Contributions to the Joint Modeling of Responses and Response Times

Sukaesi Marianti

**CONTRIBUTIONS TO THE JOINT MODELING
OF RESPONSES AND RESPONSE TIMES**

SUKAESI MARIANTI

Graduation Committee

Chairman : Prof.Dr. T.A.J. Toonen
Promotor : Prof. Dr. Ir. Jean-Paul Fox

Members : Prof. Dr. Cees Glas
Prof. Dr. Ir. Bernard Veldkamp
Prof. Dr. Adrie J. Visscher
Prof. Dr. Rob R. Meijer
Prof. Dr. Gunter Maris

Marianti, Sukaesi

Contributions to The Joint Modeling of Responses and Response Times

Ph.D. Thesis, University of Twente, Enschede, The Netherlands

ISBN: 978-90-365-3983-8

DOI: 10.3990/1.9789036539838

Copyright © 2015, S.Marianti

CONTRIBUTIONS TO THE JOINT MODELING
OF RESPONSES AND RESPONSE TIMES

DISSERTATION

to obtain
the degree of doctor at the University of Twente
on the authority of the rector magnificus,
Prof. Dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, November 18th, 2015 at 16:45

by
Sukaesi Marianti
Born on July 31st, 1980
In Sumbawa Besar, NTB, Indonesia

This dissertation has been approved by the promotor:

Prof. Dr. Ir. Jean-Paul Fox

Acknowledgements

This dissertation represents the last three years I have spent in University of Twente, the Netherlands. Foremost, I would like to thank God, the most merciful, who helps me to gain the power of knowledge.

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Ir. Jean-Paul Fox, for his supervision, continuous support, motivation, and immense knowledge. Besides my supervisor, I would like to thank Prof. Dr. Cees Glas, for his encouragement and guidance. He has been my mentor the day I stepped onto the soil of the Netherlands.

My thanks go to Directorate General of Higher Education (DGHE) of Indonesia for the scholarship which gave me opportunity to continue my study in Netherlands. I want to thank my colleagues from department of Psychology, University of Brawijaya for their support, and my colleagues from the Research Methodology, Measurement and Data Analysis (OMD) Department, University of Twente for the good time during my study there. I also want to thank my officemate Inga, for the nice cover she made for this dissertation.

It is a great privilege to thank my family, who supported me at every step of my life and helped me build my career and personality. I cannot thank enough my mother and father, who are my mentors since my birth and who helped me strengthen my roots. I would also like to thank my sisters, Parmi and Titiek, who provided me with great support throughout this venture. I owe a special thanks to Maham for having some great discussions with me which motivated me and helped me achieve my dream.

A very special thanks to my late sister Marni, for all the memories we had. I wish that you could see me today and know how blessed I feel to be your family.

Contents

Introduction.....	1
1.1 Speed	3
1.1.1 Constant working speed	3
1.1.2 Non-constant working speed.....	4
1.1.3 Dynamic process of working speed.....	4
1.2 Ability	4
1.3 Joint Modeling of Responses and Response Times	5
1.4 Outline	6
Testing for Aberrant Behavior in Response Time Modeling	9
2.1 Introduction	9
2.2 RT Modeling.....	11
2.2.1 Identification	13
2.2.2 A Bayesian Log-Normal RT Model.....	13
2.2.3 The Estimation Procedure for Log-Normal RT Models	15
2.3 Test for Aberrant RT Patterns	15
2.4 The Null Distribution	17
2.5 Bayesian Testing of Aberrant RT Patterns	18
2.6 Dealing with Nuisance Parameters.....	19
2.7 Results.....	22
2.7.1 Investigation of Detection Rates	22
2.7.2 Comparing Three Statistics.....	24
2.7.3 Model-Fitting Responses and Random Response Behavior	25
2.7.4 Test Speededness.....	26
2.7.5 One Extreme Response.....	27
2.8 Real Data Example	29
2.9 Discussion.....	31
Modeling Differential Working Speed in Educational Testing	33
3.2 Modeling Variable Speed	35
3.2.1 Measurement Model for Speed.....	35
3.2.2 Group Differences in Working Speed.....	36
3.2.3 Dynamic Factor Modeling of Response Times	36
3.2.4 Generalized Dynamic Working Speed Modeling	39

3.2.5 Non-stationary Speed Models	40
3.2.6 Joint Modeling of Ability and Speed	41
3.3 Estimation	42
3.3.1 Identification	42
3.3.2 MCMC Estimation.....	44
3.4 Empirical Illustrations	44
3.4.1 Simulation Study for Parameter Recovery and Sample Size	44
3.4.2 Detecting Aberrant Working Speed Behavior	48
3.4.3 Real Data Analysis: Dynamic Speed Modeling	50
3.5 Discussion.....	55
Latent Growth Modeling of Working Speed Measurements	59
4.1 Introduction	59
4.2.1 The lognormal random linear variable speed model.....	63
4.2.2 The lognormal random quadratic variable speed model.....	65
4.2.3 Joint model for responses and response times	66
4.2.4 Identification	67
4.2.5 Parameter Estimation	68
4.3 Simulation study.....	70
4.3.1 Simulation Design.....	71
4.3.2 Simulation Results	71
4.4 Modeling Variable Speed in the Amsterdam Chess Test Data	75
4.5 Discussion.....	81
Evaluation Tools for Joint Models for Speed and Accuracy.....	83
5.1 Introduction	83
5.2 The Joint Modeling Framework	85
5.3 Person Fit for Speed and Accuracy	87
5.4 Residual Analysis.....	91
5.5 Evaluating Distributional Assumptions.....	92
5.6 MCMC Estimation	93
5.7 Simulated Data Analysis	93
5.7.1 Parameter recovery of the joint model with guessing	93
5.7.2 Evaluate performance person-fit tests.....	94
5.8 Real Data Analysis	96
5.9 Discussion.....	103

Summary and Discussion	105
6.1 Summary.....	105
6.2 Discussion.....	106
6.3 Future Research	107
References	109
Samenvatting	117
Appendix A	
The Expected Statistic Value and Its Variance as a Function of Response Times	119
Appendix B	
Stationary Speed Process	121
Appendix C	
WinBUGS Algorithm: Two-component Mixture Model for Speed.....	123
Appendix D	
Winbugs Code for Model 1 (Equation 4.17)	125
Appendix E	
Winbugs Code for Model 2 (Equation 4.18)	127
Appendix F	
Winbugs Code for Model 3 (Equation 4.19)	129
Appendix G	
R Package: LNIRT	131
Appendix H	
R Package: LNIRTQ.....	133
Appendix I	
RSTAN Algorithm	137
Appendix J	
Simulated and Estimated Measurement Error Variance Parameters	139

Contents

x

List of Tables

Tabel 2.1 <i>Person-fit statistics for RT data under the lognormal model</i>	18
Tabel 2.2 <i>False alarm rates and detection rates of l^i for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)</i> ...	26
Tabel 2.3 <i>Detection rates of l^i for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)</i>	26
Tabel 2.4 <i>Detection rates of l^i for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)</i>	27
Table 3.1 <i>Simulated and re-estimated parameters of the dynamic speed MA(1) model for 50 data replications for different number of respondents, items, and block sizes</i>	46
Table 3.2 <i>Estimated DIC of the constant and dynamic working speed models for different block sizes</i>	52
Table 3.3 <i>Covariance and correlation estimates of person and item parameters of the joint model with constant speed and with variable speed using a random transition effect</i>	55
Table 4.1 <i>Simulated and estimated parameter values of Model 1</i>	72
Table 4.2 <i>Simulated and estimated parameter values of Model 2</i>	73
Table 4.3 <i>Simulated and estimated parameter values of Model 3</i>	74
Table 4.4 <i>ACT Chess: Covariance components and correlation estimates</i>	78
Table 5.1 <i>Detection rates of person fit tests l_i and l_s^y, in identifying aberrant response and RT patterns for $N=1,000$ and $K=20$</i>	95
Table 5.2 <i>Performance of the KS test in identifying non-normally distributed item residuals. The reported detection rates are based on one hundred replications using a significance level of .05</i>	96
Table 5.3 <i>Covariance and correlation estimates of person and item population parameters of the joint model (LNIRT) and the joint model with variable speed</i>	101

List of Figures

<i>Figure 1.1.</i> Example of a speed accuracy tradeoff	6
<i>Figure 2.1.</i> Classification probability versus probability of being flagged for the three different statistics ($N = 1,000$, $I = 10$)	25
<i>Figure 2.2.</i> The ROC curve of the I^* test for simulated data (1,000 persons, 10 items) with 10% aberrant RTs according to degrees of random response times (left subplot) and speededness (right subplot).	29
<i>Figure 2.3.</i> NAW-8 test; Estimated statistic values and corresponding posterior significance levels.....	31
<i>Figure 3.1.</i> A dynamic factor model for the stochastic speed process given four items.....	37
<i>Figure 3.2.</i> Estimated trajectories of two subjects based on five and ten blocks for a 30-item test given a dynamic speed model with an MA(1) component.	48
<i>Figure 3.3.</i> The choose-a-move chess example: The empirical mixture population density of Speed.....	49
<i>Figure 3.4.</i> Density plot of the average transition parameter μ_ρ and the population standard deviation σ_ρ of the MA(1) dynamic speed model based on 17 blocks of 10 items.	53
<i>Figure 3. 5.</i> Estimated test takers' dynamic working speed trajectories over 34, 17, and 10 blocks under the random MA(1) model.	54
<i>Figure 4.1.</i> The lognormal random linear variable speed model.....	65
<i>Figure 4.2.</i> The lognormal random quadratic variable speed model.....	66
<i>Figure 4.3.</i> Latent speed trajectories over items of those persons with the highest speed realizations. The time axis is defined as the ordered items on a scale from zero to one.	75
<i>Figure 4.4.</i> Trace plots of the ability and average speed population variance parameters, and the covariance between ability and the slope and quadratic speed components.	76
<i>Figure 4.5.</i> Item parameter estimates of the 40 chess items. The left plot shows the discrimination (diamond symbol) and difficulty (closed diamond symbol) estimates, where the right plot shows the time discrimination (diamond symbol) and time intensity (closed diamond symbol) estimates.	77
<i>Figure 4.6.</i> Random person parameter estimates; the average speed (ζ_0) , slope speed (ζ_1), and the quadratic speed (ζ_2) component plotted against ability (θ). The slope of speed is plotted against the quadratic speed component.	80
<i>Figure 4.7.</i> Fitted latent speed trajectories over items of high and low-ability participants.....	81
<i>Figure 5.1.</i> For each item, the differences between simulated and estimated values of the item parameters are plotted.	93

<i>Figure 5.2.</i> The estimated item parameters; difficulty and time intensity (left subplot) rescaled to have a mean of zero, and item discrimination and time-discrimination (right subplot) .	97
<i>Figure 5.3.</i> Estimated person fit statistics l_t with respect to the RT patterns plotted against the corresponding posterior significance probability.	98
<i>Figure 5.4.</i> Estimated person fit statistic values l_s^y with respect to response patterns plotted against the corresponding posterior significance probability.	98
<i>Figure 5.5.</i> Person fit statistic l_t (related to RTs) plotted against l_s^y (related to responses).	99
<i>Figure 5.6.</i> Estimated level of ability plotted against speed for the identified non-aberrant and aberrant test takers given the statistic values l_t .	100

CHAPTER 1

Introduction

There is no blinking the fact that beside responses, response times are also an important source of information in educational decision making. Computer based tests (CBTs) have made it possible to collect response times more easily from large numbers of test takers (Parshall, 2002). Nowadays, researchers have started to focus more on response times and have developed different models to describe these response times.

A response time is defined as the difference (often in seconds) between an item presented and answered by a test taker (Wise & Kong, 2005). Generally, response times are measured when there is a known start and end time for a test (Schnipke & Scrams, 2002). In the context of educational testing, response times reflect the time a test taker needs for completing an item (Lee & Chen, 2011).

There are several approaches for modeling response times. The first approach takes into account the accuracy of the responses. It is assumed that the correct responses reflect speed and accuracy. Thissen (1983) introduced a model that integrates response times and responses. Roskam (1997), and Wang and Hanson (2005) also proposed a model that integrates responses and response times. The second approach considers modeling response times, while ignoring the correctness of the response. It means that speed and accuracy are not two complementary aspects of a fundamental concept, labelled as mental power. Examples of this approach include research of Maris (1993), who modeled response times exclusively and scores were not taken into consideration. Schnipke and Scrams (1997) estimated rapid guessing under the assumption that responses and response times are independent of each other. There is a third approach introduced by van der Linden (2007). He modeled hierarchically the response times and responses. At the first level of this hierarchy, the model contains separate models for the responses and response times, an IRT and a lognormal model, respectively. The second level is the joint modeling of person and item parameters.

Response times can be used to detect aberrant behavior by identifying unexpected response time patterns. The difference between expected response times and observed response times is a potential source for identifying aberrant behavior (van der Linden & Guo, 2008; Marianti et al., 2014). Response times are continuous observations. Therefore, more

information is provided about possible aberrances than categorical responses (van der Linden, 2009a).

Response times, however, are not a sufficient source of information for labelling test takers. This source of diagnostic information can be used together with other sources, to build up a case related to aberrant response behavior (Meijer & Sotaridona, 2006). A second disadvantage is that response-time-based procedures for detecting aberrant behavior can only be applied to tests that are administered in a computerized mode. Another disadvantage of response times is that sooner or later, test takers would become aware of the fact that their response times are being observed and they will try to fake response times to hide their collusion. (van der Linden, 2009b).

While considering both advantages and disadvantages, response times or both response times and responses have been used in practical applications. For example, Ferrando and Lorenzo-Seva (2007a) applied an item response theory model that incorporates response times for binary personality items. A modification of the log-linear model, proposed by Thissen (1983) in the ability domain, was used in the context of personality items. Wise, Pastor and Kong (2009) used response times for identifying rapid-guessing behavior in low-stakes testing. Response times were used to differentiate between two response strategies. One type represents solution behavior (in which the examinee actively seeks to determine the correct answer to the item) and the other type represents rapid-guessing behavior (in which the examinee quickly chooses an answer without actively trying to work out the correct answer). Meng et al. (2014) proposed a general model for responses and response times in graded personality tests. In this framework, the GPCM describes the responses, while a log-normal model describes the response times.

A lognormal model is used in this dissertation since response times have values greater than zero (non-negative scale) and the distribution of response times tends to be skewed to the right (van der Linden, 2005) and in general, the lognormal model often fits the data well. A lognormal distribution to model response times have been applied in studies conducted by Schnipke and Scrams (1997), van der Linden (2006), and Entink and Herman (2009). The lognormal family is an appropriate choice because it has the positive support and a skew required for response-time distributions (van Zandt, 2000; van der Linden, 2007).

1.1 Speed

As the responses reveal information about ability, response times reveal information about the working speed. Observed responses are assumed to be indicators of a latent variable, referred to as ability, and are commonly described in an IRT framework (Rupp & Mislevy, 2007). The lognormal model (van der Linden, 2006) is applied to model the observed response times which are related to another latent variable, referred to as speed.

The working speed process of test takers can be modeled in several ways. Various ways are discussed in the upcoming chapters. For a simple introduction, a short description is given below.

1.1.1 Constant working speed

In the log-normal model for response times, it is assumed that the working speed of test takers is normally distributed and constant throughout the entire test. A log-normal distribution was proposed by van der Linden (2006, 2007). In this model, two factors, time intensity and speed, were used to describe item and individual variations in response times, respectively. The item factor represents the time intensity, and each time intensity parameter represents the population-average time needed to complete the item given a population-average level of working speed. The person factor is defined as the factor, which represents constant working speed as systematic differences in the response times in the presence of time intensities.

In this dissertation, a time-discrimination parameter is included as a slope parameter for speed (Fox et al. 2007; Klein Entink et al. 2009). The time-discrimination parameter characterizes the sensitivity of an item for different speed levels of the test takers, and allows for an additional error component.

The lognormal response time model with a constant speed can be defined as;

$$\ln T_{ik} = \lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_\varepsilon^2) \quad (1.1)$$

The T_{ik} denotes the response time of person $i (i=1, \dots, N)$ on item k ($k=1, \dots, K$). The time intensity and time-discrimination parameter of item k are represented by λ_k and ϕ_k respectively, and the speed parameter of test taker i by ζ_i . The test takers are assumed to be randomly selected from a normal population. Therefore, the speed parameter is assumed to follow a normal population distribution.

