the two cases showed only minor numerical differences, our report focuses on the first case.

The parameters were estimated using the Gibbs sampler referred to earlier. Noninformative fixed priors were chosen for the item parameters, $\mu_{\mathcal{I}}$ was set equal to

$(1, 0, 1, 0)$, and $\mathbf{\Sigma}_{\mathcal{I}}$ was taken to be a diagonal matrix with variance 10. (Note that we did not estimate the second-level distribution of the item parameters.). For the person parameters, independent, non-informative hyperpriors were specified with $\mu_{\mathcal{P}_0} = 0$ for the mean vector and $\mathbf{\Sigma}_{\mathcal{P}_0}$ a diagonal matrix with variance 10 for the covariance matrix. The Gibbs sampler was run for 10,000 iterations. A burn-in of 500 iterations was sufficient to reach convergence; autocorrelation between the draws appeared to be negligible after more than 10 iterations.

The results were evaluated using the mean square error (MSE) and bias for the EAP estimates of θ (= mean of their posterior distributions) as criteria. The MSEs for the cases without and with the RTs are defined as

$$MSE(\theta_i \mid \mathbf{y}_i) = E[(\theta_i - \hat{\theta}_i)^2 | \mathbf{y}_i] = \int (\theta_i - \hat{\theta}_i)^2 f(\theta_i | \mathbf{y}_i) d\theta_i \tag{5.18}$$

and

$$MSE(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i) = E[(\theta_i - \hat{\theta}_i)^2 | \mathbf{y}_i, \mathbf{t}_i] = \int (\theta_i - \hat{\theta}_i)^2 f(\theta_i | \mathbf{y}_i, \mathbf{t}_i) d\theta_i, \tag{5.19}$$

where, for convenience, the posterior distributions of θ_i in these two expression are denoted without the hyperparameters. Likewise, the bias for the two cases are defined as

$$Bias(\theta_i \mid \mathbf{y}_i) = E[\hat{\theta}_i - \theta_i | \mathbf{y}_i] = \int (\hat{\theta}_i - \theta_i) f(\theta_i | \mathbf{y}_i) d\theta_i \tag{5.20}$$

and

$$Bias(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i) = E[\hat{\theta}_i - \theta_i | \mathbf{y}_i, \mathbf{t}_i] = \int (\hat{\theta}_i - \theta_i) f(\theta_i | \mathbf{y}_i, \mathbf{t}_i) d\theta_i \tag{5.21}$$

For a Gibbs sampler, these expressions can easily be estimated as

$$\frac{1}{M} \sum_{m=1}^{M} (\theta_i - \theta_i^{(m)})^2, \tag{5.22}$$

respectively,

$$\frac{1}{M} \sum_{m=1}^{M} (\theta_i - \theta_i^{(m)}), \tag{5.23}$$

with $m = 1, ..., M$ denoting the iterations of the sampler after burn-in.

Figure 5.1 shows the decrease in MSE, i.e., $MSE(\theta_i \mid \mathbf{y}_i) - MSE(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i)$, due to the use of the RTs as collateral information as a function of θ for the five conditions for ρ. (Notice that the original scale of θ is not used but that for the monotone transformation of θ by its cumulative distribution function. The transformation enables us to average the MSEs over intervals with 10% of the test takers and create uniform precision for the plotted MSEs across the scale.) Clearly, $\rho = 0$ corresponds with the baseline case of fitting the IRT model without using

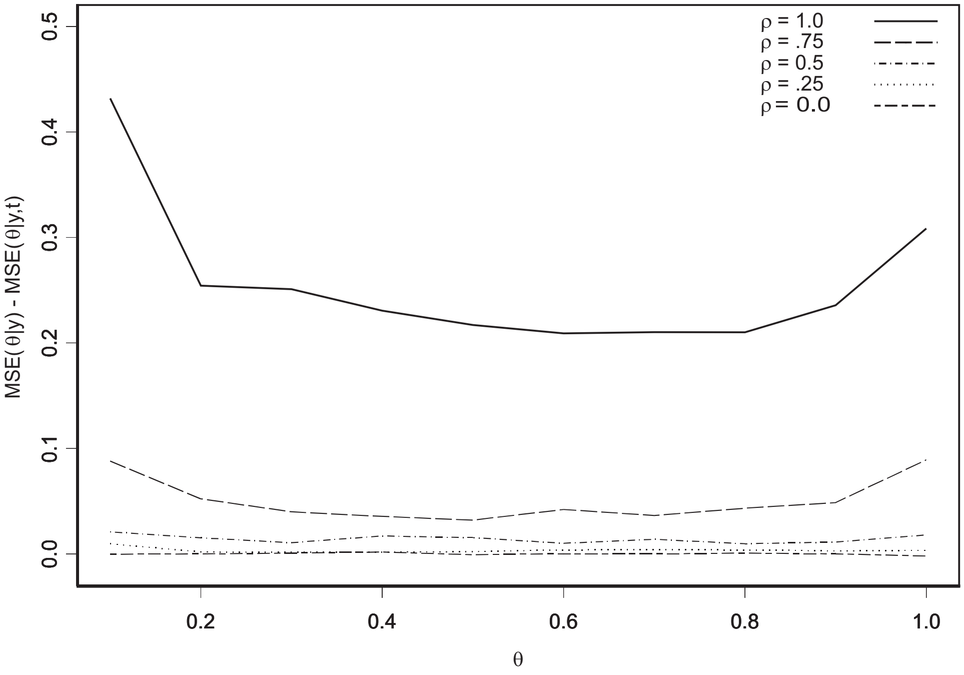


Fig. 5.1. Reduction in the MSE of the estimates of θ for different correlations ρ. (Note: for the scale of θ, see the text.)

any of the RTs. As the correlation increases, the same happens with the gain in the MSE. This trend shows that the more collateral information on the ability parameters in the RTs, the more accurate their estimates become.

Figure 5.2 shows similar plots for the decrease in bias in the estimation of θ, i.e., $Bias(\theta_i \mid \mathbf{y}_i) - Bias(\theta_i \mid \mathbf{y}_i, \mathbf{t}_i)$. For abilities below the population mean $\mu_\theta = 0$ there is a positive difference in the bias between estimation with and without RTs; above the population mean there is a negative difference. The difference is larger for larger correlations ρ. This finding indicates that when the speed parameter is informative for ability, the individual prior distributions of θ_i are pulled away from the population mean by the RTs, resulting in less estimation bias.

In conclusion, from this first study it can be seen that for test takers located near the population mean, the effects are relatively small. But for abilities toward the upper and lower end of the scale (which are hardest to measure), the accuracy of estimation improves considerably. Note that in this setup the information in the RTs improves the estimated abilities even for a moderate correlation of .5 between ability and speed. Moreover, it is obvious that these results also generalize to negative correlations between the two parameters.

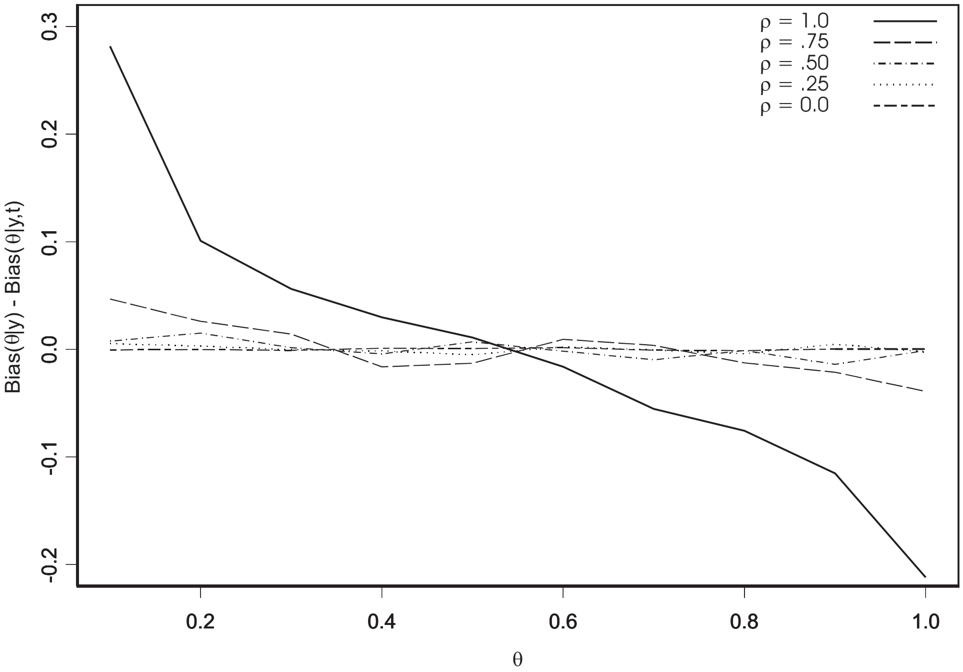


Fig. 5.2. Reduction in the bias of the estimates of θ for different correlations ρ. (Note: for the scale of θ, see the text.)