1.1.2 Non-constant working speed

Besides the variability in working speed across test takers, variability in working speed of a test taker during the test is considered. Assume a linear trend for the working speed factor. Then, the lognormal response time model is extended with a linear growth term. The lognormal response time model with a common linear trend component is represented by;

$$\ln T_{ik} = \lambda_{\mu} - (\zeta_{i0} + \gamma X_{ik}) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\omega}^2) \quad (1.2)$$

Where γ is considered the common linear trend in speed. In this model, the time intensity parameters (denoted as λ_{μ}) have equal values in order to identify the trend in speed. Here, X denotes the order in which the K items are made by person i .

In the next chapters, the linear trend component is used in combination with higher-order time components to model more complex processes of working speed.

1.1.3 Dynamic process of working speed

In another approach, test takers can increase or decrease their working speed over blocks of items. The transition of changes in working speed over blocks of items can be modeled using dynamic factor models. In general, the observational model for the response times in block c of items is represented by;

$$T_{ikc} = \lambda_{kc} - \eta_{ic} + e_{ikc} \quad (1.3)$$

Here, η_{ic} denotes the speed level of subject i in block c , referred to as the working block speed, where the speed is the average speed of a block. The time intensity of item k in block c is denoted by λ_{kc} . The item-specific errors of subject i over item k in block c are denoted by e_{ikc} , and they are independently normally distributed with a mean of zero and a variance of σ_k^2 .

1.2 Ability

Besides observed response times for measuring speed, responses are also discussed in this dissertation as the indicators of ability. A three-parameter IRT model is considered to describe the responses. The probability of a correct response is given by;

$$P(Y_{ik} = 1 | \theta_i, b_k, a_k) = c_k + (1 - c_k) \Phi(a_k \theta_i - b_k) \quad (1.4)$$

where Y_{ik} denotes the response of person i on item k , and Φ denotes the cumulative normal distribution function. The item characteristics are

described by: a_k (the discrimination parameter), b_k (the difficulty parameter), and c_k (the guessing parameter). Person characteristic is described by θ_i (the ability parameter). In the present modeling approach, the higher ability, the higher probability of an answer being correct.

1.3 Joint Modeling of Responses and Response Times

A model that takes both response and response time into account is statistically more complex than the model with a single measure of task performance (either response or response time). Data of both responses and response times is only possible to collect for tests which are administered in a computerized mode.

However, in practice, educational assessments involve cognitive processes which cannot be fully understood without taking both responses and response times into account. Modelling observations, responses and response times, at the same time can lead to a better understanding of item, and person characteristics. This way, more complete information can be obtained to improve estimates and to detect possible aberrant behaviors (van der Linden, Scrams, & Schnipke, 1999; Lee & Chen, 2011).

Responses are considered as indicators of test taker's ability which often are observed on an ordinal scale and are modeled using a two-parameter or a three-parameter IRT model. Whereas response times are considered as indicators of speed which are observed on a continuous scale, and modeled using the log-normal model which includes time-discriminations and time intensities. The two latent person variables are often jointly modeled, which supports modelling correlation between them. The dependency between speed and ability is considered to be a within person relationship.

Luce (1986) introduced a negative correlation between speed and accuracy which is a within-person phenomenon, and is known as the speed-accuracy trade-off.

A hypothetical curve which represents a speed accuracy tradeoff for a person has been plotted in Figure 1.1. The speed accuracy tradeoff theory states that a certain working speed level chosen by a test taker leads to a certain accuracy level. If the working speed is increased, then according to the speed-accuracy tradeoff, the accuracy would decrease (since more errors would be made by the test taker).

A study of van der Linden (2007) discussed a hierarchical framework that allows ability and speed to be correlated. It describes a positive correlation between ability and speed at the population level, reflecting

that the test takers with higher ability tend to work faster than these with low ability. Extensions are considered in this dissertation to improve the joint modeling of responses and response times.

Figure 1.1. Example of a speed accuracy tradeoff

1.4 Outline

This dissertation consists a collection of studies where speed modelling for describing the behavior of test takers during the test, and the measurement of complex relationships among ability and speed are introduced and investigated. The main chapters (2 to 5) were written to be self-contained. Therefore, there is some inconsistency in the notation over chapters.

Chapter 2 is focused on response times to identify aberrant behavior. Response times are independently modeled, ignoring the correctness of the responses. In this chapter, person-fit statistics, to detect aberrant response behavior of test takers given their response times, are proposed. The test statistics have been derived from a lognormal response time model. The person fit statistics are referred to as the l'_z , l'_t , and l'_s , respectively. Various simulation studies were conducted to investigate the performance of the test statistics. It was shown that different simulated types of behavior can be identified through observed response time patterns. A real data example is also given to illustrate the use of the proposed person-fit tests.

Chapter 3 presents a study about a dynamic factor model for working speed. This model describes the transition of changes in working speed

over blocks of items. The proposed model is extended to the mixture modeling of different dynamic speed models, which allows the investigation of groups of test takers who show different types of speed behaviors over a test. This modeling approach generalizes the log-normal speed model of van der Linden (2006), which assumes that test takers work with a constant speed.

The proposed mixture modeling approach, was used to identify test takers who followed a stationary speed process and who followed a non-stationary speed process. Simulation studies for parameter recovery were conducted in order to investigate the sample size from which recommendations are derived for the proposed dynamic speed models. Subsequently, two empirical examples are given to illustrate the application of the dynamic speed models.

In Chapter 4, a latent growth modeling approach to model non-constant working speed is proposed. The model is used to measure more complex relationships between ability and variable working speed. Three models are considered in this study. Models 1 and 2 introduce two growth factors, an intercept and a linear slope to model variable working speed. Model 3 introduces three growth factors, an intercept, a linear and a quadratic term with random effects. The random effects have been used to describe an individual speed process and are used to define differences in the speed process between test takers. The Amsterdam chess test (ACT; van der Maas and Wagenmakers, 2005) data was used to illustrate the application of the models.

Chapter 5 focuses on the statistical evaluation of the joint model for speed and accuracy. The performance of fit tests for the joint model has been evaluated. A Bayesian significance test based on the Kolmogorov-Smirnov test (KS test) is used to detect violations of the assumption of normality of item residuals. Simulation studies were conducted in order to evaluate the performance of person fit statistics, and the performance of the Kolmogorov-Smirnov tests. Real data examples are provided to illustrate the application of the tests on real data.

CHAPTER 2

Testing for Aberrant Behavior in Response Time Modeling¹

2.1 Introduction

Many standardized tests rely on computer-based testing (CBT) because of its operational advantages. CBT reduces the costs involved in the logistics of transporting the paper forms to various test locations, and it provides many opportunities to increase test security. CBT also benefits the candidates. It enables testing organizations to record scores more easily and to provide feedback and test results immediately. In computerized adaptive testing (CAT), a special type of CBT, the difficulty level of the items is adapted to the response pattern of the candidate; this advantage also holds for multistage testing. Multimedia tools can even be included, and automated scoring of open-answer questions and essays can be supported. CBT can be used for online classes and practice tests.

An advantage of CBT is that it offers the possibility of collecting response time (RT) information on items. RTs provide information not only about test takers' ability and response behavior but also about item and test characteristics. With the collection of RTs, the assessment process can be further improved in terms of precision, fairness, and minimizing costs.

The information that RTs reveal can be used for routine operations in testing, such as item calibration, test design, detection of cheating, and adaptive item selection. In general, once RTs are available, they could be used both for test design and diagnostic purposes.

In general, two types of test models can be recognized: (a) separate RT models that only describe the distribution of the RTs given characteristics of the test taker and test items, in other word, RTs are modeled independently of the correctness of the response. Examples of this approach are: Maris (1993) who modeled RTs exclusively, whereas accuracy scores are not taken into consideration. Schnipke and Scrams (1997), estimated rapid guessing with assumption that accuracy and RTs

¹ Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2015). Testing for Aberrant Behavior in Response Time Modeling. *Journal of Educational and Behavioral Statistics*, 39(6), 426-451.

are independent given speed and ability. (b) test models that describe the distribution of RTs as well as responses. This approach takes correctness of the response and RTs into account, the correct responses reflect both speed and accuracy. With respect to the second one, Thissen (1983) defined the timed testing modeling framework, where item response theory (IRT) models are extended to account for speed and accuracy within one model. However, these types of models have been criticized because problems with confounding were likely to occur.

Recently, there is another approach introduced by van der Linden (2006, 2007) who advocated the first type of modeling and proposed a latent variable modeling approach for both processes. He defined a model for the RTs and a separate model for the response accuracy, where latent variables (person level and item level) explain the variation in observations and define conditional independence within and between the two processes. The RT process is characterized by RT observations, speed of working, and labor intensity, which are in a comparable way defined in the RT process by observations of success, ability, and item difficulty. This framework has many advantages and recognizes two distinct processes: It adheres to the multilevel data structure, and it allows one to identify within, between, and cross level relationships.

Unfortunately, not all respondents behave according to the model. Besides random fluctuation, aberrant response behavior also occurs due to, for example, item pre-knowledge, cheating, or test speededness. Focusing on RTs might have several advantages in revealing various types of aberrant behavior. RTs are continuous and therefore more informative and easier to evaluate statistically. One other advantage, especially for CAT, is that RTs are insensitive to the design effect in adaptive testing, since the selection of test items does not influence the distribution of RTs in any systematic way. RT models are defined to separate speed from time intensities; this makes it possible to compare the pattern of time intensities with the pattern of RTs.

Different types of aberrant behavior have been introduced and studied. van der Linden and Guo (2008) introduce two types of aberrant response behavior: (a) attempts at memorization, which might reveal themselves by random RTs; and (b) item preknowledge, which might result in an unusual combination of a correct response and RTs. RT patterns are considered to be suspicious when an answer is correct and the RT is relatively small while the probability of success on the item is low. Schnipke and Scramms (1997) studied rapid guessing, where part of the items show unusually small RTs. Bolt, Cohen, and Wollack (2002) focused on test speededness

toward the end of a test. For some respondents who run out of time, one might observe unexpected small RTs during the last part of the test.

For all of these types, it holds that response behavior either conforms to an RT model representing normal behavior or it does not (i.e., it is aberrant behavior). We propose using a log-normal RT model to deal with various types of aberrant behavior. Based on this log-normal RT model, a general approach to detect aberrant response behavior can be considered in which checks can be used to flag respondents or items that need further consideration. Checks could be used routinely in order to flag test takers or items that may need further consideration or to support observations by proctors or other evidence.

After introducing the log-normal RT model, an estimation procedure is described to estimate simultaneously all model parameters. Then, person-fit statistics are defined under the log-normal RT model, which differ with respect to their null distribution. It will be shown that given all information, each RT pattern can be flagged as aberrant with a specific posterior probability, to quantify the extremeness of each pattern under the model. In a simulation study, the power to detect the aberrancies is investigated by simulating various types of aberrant response behavior. Finally, the results from a real data example and several directions for future research are presented.

2.2 RT Modeling

Van der Linden (2006) proposed a log-normal distribution for RTs on test items. In this model, the logarithm of the RTs is assumed to be normally distributed. The model is briefly discussed since it is used to derive new procedures for detecting aberrant RTs. The log-normal density for the distribution of RTs is specified by the mean and the variance. The mean term represents the expected time the test taker needs to answer the item, and the variance term represents the variance of measurement errors. In log-normal RT models, each test taker is assumed to have a constant working speed during the test. Let $p = 1, \dots, N$ be an index for the test takers, $i = 1, \dots, I$ be an index for the items, ζ_p denote the working speed of test taker p , λ_i denote the time intensity of item i , T_{ip} denote the RT of test taker p to item i . Subsequently, the logarithm of T_{ip} has mean $\mu_{pi} = \lambda_i - \zeta_p$ (see also, van der Linden, 2006). The lower the time intensity of an item, the lower the mean. In the same way, the faster a test taker operates, the lower the mean. This model can be extended by introducing a time-discrimination parameter to allow variability in the effect of

increasing the working speed to reduce the mean. Let ϕ_i denote the time discrimination of item i .

With this extension, the mean is parameterized as $\mu_{pi} = \phi_i (\lambda_i - \zeta_p)$, such that the reduction in RT by operating faster is not constant over items. The higher the time discrimination of an item, the higher the reduction in the mean when operating faster. For example, when a test taker operates a constant C faster, the mean equals $\mu_{pi} = \phi_i (\lambda_i - (\zeta_p + C)) = \phi_i (\lambda_i - \zeta_p) - \phi_i C$, such that the item-specific reduction is defined by $\phi_i C$.

Observed RTs will deviate from the mean term (i.e., expected times), and the errors are considered to be measurement errors. The response behavior of test takers can deviate slightly during the test, leading to different error variances over items. Test takers might stretch their legs or might be distracted for a moment, and so on. These measurement errors are assumed to be independently distributed given the operating speed of the test taker, the time intensities, and time discriminations. Let σ_i^2 denote the error variance of item i .

In the log-normal RT model, σ_i^2 can vary over items. The errors are expected to be less homogenous, when, for example, items are not clearly written, when items are positioned at the end of a time-intensive test, or when test conditions vary during an examination and influence the performance of the test takers (e.g., noise nuisance).

With this mean and variance, the log-normal model for the distribution of T_{ip} can be represented by

$$p(t_{ip} | \zeta_p, \lambda_i, \phi_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2 t_{ip}}} \exp \left[-\frac{1}{2\sigma_i^2} (\ln t_{ip} - \phi_i (\lambda_i - \zeta_p))^2 \right]. \quad (2.1)$$

We will refer to the time-intensity and time-discrimination parameters as the item's *time characteristics* in order to stress their connection with the definition of *item characteristics* (i.e., item difficulty and item discrimination) in IRT.

With the introduction of a time-discrimination parameter, differences in working speed do not lead to a homogeneous change in RTs over items. A differential effect of speed on RTs is allowed, which is represented by the time-discrimination parameters. The idea is that working speed is modeled by a latent variable representing the ability to work with a certain level of speed. Furthermore, it is assumed that this construct comprehends different dimensions of working speed. Depending on the

item, this construct can relate, for example, to a physical capability, a cognitive capability, or a combination of both. For example, consider two items with the same time intensity, where one item concerns writing a small amount of text and the other doing analytical thinking. Differences between the RTs of two test takers can be explained by the fact that one works faster. However, differences in RTs between test takers are not necessarily homogenous over items. One item appeals to the capability of writing faster and the other to thinking or reasoning faster, and it is unlikely that both dimensions influence RTs in a common way.

2.2.1 Identification

The observed times have a natural scale, which is defined by a unit of measurement (e.g., seconds). However, the metric of the scale is undefined due to our parameterization. First, the mean of the scale is undefined due to the speed and time intensity parameters in the mean, $\lambda_i - \zeta_p$. To identify the mean of the scale, the mean speed of the test takers is set to zero. Note that this value of zero corresponds to the population-average total test time, which corresponds to the sum of all time intensities. Second, the variance of the scale is also undefined due to the time-discrimination parameter and the population variance of the speed parameter. The variance of the scale is identified by setting the product of discriminations equal to one. It is also possible to fix the population variance of speed (e.g., to set it equal to one).

2.2.2 A Bayesian Log-Normal RT Model

Prior distributions can be specified for the parameters of the distribution of RTs in Equation (2.1). The population of test takers is assumed to be normally distributed such that

$$\zeta_p \sim N(\mu_\zeta, \sigma_\zeta^2) \quad (2.2)$$

where $\mu_\zeta = 0$ to identify the mean of the scale. An inverse gamma hyper prior is specified for the variance parameter. The prior distribution for the time intensity and discrimination parameters give support to partial pooling of information across items. When the RT information for a specific time intensity leads to an unstable estimate, RT information from other items is used to obtain a more stable estimate. This partial pooling of information within a test is based on the principle that the items in the test have an average time intensity and an average time discrimination. Each individual item can have characteristics that deviate from the average depending on the information in the RTs.

Partial pooling of information is also defined for item-specific parameters. The time intensity and discrimination parameter in Equation

(2.1) relate to the same item, and are allowed to correlate. A bivariate normal distribution is used to describe the relationship between the parameters,

$$\begin{pmatrix} \phi_i \\ \lambda_i \end{pmatrix} \sim N \left(\begin{bmatrix} \mu_\lambda \\ \mu_\phi \end{bmatrix}, \begin{bmatrix} \sigma_\phi^2 & \rho \\ \rho & \sigma_\lambda^2 \end{bmatrix} \right). \quad (2.3)$$

The mean time intensity of the test is denoted by μ_λ and represents the average time it takes to complete the test. The mean time discrimination is denoted by μ_ϕ and represents the effect of reducing the mean test time when increasing the working speed. The common covariance parameter ρ across items represents for each item the linear relation between both parameters. For example, items that are more time intensive might discriminate better between individual performances. The hyper priors will be normal distributions for the mean parameters and an inverse Wishart distribution for the covariance matrix. Although the modeling approach supports partial pooling of information, the hyper priors are specified in such a way that partial pooling of information is diminished and the within-item RT information is the most important source of information to estimate the time-intensity and time-discrimination parameters.

The measurement error variance parameters σ_i^2 are assumed to be independently inverse gamma distributed. The errors of a test taker are assumed to be independently distributed given the speed of working and the item's time characteristics.

The specification of the log-normal model leads to the following random effects model to model the logarithm of RTs:

$$\begin{aligned} \log T_{ip} &= \phi_i (\lambda_i - \zeta_p) + \varepsilon_{ip} \} \text{Modeling time observation} \\ \left. \begin{aligned} \phi_i &= \mu_\phi + r_{1i} \\ \lambda_i &= \mu_\lambda + r_{2i} \end{aligned} \right\} \text{Item specification} \\ \zeta_p &= \mu_\zeta + e_p \} \text{Test-taker specification,} \end{aligned} \quad (2.4)$$

where three levels can be recognized. At Level 1, time observations are modeled using a normal distribution for the logarithm of RTs and three random effects to address the influence of the test taker's speed of working and of the item's time characteristics. The test item's properties are modeled as multivariate normally distributed random effects and are modeled at the level of items. Finally, the test taker's working speed is modeled at the level of persons.

2.2.3 The Estimation Procedure for Log-Normal RT Models

The model parameters and the test statistics are computed using a Bayesian estimation procedure. With the Markov chain Monte Carlo (MCMC) method referred to as Gibbs sampling, samples are obtained from the posterior distributions of the model parameters. Gibbs sampling is an iterative estimation method where, in each iteration, a sample is obtained from the full conditional distributions of the model parameters. To apply Gibbs sampling, the full conditional distributions of the model parameters need to be specified. For the log-normal model, the technical details of the estimation method are given by Klein Entink, Fox, and van der Linden (2009a), van der Linden (2007), and Fox, Klein Entink, and van der Linden (2007).

2.3 Test for Aberrant RT Patterns

One of the most popular fit statistics in person-fit analysis is the l_z statistic (Drasgow, Levine, & Williams, 1985), which is the standardized likelihood-based person-fit statistic l_o of Levine and Rubin (1979). This person-fit statistic has received much attention in educational measurement. Studies have shown that it almost always outperforms other person-fit statistics, and it is commonly accepted as one of the most powerful person-fit statistics to detect aberrant response patterns. With this in mind, we propose a person-fit statistic for aberrant response behavior for RT patterns.

The log-likelihood of the RTs is used to evaluate the fit of a response pattern consisting of RTs. We will use $t_{ip}^* = \ln(t_{ip})$ to denote the logarithm of the RT of test taker p on item i . Our likelihood-based person-fit statistic for RTs requires knowledge of the density of the response pattern. This follows directly from the normal model for the logarithm of RTs; that is,

$$l_o(\zeta_p, \lambda, \phi, \sigma^2; \mathbf{t}_p^*) = -2 \log p(\mathbf{t}_p^* | \zeta_p, \lambda, \phi, \sigma^2) = \sum_{i=1}^I l_{oi}. \quad (2.5)$$

The l_o statistic can be evaluated over all items in the test, but it is also possible to consider a subpart of the test. An unusually large value indicates a misfit, since it represents a departure of the RT observations from expected RTs under the model. The posterior distribution of the statistic can be used to examine whether a pattern of observed RTs is extreme under the model.

Given the model specification in Equation (2.1), the probability density function of a response pattern is represented by the product of individual RTs. The probability density of response pattern $\mathbf{t}_p^* = (t_{1p}^*, \dots, t_{Ip}^*)$ is given by

$$\begin{aligned}
 -2 \log p(\mathbf{t}_p^* | \zeta_p, \lambda, \phi, \sigma^2) &= -2 \sum_{i=1}^I \log p(t_{ip}^* | \zeta_p, \lambda_i, \phi_i, \sigma_i^2) \\
 &= \sum_{i=1}^I \left(\left(\frac{t_{ip}^* - \mu_{ip}}{\sigma_i} \right)^2 + \log(2\pi\sigma_i^2) \right) \\
 &= \sum_{i=1}^I (Z_{ip}^2 + \log(2\pi\sigma_i^2)),
 \end{aligned} \tag{2.6}$$

where Z_{ip} is standard normally distributed, since it represents the standardized error of the normally distributed logarithm of RT.

The test statistic l_0 depends on various model parameters. It is possible to compute statistic values given values for the model parameters or given posterior distributions of the model parameters. In the last case, the posterior mean statistic value is estimated by integrating over the posterior distributions of the model parameters.

In the person-fit literature, the standardized person-fit statistic, which is usually denoted as l_z , receives much attention because it has an asymptotic standard normal distribution. Drasgow et al. (1985) showed that for tests longer than 80 items, the l_z statistic is approximately normally distributed. Other studies (e.g., Meijer & Sijsma, 1995; Molenaar & Hoijtink, 1990) showed that for shorter tests the distribution of the test statistic was negatively skewed, violating the assumption of symmetry of the normal distribution. Snijders (2001) proposed an adjustment to standardize the l_z statistic, thereby accounting for the fact that parameter estimates are used to compute the statistic value.

The standardized version of the l_0^t for RTs, denoted as l_z^t , requires an expression for the expected value and the variance of the statistic in Equation (2.5). In Appendix A, it is shown that the conditional expectation is given by

$$E[l_0(\zeta_p, \lambda, \phi, \sigma^2) | \mathbf{t}_p^*, \zeta_p, \lambda, \phi, \sigma^2] = \sum_i (1 + \ln(2\pi\sigma_i^2)) \tag{2.7}$$

and the variance is given by

$$Var[l_0(\zeta_p, \lambda, \phi, \sigma^2) | \mathbf{t}_p^*, \zeta_p, \lambda, \phi, \sigma^2] = 2I, \tag{2.8}$$

where I is the total number of test items. Subsequently, the standardized version, l_z^t , is derived by standardizing the statistic in Equation (2.5) using the terms in Equations (2.7) and (2.8). It follows that

$$l_z^t(\zeta_p, \lambda, \phi, \sigma^2; \mathbf{t}_p^*) = \frac{\left(\sum_{i=1}^I Z_{ip}^2 + \log(2\pi\sigma_i^2) \right) - \left(\sum_{i=1}^I 1 + \log(2\pi\sigma_i^2) \right)}{\sqrt{2I}} = \frac{\sum_{i=1}^I Z_{ip}^2 - I}{\sqrt{2I}} \quad (2.9)$$

To ease the notation, the statistic's dependency on the model parameters is ignored, leading to $l_z^t(\zeta_p, \lambda, \phi, \sigma^2; \mathbf{t}_p^*) = l_z^t(\mathbf{t}_p^*)$. In the computation of l_z^t , model parameters are assumed to be known, or the posterior expectation is taken over the unknown model parameters.

2.4 The Null Distribution

In order to come to a person-fit statistic, the null distribution of l_z^t has to be derived. First we introduce some notation. The logarithm of RTs is represented by a random variable T_{pi}^* , which is normally distributed, where the observed values are denoted by t_{pi}^* . An RT pattern of test taker p is represented by $\mathbf{T}_p^*$. Given this notation, the null distribution of $l_z^t(\mathbf{T}_p^*)$ can be derived in three different ways, resulting in three different person-fit statistics for $\mathbf{T}_p^*$ under the log-normal model.

First, the null distribution of the $l_z^t(\mathbf{T}_p^*)$ follows from the fact that the errors Z_{ip} (see Equation (2.9)) are standard normally distributed. The sum of squared errors, which are standard normally distributed, is known to be chi-squared distributed with I degrees of freedom. Box, Hunter, and Hunter (1978, p. 118) showed that a chi-squared distributed variable T with I degrees of freedom, the distribution of $(T - I)/\sqrt{2I}$ is approximately standard normal. Therefore, the null distribution of the $l_z^t(\mathbf{T}_p^*)$ can be considered to be approximately standard normal.

Second, an exact null distribution can be obtained by considering a nonstandardized version of the $l_z^t(\mathbf{T}_p^*)$, which is the sum of squared standardized errors:

$$l^t(\mathbf{T}_p^*) = \sum_{i=1}^I Z_{ip}^2. \quad (2.10)$$

This sum of squared errors, which are standard normally distributed, is known to be chi-squared distributed with I degrees of freedom.

Third, the Wilson–Hilferty transformation can be used to standardize the person-fit statistic $l^t(\mathbf{T}_p^*)$ in such a way that it is approximately standard normal distributed. This leads to

$$l'_s(\mathbf{T}_p^*) = \frac{\left(\sum_{i=1}^I Z_{ip}^2 / I \right)^{1/3} - (1 - 2/(9I))}{\sqrt{2/(9I)}}. \quad (2.11)$$

Summarized, three person-fit statistics for RTs are considered that differ in the way the null distribution is derived. An overview of the tests is given in Table 2.1.

Tabel 2.1
Person-fit statistics for RT data under the lognormal model

Statistic	Type Null Distribution	Exact or Approximation	Probability of Significance
l'_z	Normal	Approximation	$P(l'_z(\mathbf{T}_p^*) > C) \approx \Phi(l'_z(\mathbf{T}_p^*) > C)$
l'_t	Chi-squared	Exact	$P(l'_t(\mathbf{T}_p^*) > C) = P(\chi_I^2 > C)$
l'_s	Normal	Approximation	$P(l'_s(\mathbf{T}_p^*) > C) \approx \Phi(l'_s(\mathbf{T}_p^*) > C)$

2.5 Bayesian Testing of Aberrant RT Patterns

To assess the extremeness of the pattern of RTs, the posterior probability can be computed such that the estimated statistic value, say $l'(\mathbf{t}_p^*)$, is greater than a certain threshold C . This threshold C defines the boundary of a critical region, which is the set of values for which the null hypothesis is rejected if the observed statistic value is located in the critical region. The critical value C can be determined from the null distribution; that is,

$$P(l'(\mathbf{T}_p^*) > C) = P(\chi_I^2 > C) = \alpha, \quad (2.12)$$

since the null distribution is a chi-squared distribution with I degrees of freedom, where α is the level of significance. When the observed statistic value, $l'(\mathbf{t}_p^*)$, is larger than C , the RT pattern will be flagged.

Given the sampled parameter values in each MCMC iteration, it is also possible to compute a function of the model parameters (e.g., a probability statement). To illustrate this, consider the tail-area event as specified in Table 2.1. Given sampled values from the posterior distribution of the model parameters, the posterior probability can be computed as

$$\begin{aligned} P(l'(\mathbf{T}_p^*) > C) &\approx \sum_{m=1}^M P(l'(\mathbf{T}_p^*) > C) p(\zeta_p^{(m)}, \lambda^{(m)} | \mathbf{t}_p^*) \\ &= \sum_{m=1}^M \Phi(l'(\mathbf{T}_p^*) > C) p(\zeta_p^{(m)}, \lambda^{(m)} | \mathbf{t}_p^*) \end{aligned} \quad (2.13)$$

where m denotes the MCMC iteration number. The terms to standardize the test statistic depend on the model parameters. In each iteration, the test statistic is computed using the sampled model parameters, and the average posterior probability approximates the marginal posterior probability of obtaining a test statistic larger than a criterion value C . The uncertainty in the parameters is taken into account in the computation of the posterior probability.

Note that in Equation (2.13), draws are used from the posterior distribution to compute the marginal posterior probability. When using posterior draws, the posterior distribution of the model parameters might be distorted by RT data that do not fit the model. An alternative would be to use draws from the prior distribution. Then, most often a much larger number of draws will be required to obtain an accurate estimate of the marginal posterior probability. Moreover, a misspecification of the priors might lead to a biased posterior probability estimate.

Besides testing whether a pattern of RTs is in a critical area defined by a threshold C , it is also possible to quantify the extremeness of the observed RT pattern by computing the right-tail area probability under the model. This right-tail probability represents the posterior probability of observing a more extreme statistic value under the model. The estimated statistic value is constructed from the sum of squared errors, and an extreme statistic value indicates that the RT pattern is not likely to be produced under the log-normal model. When the posterior probability is close to zero, it can be concluded that the pattern is unlikely under the posited log-normal model and the pattern is considered to be aberrant given the observed data.

Note that the decision to flag an RT pattern as extreme depends on the size of the statistic value but also on the posterior uncertainty. When the distribution of the test statistic is rather flat, it is less likely to conclude with high posterior probability that an RT pattern is extreme in comparison to a highly peaked distribution. Given accurate information, a more definitive decision can be made about the extremeness of the RT pattern.

2.6 Dealing with Nuisance Parameters

The test statistic depends on the model parameters, which follows directly from the definition of Z_{pi} . To compute the marginal posterior probability of observing a more extreme value than the observed one, an integration needs to be performed over all model parameters:

$$P(l^t(\mathbf{T}_p^*) > C) = \int \int_{\lambda, \zeta_p} P(l^t(\mathbf{T}_p^*) > C | \zeta_p, \lambda) p(\zeta_p, \lambda) d\zeta_p d\lambda. \quad (2.14)$$

The marginal posterior probability is obtained by integrating over the model parameters. MCMC can be used to obtain draws from the posterior distribution of the model parameters. For each draw, the probability that the computed statistic value is above a threshold value C can be computed. The average posterior probability over MCMC iterations is an estimate of the marginal posterior probability as specified in Equation (2.12).

In Equation (2.14), the distribution of the statistic is assumed to be known, and the assessment of the test statistic is known as a *prior predictive test* (Box, 1980). Given prior distributions for the model parameters, it is assessed how extreme the observed statistic value is. Prior predictive testing is usually preferred, since the double use of the data in posterior predictive assessment is known to bias the distribution of estimated tail-area probabilities. When the data are used to estimate the model parameters and to assess the distribution of the test statistic, the tail-area probabilities are often not uniformly distributed. This makes it more difficult to interpret the estimated probabilities. In the prior predictive assessment approach, as stated in (12) and (14), the double use of the data is avoided and the tail-area probability estimates can be correctly interpreted.

To assess whether an RT pattern is extreme, a classification is made based on the value of the test statistic. The exact or an accurate approximation of the null distribution of the statistic is known but depends on unknown model parameters. When the statistic is computed by plugging in parameter estimates, the corresponding tail-area probability might be biased. Therefore, the probability that an RT pattern will be flagged as extreme is evaluated in each MCMC iteration. An accurate decision can be made in each MCMC iteration given values for the model parameters. Let random variable F_p take on a value of one when the RT pattern of test taker p is flagged, or a value of zero otherwise. Thus,

$$F_p = \begin{cases} 1 & \text{if } P(l'(\mathbf{T}_p^*) > l'(\mathbf{t}_p^*)) < \alpha \\ 0 & \text{if } P(l'(\mathbf{T}_p^*) > l'(\mathbf{t}_p^*)) \geq \alpha. \end{cases} \quad (2.15)$$

Interest is focused on the marginal posterior probability that the RT pattern of test taker p will be flagged, which is computed by

$$\begin{aligned} P(F_p = 1 | \mathbf{t}_p^*) &= \int \int_{\lambda, \zeta_p} I(F_p = 1 | \mathbf{t}_p^*, \zeta_p, \lambda) p(\zeta_p, \lambda) d\zeta_p d\lambda \\ &\approx \sum_{m=1}^M I(F_p^{(m)} = 1 | \zeta_p^{(m)}, \lambda^{(m)}) / M, \end{aligned} \quad (2.16)$$

where in MCMC iteration m , $F_p^{(m)} = 1$ when $P\left(\chi^2 > l^t(\mathbf{t}_p^*) \mid \zeta_p^{(m)}, \lambda^{(m)}\right) < \alpha$. So, the probability that a pattern will be flagged is evaluated in each iteration. The average probability over iterations approximates the marginal probability of a flagged RT pattern. The extremeness of the pattern can be quantified, since the posterior probability in Equation (2.16) states how likely it is that the pattern will be flagged under the log-normal model. It can be decided that only patterns that have a posterior probability of .95 or higher will be flagged under the model. This reduces the probability of making a Type I error, since the posterior probability quantifies the extremeness of each RT pattern, instead of classifying the pattern based on a chosen significance level α .