5.4.2 Study 2: Item Calibration

The second study was to evaluate the use of the collateral information in the RTs in item calibration. Its setup had to be more complicated because we wanted to isolate the effects of the RTs on the item parameters a_k and b_k. The trick we used was to first obtain the posterior distributions of θ for the persons in the sample for known item and speed parameters and then estimate the items parameters using a version of the Gibbs sampler in which the values for the θs were drawn directly from their known posterior distributions. First, we held the parameters b_k constant and varied the parameters a_k. Second, the role of these two kinds of parameters was reversed. Finally, to avoid the effects of possible tradeoffs between the a_k and c_k parameters - known to exist due to occasional weak identifiability of the 3PL model -, the simulation was conducted with c_k fixed at .25 for all items.

More specifically, the setup was the following. First, a 10-item test with known item parameters was used to obtain the posterior densities of the ability parameters of $N = 300$ test takers. Item parameters b_k were taken to be equally spaced on $[-1.8, 1.8]$ with steps of size 0.4; parameters a_k were randomly drawn from $U(.8, 1.2)$. The 300 test takers were selected to have ability parameters corresponding to the 300 quantiles of the $N(0, 1)$ distribution. The speed parameters were randomly drawn from the conditional distributions of $\zeta|\theta$ for the appropriate correlation ρ. For these parameters, response patterns $\mathbf{y}_1$ were simulated for the test takers. The Gibbs sampler was then run with known item parameters and speed parameters to obtain 10,000 draws (after burn in) from the posterior distributions of θs. The procedure was repeated for four sizes of correlation between the ability and speed parameters: $\rho = 0, .5, .7$, and .9.

Second, the posterior densities of the ability parameters obtained in the first step were used to calibrate five new items. From Study 1, we knew that the higher the correlation ρ, the more accurate the ability estimates. As a result, we also expected more accurate estimates of the item parameters. To confirm this expectation, for the same population of test takers, response patterns $\mathbf{y}_2$ were generated for two different versions of a 5-item test: (i) $a_k = .5$, .7, .9, 1.1, and 1.3 but $b_k = 0$ for all items and (ii) $b_k{=}-2$, -1, 0, 1, and 2 but $a_k = 1$ for all items. By first varying the a_ks while holding the b_ks constant and vice versa, we were able to evaluate the effects of the correlation ρ on the two kinds of item parameters in isolation. For both cases, the item parameters were calibrated using the version of the Gibbs sampler with the earlier obtained draws from the posterior distributions of the ability parameters.

Again, the MSE was used as a criterion to evaluate the parameters estimates. For example, for the parameters a_k we compared

$$MSE(a_k|\mathbf{y}_2, \mathbf{y}_1) = E_{a|\mathbf{y}_2}[E_{\theta|\mathbf{y}_1}[(a_k - \hat{a}_k)^2|\mathbf{y}_2, \mathbf{y}_1]] \tag{5.24}$$

with

$$MSE(a_k|\mathbf{y}_2, \mathbf{y}_1, \boldsymbol{\zeta}) = E_{a|\mathbf{y}_2}[E_{\theta|\mathbf{y}_1,\boldsymbol{\zeta}}[(a_k - \hat{a}_k)^2|\mathbf{y}_2, \mathbf{y}_1, \boldsymbol{\zeta}]], \tag{5.25}$$

where the inner expectation involves integration over θ given the calibration data, respectively the calibration data and the RTs. The MSEs for the parameters b_k were defined analogously. Again, these statistics can be estimated from the MCMC chain in a similar fashion to (5.22).

The results are presented in Figures 5.3 and 5.4. Both plots show the decrease in MSE relative to the condition without the use of RTs, e.g., for the parameters a_k, $MSE(a_k|\mathbf{y}_2, \mathbf{y}_1) - MSE(a_k|\mathbf{y}_2, \mathbf{y}_1, \boldsymbol{\zeta})$. It is well known that estimation error tends to be higher for the larger parameters a_k. Therefore, the improvement in accuracy of the estimation of these parameters is mainly found for the higher discriminating items. For the parameters b_k, the improvement in estimation is mainly found at the lower and upper ends of the ability scale, similar to that for the ability parameters in Study 1.

An interesting result was obtained for the information functions of the items. In Figure 5.4, the result of an integration of the information functions over θ is plotted as a function of the difficulty parameter of the items. The plot shows that the impact of the RTs is most effective in the areas where the difficulty parameters are hardest to estimate. Finally, we emphasize again that the improvements in the estimation of the item parameters comes only from the improved accuracy of ability parameters estimates in the presence of the collateral information in the RTs. Additional beneficial effects through the correlation between the item parameters in the response and RT models were not included in these studies.

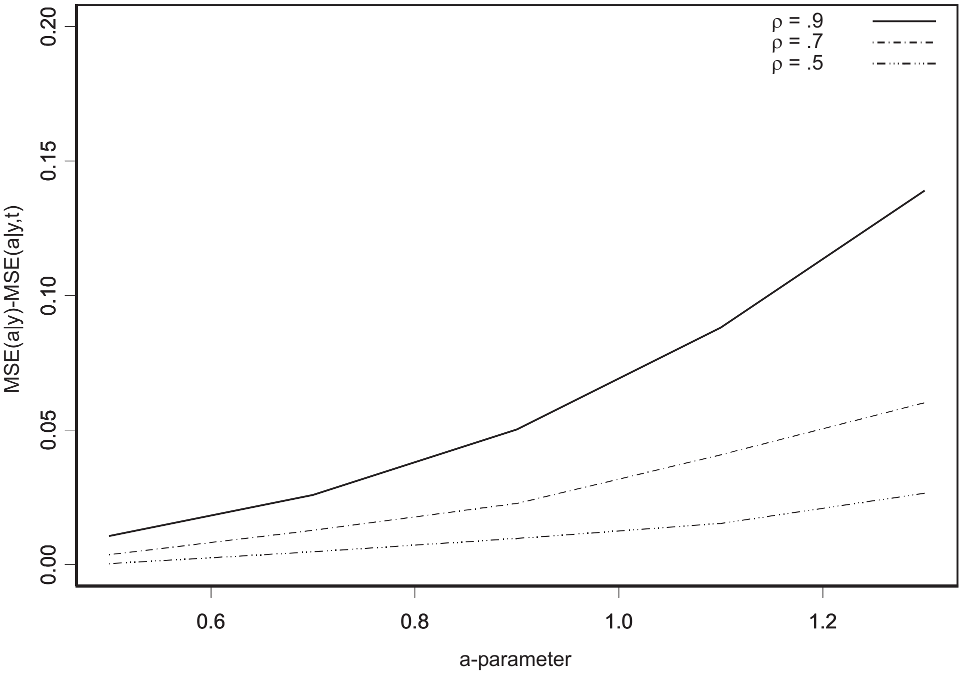


Fig. 5.3. Reduction in the MSE of the estimates of a_k for different correlations ρ.

5.5 Discussion

Since its inception, test theory has been hierarchical; the randomness of an observed score of an individual test taker has always been distinguished from that of his or her true score due to sampling from a population. In addition, for its statistical inference, test theory has been an early adopter of the Bayesian methodology. It therefore seems natural to broaden the traditional hierarchical (vertical) type of modeling of responses in IRT with the horizontal extension of RT modeling in this paper.

Further improvement of parameter estimation has always been a concern of the testing industry; it makes test scores more informative and reduces the costs of item calibration. But there has also been a general reluctance to use other information than the test takers' performances on the test items, especially, when the information is population dependent . We respect this reluctance but add the following elements to the discussion. First, RTs *are* part of the test takers' performances on the test items. Using them is not the same as, for example, the practice of regressing the test takers' abilities on background variables or any other type of information with only an indirect relation to the test performance. Therefore, we expect less objection against the use of RTs, particularly when estimating item parameters.