The posterior probability of the extremeness of the response pattern in Equation (2.14) can also be defined from a posterior predictive perspective. Given the model parameters, the posterior probability of the test statistic is evaluated given its sampling distribution. When the distribution of the statistic is unknown, the posterior predictive distribution of the data can be used to assess the distribution of the test statistic. In that case, the extremeness of the estimated test statistic is evaluated using the posterior predictive distribution of the data. This is shown by

$$P\left(l^t(\mathbf{T}_p^{*rep}) > l^t(\mathbf{t}_p^*)\right) = \int_{\mathbf{t}_p^{*rep}} P\left(l^t(\mathbf{T}_p^{*rep}) > l^t(\mathbf{t}_p^*)\right) p(\mathbf{T}_p^{*rep} \mid \zeta_p, \lambda) d\mathbf{T}_p^{*rep}, \quad (2.17)$$

where $\mathbf{T}_p^{*rep}$ denotes the replicated data under the model and the left-hand side of Equation (2.17) represents the posterior predictive probability of observing a statistic value that is greater than the statistic value based on the observed data.

Posterior predictive tests have been suggested in many different applications to evaluate the fit of models. Rubin (1984), among others, advocated the use of posterior predictive assessment to evaluate the compatibility of the model to the data. Box (1980) recommended the use of the marginal predictive distribution of the data to evaluate the fit of the model, which is also known as prior predictive assessment.

van der Linden and Guo (2008) also suggested using a predictive distribution to evaluate RTs. In their approach, a cross-validation predictive residual distribution is used to evaluate the extremeness of the remaining RTs. Furthermore, the predicted response is compared to the observed response in an adaptive test application. The normal distribution of the logarithm of RTs is used to calculate the power of identifying aberrant RTs. They also used a less accurate method, which was based

on classifying estimated residuals. Ignoring the uncertainty of the estimates, RTs were flagged as aberrant when the corresponding estimated standardized residuals were larger than 1.96 or smaller than -1.96. In the present approach, the posterior uncertainty is taken into account, and RTs are flagged to be aberrant with a certain posterior probability.

2.7 Results

Through simulation studies, the performance of the person-fit statistics for RT patterns is evaluated. A comparison is made between three different programs for estimating the model parameters. The detection rates of the l' statistic are evaluated for different types of misfit. Different conditions are simulated to investigate the performance of the statistic. The MCMC method for estimating the model parameters of the log-normal model was implemented in R² and is referred to as Log Normal Response Times (LNRT).

2.7.1 Investigation of Detection Rates

Data sets were generated under different types of response behavior to simulate aberrant responses. Different data specifications were considered: sample sizes of 500 and 1,000 test takers, and test lengths of 10 and 20 items. For each type of aberrant response behavior, 5%, 10%, or 20% of the test takers responded in this way. The remaining response patterns were generated according to the log-normal model. The specification of the log-normal model was equal to the setting in the parameter recovery study, except that time-discrimination parameters were generated from a normal distribution with mean = 1 and variance = .17. Three types of aberrant behavior were simulated:

Random response behavior. The first type of aberrant RTs represented test takers who responded to the test items with random RTs on a subset of items. The simulated aberrant RTs did not correspond with the time intensities of the items. Much faster or slower times were simulated given the time intensities of the items. For half of the test items, aberrant RTs were generated from a log-normal distribution with the mean equal to the average item time and three times the average standard deviation of the RTs. The average test times for the aberrant RT patterns were similar to those for the nonaberrant RT patterns. This corresponds to the strategy that a test taker might know the average time to complete the test but not the average time to complete each item.

² The LNRT program written in R will be made available by the authors.

Test speededness or variant working speed. Test takers with an invariant working speed will work with a constant level of speed. The assumption of conditionally independently distributed RTs given working speed is violated when the working speed is variant. This can occur when, for example, the test taker is not concentrating, has pre-knowledge of some items, or operates under higher time pressure than others. In this second type of aberrant pattern, half of the test items were answered much faster than expected under the log-normal model. For half of the test items, working speed of (aberrant) test takers with a variant working speed were simulated to be 1.5 standard deviations faster than the population average working speed.

One extreme RT. Test takers are assumed to work with a constant speed such that the total test time is assumed to reflect the total amount of time required to produce all answers. The total test time will be biased when test takers are interrupted or distracted while taking the test. When a test taker is taking a break (e.g., getting coffee) and is not working on the test, the next observed RT will not reflect the time spent on producing an answer. This will also bias the total test time. In this third condition, extreme RTs were simulated from a log-normal distribution with a mean equal to at least twice the maximum time intensity of the items in the test. Each aberrant RT pattern consisted of only one extreme RT.

The detection and false alarm rates were investigated under the log-normal model for the different types of violations. In this study, item parameters were assumed to be known, but the working speed and other model parameters were estimated from the data using the LNRT program. Note that the posterior uncertainty in the model parameters were taken into account in the estimation of the test statistics and the flagging of RT patterns. RT patterns were flagged to be aberrant in different ways. First, following Equation (2.16), each test taker's probability of a flagged pattern was computed. Subsequently, the average posterior probability was computed from the individual posterior probabilities of a flagged pattern, thus representing the average posterior probability of flagged patterns in the population. Under the model, this average probability of flagged patterns represents the Type I error. Furthermore, for RTs generated under the model, patterns were approximately flagged to be aberrant with probability .05, when using the significance level $\alpha = .05$. Second, patterns were flagged to be aberrant when the posterior probability of an aberrant pattern was at least .80 or .90 (according to Equation (2.16)), which will be referred to as the classification probability.

2.7.2 Comparing Three Statistics

Before looking into detail at the false alarm rates and detection for the various conditions, the three statistics in Table 2.1 were compared. For data simulated under the log-normal model, the classification probability of being assigned to the class of patterns included in the estimation of item parameters (according to Equation (2.19)) and the probability of a flagged pattern (according to Equation (2.16)) were computed for the three statistics. In Figure 2.1, for each statistic the probabilities of each pattern are plotted against each other and a smoothing curve is drawn through the points to represent the relationship. For the curve of l^t and l_s^t , patterns with a classification probability less than 5% are most likely to be flagged as aberrant, since a significance level of 5% was used. Both statistics give a similar picture, and the curves are almost equal. Therefore, it can be concluded that the approximate null distribution of l_s^t is nearly as accurate as the exact null distribution of l^t .

The curve of the approximate null distribution of l_z^t shows a shift to the left for low classification probabilities. These posterior classification probabilities are too conservative, which leads to lower probabilities of being flagged for l_z^t compared to l^t . This makes l_z^t not very useful for the detection of aberrant patterns.

For each RT pattern, a probability of being flagged and a classification probability are computed. In Figure 2.1, each point of the curve represents an RT pattern. The location of the point in the curve shows whether it is a regular or a suspicious pattern. The Type I error is equal to the expected probability of being flagged in the population. Subsequently, patterns can be marked as aberrant with a specific posterior probability, which represents the accuracy of making the right decision. However, increasing the accuracy of correctly identifying an aberrant pattern is accompanied with a decrease in the probability of identifying all aberrant patterns.

Since l_z^t is not very useful for the detection of aberrant patterns and the approximate null distribution of l_s^t is nearly as accurate as the exact null distribution of l^t , attention will be focused on l^t in the simulation study.

Figure 2.1. Classification probability versus probability of being flagged for the three different statistics ($N = 1,000$, $I = 10$)

2.7.3 Model-Fitting Responses and Random Response Behavior

In Table 2.2, the false alarm rates and detection rates, averaged over 50 replicated data sets, are given for the l^t statistic for different sample sizes and for model-fitting responses and responses with 5%, 10%, and 20% of the RT patterns generated under random response behavior.

In the model-fitting condition, differences in false alarm rates were found. The false alarm rate is slightly lower for a population size of 500 compared to a size of 1,000. When flagging patterns with a posterior classification probability of at least .80, the false alarm rate is much lower than the results for the average posterior probability flagging and decreases slightly more for a classification probability of .95. In that case, only the most extreme patterns are classified.

With respect to aberrant response types, the aberrant patterns were detected in all cases under all classification probabilities (under the heading “Aberrant” in Table 2.2). Given the specifications of random response behavior, the patterns were detected as significantly different from patterns that can be expected under the model. When 5% was simulated to be aberrant, then this 5% was also identified in the population (under the heading “Aberrant”). Under the different percentages, the percentage of aberrant patterns was still detected in the population.

Tabel 2.2

False alarm rates and detection rates of l^i for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)

		Random Response Behavior						
		Model Fit	5%		10%		20%	
	Post. Class	Pop.	Aberrant	Pop.	Aberrant	Pop.	Aberrant	Pop.
$N=500$ $l=10$	No	0.044	1.000	0.052	1.000	0.102	1.000	0.201
	.80	0.025	1.000	0.050	1.000	0.100	1.000	0.200
	.95	0.021	0.999	0.050	1.000	0.100	1.000	0.200
$N=1,000$ $l=10$	No	0.056	1.000	0.051	1.000	0.101	1.000	0.201
	.80	0.035	1.000	0.050	1.000	0.100	0.999	0.200
	.95	0.030	1.000	0.050	0.999	0.100	0.999	0.200
$N=500$ $l=20$	No	0.035	1.000	0.050	1.000	0.100	1.000	0.200
	.80	0.024	1.000	0.050	1.000	0.100	1.000	0.200
	.95	0.019	1.000	0.050	1.000	0.100	1.000	0.200
$N=1,000$ $l=20$	No	0.047	1.000	0.050	1.000	0.100	1.000	0.200
	.80	0.033	1.000	0.050	1.000	0.100	1.000	0.200
	.95	0.029	1.000	0.050	1.000	0.100	1.000	0.200

Note. Post.Class = Posterior Classification; Pop. = Population

2.7.4 Test Speededness

In Table 2.3, detection rates are given for the l^i statistic for different sample sizes and responses simulated under test speededness or variant working speed. In the same way, data sets were simulated with 5%, 10%, and 20% of the RT patterns generated under test speededness, and patterns were flagged to be aberrant with a significance level of .05.

Tabel 2.3

Detection rates of l^i for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications)

		Test Speededness					
		5%		10%		20%	
	Post. Class	Aberrant	Pop.	Aberrant	Pop.	Aberrant	Pop.
$N=500$ $l=10$	No	0.888	0.078	0.885	0.116	0.850	0.192
	.80	0.859	0.060	0.855	0.097	0.800	0.166
	.95	0.848	0.056	0.836	0.093	0.771	0.159
$N=1,000$ $l=10$	No	0.929	0.093	0.917	0.131	0.878	0.205
	.80	0.910	0.073	0.894	0.110	0.836	0.176
	.95	0.899	0.068	0.880	0.105	0.816	0.170
$N=500$ $l=20$	No	0.991	0.074	0.990	0.121	0.979	0.213
	.80	0.987	0.063	0.986	0.110	0.813	0.167
	.95	0.986	0.060	0.982	0.107	0.807	0.164
$N=1,000$ $l=20$	No	0.995	0.085	0.994	0.131	0.988	0.224
	.80	0.993	0.072	0.992	0.117	0.981	0.205
	.95	0.991	0.069	0.990	0.114	0.978	0.202

Note. Post.Class = Posterior Classification; Pop. = Population

For different percentages, with patterns showing test speededness, the detection rate is around .90 for a test of 10 items and approximately .99 for a longer test of 20 items. The detection rates are only somewhat smaller when they are computed using a classification probability of at least .80 or .90. In the worst case of 20% aberrant patterns, the detection rate is around 77% of the simulated aberrant patterns. When looking at the percentage of detections in the population, slightly more patterns are flagged than the simulated percentage of aberrant patterns.

2.7.5 One Extreme Response

In Table 2.4, averaged over 50 replicated data sets, detection rates are given for the l' statistic for different sample sizes and RT patterns including an extreme response for the first item. The detection rates are somewhat acceptable, when only 5% of the patterns include an extreme response. When the test length increases, the detection rates decrease, since it becomes more difficult to identify the longer RT patterns with just one extreme RT. When the sample size increases, the detection rates also increase. A distortion in detection rates became visible when the percentage of aberrant patterns increased. In that case, the measurement error variance increased, which simply adjusted the range of possible RTs. Thus, the variability in RTs for the first item was increased by an increase in the estimated measurement error variance for the first item. The detection rates were much better when the extreme response was randomly assigned across patterns to one of the test items.

Tabel 2.4

Detection rates of l' for a 10- and 20-item test and 500 and 1,000 examinees using a significance level of .05 (50 replications).

		An Extreme RT					
		5%		10%		20%	
	Post. Class	Aberrant	Pop.	Aberrant	Pop.	Aberrant	Pop.
N=500 l=10	No	0.830	0.072	0.732	0.101	0.314	0.088
	.80	0.782	0.055	0.664	0.081	0.251	0.064
	.95	0.738	0.049	0.604	0.072	0.219	0.055
N=1,000 l=10	No	0.858	0.083	0.741	0.111	0.380	0.108
	.80	0.824	0.065	0.688	0.090	0.320	0.083
	.95	0.788	0.06	0.636	0.081	0.288	0.073
N=500 l=20	No	0.676	0.057	0.473	0.072	0.137	0.048
	.80	0.606	0.044	0.396	0.056	0.105	0.034
	.95	0.554	0.039	0.352	0.049	0.089	0.028
N=1,000 l=20	No	0.811	0.077	0.555	0.090	0.175	0.064
	.80	0.766	0.063	0.490	0.073	0.141	0.047
	.95	0.715	0.058	0.446	0.065	0.127	0.042

Note. Post.Class = Posterior Classification; Pop. = Population

In Figure 2.2, the receiver operating characteristic (ROC) curves of the f' test illustrate the performance for artificial data generated for 1,000 students and 10 items, where 10% of the students show aberrant behavior on five items. The x-axis, referred to as the false alarm rate, represents the percentage of incorrectly identified aberrant RT patterns and the y-axis, referred to as the hit rate (i.e., sensitivity), represents the percentage of correctly identified aberrant RT patterns.

For the left plot, random RTs were generated for five items, where the variability in random RTs was equal to, or one and a half or two times larger than the variability in RTs generated under the log-normal model. It follows that for small threshold values, accurate decisions can be made when the variance of random RTs is larger than the variance of the model generated RTs. In that case, with a significance level of .1, more than 80% of the patterns can be correctly classified.

For the speededness condition, the performance of the f' test was less good. In this condition, 10% of the students worked slower on the first five items. Their speed levels were one, two, or three standard deviations lower compared to the last five items of the test. An increase of one standard deviation in working speed means for a student who was working with a population average speed level increased his/her speed level to work faster than 84% of the students in the population. It follows from the ROC curves that even in the extreme situation, only 75% of the RT patterns of students who increased their speed levels with three standard deviations were detected, given a false positive rate of less than 10%.

The difference in test performance between the two conditions can be explained by the fact that the two conditions, random RTs and speededness, induce misfit at different levels of analysis. The condition random RTs induce a misfit at the level of observations, and the f' test is designed to detect misfits at this level. The speededness condition implies a violation at the level of students, since students were assumed to work with a constant speed level. Thus, the f' test can only pick-up the implied residual deviations due to a change in working speed at level 2, and this decreased the performance of the test.

Figure 2.2. The ROC curve of the l' test for simulated data (1,000 persons, 10 items) with 10% aberrant RTs according to degrees of random response times (left subplot) and speededness (right subplot).

2.8 Real Data Example

The data of Wise, Pastor, Kong (2009) was investigated using the l' person-fit statistic. The data set included 329 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. Wise et al. tried to identify item and examinee characteristics to identify rapid guessing behavior of test takers with motivation problems. van der Linden (2009a) investigated the data for a possible collusion between RT patterns of test takers. However, the main purpose of the current study is to investigate the extremeness of RT patterns under the general log-normal RT model using the proposed person-fit statistic. This example will illustrate the ease

of computing person-fit statistics given RT patterns and further relevant quantities.

First, the log-normal RT model was fitted using 15,000 MCMC iterations and 5,000 iterations as the burn-in. The average time intensity was around 2.70 on a logarithmic scale (around 14.88 seconds), with a posterior standard deviation of .08. The variability in time intensities was around .39. The variability in test takers' working speed was around .08, where the average level of working speed in the population was fixed at zero. So, most of the variability between RTs was explained by the differences in time intensities.