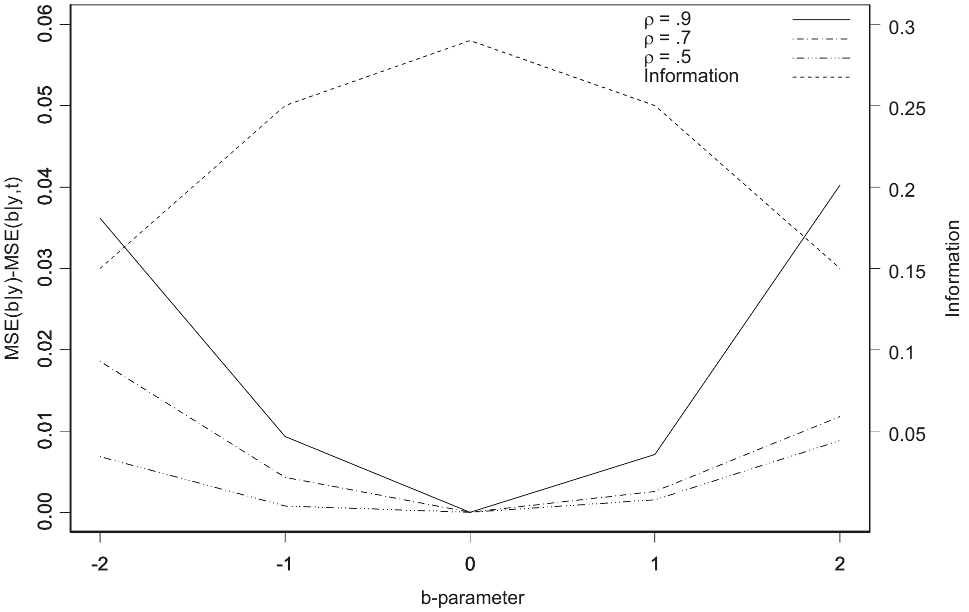


Fig. 5.4. Reduction in the MSE of the estimates of b_k for different correlations ρ.

Second, the use of RTs does not change the construct or dimension measured by the test in any way. The same parameter θ is estimated with and without the use of RTs as collateral information. In both cases, θ is the only person parameter that explains the probability distributions of the responses on the items. Besides, it does not have any impact on the RT distributions. This lack of impact is reflected in the assumption of local independence between the responses and RTs in the hierarchical frameworks used in this paper. Thus, the difference between estimating ability parameters with and without the collateral information in the RTs is *not* in the parameter that is estimated, only in the increase in accuracy with which this is done.

Third, the modeling framework does require the specification of a second-level population distribution and may therefore seem to suggest some form of population-dependent test scoring. However, the role of the second-level distribution is different from that in traditional hierarchical estimation. For example, in Kelley's regression function and in (5.2), the estimates are drawn to the mean true score in the population of test takers, and different estimates are obtained for different populations. On the other hand, the θ estimate from (5.16) is dependent only on the *conditional* distribution of θ given the test taker's speed ζ. It is a general statistical finding that such conditional distributions tend to be invariant with respect to their marginal distribution, i.e., the population distribution of θ. (For example, the same invariance drives the equating of observed test scores with poststratification in a nonequivalent-groups equating design; (e.g. Kolen & Brennan, 2004, sect. 5.1)).

For the current application, we expect the conditional distributions of θ given ζ to be dependent on such factors as the type of skill/knowledge measured by the test, the test instructions, and the time limit, but not on any differences between populations of test takers once these factors have been standardized.

Fourth, we expect the use of RTs as collateral information not to be an issue for ability estimation in low-stakes testing (e.g., diagnosis for remedial instruction in education). If it would be an issue in the more controversial area of high-stakes testing, we could still use the RTs jointly with the responses to optimize the test but produce a final score based on the responses only. An example is adaptive testing, where the items during the test can be selected using the θ estimates in this paper but the final estimate could be inferred from the responses only. For this application, roughly the same reduction of test length has been found as for the MSEs of the θ estimates in the empirical example above (van der Linden, 2008).

Finally, the main conclusion from this paper can be summarized by stating that in order to get better estimates of the test takers' abilities, the speed at which they have responded should be estimated as well. We did not discuss the reverse problem of estimating the test takers' speed in this paper. Since the modeling framework is symmetrical with respect to the two estimation problems, the reverse conclusion holds, too: in order to efficiently estimate how fast test takers respond to the items in the test, their abilities should be estimated.

At first sight, both conclusions seem counterintuitive. But they follow directly from the Bayesian principle of collateral information for the joint hierarchical modeling used in this research.

6

Multivariate Generalized Linear Mixed Models for Responses and Response Times

Summary. Computerized testing makes it straightforward to collect both responses and response times on test items. To make optimal use of both data sources, multivariate methods are required that also take account of the nesting of observations within test takers. The class of multivariate generalized linear mixed models is well suited for these kind of problems. It allows the user to specify different link functions for several data types and to model dependencies between multivariate observations via the random effects structure. Moreover, this modeling framework can be treated using standard available software. Two empirical examples illustrate the advantages of making joint inferences from responses and response times.

6.1 Introduction

Computerized testing easily allows for the collection of both responses and response times (RTs) on test items. The RTs are additional data on the test items that come for free and it is of interest to take them into account, since they might reveal information on test takers and/or the items in the test. For example, RTs can help to reveal rapid-guessing behavior of test takers (Schnipke & Scrams, 1997), which is more difficult to infer from the responses alone.

To be able to make joint inferences from the responses and RTs, we require a statistical framework that can deal with the observed data. First, that requires models that can handle mixed binary and continuous responses. Second, the models should incorporate the nesting of responses and RTs within persons. That is, we have to model dependencies *within* the two data sources to account for correlations between measurements within subjects. Third, the models should allow for modeling possible dependencies *between* the two data sources, since we are interested in making inferences about the relationships between the responses and RTs.

The first requirement can be accommodated by the class of generalized linear models (GLMs) (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). The class of GLMs links non-normal data to a linear predictor via a link function. This allows, for instance, to relate binary data via a probit or logit link to a linear model.

A strong feature of GLMs is that by specifying different link functions it is possible to deal with various types of data. Also, these models are well developed and can be evaluated easily within a likelihood framework. Most statistical software packages have routines that are able to deal with these models.

The nesting of observations within subjects is commonly modeled at a second level of modeling. By assuming that the same parametric structure holds for all subjects, but that (some) parameters are allowed to vary randomly over the subjects, the heterogeneity between subjects can be taken into account (Snijders & Bosker, 1999). These types of models are know as mixed models, since they allow covariates to enter the model as fixed (general) effects and/or as subject-specific random effects. Also the class of GLMs allows easily for a generalization to mixed models, known as generalized linear mixed models (GLMMs). As with GLMs, GLMMs relate non-normal data via a link function to a linear predictor, but now the linear predictor consists of both a fixed and a random part. Thereby, GLMMs fulfil our first two requirements and allow us to model univariate discrete and continuous responses, taking the nesting of the data within subjects into account. Still, these models can be treated in a likelihood framework but are more difficult to deal with. Estimation requires integrating out the random effects and relies either on numerical approximation of the integral, like Gaussian quadrature, or analytical approximations of the integrand, like Taylor series expansions (Tuerlinckx, Rijmen, Verbeke, & De Boeck, 2006).

Models for multivariate responses have not had much attention in the psychometric literature, an exception being Liu and Hedeker (2006) who developed a model for multiple ordinal outcomes. (With multivariate responses is meant here that we observe responses and response times, of course, an IRT model is already multivariate in itself.) However, multiple responses nested within subjects often arise in medical studies. For example, bivariate mixed effects models have been employed for the joint modeling of CD4 and CD8 lymphocytes, to follow the evaluation of these markers in HIV infection studies (Shah et al., 1997; Thiébaut, Jacqmin-Gadda, Chêne, Leport, & Commenges, 2002). Models for mixed continuous and discrete/polytomous data in biological applications were discussed in, for example, S.-Y. Lee and Shi (2001), Gueorguieva (2001) and Fieuws, Verbeke, and Molenberghs (2007). Within the mixed model framework, dependencies between multiple outcome variables can be modeled via the random effects structure. Possibilities to do so include that of correlated error terms, shared random effects or correlated random effects. The shared random effects approach has been popular in the biostatistics field. There, interest is commonly focussed on the outcome variables itself and the shared random effect merely serves as a tool to analyze dependencies between them. Moreover, a shared random effect only allows for the modeling of a positive dependency between variables and is therefore less flexible. A recent paper by McCulloch (2008) discusses these different techniques of modeling dependencies between multiple outcomes via the random effects structure. This extends the GLMMs to the class of multivariate generalized linear mixed models (MGLMMs).

Thereby, the MGLMMs fulfil all three requirements, which links our problem of modeling mixed response data to a broad class of statistical models. The advantages that Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) mentioned in their paper for modeling item response theory (IRT) models in a GLMM framework also generalize to the class of MGLMMs. They advocated that an advantage of the mixed model approach to IRT modeling is that it relates the latter to a broad statistical literature. Also, using a standard framework of modeling easily accommodates for adaptions or extensions of the model.

In educational testing, the random effect is usually assumed to be the underlying ability construct for the responses. Similarly, the heterogeneity between test takers on the RTs will be assumed to result from differences in speed, which gives us a sensible interpretation of the random effect for the response times. Separate random effects for each outcome variable allow us to model the dependencies between the two outcomes by assuming a common distribution for the random effects, an approach followed earlier by van der Linden (2007).

In this chapter, we will exploit the class of MGLMMs to jointly model responses and response times on test items. It provides us with a flexible framework for analysis that can be treated in a likelihood framework using standard software. Two examples are provided for illustration of the use and usefulness of the methods.

6.2 A Generalized Mixed Model for Multivariate Item Response Patterns

6.2.1 A Multivariate Mixed Effects Model

A generalized mixed model assumes that, conditional on a vector of random effects $\boldsymbol{\theta}$, the observations Y_{ik} are independent with means $\mu_{ik} = E(Y_{ik}|\boldsymbol{\theta})$. Subsequently, the conditional mean μ_{ik} is related to a linear predictor η_{ik} via a link function $g(\cdot)$. For the kth observation Y_{ik} on subject i this gives

$$E(Y_{ik}|\theta_i) = \mu_{ik} = g_1^{-1}(\eta_{ik}) = g_1^{-1}(\mathbf{x}'_{ik}\boldsymbol{\beta} + \mathbf{z}'_{ik}\boldsymbol{\theta}_i), \tag{6.1}$$

where $\mathbf{x}_{ik}$ is a vector of covariates of length p, $\boldsymbol{\beta}$ is a vector of fixed effect regression parameters, $\mathbf{z}$ is a design matrix (the equivalent of $\mathbf{x}$) for the random subject components $\boldsymbol{\theta}_i$. Further, $g(\cdot)$ denotes a link function and $g(\cdot)^{-1}$ its inverse. When y_{ik} is a realization of an indicator variable denoting a correct (1) or incorrect (0) response of subject i on item k, a Bernoulli sampling model is used and the logit-link can be chosen for $g(\cdot)$. When $\mathbf{x}$ and $\mathbf{z}$ are indicator matrices denoting if a subject i answered item k, it follows that (6.1) is equivalent to the Rasch measurement model as used in IRT (Kamata, 2001; Rijmen et al., 2003). Then, the random effect θ_i represents the latent ability construct and the fixed effects parameters $\boldsymbol{\beta}$ denote the item parameters.

Within the generalized mixed model framework it is just as well possible to model continuous response data. In that case, an appropriate link function has to be chosen, for instance the identity link for normally distributed data or a log-link in case of positively skewed data. For the modeling of response times on test items the latter is a convenient option to account for their skewness. Let T_{ik} denote the response time of person i on item k, then similarly a mixed effects model can be specified as:

$$E(T_{ik}|\boldsymbol{\zeta}_i) = \mu_{2ik} = g_2^{-1}(\eta_{2ik}) = g_2^{-1}(\mathbf{x}'_{ik}\boldsymbol{\lambda} + \mathbf{z}'_{ik}\boldsymbol{\zeta}_i), \tag{6.2}$$

where $\boldsymbol{\lambda}$ is a vector of fixed effects parameters and $\boldsymbol{\zeta}_i$ is a vector of random subject effects. The covariate matrices $\mathbf{x}$ and $\mathbf{z}$ fulfil the same function as in (6.1), but it is not necessary that these matrices are equivalent. Again, when $\mathbf{x}$ and $\mathbf{z}$ are indicator matrices, denoting if a subject i answered item k, the fixed effects can be interpreted as item effects, representing the expected RTs on the items on the log-time scale. Then the random subject effect represents a person slowness parameter. That is, higher a ζ_i leads to higher expected RTs. We will therefore interpret $-\zeta$ as speed, this to maintain the common notation in mixed effect models.

Commonly, the random effects parameters $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are assumed to be normally distributed with a mean of zero. Possible dependencies between responses and RTs can be modeled by allowing these random effects to correlate. For the response Y_{ik} and RT T_{ik} of test taker i on item k, let $\mu_{1ik} = p(Y_{ik} = 1|\theta_i)$ and $\mu_{2ik} = E(T_{ik}|\zeta_i)$. Then, for the special case that $\mathbf{x}$ and $\mathbf{z}$ are indicator matrices for person i answering item k the multivariate model is given by

$$\text{logit}(\mu_{1ik}) = \eta_{1ik} = \beta_k + \theta_i \tag{6.3}$$

$$\log(\mu_{2ik}) = \eta_{2ik} = \lambda_k + \zeta_i \tag{6.4}$$

where $(\theta_i, \zeta_i) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{bmatrix}. \tag{6.5}$$

This multivariate model simultaneously models binary and continuous responses on test items, assumed to be nested within test takers. Different link functions can be specified for each outcome variable. For instance, the logit-link for the binary data and the log-link for the response times. It treats the responses as nested within subjects by assuming fixed item effects and random subject components. The covariance matrix $\boldsymbol{\Sigma}$ models the variability in ability and speed $(-\zeta)$ of the subjects and allows for dependencies between these constructs via the covariance component ρ. It is possible to assume independence between the two data sources by specifying a diagonal covariance matrix for the random effects.

6.2.2 Extensions of the Model

The mixed-effects model easily allows for extensions to account for grouped data structures or the inclusion of covariates. We will discuss these possibilities in this section.

Including Person Covariates

In the presence of covariates for the person parameters, the analysis might benefit in two ways. First, it allows us to assess where the differences between persons originate from. In (6.5), a simple structural model was specified for the random effects. It models the variability in the latent traits of the persons, thereby assuming they originate from a common population. However, this model does not explain the differences between persons. Including person covariates into the model might help to understand these differences. Second, as was shown by Mislevy (1987), the incorporation of covariates can reduce the uncertainty in the person parameters and lead to an increase in estimation precision of person and item parameters.

Univariate approaches to explain differences in ability levels of test takers have been presented earlier by Mislevy (1987); Adams, Wilson, and Wu (1997); Rijmen et al. (2003) and Fox (2005). An example of regression on the random effects structure in multivariate models can be found in Snijders and Bosker (1999). The regression model is given by:

$$\theta_i = \mathbf{w}_i^t \boldsymbol{\gamma}_1 + e_{1i}, \tag{6.6}$$

$$\zeta_i = \mathbf{w}_i^t \boldsymbol{\gamma}_2 + e_{2i}, \tag{6.7}$$

where $(e_{1i}, e_{2i}) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and we assumed a common covariate matrix $\mathbf{w}$ for both random effects. This for notational simplicity only, of course it is possible to use different covariate matrices for ability and speed if this is required. Substitution of the above into (6.3) and (6.4) gives:

$$\eta_{1ik} = \beta_k + \mathbf{w}_i^t \boldsymbol{\gamma}_1 + e_{1i}, \tag{6.8}$$

$$\eta_{2ik} = \lambda_k + \mathbf{w}_i^t \boldsymbol{\gamma}_2 + e_{2i}. \tag{6.9}$$

A Multilevel Approach for Grouped Subjects

Grouped data structures like pupils nested in schools, or schools nested in countries are frequently encountered in survey research. Mixed effects models are well suited to account for such data structures, see, for instance, Fox and Glas (2001). The model structure given in (6.3) and (6.4) accounts for multivariate responses on fixed items that are nested within subjects. A generalization that models subjects $i = 1, \ldots, n_j$ nested in groups $j = 1, \ldots, J$ can be specified as:

$$\theta_{ij} = \gamma_{1j} + e_{1ij}, \tag{6.10}$$

$$\zeta_{ij} = \gamma_{2j} + e_{2ij}, \tag{6.11}$$

where $(e_{1ij}, e_{2ij})^t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_j)$ and

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma^2_{\theta_j} & \rho_j \\ \rho_j & \sigma^2_{\zeta_j} \end{bmatrix}. \tag{6.12}$$

That is, the random person effects(θ, ζ) have a group specific intercept plus a random subject component. It is possible to model the random components as unequal across groups. However, sometimes it is desired to assume that $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$ for numerical stability. Substitution of the above into (6.3) and (6.4) gives:

$$\eta_{1ijk} = \beta_k + \gamma_{1j} + e_{1ij}, \tag{6.13}$$

$$\eta_{2ijk} = \lambda_k + \gamma_{2j} + e_{2ij}. \tag{6.14}$$

6.3 Estimation

The above models can be fit using standard available software. We used the SAS PROC GLIMMIX macro under SAS/STAT 9.1.3 to estimate these models. This macro allows the modeling of multiple mixed outcomes in a mixed model framework, where dependencies are modeled via the random effects structure. An important feature is that the GLIMMIX macro allows for the specification of different link functions for multivariate outcome variables. Thereby, the modeling framework becomes very flexible, allowing the user to study multiple data types.