For each test taker, a person-fit statistic value of l' and posterior probability of the extremeness of the RT pattern was computed. In Figure 2.3, the estimated statistic values (x-axis) are plotted against the posterior probability of significance. The statistic values were assumed to be chi-square distributed with 65 degrees of freedom, which marks the point of a observing statistic value with 50% probability. When considering a significance level of .05, estimated statistic values higher than 84.82 were located in the critical region. The estimated number of aberrant patterns was around 20%, which means that one out of five patterns were flagged as aberrant, and when including variable time discriminations around 34% was identified as aberrant. The students in the test had no stakes whatsoever in the test and were not motivated to give their best effort. Wise et al. estimated the proportion of rapid-guessing to be around 10%. Around 25% of the students showed rapid-guessing behavior on more than 10% of the items. However, only 7% of the rapid-guessers were also flagged by the l' test. Our test accounts for variable working speed and item characteristics, where the diagnostic of Wise et al. is based on a known item time threshold to identify rapid guessing. Furthermore, other types of aberrant response behavior might be responsible for the 20% RT patterns that were flagged in the current study.

This study stresses the importance to identify non-effortful responses, which would otherwise undermine the success of low-stakes achievements tests. A good test is of little value when students are not willing to cooperate and to put effort in their work. Therefore, it is important to have a person-fit test for RTs that can be used to check patterns and to identify aberrant response behavior.

Figure 2.3. NAW-8 test; Estimated statistic values and corresponding posterior significance levels.

2.9 Discussion

The response behavior of test takers needs to be checked in order to assess the quality of tests. Aberrant response behavior will bias the test results, represented by biased parameter estimates and incorrect statistical inferences. RT patterns can be checked by evaluating the residuals given a model that explains the variability of patterns of a population of regular test takers. As an analogue to the likelihood-based statistic in person-fit testing to evaluate response patterns, usually denoted as l_z , a likelihood-based person-fit statistic for RT patterns was proposed, denoted as l' . In total, three versions of this statistic were considered: l'_z and l'_s have approximately normal sampling distributions, and l' has an exact chi-squared distribution.

RT checks are meant to identify aberrant patterns, which can appear for several reasons. The proposed checks can be used to flag patterns, and adjustments can be made to flag items as well. Further investigations are required to analyze flagged patterns more thoroughly using possibly additional information. Other types of residual checks can be defined. For example, statistics based on residuals can be used to investigate RT differences between groups of test takers. Item-specific between-group differences in RTs can indicate differential item functioning; that is, an item's time intensity differs across groups. Between-group differences in RTs can also indicate group-specific distributions of working speed.

More research is needed to include response information in the detection of aberrant response behavior. The connection of RT patterns with patterns of accuracy (correct/incorrect) will certainly increase the power of detecting aberrant behavior (van der Linden & Guo, 2008).

CHAPTER 3

Modeling Differential Working Speed in Educational Testing³

3.1 Introduction

Observed response times on a test are assumed to be indicators of an underlying latent variable that represents the test taker's working speed. Individual working speed in educational assessments is commonly modeled as fixed throughout the entire set of items. Van der Linden (2006; 2007) introduced a model where a constant working speed is assumed, and brought that into relation with a certain accuracy level. More recent work in response time modeling also considered a time-invariant working speed variable for each person (Molenaar, Tuerlinckx, & van der Maas, 2015a; Ranger & Kuhn, 2012b, 2014; Wang, Chang, & Douglas, 2013).

In practice, the assumption of constant working speed might be too restrictive. Several factors during a test can influence the behavior of a student and influence his/her performance and working speed. Time mismanagement can lead to an increase in working speed to finish the test in time. A straightforward example of a violation of constant speed is speededness toward the end of a power test, where a time limit affects the behavior of the test takers, and most often in different ways (Evans & Reilly, 1973; Lawrence, 1993; van der Linden, 2011). However, when more test time is available than expected at the beginning of the test, a student might decrease her/his working speed to improve her/his accuracy level. It is also possible that a person's motivation changes during the test leading to a decrease in speed. Students may show decreasing speed over time due to fatigue (Ackerman & Kanfer, 2009), such that the average speed level is merely a rough representation of the actual working speed during the test. When the test consists of different blocks of items, a student might show different working speed over blocks since s/he is familiar with the content of several item blocks and not with the others.

Usually a random sample of subjects from a population is considered, and the speed parameter is modeled as a random person parameter.

³ Fox, J.-P., Marianti, S., & Klein Entink, R. H. (2015). Differential Working Speed in Responses Time Modeling. Manuscript submitted for publication.

Persons can differ in their speed values but these are characterized as between-person differences and each person is assumed to operate at a constant speed throughout the test. To account for subject-specific changes in working speed the log-normal speed model of van der Linden (2006) is generalized. The working speed process of each subject will be modeled by introducing a dynamic factor component in the log-normal speed model. Following the seminal work of Molenaar (1985) and West (1993) on dynamic factor models, the observation model for the response times is extended with a dynamic component or evolution model that describes the change in working speed over the test.

Blocks or clusters of items will be considered, possibly defined by their content or the design of the test, where for example questions within an item block relate to a reading passage or assess a comprehension process (e.g., Rijmen, 2011). Subsequently, the variability in working speed is modeled over blocks of items assuming an underlying average working speed for the entire test.

A general hierarchical dynamic modeling structure is defined, where the average speed level underlies the block-specific speed levels. The test taker's speed levels are serially correlated over item blocks through error components that depend on the difference between the average speed and the working speed of the former blocks of items. This dynamic modeling of working speed can include a moving average component, defined at the item-block level, such that the working speed can vary over blocks of items.

In the proposed modeling framework, stationary and nonstationary speed models are considered, where a stationary speed process has an invariant mean and variance over item blocks. For the stationary speed model, the average (invariant) speed can be related to ability, since they are both assumed to be constant during the test. A speed-accuracy trade-off can be assumed, where a test taker chooses to work at an average speed level, which corresponds to a certain accuracy level. A joint model is proposed for responses and response times, where average speed is related to ability, while accounting for differential working speed using a moving average component.

A mixture modeling approach can be used to identify test takers following a stationary speed process, where the mean and variance do not change over time, and those following a nonstationary speed process, with parameters changing over time. Response time patterns considered to be generated under a nonstationary working speed process can indicate

aberrant response behavior (e.g., item pre-knowledge, guessing, speededness).

Besides the variability in working speed over the test, heterogeneity in working speed across test takers is also considered. Common working speed models allow for variation in working speed across individuals and groups of individuals when covariates are available (Klein Entink, Fox, & van der Linden, 2009). In the proposed modeling framework, a mixture modeling approach is considered to differentiate latent groups of respondents with different mean levels of working speed. This mixture response time model allows for various latent classes with different mean speed levels.

In the following section, the dynamic factor speed models are introduced. A discussion of estimation and identification of the models in a Bayesian framework is then presented. In a simulation study, it is shown that the parameters of the model can be recovered. Two empirical examples illustrate the application and applicability of the dynamic speed models for response times and for responses and response times. Finally, conclusions and suggestions for further research are proposed.

3.2 Modeling Variable Speed

In this section the model is developed, starting from a log-normal measurement model for working speed. It will be shown how mixture models can be applied to model subpopulations with different speed levels. Subsequently, a dynamic factor model is introduced to model changes in working speed during the test. Finally, the dynamic speed model is integrated into a larger framework for the joint modeling of ability and variable speed.

3.2.1 Measurement Model for Speed

In the log-normal working speed model of van der Linden (2006), the logarithm of the response time of subject i to item k , denoted as T_{ik} , is assumed to be normally distributed,

$$T_{ik} = \lambda_k - \zeta_i + e_{ik}, \quad (3.1)$$

Where $e_{ik} \sim N(0, \tau_k^2)$. The time intensity of item k is denoted by λ_k and ζ_i denotes the speed parameter for subject i . The continuous response-time observations allow item-specific error variances denoted as τ_k^2 . The working speed parameter is defined as a random effect such that each test taker is assumed to work at a specific level of speed. This model specification is represented by the population distribution of speed,

$$\zeta_i \sim N(\mu, \sigma^2), \quad (3.2)$$

where μ and σ^2 are the average working speed of the population and the variability between working speed levels within the population, respectively.

3.2.2 Group Differences in Working Speed

The population distribution of speed defined in Equation (3.2) indicates that test takers are working independently of one another with respect to working speed. However, test takers who followed similar training programs or used similar answering strategies may show similar behavior in their working speed. A latent class population distribution is defined to account for the grouping of test takers and to model the correlation between test takers assigned to the same group. The generalization to a mixture response-time model can handle latent groups of respondents that differ in their mean speed levels. Assume that there are Q latent subpopulations that differ in average working speed. The mixture population model for speed is defined as

$$\zeta_i \sim \sum_q \pi_q p(\zeta_i | \mu_q, \sigma_q^2), \quad (3.3)$$

where index q defines the subpopulation with mean μ_q and variance σ_q^2 .

The parameter π defines the mixture proportions.

3.2.3 Dynamic Factor Modeling of Response Times

For each test taker, the pattern of observed response times can be treated as a sequence of realizations. A discrete stochastic process can be defined for the sequence of response times. This stochastic process is discrete, since observations can only vary across the discrete number of test items. Instead of an underlying subject-specific speed parameter, a dynamic process is defined such that the working speed can vary across items.

Let $\eta_i(t), t \in (t_1, \dots, t_n)$ denote the stochastic speed process for person i . The state of the random variable $\eta_i(t)$ represents the level of working speed at time point t . Given the order in which the items are solved, the item response times are informative about the speed levels. The random variable $\eta_i(t)$ represents the stochastic process and is treated as a factor variable. Each subset of consecutive items represents a time frame, and the speed factor can be measured within each time frame.

This stochastic speed process is illustrated in Figure 3.1 for four items. The observed times $T_1, T_1 + T_2, \dots$, represent the real testing time, when test taker i solves item 1 to 4. The item response times $(T_1, \dots, T_4)$ are modeled and used to measure the levels of working speed. A stochastic error

component, ε , influences the measurement of working speed. In this case, each consecutive subset of two items defines a time frame. A state of the variable $\eta_i(t)$ is measured in each time frame using the two observed response times. In each time frame, the response times are assumed to be independently distributed given the level of working speed. The working speed level of the first time frame is assumed to influence the working speed of the subsequent time frame. This is illustrated by the arrow pointing from $\eta_i(t_1)$ to $\eta_i(t_2)$. Finally, an average level of working speed across time frames, ζ_i , relates to the time-point specific working speed measurements. Without the introduced dynamic speed process, represented by $\eta_i(t_1)$ and $\eta_i(t_2)$, the response times can be assumed to be independently distributed given the constant speed level ζ_i as defined in Equation (3.1).

A subset of items will also be referred to as a cluster or block, and the $\eta_i(t)$ will also be denoted as η_{ic} for clusters $c=1,...,C$. A dynamic factor model will be introduced to model each individual trajectory (or evolution process) defined by the stochastic speed process. Note that it is impossible to estimate a state of the stochastic process with one item, since more items are required to estimate a factor-variable level.

Figure 3.1. A dynamic factor model for the stochastic speed process given four items.

So, the set of items is partitioned into $c=1,...,C$ blocks. For each group of items, a cluster-specific speed parameter defines the average working speed in the time frame defined by the set of items. Large clusters are

very specific about each average speed level, but large clusters lead to fewer "realizations" of the stochastic speed process, when considering the entire test. In the dynamic factor modeling framework, the uncertainty of each speed factor level is taken into account in estimating the process parameters. So a smaller number of clusters of a greater number of items within each cluster reduces the uncertainty in the trajectory estimation, whereas a greater number of clusters of a smaller number of items within each cluster increases the uncertainty but leads to more detailed trajectories.

In this modeling framework, the items assigned to the same block are also consecutively placed in the test. Furthermore, the order in which the items are solved is assumed to be known. The working speed process applies to the clusters of items in the order they were solved.

When considering a stationary speed model for test taker i , the average level of speed represents the expected working speed used to solve the item. However, within each block of items, the speed level can be differentiated from the average level. With the η_{ic} representing the working block speed of test taker i in block c and ζ_i the average speed level, the observational model for the response times in block c is represented by

$$T_{ik_c} = \lambda_{k_c} - \eta_{ic} + e_{ik_c}, \quad (3.4)$$

where k_c denotes item k in block c . The item-specific errors e_{ik_c} are independently normally distributed with a mean of zero and a variance of σ_k^2 . At Level 2, the dynamic process that governs the change in working speed over blocks of items is defined.

Consider a moving average (MA) process for speed; then the working speed in block c of a test taker depends on former speed levels through a dependence on the centered speed levels in the former blocks. Let the speed level in block c depend on past speed levels $\eta_{i1}, \dots, \eta_{i(c-1)}$ as follows:

$$\begin{aligned} \eta_{ic} &= \zeta_i + \sum_{j=1}^{c-1} \rho_j (\eta_{ij} - \zeta_i) + r_{ic} \\ &= \zeta_i + \sum_{j=1}^{c-1} \rho_j r_{ij} + r_{ic}, \end{aligned} \quad (3.5)$$

The serial correlation between speed levels over blocks is modeled through a moving-average structure, represented in a mean-adjusted form, where ρ_j represents the transition parameter. The within-block errors, represented by r_{ic} , are assumed to be normally distributed with a

mean of zero and a variance of σ_r^2 . The average working speed of test takers is assumed to be normally distributed,

$$\zeta_i = \mu_\zeta + \nu_i, \quad (3.6)$$

where errors ν_i are independently and normally distributed with a mean of zero and a variance of σ_ζ^2 .

The dynamic response time model is defined at three different levels. At Level 1, the response time observations are related to the time intensities and subject-specific block-speed parameters. At Level 2, the block-speed parameters are modeled via a dynamic MA process, where the block-invariant part of the mean term, ζ_i represents the subject-specific average working speed. This dynamic component, at Level 2, describes the evolution equation or the time trend. At Level 3, the average working speed of the test takers are independently normally distributed.

Under the measurement model defined in Equation (3.4), with an evolution equation defined in Equation (3.5), the sequence of response times follows a stationary speed process under certain conditions. This speed process is stationary when the expected response times and covariances are only influenced by the average working speed and the time intensities. In Appendix B, it is demonstrated that the stochastic process can be characterized as speed covariance stationary, by showing that the mean and covariances are not influenced by block-specific speed parameters when the start of the sequence is changed.

The characteristics of the dynamic linear component are well known from the time series literature. From the theory of dynamic linear models, it is possible to extend the model for speed in various directions, for instance with time-varying effects. Moreover, estimation techniques based on sequential updating (known as Kalman filtering equations) can be used to estimate those models based on standard normal theory. West and Harrison (1997) discuss the theory of dynamic linear models from the Bayesian perspective.

3.2.4 Generalized Dynamic Working Speed Modeling

The response time model with a variable speed process can be generalized in different ways. The dynamic model for speed can be made more flexible by accounting for a subject-specific MA process. This is accomplished by defining the transition parameter ρ to be a person-specific parameter. Consider a first-order moving-average evolution equation to model the serial correlation of speed over blocks of items. This evolution equation for the latent variable corresponds to the special case

considered by Dunson (2003) and the quasi-Markov model of Joreskog (1973). When introducing a random transition parameter, the evolution equation represents a MA process of order one, referred to as

$$\begin{aligned} T_{ik_c} &= \lambda_{k_c} - \eta_{ic} + e_{ik_c} \\ \eta_{ic} &= \zeta_i + \rho_i r_{i(c-1)} + r_{ic} \\ \rho_i &= \mu_\rho + \omega_i, \\ \zeta_i &= \mu_\zeta + v_i \end{aligned} \quad (3.7)$$

where μ_ρ represents the average lag-1 dependence and the between-subject variation of lagged residual dependence is represented by σ_ρ^2 . The evolution component is a stationary speed model that allows test takers to vary their speed over blocks of items, described by a person-specific MA process of order one.

3.2.5 Non-stationary Speed Models

Test takers do not have to follow the model for stationary speed. Instead, they can, for example, increase their working speed during a test, or decrease it and work more slowly toward the end of the test. Such behavior can be modeled using an autoregressive (AR) component in the model. Let the AR-coefficient be set to 1 to model nonstationary speed. Subsequently, a mixture distribution is assumed to describe two groups of test takers: (a) the stationary group, with probability π_{i1} and speed modeled by the stationary evolution model given in Equation (3.7); or (b) the nonstationary group, with probability π_{i2} and speed following a nonstationary AR(1) model. Therefore, assume a mixture of the AR and MA processes, both of order one, such that the model is represented by

$$\begin{aligned} T_{ik_c} &= \lambda_{k_c} - \eta_{ic} + e_{ik_c} \\ \eta_{ic} &= \zeta_i + \pi_{i1} \rho_i r_{i(c-1)} + \pi_{i2} \eta_{i(c-1)} + r_{ic} \\ \rho_i &= \mu_\rho + \omega_i, \\ \zeta_i &= \mu_\zeta + v_i \end{aligned} \quad (3.8)$$

where π_{i1} is the probability that a test taker will follow the stationary model. When the AR coefficient is fixed to 1, as in Equation (3.8), the AR structure corresponds to a first-order differencing step. When $\pi_{i2} = 1$, it follows that $E(\eta_{ic} - \eta_{i(c-1)}) = E(\zeta_i + r_{ic}) = \zeta_i$; this resembles either an increasing or a decreasing trend, depending on the sign of ζ_i .