The estimation is a doubly iterative process. First, Taylor series expansions are used to the approximate nonlinear models with random effects by a linear mixed model. Subsequently, the linear mixed model is fitted, which is an iterative step itself. The default estimation technique for fitting models with random effects is know as restricted pseudo-likelihood (Wolfinger & O'Connell, 1993). For more details on the estimation methods, see the SAS user's manual (SAS Institute Inc., 2008). The SAS code for fitting the various models discussed in the empirical examples below is given in the Appendix.

The treatment of missing data by PROC GLIMMIX depends on the kind of observation that is missing. If an outcome observation is missing, either a response or an RT on person-item ik, that case is ignored in the analysis. For only a few missing responses this should not constitute serious problems. However, if there is missing data in one of the covariates for a person i, this person is deleted entirely from the analysis by the routine. Therefore, when estimates of the latent traits of all test takers are required, the user might want to use imputation methods to avoid deletion of a test taker from the analysis.

6.4 Testing Hypotheses

For parameters entering the model as fixed effects, the programm reports a t-test that can be used to evaluate a covariate. Note, however, that the degrees of freedom reported by SAS Proc Glimmix are not correct. The estimated degrees of freedom is based on $N \times K \times 2$ observations, due to the way the data set is constructed (see Appendix). The correct degrees of freedom would be $N - 1$. As an alternative, the Wald test can be used, which relies upon the standard normal distribution as the reference distribution. However, since N is usually large, the t-distribution is quite close to the standard normal one. Testing hypotheses about covariance components was not yet available for the GLIMMIX procedure under SAS/STAT

9.1.3. However, proper tests for the covariance components are difficult to obtain, since it involves evaluating hypotheses on the boundary of the parameter space (Self & Liang, 1987; Stram & Lee, 1994).

A Test For Conditional Independence

Conditional independence is an important assumption in latent variable modeling. This is the assumption that, conditional on the random effect, there is no dependency between items. Otherwise stated, the latent variable(s) should explain all the relationships between items. For the multivariate model above, this assumption also extends to conditional independence between the two outcome variables, given the random effects. Therefore, it is of interest to test the validity of the conditional independence assumptions.

To test the conditional independence assumption between the responses and response times, we incorporate conditional dependence on the responses as a covariate into the RT model. This approach is not new and can be found in Glas (1999) and Glas and Suarez-Falcon (2003), who proposed Langrange Multiplier tests for testing conditional independence assumptions within IRT models. In a multivariate setting, this kind of test has been proposed by, for instance, Gueorguieva (2001) and van der Linden and Glas (2008) to test for conditional independence between the outcome variables.

Let $Y^*_{ik} = 2Y_{ik} - 1$, which takes a value of 1 for a correct response and value -1 for an incorrect response. Subsequently, this variable is incorporated into the linear predictor for the RTs as:

$$\eta_{2ik} = \lambda_k + \zeta_i + \gamma_k Y^*_{ik} \tag{6.15}$$

Now the test for conditional independence between the responses and response times on item k reduces to testing the hypothesis for the fixed effect $H_0 : \gamma_k = 0$ versus the alternative $H_1 : \gamma_k \neq 0$. The same kind of test can be used to test the conditional independence assumption between items. Within SAS, these hypotheses can be evaluated with a t-test or F-test and confidence intervals of level $1 - \alpha$ can be obtained for γ, as will be shown below.

Testing for DIF

Differential item functioning (DIF) is an important topic for test administrators for reasons of fairness (Penfield & Camilli, 2007). An item is said to exhibit DIF when the probability of a correct response differs for two (or more) groups of test takers who have the same ability level. A good item should be equally difficult for all groups of test takers in the test and not differ in difficulty for, say, whites and Hispanics because of differences in cultural background. So far, DIF has been assessed on the response patterns only. However, with the availability of RTs on test items, it is possible to address the question if items are not only non-DIF with respect to difficulty, but also with respect to time intensity of the item. If two groups of test takers seriously differ in their time needed to complete the item this can be considered unfair, especially for tests with a strict time constraint. Time-DIF could be defined as *an item for which the expected response time differs across groups of test takers, conditional on their level of speed*. It has been noted that difficulty and time intensity of an item are not necessarily related to each other (van der Linden, 2007), therefore, although they may coincide, DIF and time-DIF are two different measures.

A simple test for DIF that has often been used makes use of an indicator variable that takes the value of 1 for group 1 and the value of 0 for group 2 on item k. Then the time intensity for item k equals $\lambda_k + \gamma_k$ for group 1 and equals λ_k for group 2. It is easily seen that in the case of DIF $\gamma_k \neq 0$, so a test for DIF then follows from evaluating the hypothesis $H_0 : \gamma_k = 0$ versus $H_1 : \gamma_k \neq 0$. This test can be evaluated using a t-test and a confidence interval of level $1-\alpha$ for γ can be obtained. Another possibility that basically gives the same result, is to use the indicator variable to split the test takers in two groups, so that estimates of the difficulty parameter $b_k^{(1)}$ and $\lambda_k^{(2)}$ are obtained for group 1 and group 2, respectively. Subsequently, one can evaluate if the contrast $\phi = \lambda_k^{(1)} - \lambda_k^{(2)}$ is zero. Both ways are easily implemented in SAS.

6.5 Illustrative Examples

6.5.1 Example 1

When the random effects correlate, we would expect to gain in estimation precision in the model parameters. That is, we can borrow information on the response model parameters from the response times via the correlated random effects structure. As shown earlier by Mislevy (1987) who used collateral information from person covariates, especially for short tests improvements can be expected. To illustrate this, we used a 12 item figural reasoning ability test that was presented to 356 German army recruits, a test developed by Hornke and Habon (1986). The item parameters for the response model were estimated twice. First using the responses only, the second time the response times were taken into account too.

There was a moderate correlation between the random effects, $cor(\theta, \zeta) = .65$. The item parameter estimates are presented in Table 6.1. In the last column of this table, the relative increase in estimation precision is given, which was estimated as the ratio of squared standard errors. It can be seen that for this data set, with a moderate correlation between the random effects, the gains in precision are considerable for most items. Only for item 9 (and to lesser extent item 5) the gain was low, but this can be attributed to this item being extremely difficult (only 2 persons answered it correctly).

6.5.2 Example 2

In this example we analyze a data set studied earlier by Wise et al. (2007). The data were collected on a low stakes test, the Natural World Assessment Test (NAW-8).

Table 6.1. Item parameter estimates for the German army data

	Responses		Responses and RTs		
item	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	% gain in precision
01	0.4154	0.1248	0.4078	0.1198	8.52
02	0.0430	0.1233	0.0463	0.1183	8.63
03	-0.0716	0.1234	-0.0649	0.1184	8.62
04	-0.6620	0.1277	-0.6398	0.1230	7.79
05	3.1724	0.2508	3.1080	0.2476	2.60
06	0.8084	0.1293	0.7893	0.1243	8.21
07	-1.1495	0.1365	-1.1169	0.1321	6.77
08	-1.0394	0.1340	-1.0089	0.1296	6.91
09	5.4408	0.7123	5.3676	0.7111	0.34
10	0.1448	0.1235	0.1452	0.1185	8.62
11	0.6966	0.1277	0.6808	0.1227	8.32
12	-0.0843	0.1234	-0.0773	0.1185	8.44

A computerized version consisting of 65 items was administered to 396 examinees, 2nd year students who were required to participate in the university's educational assessment. Some additional measures were collected on the students: citizenship (CS) was a self-report measure of the willingness to cooperate with the testing programm of the university and test importance (TI)was a self-report scale measuring how important the test was to the test taker. Also gender (GE) and SAT scores were available. However, from 10 persons the SAT scores were missing and these students were ignored in the following analyses. Aim of this example is to evaluate to what extend the person covariates explain the differences in ability and speed of working. Moreover, it is interesting to see what additional information the response times present to us over the responses alone. However, first the conditional independence assumption between the responses and response times was checked and we tested for DIF using gender as a grouping variable.