3.2.6 Joint Modeling of Ability and Speed

In Equation (3.2), persons are considered to be sampled from a population, which also implies that persons are assumed to work with a different level of speed. This relates to the commonly assumed within-person relationship between speed and ability. It is to be expected that the accuracy level decreases when a person starts to work faster. This within-person relationship between speed and ability (i.e., speed-accuracy trade-off) cannot be observed when speed and ability are constant. However, the between-person relation between speed and ability can still be studied for the population.

When the dynamic factor model is considered for the response times, an underlying average working speed and block-specific speed parameters are measured. A relationship between average speed and ability can be assumed. Then, the average working speed is assumed to relate to the ability level, while accounting for variable working speed of a person. The relationship between ability and average working speed can be considered to be a between-person relationship. Persons can differ with respect to their average level of working speed, where the persons working fast on average may show more errors in their responses than those working on average more slowly. Consider a two-parameter item response model, where the probability of a correct response is given by

$$P(Y_{ik} = 1 | \theta_i, a_k, b_k) = \Phi(a_k \theta_i - b_k), \quad (3.9)$$

and let θ_i denote the ability of person i , and a_k and b_k the item discrimination and difficulty parameter, respectively.

The hierarchical framework for ability and speed of van der Linden (2007), Klein Entink et al. (2009), and van der Linden and Fox (2015) is considered to model responses and response times. In this framework, the unidimensional IRT model, Equation (3.9), serves as the measurement model for ability, denoted by θ . The lognormal response time-model is used to measure a constant speed parameter. At a second level, hierarchical population models describe the dependencies between the item parameters on the one hand and the person parameters on the other.

When assuming a dynamic speed process according to Equation (3.7) or (3.8), a relationship can be defined between the average speed ζ_i and ability θ_i . Therefore, a multivariate normal population model can be defined,

$$(\theta_i, \zeta_i) \sim MVN(\mu_p, \Sigma_p), \quad (3.10)$$

where $\mu_P = (\mu_\theta, \mu_\zeta)$ and

$$\Sigma_P = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta,\zeta} \\ \sigma_{\theta,\zeta} & \sigma_\zeta^2 \end{bmatrix} \quad (3.11)$$

The covariance structure given by Equation (3.11) represents the population model for ability and speed for the population (Klein Entink et al., 2009; van der Linden, 2007).

When considering an MA(1) model for speed, it can be shown that the expected working speed in each block is equal to ζ_i . For the block-specific speed component, η_{ic} ($c=1, \dots, C$) it follows that

$$E(\eta_{ic} | \zeta_i, \rho_i) = E(\zeta_i + \rho_i r_{i(c-1)} + r_{ic}) = \zeta_i$$

$$Var(\eta_{ic} | \zeta_i, \rho_i) = Var(\zeta_i + \rho_i r_{i(c-1)} + r_{ic}) = \sigma_\rho^2 (1 + \rho_i^2) = \sigma_\rho^2.$$

Thus, speed is allowed to vary over blocks but within each block the expected speed corresponds to the average speed ζ_i , which is assumed to relate to ability. Although persons can work with different speed during the test, the speed-accuracy trade-off can be assumed. A person can choose to work at a certain speed, and according to the (stationary) MA(1) process the expected working speed is constant and relates to ability. When a person increases his/her average working speed, his/her accuracy level is expected to decrease. In conclusion, a person can choose to work at a different speed, while assuming a relationship between the expected speed and ability. In general, for stationary speed models an underlying speed-accuracy trade-off can be assumed for each person. For these models, the expected speed is constant and can be assumed to be related to ability. For non-stationary models this is not possible. In that case, ability cannot be assumed to be constant in relation to a variable speed component, while assuming a speed-accuracy trade-off.

3.3 Estimation

This section discusses the necessary steps to identify the model, the choices for the prior distributions, and the Markov chain Monte Carlo (MCMC) algorithm.

3.3.1 Identification

The model has to be identified on several levels. First, a restriction is placed on the population-average speed parameter, which is restricted to zero, $\mu_\zeta = 0$. A similar restriction can also be placed on the time

intensities, $E(\lambda_k)=0$ to avoid the possibility that a shift in the time intensities is compensated by a similar shift in the person-speed parameters. Second, the mixture model is not identified, which can give rise to label-switching problems in an MCMC algorithm. To solve that, an ordering on the variance components is implied such that $\sigma_2^2 > \sigma_1^2$, which uniquely identifies the mixture. The sum of the mixture proportions is restricted to 1.

Furthermore, there is an identification problem with the first-order MA component. Consider the MA(1) process defined in Equation (3.7) with a random transition parameter ρ_i , an average lag-1 effect μ_ρ and σ_ρ^2 the variability in lag-1 effects over persons. Then, the autocorrelation function $\gamma(h)$, representing the average correlation between speed of persons in block c and block $c+h$ of the MA(1) model is given by:

$$\gamma(h) = \begin{cases} 1 & h=0 \\ \frac{\mu_\rho}{1+\mu_\rho^2} & h=1 \\ 0 & \text{otherwise,} \end{cases} \quad (3.12)$$

and thus, the MA(1) model has the same autocorrelation function for coefficient μ_ρ as for coefficient $\frac{1}{\mu_\rho}$. The constraint is imposed that $|\mu_\rho| < 1$, which is referred to as an invertible MA(1). It is also possible to impose that $|\mu_\rho| > 1$, which is referred to as a noninvertible MA(1). A truncated normal prior is defined for μ_ρ , with a mean of zero and variance one, with an upper bound of 1 and a lower bound of -1. In WinBUGS it is more efficient to use a parameter transformation technique to restrict the μ_ρ to the interval $[-1,1]$. Therefore, parameter μ_ρ is defined as $\mu_\rho = 2\rho_2 - 1$, where $\rho_2 \sim \text{Beta}(a,b)$, and hyper-parameters a and b are given flat uniform hyper-priors, $a, b \sim U(0,100)$. This prior is only used for the non-stationary speed model in Equation (3.8). The following priors complete the Bayesian formulation of the models. The mean parameter of a normal distribution is given a vague normal prior such that $\mu_\rho \sim N(0,V)$. The mixture proportion π_{i1} is given a $\text{Beta}(1,1)$ prior, where $\pi_{i1} = 1 - \pi_{i2}$. The time-intensity parameters λ_k follow a normal prior, $\lambda_k \sim N(\mu_\lambda, \sigma_\lambda^2)$.

For the joint model, the mean and variance of the ability scale also needs to be identified. Therefore, the population mean is restricted to zero, $\mu_\theta = 0$, and the product of the item discriminations is restricted to one.

3.3.2 MCMC Estimation

With the specification of the priors, an MCMC algorithm can be defined. For details on sampling the parameters of the basic measurement model given in (3.1), as well as the joint model for ability and speed, the reader is referred to Klein Entink et al. (2009), who describe an efficient Gibbs sampling scheme. For the proposed dynamic speed models, various WinBUGS implementations have been developed. The R-program *cirt* of Fox, Klein Entink, and van der Linden (2007) was modified to estimate the dynamic factor speed model with an MA(1) component with a fixed and with a random transition effect according to Equation (3.7). Appendix C presents the WinBUGS implementation of the two-component mixture, AR(1) and MA(1), as the evolution model for speed, with a log-normal observation model for the response times according to Equation (3.8). In Appendix I, Rstan code (Stan Development Team, 2014) is given of the working speed model represented in Equation (3.7).

3.4 Empirical Illustrations

First, results of simulation studies for parameter recovery are discussed. Subsequently, two empirical examples are given to illustrate the application of the dynamic speed models on real data.

3.4.1 Simulation Study for Parameter Recovery and Sample Size

This section presents the results of a simulation study that had two goals. The first goal was to show that parameter recovery was satisfying. The second goal was to investigate the influence of the chosen block sizes on the parameter estimates. The two goals can be addressed by the more specific question of the necessary size of the number of items, K , and the number of item blocks, C .

Therefore, several simulations were conducted with different sample sizes of N , K and C . Data were simulated under the dynamic speed model with an MA(1) component in the evolution equation and a subject-specific transition parameter; see Equation (3.7). For each condition, 50 data sets were generated and the average parameter estimates of the evolution equation were reported in Table 3.1.

The time intensity parameters were chosen to be equally spaced from 2 to 4.2 with a step size of $(4.2-2)/(K-1)$. Test-taker working speed parameters ($N=300$) and ($N=500$) were simulated from normal distributions, where the population average AR effect equaled $\mu_\rho = .5$ and

the variability in AR effects over test takers was set at $\sigma_{\rho}^2 = .15$. The unexplained variability in working speed in each block of items was set at $\sigma_r^2 = 1$. The population variance of speed was set at $\sigma_{\zeta}^2 = 1$. The developed R-program (a modified version of the R-package *cirt* of Fox et al. (2007)) was used to estimate all model parameters using MCMC. For each replicated data set, a total of 5,000 iterations was made, where the first 1,000 iterations were considered as the burn-in. Trace plots of the model parameters did not show convergence issues.

Table 3.1 presents a selection of the results. It can be seen that, for all conditions, the variance of the within-block errors and the population variance of speed were accurately estimated. When increasing the number of persons from 300 to 500, the accuracy of the population variance of speed increased. However, when decreasing the number of blocks from $C=10$ to $C=6$, the standard deviation increased, since less block measurements per subject were available to estimate the population variance of speed.

The true mean of the random transition effect μ_{ρ} was around .50 and the average estimated values are close to the true values for each condition. For a block size of two ($K=20, C=10$) and three items ($K=30, C=10$), the population variance of the transition parameter was estimated slightly below the true value, where the largest average difference was .04. The block-specific speed and average speed parameters are both latent variables, and the estimates of the latent variables showed shrinkage towards the population averages. As a result, the estimated block-specific errors at level 2, r_{ic} , were also biased due to effects of shrinkage, which led to less variance in the estimated transition effects. For the speed measurements based on two or three item observations, the amount of shrinkage was relatively high but decreased for blocks with more items. It can be seen that for a block size of 10 items ($K=60, C=6$), the estimated random effect variance is even slightly above the true value of .15.

The expected a posteriori (EAP) estimates of the time intensity parameters were close to the simulated values of the time-intensity parameters, and it was concluded that those were recovered well for the different conditions.

A similar simulation study was done for the dynamic speed model with a fixed transition effect (results are not shown here). The setup of the simulation study was the same. For all conditions, the average transition

effect was around the true value of .50 and the other estimates were also close to their true values.

Table 3.1

Simulated and re-estimated parameters of the dynamic speed MA(1) model for 50 data replications for different number of respondents, items, and block sizes.

N	K	C	Random			
			Par.	Sim.	EAP	SD
300	20	10	μ_p	0.50	0.47	0.04
			σ_ρ^2	0.15	0.12	0.01
			σ_ζ^2	1.00	1.00	0.10
			σ_r^2	1.00	1.02	0.03
	30	10	μ_p	0.50	0.48	0.05
			σ_ρ^2	0.15	0.13	0.01
			σ_ζ^2	1.00	1.01	0.10
			σ_r^2	1.00	1.02	0.03
	60	6	μ_p	0.50	0.55	0.08
			σ_ρ^2	0.15	0.16	0.01
			σ_ζ^2	1.00	0.93	0.16
			σ_r^2	1.00	1.08	0.08
500	20	10	μ_p	0.50	0.47	0.04
			σ_ρ^2	0.15	0.11	0.00
			σ_ζ^2	1.00	1.01	0.08
			σ_r^2	1.00	1.02	0.02
	30	10	μ_p	0.50	0.48	0.05
			σ_ρ^2	0.15	0.12	0.00
			σ_ζ^2	1.00	1.00	0.08
			σ_r^2	1.00	1.03	0.02
	60	6	μ_p	0.50	0.53	0.09
			σ_ρ^2	0.15	0.15	0.01
			σ_ζ^2	1.00	0.97	0.18
			σ_r^2	1.00	1.06	0.11

It was concluded that the estimation algorithm performed well. Different sample sizes were considered, and for 300 persons, a two-item block size with a test length of 20 items, accurate results were obtained. The accuracy of the estimation results can be improved by increasing the block sizes and the number of blocks. It was shown that for this moderate sample size, a dynamic MA(1) speed model with a random transition effect can be fitted to explore variable speed behavior of respondents.

With respect to the number of blocks and the number of items per block, the following statements can be made. When increasing the number of blocks, more information is available to estimate the speed trajectory, since more block-specific speed measurements are available. Subsequently, the average transition effect and population variance can be more accurately estimated. However, when decreasing the number of blocks, less information becomes available about the speed trajectory, since only a few measurements of the trajectory will be obtained. Furthermore, the accuracy of the average speed and population variance will also decrease when decreasing the number of blocks. When increasing the number of items per block, the accuracy of the block-specific speed measurements and block-specific error variances will improve. A decrease in the number of items per block will lead to more biased block-specific estimates due to shrinkage.

In Figure 3.2, the estimated trajectories of speed are plotted for two subjects, one represented by dotted lines and another by solid lines. For each subject, two trajectories are plotted one based on 5 blocks (filled circles), each of 6 items, and another based on 10 blocks each of 3 items (open circles). The x-coordinate of each point represents the middle of the corresponding set of clustered items used to compute the block speed. Thus, for the trajectory based on 10 blocks, the first measurement's x-coordinate is 2, which represents the block represented by items 1 to 3. It can be seen that the speed trajectory based on 10 blocks shows more variability, since more measurements are involved compared to the trajectory based on 5 blocks. However, both trajectories of each subject show a similar pattern, where the number of blocks influence the smoothness of the estimated trajectory. For a fixed test length, there is a trade-off between the number of blocks and the number of items per block. More (less) blocks lead to a more (less) detailed trajectory but based on measurements with a higher (smaller) standard deviation.

Figure 3.2. Estimated trajectories of two subjects based on five and ten blocks for a 30-item test given a dynamic speed model with an MA(1) component.

3.4.2 Detecting Aberrant Working Speed Behavior

A mixture population model for the speed parameter was defined to identify extreme speed behavior. Extreme speed behavior was characterized by extremely fast or slow response behavior over items in comparison to other test takers in the population. At the individual level, the speed parameter was modeled by a two-component mixture, where one component had a much higher population variance compared to the other component; that is,

$$\zeta_i \sim (1 - \pi)\phi(\mu_\zeta, v^2) + \pi\phi(\mu_\zeta, Mv^2), \quad (3.13)$$

where μ_ζ is the mean level of speed and M a (large) fixed constant. This mixture speed model is a special case of the mixture population model given in Equation (3.3). The general idea is based on the automated variable selection modeling procedure of George and McCulloch (1993), who used a mixture prior to separate active regressor variables from nonactive regressor variables by defining a mixture prior for the regression coefficients located around zero. One component is defined with a very small variance, such that nonrelevant regressors were assigned regression effects near zero.

In this example, data of van der Maas and Wagenmakers (2005) were used from a 40-item choose-a-move chess test taken by 259 test takers. Since all were competitive chess players, it was assumed that not more than 1% of the test takers would show aberrant speed behavior. Therefore, $\pi = .01$ was chosen a priori, and $M=10$ to allow for wider tails in the extreme-speed group (note that groups with both extremely low speed and extremely high speed are covered by this mixture distribution). The

MCMC algorithm was run for 5,500 iterations, and the first 500 were discarded as the burn-in.

Test takers who responded extremely fast to all items were considered to be outliers in the scale-restricted mixture component characterized by a small variance parameter. Therefore, players showing extreme speed behavior were classified to the less-restricted mixture component characterized by a huge variance, where $M=10$. The majority of the test takers exhibited regular speed behavior and were classified to the subpopulation with a restricted variance component. The outliers were classified to the less-restricted subpopulation model with a large variance component.

In Figure 3.3, the estimated posterior density for speed is represented, which shows the estimated two-component mixture distribution. The mixture model identified six test takers with aberrant speed behavior who had a posterior probability of $\pi_i|T > .05$. Three of them had a classification probability of $\pi_i|T > .99$. and they defined a second subpopulation of speed characterized as extremely fast.

Figure 3.3. The choose-a-move chess example: The empirical mixture population density of Speed

3.4.3 Real Data Analysis: Dynamic Speed Modeling

Responses and response times of 170 items from 723 test takers were used to illustrate the performance of the dynamic factor model. A modified version of the cirt R-program of Fox et al. (2007) was used to estimate the parameters of different dynamic speed models. The WinBugs program in Appendix C was used to estimate the dynamic speed model with an ARMA(1,1) component. The R-program was used to estimate the parameters of the MA(1) component with a fixed and random transition parameter. Joint models for responses and response times with constant speed and an MA(1) component were used to investigate the relationship between speed and ability. The joint models were identified by restricting the population means of ability and speed to zero and by restricting the product of item discriminations to one.

For the MA(1) models, Equation (3.4) and (3.7), the MCMC algorithm (using the modified cirt program) was run for 20,000 iterations and the burnin-period was set at 5,000 iterations. A visual inspection of the chains did not show any convergence problems. For the ARMA-process model, Equation (3.8), estimation was performed using the R2WinBUGS package and the model description given in Appendix C. In WinBUGS, a total of 10,000 iterations was made, and the first 5,000 samples were discarded as the burn-in.

The main goal was to model the dynamic process of working speed during the test, and to investigate differences in working speed processes between test takers. The 170 items were divided into 5 blocks of 34 items, 10 blocks of 17 items, and 17 blocks of 10 items, to measure the speed changes of test takers during the test. The items were consecutively placed in blocks, where the order of classifying the items resembled the order of the items in the test. As a result, it was assumed that working speed possibly changed over the consecutive blocks of items analogue to the order in which they were assessed. Each chosen number of blocks provide information about speed changes over the corresponding clustered set of items. For a small number of blocks, each block contains sufficient items to measure accurately the block-specific speed level. However, the few average block-specific measurements do not provide detailed information about the working speed trajectory, since only a few points are measured for the trajectory. With more block-specific speed measurements a more detailed trajectory estimate can be obtained, but these block-measurements contain more variance since they are based on less items.

The choice of the number of blocks and the classification of items can also be based on contextual information (e.g., the kind of questions, the test structure), assuming that speed within a block does not change much but changes are more likely over blocks. This information was not available. As will be shown below, the defined item blocks gave support to a dynamic speed analysis, where the transition of working speed was explicitly modeled through different evolution models. The different block sizes led to comparable trajectories of working speed.

Model fitting was done progressively, starting with the basic log-normal speed model for response times, as given in Equation (3.1), and assuming a homogenous population of speed according to Equation (3.2). This model served as the reference null model. According to Equation (3.7), the MA(1) evolution model with a fixed transition effect (i.e., $\sigma_p^2 = 0$), referred to as fixed MA(1), was fitted. Subsequently, the evolution model with a random transition effect $\sigma_p^2 > 0$, referred to as random MA(1), was estimated. In the null model, a fixed constant working speed was assumed. Both MA(1) models assumed a stationary speed process, where the working speed process was described through an MA process of order one. For each model, different block sizes were used. The number of items per block was chosen to be 5, 10, and 17, corresponding to 34, 17, and 10 blocks, respectively.

In Table 3.2, the estimated DIC is given for each model and each block size. For the constant speed model, Equation (3.1), the block size is 170 items, since that was the total number of items in the test. The DIC represents the fit of the lognormal speed model, which was compared to lognormal speed models with dynamic speed components. The DIC of the speed model with a fixed MA(1) component showed an improvement over the constant speed model for the different block sizes. For a block size of 5 items, a total of 34 random block-speed parameters were estimated, which partly explained the huge penalty term (pD). This penalty term was reduced when less blocks with more items were used, which also led to a lower DIC value. The DIC favored the random MA(1) model. The random MA(1) model with a block size of 17 has the lowest DIC value. Although the differences in DIC values are not large, it indicates that the test takers did indeed work at variable speed levels.

Table 3.2
Estimated DIC of the constant and dynamic working speed models for different block sizes.

Model		Block Size	$\bar{D}$	D	pD	DIC
Constant MA(1)	Fixed	170	240,552.3	239,859.9	694.15	241,246.5
		5	236,330.2	232,842.7	3,486.26	239,816.5
		10	236,340.1	233,533.2	2,800.58	239,140.7
		17	236,345.0	234,011.0	2,327.40	238,672.4
MA(1)	Random	5	236,295.8	234,359.5	1,847.00	238,142.8
		10	236,269.4	234,366.4	1,756.18	238,025.6
		17	236,234.8	234,360.3	1,683.02	237,917.8

The dynamic speed model with a mixture component, where one component defined the MA(1) process and the other component the AR(1) process, according to Equation (3.8), was estimated. However, WinBUGS could not provide an estimate of the deviance information criterion (DIC) for the ARMA(1,1) model. This evolution model assumed a mixture of the stationary MA(1) and a nonstationary AR(1) model, where students' working speed behavior was classified as either stationary or nonstationary. The estimated classification probabilities showed that all students were assigned to the MA(1) model. It was concluded that none of the response-time patterns of the students showed support for nonstationary speed behavior.

Figure 3.4 shows the density plot of the population-average transition effect of the moving-average component. It can be seen that the effect is small, .034, and not significantly different from zero. The posterior standard deviation of the random transition effect is around .12, and indicates that there is a small variation over test takers in the transition effect. The estimates show that test-takers worked at different speeds, but the differences were small and the test takers show very similar working speed behavior. The items also did not show much variation in the time to be completed. The average time intensity was around 3.99 (54.6 s.) and the variation in time intensities was only .102. This showed that all items required more or less the same amount of time.

Figure 3.4. Density plot of the average transition parameter μ_ρ and the population standard deviation σ_ρ of the MA(1) dynamic speed model based on 17 blocks of 10 items.

In Figure 3.5, the speed processes of subject 130 for four different block sizes are plotted. The average speed of this subject was around -.10, which is represented by the dotted line, referred to as constant and one block represents all items. When considering the estimated speed process based on 34 blocks of 5 items, represented by the straight line, relatively small fluctuations in working speed around the average speed value were estimated. Each estimated block speed based on 5 response times was shrunk towards the average speed, which led to small deviations from the average speed level. When considering the estimated speed trajectory based on 17 blocks of 10 items, the amount of shrinkage was smaller. Subsequently, the estimated speed trajectory shows more variation over the blocks. The speed trajectory of 10 blocks of 17 items shows a similar pattern. It can be concluded that for different block sizes the estimated speed trajectories show a similar pattern, but smaller block

sizes leads to a more smoothly speed trajectory due to effects of shrinkage in the block speed estimates.

Figure 3. 5. Estimated test takers' dynamic working speed trajectories over 34, 17, and 10 blocks under the random MA(1) model.

Finally, two joint models for responses and response times were estimated using the modified *cirt* R-program. Both models were identified by fixing the average population level of speed to zero, and fixing the product of discriminations to one. In Table 3.3, under the label LNIRT the covariance estimates of the person and item parameters are given of the joint model with constant speed. The estimates of the joint model with a random MA(1) component, using 10 blocks of 17 items, are given under the label LNIRT MA(1). For the dynamic speed model, the covariance structure between the average speed and ability was modeled according to Equation (3.10) and (3.11). For both models, it can be seen that the population variance of ability and speed are both small and do not differ much. The correlation between ability and constant speed is around .484. It shows that high-ability students were working faster than low-ability students. This estimated correlation between ability and the average speed is slightly smaller, around .426. Accounting for variable speed behavior led to a small decrease in the correlation between speed and ability.

The estimated average transition effect is .034 and not significantly different from zero, but the variance of the random transition effect is .116 and shows a significant variability in the lag-1 correlation between block speeds. The item covariance structure shows similar estimates under both models. The introduction of the dynamic speed component influenced the relationship between speed and ability, but did not led to different item parameter estimates.

Table 3.3

Covariance and correlation estimates of person and item parameters of the joint model with constant speed and with variable speed using a random transition effect.

Variance Components		LNIRT			LNIRT MA(1)		
		Mean	SD	Cor.	Mean	SD	Cor.
Person Covariance Matrix							
Ability	σ_{θ}^2	0.094	0.006		0.092	0.006	
	$\sigma_{\theta,\zeta}$	0.022	0.002	0.484	0.020	0.002	0.426
Speed	σ_{ζ}^2	0.022	0.001		0.024	0.001	
	μ_{ρ}	-			0.034	0.028	
	σ_{ρ}^2	-			0.116	0.004	
Item Covariance Matrix							
Discrimination	$\sum_{11}$	0.280	0.042		0.283	0.042	
	$\sum_{12}$	-0.063	0.022	-0.242	-0.062	0.022	-0.238
	$\sum_{13}$	0.012	0.014	0.071	0.012	0.014	0.071
Difficulty	$\sum_{22}$	0.242	0.027		0.239	0.027	
	$\sum_{23}$	0.073	0.014	0.462	0.072	0.014	0.461
Time Intensity	$\sum_{33}$	0.103	0.011		0.102	0.011	

3.5 Discussion

Dynamic factor models for modeling working speed have been proposed. The dynamic factor models describe the transition of changes in working speed over blocks of items. The observational component is the common log-normal distribution to measure the block-specific working speed given the response times. A mixture modeling approach was introduced to model unobserved heterogeneity in the working speed process of test takers. This modeling approach supports identifying test takers working under a stationary or a nonstationary process.

The mixture modeling combined with dynamic speed models allows the investigation of groups of test takers who show different speed behavior over a test. In the first empirical example, test takers who showed extreme speed levels compared to the general population could be identified by means of a mixture approach. In a second empirical example, variable working speed trajectories were estimated using a random MA(1) model. The test takers worked at different speeds, but the differences in working speed were small.

In the log-normal speed model of van der Linden (2006), it is assumed that test takers work at fixed speed. In some situations, it can be more likely that speed changes depending on several variables such as a time

limit, the chosen test strategy to solve the items, the context of the items, and the structure of the test. The proposed modeling approach generalizes the fixed-speed modeling approach by modeling the change in speed with a dynamic factor model. The dynamic process of speed is defined over blocks of items, which is explicitly modeled to gain insight into the working speed behavior of test takers. The speed-accuracy trade-off can be assumed, when a stationary speed process is modeled. In that case, changes in speed are allowed but each test taker is assumed to work with an average speed level during the test. This average speed can be related to ability, and working faster on average can be assumed to be negatively related to ability. Therefore, a joint model for responses and response times can be defined, where the average speed relates to ability. In this joint model, working speed can vary over blocks of items to explore changes in speed, while the average speed correlates with ability and the item parameters are also allowed to correlate.

Results of the simulation study showed that for a 20 item test, a total of 10 blocks, with each two items, led to reasonable parameter estimates of the dynamic speed model. So, in practice, the dynamic speed processes can already be investigated using data from tests with 20 items. It was shown that the number of blocks and the number of items per block has an influence on the parameter accuracy. The accuracy of the block-specific estimates can be improved by increasing the number of items per block, and of the population parameters by increasing the number of blocks.

Some caveats to the presented modeling approach should be considered. First, to define the stochastic speed process correctly, the order in which the items were solved needs to be known. Most often, all test takers solve the items in the same order and the items can be consecutively stored in blocks. However, in an adaptive testing setting, this is not feasible. Furthermore, the practical applicability of the dynamic speed models requires a reasonable amount of blocks of items to allow the modeling of variable speed and is, therefore, not very suitable for short tests (e.g., fewer than 20 items).

The dynamic factor model for speed can also be used to evaluate test performance. When test takers are hindered by a time limit, their test performance is often negatively biased. When the test length does not correspond with the fixed total test time, or when time-intensive items are placed at the end of test, the dynamic factor model can be used to explore changes in speed trajectories. Students increasing their working speed at the end of the test often indicate that they feel time pressure and therefore

change their strategy with respect to time management. The estimated trajectories can highlight changes in time management strategies. It will be of interest to relate estimated trajectories to specific test design features, as in rule-based item-generation methods (Geerlings, van der Linden, & Glas, 2013) to make more profound inferences.

CHAPTER 4

Latent Growth Modeling of Working Speed Measurements⁴

4.1 Introduction

Responses to items of an ability test will reveal information about the accuracy of the responses (i.e., the degree of correctness), which are related to ability. With the introduction of computer-based testing, responses and responses times can be collected. The response times to produce the answers will reveal information about the speed of working of the respondent. Traditionally, in psychological research, a speed-accuracy trade-off applies, and fast-working subjects often produce more incorrect responses than subjects who work slower. In educational research, this within-person relationship between speed and ability is also assumed (e.g., van der Linden, 2007). The within-person relationship cannot be studied when speed and ability are assumed to be constant. The between-person relationship between ability and speed has been studied, building on the information that test takers differ in ability and working speed. Various studies report about a negative correlation, estimated at the population, between speed and ability of test takers. Empirical examples of Klein Entink, Fox, and van der Linden (2009) showed that higher-level ability students were tended to work at a slower speed than low-level ability students. Klein Entink, Kuhn, Hornke, and Fox (2009), Roberts and Stankov, (1999), and van der Linden and Fox (2015) also report about a negative correlation between ability and speed in their empirical examples.

In the common log-normal response time model of van der Linden (2006) it is assumed that the working speed of a subject is constant throughout the test. The general item response theory models are based on the principle that a test taker will use her/his cognitive knowledge to respond to the test items. Therefore, the relationship between ability and speed is assumed to be constant for each subject working with a constant speed level.

⁴ Fox, J.-P., & Marianti, S. (2015). Latent Growth Modeling of Speed in Joint Modeling of Responses and Response Times. Manuscript submitted for publication.

The assumption of a constant (latent) speed parameter is in correspondence to the assumption of a constant (latent) ability. However, it is reasonable to assume that a test taker varies her/his working speed during the test. Changes in time management could be required to finish the test in time or a test taker could decide to work slower to improve her/his level of accuracy. The working speed can also vary when test takers show aberrant response behavior (e.g., cheating, guessing) (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014).

Evidence of variable working speed can also be found in psychological testing, where subjects are asked to do different performance tasks. By manipulating the experimental conditions, the subject will change her/his response behavior. For example, the subject will work faster when the time pressure is increased but the level of accuracy might not change. It might also be possible that the level of accuracy decreases due to time pressure, but the speed-accuracy trade-off is different over the levels of the time-pressure condition. For example, Vandekerckhove, Tuerlinckx, and Lee (2011) defined a hierarchical diffusion model for two-choice response times, where the parameters of the response process can vary over persons, items and experimental conditions to model the underlying response process. Assink, van der Lubbe, and Fox (2015) used the hierarchical drift diffusion model to identify tunnel vision (i.e., tendency to focus exclusively on a limited view) due to time pressure. In an experiment, they found an interaction effect of the time pressure condition with the response time but not with the response accuracy.

In educational testing, different joint models for ability and speed assume a constant speed parameter for persons. The hierarchical latent variable modeling of responses and response times (Fox et al., 2007; Klein Entink et al., 2009; van der Linden & Glas, 2010; van der Linden, 2007) and the generalized linear IRT approach (Molenaar, Tuerlinckx, & van der Maas, 2015b) both assume a constant latent working speed parameter for each individual. The constant speed parameter is also assumed in the IRT modeling approach of categorical response times (e.g., DeBoeck & Partchev, 2012; Partchev & DeBoeck, 2012; Ranger & Kuhn, 2012a) and the nonlinear regressions of IRT parameters on response times (Ferrando & Lorenzo-Seva, 2007a, 2007b).

To model non-constant working speed, a latent growth modeling approach is defined for the speed parameter. For each test taker, the within-subject systematic differences in observed response times conditional on the time intensities (i.e., the population average time needed to complete each item) are modeled using latent variable

modeling. An individual speed process is assumed, describing the changes in speed over items. So, individuals can work with different levels of speed during the test. Each individual speed process will be defined using random effects to model correlations between response times of each test taker. A linear (within-person) relationship is defined between the individual response times and the random effects.

Furthermore, the random effects are also used to define differences in the speed process between test takers. This will generalize the common log-normal speed model, where a random intercept is used to define differences in speed across test takers. The latent growth speed process will be a second level of the log-normal speed model.