To test for conditional independence, the proposed method above was used. That is, we estimated (6.15) as the RT model and subsequently evaluated the estimated effects for γ_{ik} using a t-test. Interestingly, for 15 of the 65 items we found a statistically significant result at the $\alpha = .05$ level. However, upon closer inspection, the size of the effects appeared to be very small for most items. On the time scale, the differences between people who answered an item incorrectly or correctly was about 2 seconds, a difference that can safely be ignored. Except for two items. For item 44, the difference on the time scale was substantial. Test takers who answered incorrectly or correctly completed that item in general in 25 and 43 seconds, respectively. It appeared that among test takers who answered that item incorrectly, there were a lot who answered within a few seconds, which is indicative for guessing behavior. Similarly, but with a smaller difference, for item 57 the difference between test takers who answered correct and incorrect was estimated to be 12 seconds.

The DIF analysis using gender as the classification variable revealed some items for which a significant ($p < .05$) difference between males and females was found. For 8 items, there appeared to be a difference in difficulty of that item. The biggest effects were found for items 28 and 29, for which the estimated effects were $-.93(.24)$ and $-.74(.25)$, denoting that these items were relatively easier for males. Regarding the RTs, items 8 and 31 showed a statistically significant difference, but on the time scale these effects amounted less than 2 seconds.

Next, the person covariates where evaluated. Item 44 and 57 were left out of the analysis since for these items a serious violation of conditional independence was detected. However, for the remainder of this example we were not interested in DIF so all other items were maintained. The following full model, including all covariates, was formulated:

$$\eta_{1ik} = \beta_k + CS_i\gamma_{11} + TI_i\gamma_{12} + GE_i\gamma_{13} + SAT_i\gamma_{14} + e_{1i}, \tag{6.16}$$

$$\eta_{2ik} = \lambda_k + CS_i\gamma_{21} + TI_i\gamma_{22} + GE_i\gamma_{23} + SAT_i\gamma_{24} + e_{2i}. \tag{6.17}$$

The first terms on the right hand side of the above equations denote the item effects, representing the difficulty and time intensity of the items. The remaining terms on the right hand side denote the latent regression on the random person effects θ and ζ. In this case the covariates enter the model as fixed effects as in (6.6) and (6.7), and therefore a t-test can be used to evaluate their estimated coefficients at a significance level of $\alpha = .05$.

From fitting the full model, the following results were obtained. There were no gender differences observed, neither for ability ($p = .54$), nor for speed ($p = .74$). Also citizenship was not related to ability or speed of the test takers ($p = .25$ and $p = .17$). As expected, the SAT scores were positively related to ability ($p < .0001$) but they were not significantly related to the speed of working of the test takers ($p = .15$). However, test importance explained a proportion of variance in both traits ($p < .0001$). For ability this was a positive relationship, while for speed it was negative. This led us to fitting a reduced model, leaving out gender and citizenship for both dimensions and maintaining SAT and test importance in the latent regressions:

$$\eta_{1ik} = \beta_k + TI_i\gamma_{12} + SAT_i\gamma_{14} + e_{1i}, \tag{6.18}$$

$$\eta_{2ik} = \lambda_k + TI_i\gamma_{22} + SAT_i\gamma_{24} + e_{2i}. \tag{6.19}$$

For this reduced model the parameter estimates were practically unchanged up to some decimals, except for γ_{24} which was now significant at $\alpha = .10$. The estimated effects for this model can be found in Table 6.2. Furthermore, the correlation between the random terms was $\text{cor}\ (e_1, e_2) = .75$. The latter means that test takers working at a lower speed (speed $= -\zeta$) in general were also the ones with a higher score on the test. This might suggest that the quality of the test results is dubious. Namely, a requirement for a test to be a valid measurement instrument is that students take it seriously and do their best to solve the tasks. However, the strong negative correlation between ability and speed hints that only students who took their time were serious about the test. The positive and negative loadings of test

Table 6.2. Parameter estimates for the latent regression

Parameter	Estimate	S.E.	$p(df = 385)$
γ_{12}	0.0509	0.0101	$< .0001$
γ_{14}	0.0580	0.0076	$< .0001$
γ_{22}	0.0027	0.0002	$< .0001$
γ_{24}	0.0003	0.0001	0.0918

importance on ability and speed, respectively, and that we deal with a low stakes test, are in line with this interpretation. Although care has to be taken to draw strong conclusions from this one example, it shows us that response times might reveal information about test taker behavior that would be missed when analyzing the responses only.

6.6 Discussion

It was shown that multivariate generalized linear mixed models provide a suitable framework for making joint inferences from responses and response times on test items. Working within the MGLMM framework allowed us to use standard avaliable software for fitting these models. Two empirical examples illustrated how response time information may enhance the analysis of test data.

However, the use of standard methods also imposes restrictions on the analyses. It was not possible, for instance, to implement more advanced measurement models for the responses in the procedures used here. Within the mixed model framework, a 2-parameter IRT model was implemented in the NLMIXED procedure of SAS. With the programming statements in this routine it is possible to use customized log-likelihood specifications for different data sources. However, we were not successful to obtain parameter estimates for the multivariate model for (well constructed) simulated examples. (Also, strictly speaking the 2-parameter IRT models (and its generalizations) are not linear anymore, since effects enter the model as a product of parameters.) Other difficulties we ran into were when the number of observations became large or in the case of many missing observations. Then the GLIMMIX procedures sometimes did not converge.

Bayesian MCMC methods can overcome these limitations, as was illustrated in the work of van der Linden (2007) and Fox et al. (2007). These authors developed a Gibbs sampling approach to estimate all model parameters simultaneously, also for 2 and 3-parameter IRT models. However, a disadvantage of these methods is that it requires more expertise from the user to fit the models. Also, computation times are substantially larger.

Therefore, the methods outlined in this chapter provide the user with fast and easy to use tools to make inferences from multivariate response data. Eventually, using these analyses as a guideline, then in subsequent steps more elaborate tools could be used.

6.7 Appendix: SAS code for fitting the various models

The following code was used to fit the several models using the GLIMMIX macro under SAS/STAT 9.1.3.
The code for Example 1:

```
proc glimmix data= datafile;
class dist person item;
model response(event=’1’) = dist*item01 dist*item02 ... dist*item12 / solution s
noint dist=byobs(dist);
random dist / sub = person solution s type=chol;
run;
```

The code for fitting a multilevel model:

```
proc glimmix data= dataml;
class dist person group item;
model response(event=’1’) = dist*item01 dist*item02 ... dist*item20 / solution s
noint dist=byobs(dist);
random dist / sub = group solution s type=chol;
random dist / sub = person(group) solution s type=chol;
run;
```

Table 6.3. Example of a data set structured for use in SAS GLIMMIX with the above code. Note that the log of the RTs has been used.

dist	person	item	group	response	item01	item02	item03	...	item20
Binomial	1	1	1	0.00	-1	0	0		0
Normal	1	1	1	2.63	1	0	0		0
Binomial	1	2	1	1.00	0	-1	0		0
Normal	1	2	1	3.62	0	1	0		0
Binomial	1	3	1	0.00	0	0	-1		0
Normal	1	3	1	0.86	0	0	1		0
⋮									⋮
Binomial	1	20	1	0.00	0	0	0		-1
Normal	1	20	1	6.24	0	0	0		1
Binomial	2	1	1	0.00	-1	0	0		0
Normal	2	1	1	1.13	1	0	0		0
Binomial	2	2	1	0.00	0	-1	0		0
Normal	2	2	1	1.73	0	1	0	...	0

Epilogue

Response times on test items are a potential source of information about test taker behavior as well as item characteristics. Item Response Theory (IRT) for making inferences from the responses on test items has been well developed, but including response times into this framework has not received much attention. In this thesis, statistical methods are proposed for making simultaneous inferences from responses and response times. The response times are modeled in a way analogous to IRT: heterogeneity between test takers is described by latent variables (speed) and also item characteristics are accounted for. In the Introduction, joint measurement models for ability and speed that incorporate assumptions of conditional independence that form the cornerstone of this thesis are discussed.

Chapter 2 presents a structural multivariate multilevel modeling framework to investigate differences in ability and speed between (groups of) test takers as a function of covariates. Hypotheses about the structural model can be tested using a model specific implementation of the deviance information criterium. Moreover, an extensive discussion of the correlation structure between the person parameters is given. A Bayes factor test provides the means to evaluate the strength of the dependency between speed and accuracy within the population of respondents.

In Chapter 3, the population model for the item parameters is extended with a structural component. With this extension, content specific information about the items can be used to explain the differences in their difficulties and time intensities. This can improve the understanding of the item characteristics, which can again lead to improvements in test development.

Transformations to normality have obvious and much exploited advantages for the statistical modeling of non-normal data. For modeling response times in a psychometric application, the log-transform has proven to be useful. However, motivated by a data set for which the lognormal model was not able to capture certain aspects of the data, in Chapter 4, the class of Box-Cox transformations was considered, which allows for more flexibility in the description of the data. Also, its conjugacy with the multivariate normal level-2 models for the person and item parameters allows for straightforward modeling of the dependencies between the parameters in the level-1 models (van der Linden, 2007; Fox et al., 2007).

Chapter 5 discusses the role of response times as collateral information in estimating IRT model parameters. The hierarchical structure of the modeling framework allows for borrowing strength from the response times to improve estimation of the IRT model parameters. The main conclusion from this chapter can be summarized by stating that in order to get more accurate estimates of the test takers' abilities, the speed at which they have responded should be estimated as well. Moreover, since the modeling framework is symmetrical with respect to the two estimation problems, the reverse conclusion holds too: in order to efficiently estimate how fast test takers respond to the items in the test, their abilities should be estimated, too.

A subclass of the models proposed in this thesis consists of the Rasch model and assumes all discrimination parameters in the response time model to be equal. It is shown in Chapter 6 that this subclass fits well into the framework of multivariate generalized linear mixed models. Multivariate generalized linear mixed models can handle different types of data by relating the observations to a linear model via a (nonlinear) link function. Also, these models are suited for nested data structures and can take account of possible dependencies between multivariate outcomes. An advantage is that these models can be estimated with standard commercial software. Further, a test for evaluating the conditional independence assumption between the responses and response times is proposed. However, the necessary restrictions to equal item discrimination parameters for both the IRT and the response time model might prove unrealistic.

Except for the last chapter, the models in this thesis were all discussed within the Bayesian statistical framework. The reason for the Bayesian choice resides in its flexibility to handle complex estimation and testing problems using Markov Chain Monte Carlo (MCMC) methods. These sampling based algorithms are computationally intensive, but their advantage is that they easily deal with high-dimensional problems where frequentist methods are often limited in the number of groups, persons and/or items that they can handle.

Another advantage of using MCMC methods for the models discussed in this thesis is that extensions or adjustments can be easily implemented by replacing one sampling step in the algorithm for another. For example, the response model can easily be adjusted to deal with polytomous response data. MCMC algorithms for polytomous IRT models can be found in Fox (2005); Patz and Junker (1999), and Johnson and Albert (1999), among others. The necessary adjustment of the MCMC algorithm consists of replacing the random draws from the parameters in the three-parameter normal-ogive IRT model with those in a polytomous model. If subpopulations of test takers follow different strategies to solve the items, differences in the joint distribution of accuracy and speed can be expected. To model them, a mixture modeling approach with different latent classes for different strategies can be used (see, for instance, Rost, 1990). Currently, the models proposed assumed the underlying constructs of the test to be unidimensional. However, these models can also be extended to deal with multidimensional problems (e.g., Adams, Wilson, & Wang, 1997; Embretson, 1997). An interesting opportunity would be to investigate the possibility of multidimensional speed as well.

References

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

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd international symposium on information theory* (p. 267-281). Budapest: Akademiai Kiado.

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

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

Barnard, J., McCullogh, R., & Meng, X.-L. (2000). Modelling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281-1311.

Becker, P. (1999). Beyond the big five. *Personality and Individual Differences*, *26*, 511-30.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

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

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317-352.

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.

Boscardin, W. J., & Zhang, X. (2004). Modeling the covariance and correlation matrix of repeated measures. In A. Gelman & X.-L. Meng (Eds.), *Applied bayesian modeling and causal inference from incomplete-data perspectives* (p. 215-226). New York: Wiley.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, series B*, *26*, 211-252.

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

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

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*, 137148.

Browne, S. D., & Heathcote, A. (2008). The simplist complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology.* (doi:10.1016/j.cogpsych.2007.12.002)

Browne, W. J. (2006). MCMC algorithms for constrained variance matrices. *Computational Statistics & Data Analysis*, *50*, 1655-1677.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis.* Chapman & Hall.

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: Duxbury.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*, 347-361.

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.

DeIorio, M., & Robert, C. P. (2002). Discussion of spiegelhalter et al. *Journal of the Royal Statistical Society, series B*, *64*, 629-630.

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.

Fieuws, S., Verbeke, G., & Molenberghs, G. (2007). Random-effects models for multivariate repeated measures. *Statistical Methods in Medical Research*, *16*, 387-397.

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

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145-172.

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.

Freeman, J., & Modarres, R. (2006). Inverse Box-Cox: The power-normal distribution. *Statistics & Probability Letters*, *76*, 764-772.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

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. (1999). Modification indices for the 2PL and the nominal response model. *Psychometrika*, *64*, 159-172.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 81-100.

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

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

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.

Gueorguieva, R. (2001). A multivariate generalized mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, *1*, 177-193.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and related problems. *Journal of the American Statistical Association*, *72*, 320-338.

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. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577-601.

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.

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.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling.* 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.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

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

Kennedy, M. (1930). Speed as a personality trait. *Journal of Social Psychology*, *1*, 286-298.

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*.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (in press). Evaluating cognitive theorie: A joint modeling approach using responses and response times. *Psychological Methods*.

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (in press). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*.

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking.* Spinger.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-74.

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.

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*, 119-122.

Lee, P. M. (2004). *Bayesian statistics: an introduction* (2nd ed.). New York: Arnold.

Lee, S.-Y., & Shi, J.-Q. (2001). Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics*, *57*, 787-794.

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

Liu, L. C., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*, 261-268.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Lawrence 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.

McCrea, R. R., & Costa, P. T. (1997). Peronality trait structure as a human universal. *American Psychologist*, *52*, 509-516.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models.* London: Chapmann and Hall.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, *17*, 53-73.

McCulloch, R. E., Polson, N. G., & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, *99*, 173-193.

McCulloch, R. E., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64*, 207-240.

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223-236.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81-91.

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., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661-679.

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.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, *135*, 370-384.

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

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelhood bootstrap. *Journal of the Royal Statistical Society. Series B*, *56*, 3-58.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* McGraw-Hill.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in m groups: A cross validation study. *British Journal of Mathematical and Statistical Psychology*, *25*, 33-50.

Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The estimation of proportions in m groups. *Psychometrika*, *38*, 19-46.

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.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26* (p. 125-167). Amsterdam: Elsevier.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, *68*, 35-43.

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

Rabe-Hesketh, S., & Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, *57*, 1256-1264.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347-356.

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

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

Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika*, *70*, 145-156.

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

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185-205.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods.* New York: Springer-Verlag.

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* (p. 187-208). New York: Springer-Verlag.

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. (in press). A hierarchical process dissociation model. *Journal of Experimental Psychology: General.*

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.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*, 377-401.

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: Cambridge University Press.

SAS Institute Inc. (2008). *Sas/stat 9.2 users guide.* Cary, NC: SAS Institute Inc.

Schafer, J. L., & Yucel, R. C. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437-457.

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.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. Mills, M. P. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (p. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: John Wiley & Sons.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605-610.

Shah, A., Laird, N., & Schoenfeld, D. (1997). Random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, *92*, 775-779.

Shi, J. Q., & Lee, S. Y. (1998). Bayesian sampling based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *51*, 233-252.

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.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*, 196-201.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis, an introduction to basic and advanced multilevel modeling.* London: Sage Publishers.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, *64*, 583-639.

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.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*, 1171-1177.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement*, *8*, 353-374.

Thiébaut, R., Jacqmin-Gadda, H., Chêne, G., Leport, C., & Commenges, D. (2002). Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine*, *69*, 249-256.

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.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225-255.

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

van Breukelen, G. J. P. (1995). Psychometric and information processing properties of selected repsonse 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. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *32*, 5-20.

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

van der Linden, W. J., & Glas, C. A. W. (2008). Statistical tests of conditional independence between responses and response times on test items. *Psychometrika.* (Submitted)

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., Klein Entink, R. H., & Fox, J.-P. (in press). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement.* (conditionally accepted)

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speedeedness 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 der Lubbe, R. H. J., Jaśkowski, P., Wauschkuhn, B., & Verleger, R. (2001). Influence of time pressure in a simple response task, a choice-by location task, and the Simon task. *Journal of Psychophysiology*, *15*, 241-255.

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. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York: Springer.

Wagenmakers, E. J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 3-22.

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.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Wise, S. L., Kong, X. J., & Pastor, D. A. (2007, April). *Understanding correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice.* (Paper presented at the 2007 anual meeting of the National Council on Measurement in Education, Chicago, IL)

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, *48*, 233-243.

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.

Samenvatting

De psychometrie is het vakgebied dat zich bezighoudt met de theorie en methodes voor het meten van vaardigheden en attitudes. Omdat vaardigheden niet direct observeerbaar latent zijn (je kunt niet aan iemand zien wat hij/zij weet van quantum-mechanica), wordt er voor het meten van vaardigheden en attitudes vaak gebruik gemaakt van testen of vragenlijsten. Item response theorie (IRT) beschrijft hoe met statistische modellen uit de geobserveerde antwoorden op de test conclusies kunnen worden getrokken over de door de test te meten vaardigheid. Deze IRT modellen beschrijven de kans op het geven van een goed antwoord op een vraag als een functie van karakteristieken van het item en de latente vaardigheid van de kandidaat. (In de psychometrie spreekt men van items en niet van vragen omdat een item niet noodzakelijkerwijs een vraag is, maar ook een waar/onwaar stelling kan zijn of een taak die moet worden uitgevoerd.)

Conclusies over de kwaliteit van de items in de test of het vaardigheidsniveau van een kandidaat zijn voornamelijk gebaseerd op de antwoorden op de items. De reaktietijden op de items werden tot voor kort vaak niet meegenomen. Dat had vooral een praktische reden. Bij toetsafname met pen en papier is het erg lastig om te registreren hoe lang een kandidaat doet over het beantwoorden van een vraag. Sinds toetsen en examens steeds vaker op de computer worden afgenomen, is dit aanzienlijk vereenvoudigd.

Daarmee wordt het interessant om de IRT modellen uit te breiden en ook reaktietijden mee te modelleren. Reactietijden kunnen bijvoorbeeld een extra bron van informatie zijn voor het analyseren van het responsgedrag van kandidaten of gebruikt worden voor het evalueren van de test condities met betrekking tot tijdslimitieten (van der Linden et al., 1999). Sectie 6.5.2 in dit proefschrift geeft een illustratie van de additionele waarde van reactietijden aan de hand van een praktisch voorbeeld.

Dit proefschrift behandelt statistische methoden voor het maken van inferenties van antwoord- en reactietijdenpatronen op toetsen en examens. In de inleiding wordt het raamwerk uiteengezet waar de rest van dit proefschrift op bouwt. Waar wordt aangenomen dat de latente vaardigheid van de kandidaat ten grondslag ligt aan de geobserveerde antwoordpatronen wordt op analoge wijze verondersteld dat snelheid van werken het onderliggende construct voor de reactietijden is. Voor zowel vaardigheid als snelheid wordt een meetmodel beschreven dat de geobserveerde antwoorden/reactietijden linkt aan deze latente constructen, daarbij rekening houdend met de karakteristieken van de items.

Het tweede hoofdstuk behandelt een structureel model dat mogelijkheden biedt voor het verklaren van variantie in vaardigheid en snelheid van kandidaten. Deze modeluitbreiding wordt begeleid door modelspecifieke toetsen waarmee hypotheses kunnen worden gevalueerd met betrekking tot de relatie tussen covariaten en vaardigheid en snelheid als mede hypotheses aangaande verschillen tussen groepen van personen. Met behulp van een Bayes factor toets kunnen mogelijke afhankelijkheden tussen de antwoorden en reactietijden getoetst worden.

De itemparameters in de twee meetmodellen voor vaardigheid en snelheid beschrijven verschillen tussen de items in moeilijkheidsgraad en tijdsintensiviteit. Echter, daarmee zijn de onderliggende oorzaken voor die verschillen nog niet duidelijk. Om daar dieper op in te gaan behandelt hoofdstuk 3 een structureel model voor de itemparameters. Met dit model kan de relatie tussen de structuur- en inhoudskenmerken van het item en zijn moeilijkheid/tijdsintensiviteit onderzocht worden. Een beter begrip van deze relaties is interessant voor het ontwikkelen van nieuwe items. De methoden worden geillustreerd met een analyse van een grootschalige intelligentietest waarbij een cognitief ontwerp aan de items ten grondslag ligt.

Omdat reaktietijden een ondergrens bij nul hebben is hun verdeling niet symmetrisch (als bij een normale verdeling) maar wordt gekarakteriseerd door een langere staart aan de rechterzijde van de verdeling. De log-transformatie is veelvuldig toegepast voor het modelleren van reactietijden, waarna de getransformeerde tijden een meer symmetrische verdeling hebben, zodat vervolgens een normale verdeling verondersteld kan worden. Het gebruik van een normale verdeling heeft vele voordelen, omdat de matematische eigenschappen goed bekend zijn en er lineaire modellen gebruikt kunnen worden. Daardoor vereenvoudigen verdere analyses sterk ten opzichte van niet-lineaire alternatieven. Echter, de log-transformatie levert niet per definitie normaal verdeelde data op, zoals een voorbeeld in hoofdstuk 4 illustreerd. Om die reden wordt er in dat hoofdstuk gekeken naar de klasse van Box-Cox transformaties. Het gebruik van een volledige klasse van transformaties geeft meer flexibiliteit in het beschrijven van reactietijdenverdelingen. Tevens poogt de Box-Cox transformatie de normale verdeling zo goed mogelijk te benaderen, waaraan eerdergenoemde voordelen verbonden zijn.

Zoals in een aantal voorbeelden in dit proefschrift van analyses met echte test gegevens blijkt, kunnen er (sterke) correlaties tussen de antwoorden en reaktietijden gevonden worden. Deze correlaties kunnen resulteren vanuit de personen, omdat er mogelijk een verband tussen vaardigheid en snelheid bestaat, ofwel omdat er relaties op item niveau zijn, bijvoorbeeld tussen moeilijkheid en tijdsintensiviteit van de items. Hoofdstuk 5 laat zien dat het gebruik van deze correlaties tussen de persoons- of itemparameters in de meetmodellen leidt tot accuratere schattingen van deze modelparameters. Dat wil zeggen, bij correlaties ongelijk aan nul bevatten de reaktietijden informatie over de IRT model parameters en vormen daarbij een extra bron van informatie naast de al beschikbare antwoorden. Hoofdstuk 5 laat zien hoe de opbouw van het grotere model voor simultane analyse van antwoorden en reactietijden zich goed leent om gebruik te maken van deze mogelijke correlaties. Vanzelfsprekend geldt dit principe van extra informatie ook voor het schatten van de parameters in het reactietijden model. Vanwege de symmetrie van het schattingsprobleem bevatten de antwoorden net zo goed informatie over de parameters in het reactietijden model als vice versa.

Hoofdstuk 6 tot slot, bespreekt een subklasse van de in dit proefschrift besproken modellen. Deze subklasse restricteert de meetmodellen waarbij wordt aangenomen dat er geen verschillen in de discriminatieparameters van de items zijn en ook de gokkans in het meetmodel voor vaardigheid genegeerd wordt. Deze subklasse blijkt goed te passen in een ruime klasse van statistische modellen die bekend staan als multivariaat gegeneraliseerde lineaire gemixte modellen (multivariate generalized linear mixed models). De voordelen van het gebruik van een dergelijke klassen van modellen voor het simultaan analyseren van antwoorden en reactietijden is dat er een directe link is met een brede statistische literatuur en tevens dat er commerciele software gebruikt kan worden voor het verkrijgen van parameterschattingen. Ondanks dat in de praktijk de genoemde vereenvoudigen te restrictief kunnen blijken, geven deze methodes toch een eenvoudig instrument om een aantal verkennende onderzoeken uit te voeren.

Met uitzondering van hoofdstuk 6 zijn alle statistische methodes in dit proefschrift ontwikkeld vanuit de Bayesiaanse statistiek. De redenen daarvoor zijn praktisch van aard. Een Bayesiaanse benadering maakt het gebruik van stochastische simulatietechnieken, beter bekend als Markov Chain Monte Carlo (MCMC) technieken, mogelijk voor het verkrijgen van parameterschattingen. Deze MCMC methodes zijn zeer geschikt voor hoogdimensionele schattings- en testproblemen, waar frequentistische methodes al snel gerestricteerd blijken wat betreft het aantal groepen, personen en items dat zij aankunnen. Een ander voordeel van de Bayesiaanse benadering en MCMC is dat modeluitbreidingen en teststatistieken eenvoudig in de schattingsprocedures ingebouwd kunnen worden, wat de onderzoeker flexibiliteit geeft voor aanpassingen.