In latent growth curve analysis, a time scale is needed to model the speed process and to define the individual variation in initial status and growth rate. In the present approach, the order in which the items are solved will define the underlying time scale describing the sequence of observed item response times. Each item functions as a measurement occasion for speed and each pattern of response times are treated as longitudinal response time data with respect to the speed process. The measurement occasions (defining the time scale of the speed process) are defined on a scale from zero to one. The chosen time scale values are arbitrary and only represent the order in which the items are solved and that the observations are made at equidistant time points. The time variable will be defined on this scale, where the first (last) measurement corresponds to responding to the first (last) item.

It will be shown that the latent growth model for working speed can be integrated with an item response theory model for ability. Under the variable speed model, the ability parameter is influenced by the speed process parameters. This generalizes the univariate relationship between ability and a single speed variable within a test, since multiple speed components are involved in this relationship. In this approach, ability will be influenced by a weighted average of the subject-specific speed process parameters.

MCMC will be used for parameter estimation, which enables joint estimation of all model parameters. The developed MCMC method is built on the estimation methods of Klein Entink et al. (2009) and Fox (2010), who developed MCMC schemes for joint models for responses and response times assuming a constant working speed model.

Simulated and real data examples will be given to illustrate the modeling framework. The Amsterdam chess test (ACT; van der Maas and Wagenmakers, 2005) data were used to model variable working speed

using a linear and a quadratic speed component. A direct comparison is made with the hierarchical model of van der Linden (2007) and Klein Entink, Fox, and van der Linden (2009).

4.2 The variable working speed model

Van der Linden (2006, 2007) proposed a log-normal model for the response times using two parameters to describe item and individual variations in response times. An item factor is defined which represents the time intensity of an item, and each time intensity parameter represents the population-average time needed for completing the item. A person parameter is defined which represents the constant working speed as the systematic differences in response times given the time intensities. For example, a test taker works more slowly (faster) than the average level in the population when the differences between response times and time intensities are all positive (negative), since over items more (less) time is needed than the population-average time.

Let T_{ik} denote the response time of person i ($i=1, \dots, N$) on item k ($k=1, \dots, K$). A lognormal response time distribution is assumed, to account for the positively skewed characteristic of response time distributions, which leads to

$$\ln T_{ik} = \lambda_k - \zeta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2). \quad (4.1)$$

The time intensity parameter is represented by λ_k and the common speed parameter by ζ_i . The test takers are assumed to be randomly selected from a population. Therefore, the speed parameter is assumed to follow a normal population distribution

$$\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2). \quad (4.2)$$

In Fox et al. (2007) and Klein Entink et al. (2009), a time-discrimination parameter has been included as a slope parameter for speed. The time-discrimination parameter characterizes the sensitivity of the item for different speed-levels of the test takers. This leads to the following specification of the lognormal speed model,

$$\ln T_{ik} = \lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2). \quad (4.3)$$

From Equation (4.3), it follows that the time-discrimination parameter is also used to model the unexplained heterogeneity between each time-pattern responses. This follows from the fact that the covariance between the response time to item k and l of person i includes the time discriminations, which is given by

$$\begin{aligned}
 \text{cov}(T_{ik}, T_{il}) &= \text{cov}(\lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \lambda_l - \phi_l \zeta_i + \varepsilon_{il}) \\
 &= \text{cov}(\phi_k \zeta_i + \varepsilon_{ik}, \phi_l \zeta_i + \varepsilon_{il}) \\
 &= \text{cov}(\phi_k \zeta_i, \phi_l \zeta_i) \\
 &= \phi_k \text{var}(\zeta_i) \phi_l = \phi_k \sigma_\zeta^2 \phi_l.
 \end{aligned} \tag{4.4}$$

Van der Linden (2015) defined the time-discrimination parameter to be a measurement error variance parameter such that $\sigma_\varepsilon^2 = 1/\phi_k$. In that case, the time discrimination (or error variance parameter) will not influence the covariance between response times.

A time variable, representing the order of the items, is defined to model the speed process over time on a convenient scale. Therefore, let $\mathbf{X}_i = X_{i1}, X_{i2}, \dots, X_{iK}$, where $X_{i1} = 0$. The measurement of speed from the first item observation is defined as the intercept, and subsequent item observations can be used to model change in speed. Let $\mathbf{X}_{(i)} = X_{(i1)}, X_{(i2)}, \dots, X_{(iK)}$ denote the order in which the K items are made by person i . Then, a convenient time scale is defined by $X_{ik} = (X_{(ik)} - 1)/K$.

The times are defined on scale from zero to one, where one is the upperbound representing an infinite number of items. Another property of this time scale is that the items, which are subsequently solved, are placed on equal distances. More formally it means that the time between measurement occasions is equal. This assumption would be violated when the test taker would take a break after finishing an item before moving on to the next question. However, information about the time used between items is usually unknown, since it is generally assumed that a test taker starts with the next item when finishing the former one. This time scale can be improved when more information is available. Note that the scale on which the latent variable working speed is measured is arbitrary. Therefore, the numerical values of the time scale for the speed process only need to address the order in which the items were solved and the assumed equidistant property of the measurements.

4.2.1 The lognormal random linear variable speed model

To introduce the latent growth model for speed, the lognormal response time model is extended with a linear growth term. This model will be in itself not of particular interest, since it is not realistic to assume that persons will accelerate their speed of working in a linear way. However, the linear trend component can be used in combination with higher-order time components to model more complex processes of working speed.

The lognormal response time model with an linear trend for speed can be defined using the time variable $\mathbf{X}$, it follows that

$$\ln T_{ik} = \lambda_k - \phi_k (\zeta_{i0} + \zeta_{i1} X_{ik}) + \varepsilon_{ik}$$

$$\begin{pmatrix} \zeta_{i0} \\ \zeta_{i1} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\zeta_0}^2 & \rho_{\zeta_{01}} \\ \rho_{\zeta_{01}} & \sigma_{\zeta_1}^2 \end{pmatrix} \right). \quad (4.5)$$

The parameter ζ_{i0} represents the value of speed measured with the first item solved, also referred to as the initial value of speed. The parameter ζ_{i1} represents the random slope in speed, which means that persons can differ in their growth rate of speed. Note that both random effects have a mean of zero. This means that the average of time intensities defines the average time to complete the test. Furthermore, the population-average speed trajectory is constant, and shows no changes in speed, since the means of the random effects are zero. So, the population-average trajectory with zero values for the random speed variables represents a constant population-average level of speed throughout the test. Persons can work faster than this average level, which corresponds to a positive initial speed value. Furthermore, persons can show an increasing or decreasing trend in their speed rate, which is represented by a positive or negative growth rate, respectively.

Figure 4.1 represents this (random linear) variable speed model. The bottom scale represents the order in which the items are solved. The upper axis represents the real time scale. For each item, a response time is observed to measure speed, and for each response time observation, an error term represents the measurement error which is involved in measuring speed. The latent speed measurements are modeled using a random intercept, referred to as I , and a random growth rate, referred to as S . The average level of speed I is measured by all item observations, where the growth rate is measured by all item observations excluding the first item. The variances of the growth model variables, I and S , define the variability between persons in their initial speed value and their growth rates. A covariance term is specified between the growth model variables. Persons who worked too slowly at the start of the test might improve their speed to finish the test in time. Persons who started working very fast might decrease their speed (possibly improving their accuracy level) since they have sufficient time to finish the test in time. This corresponds to a negative correlation between the growth model parameters.

Figure 4.1. The lognormal random linear variable speed model.

4.2.2 The lognormal random quadratic variable speed model

As said before, to define a more complex speed process, the linear trend component is extended with a quadratic term. The linear trend can be used to model the speed processes of a test taker who starts to work faster and remains to do so until the end of the test. However, a quadratic term can be used to decelerate or accelerate this linear trend. For example, a positive linear trend for speed can be decelerated by a negative quadratic term.

A random quadratic time component is included to define person-specific growth parameters. Then, each trajectory of working speed is modeled by an intercept, a linear trend, and quadratic time component using individual parameters. The lognormal model with a random quadratic time variable is represented by,

$$\ln T_{ik} = \lambda_k - \phi_k \left(\zeta_{i0} + \zeta_{i1} X_{ik} + \zeta_{i2} X_{ik}^2 \right) + \varepsilon_{ik}$$

$$\begin{pmatrix} \zeta_{i0} \\ \zeta_{i1} \\ \zeta_{i2} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\zeta_0}^2 & \rho_{\zeta_{01}} & \rho_{\zeta_{02}} \\ \rho_{\zeta_{01}} & \sigma_{\zeta_1}^2 & \rho_{\zeta_{12}} \\ \rho_{\zeta_{02}} & \rho_{\zeta_{12}} & \sigma_{\zeta_2}^2 \end{pmatrix} \right). \quad (4.6)$$

In Figure 4.2, a graphical representation of the model is given. The bottom scale represents the order in which the items are responded. The upper scale is the true time scale. The random intercept refers to the initial or average speed level, the linear trend is given by S, and a quadratic time component is given by Q. The random growth components are assumed to be correlated with common covariances across persons, according to

the covariance matrix in Equation (4.6). In this model, the individual speed trajectories are modeled using three random effects, each with a mean of zero such that the average time intensities define the average time to complete the test.

Figure 4.2. The lognormal random quadratic variable speed model.

4.2.3 Joint model for responses and response times

Besides observing response times, let Y_{ik} denote the response of person i ($i = 1, \dots, N$) on item k ($k = 1, \dots, K$). An item response theory model is considered to model the item responses and to measure ability of each test taker. When considering binary response data, a two-parameter normal ogive model with item discrimination parameter a_k and difficulty parameter b_k . Using the underlying latent response formulation, a latent response Z_{ik} is used, which is normally distributed with mean $a_k \theta_i - b_k$ and variance 1, and truncated from below (above) by zero when the response is correct (incorrect). The joint model for responses and response times, allowing for variable speed is given by,

$$Z_{ik} = a_k \theta_i - b_k + \omega_{ik}, \quad \omega_{ik} \sim N(0, 1)$$

$$\ln T_{ik} = \lambda_k - \phi_k \left(\zeta_{i0} + \zeta_{i1} X_{ik} + \zeta_{i2} X_{ik}^2 \right) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2) \quad (4.7)$$

$$\begin{pmatrix} \theta_i \\ \zeta_{i0} \\ \zeta_{i1} \\ \zeta_{i2} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_\theta \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\zeta_0} & \rho_{\theta\zeta_1} & \rho_{\theta\zeta_2} \\ \rho_{\theta\zeta_0} & \sigma_{\zeta_0}^2 & \rho_{\zeta_01} & \rho_{\zeta_02} \\ \rho_{\theta\zeta_1} & \rho_{\zeta_01} & \sigma_{\zeta_1}^2 & \rho_{\zeta_12} \\ \rho_{\theta\zeta_2} & \rho_{\zeta_02} & \rho_{\zeta_12} & \sigma_{\zeta_2}^2 \end{pmatrix} \right).$$

The prior distribution of the person parameters $(\theta_i, \zeta_i) = (\theta_i, \zeta_{0i}, \zeta_{1i}, \zeta_{2i})$ can be given as,

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_\theta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_\zeta \end{pmatrix} \right). \quad (4.8)$$

The relationship between speed and ability is defined by the covariance between ability and the speed components and given by $\Sigma_{\theta\zeta}$. The ability parameter is influenced by the different speed components. This follows directly from the conditional distribution of ability given the speed variables. This distribution is given by,

$$\theta_i | \zeta_i \sim N \left(\mu_\theta + \Sigma_{\theta\zeta} \Sigma_\zeta^{-1} (\zeta_i - \mu_\zeta), \sigma_\theta^2 - \Sigma_{\theta\zeta} \Sigma_\zeta^{-1} \Sigma_{\zeta\theta} \right) \quad (4.9)$$

Ability is influenced by the weighted average of the speed components, where the weights are defined by the covariance matrix $\Sigma_{\theta\zeta}$ times the inverse of the variance of speed components. When test takers do not vary their speed, only the first diagonal component of Σ_ζ will be larger than zero, showing the variability in constant speed values across test takers. The linear trend and quadratic change in speed will be around zero, which leads to negligible influence of the remaining variable speed components on ability. When test takers vary their speed according to the quadratic variable speed model, the diagonal components of Σ_ζ will be larger than zero and, together with the covariance matrix $\Sigma_{\theta\zeta}$, defines the relation with ability. It follows that the constant speed model is generalized by allowing variable speed components to influence ability.

It will depend on the application whether changes in working speed will improve the accuracy of the responses. By measuring changes in working speed and modeling the relationship between speed and ability, it is possible to estimate speed trajectories of students with different levels of ability. High-ability students may have different speed trajectories than low-ability students. The speed trajectories of students may also differ over tests. It will be possible to investigate the effects of time limits on students' speed changes but also to investigate the speed changes of proficient students. However, the benefits of estimating speed trajectories in relation to ability will depend on the application.

4.2.4 Identification

The joint model parameters are not identified, and to obtain identification the scale of ability and speed needs to be defined. When restricting the mean and variance of the ability scale, this scale is

identified. This can be accomplished by restricting the sum of item difficulties and product of discriminations, or by restricting directly the mean and variance of the ability parameter.

Subsequently, for the variable speed model the scale of the latent speed variable needs to be identified. This can also be accomplished by two restrictions. In the present description of the model, the mean of speed parameter is set to zero to identify the mean of the speed scale. The average of the time intensity parameters represents the population-average time needed to finish the test given an average working speed of zero. The variance of the speed scale is identified by fixing directly this variance or by restricting the product of time-discriminations to one. For the joint model, the mean of each person parameters is restricted to zero, and the product of discriminations and time-discriminations are restricted to one. These identification restrictions are also used by Klein Entink et al. (2009) and Fox (2010).

In the variable working speed model, an additional restriction is required since the covariance between speed components is modeled by the time intensity parameters and by the covariance matrix of the speed components, Σ_{ζ} . As mentioned, the time intensity parameters will influence the correlation between the response times which leads to an indeterminacy between the covariance parameters of speed and the time intensity parameters. Therefore, as an additional constraint, the covariance matrix of the speed components is restricted to have zero non-diagonal terms and the covariance between speed components is modeled by the time intensity parameters. When the time intensity parameters are all fixed to one, the covariance matrix of the speed components is a free matrix and no additional restriction is required. The residual errors are assumed to be independently distributed and do not influence the covariance modeling structure. When the ability and speed scale are identified, all higher-level model parameters will also be identified.

4.2.5 Parameter Estimation

The model parameters can be estimated using MCMC. The MCMC algorithms for the joint model with variable speed will follow the algorithms for the constant speed-ability joint models. In Fox (2010), the MCMC steps are fully explained for the so-called (constant speed) RTIRT model. The following sampling steps are required. The MCMC method was implemented in a modified version of the cirt R-program of Fox et al. (2007). At iteration $m=1, \dots, M$,

1. For $k=1, \dots, K$, sample item parameters from $p(\phi_k, \lambda_k, a_k, b_k | \mathbf{z}_k, \mathbf{t}_k, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, using a multivariate normal prior with mean $\boldsymbol{\mu}_I$ and covariance matrix $\boldsymbol{\Sigma}_I$.
2. For $k=1, \dots, K$, sample the residual variance in the log-normal model from $p(\sigma_{\varepsilon_k}^2 | \mathbf{t}_k, \boldsymbol{\zeta}, \lambda_k)$.
3. For $i=1, \dots, N$, sample the ability parameter from $p(\theta_i | \zeta_i, \mu_\theta, \boldsymbol{\Sigma}_{\theta\zeta}, \mathbf{z}_i, \mathbf{t}_i)$.
4. The hyperparameters $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$ are sampled from $p(\boldsymbol{\mu}_I | \phi, \lambda, \mathbf{a}, \mathbf{b}, \boldsymbol{\Sigma})$ and $p(\boldsymbol{\Sigma}_I | \phi, \lambda, \mathbf{a}, \mathbf{b})$.
5. For the constant speed model, the hyperparameter $\boldsymbol{\Sigma}_{\theta\zeta}$ is sampled from $p(\boldsymbol{\Sigma}_{\theta\zeta} | \boldsymbol{\theta}, \boldsymbol{\zeta}, \phi, \lambda, \mathbf{a}, \mathbf{b})$.

For the variable speed model, several additional sampling steps are required. With an identification restriction on the covariance matrix of the person parameters, the sampling of the speed components $\boldsymbol{\zeta}$ and the free parameters of the covariance matrix requires a stepwise approach. The speed components are a priori independently and normally distributed. Each diagonal component of the covariance matrix $\boldsymbol{\Sigma}_\zeta$ is inverse-gamma distributed with an inverse gamma prior with parameters g_1 and g_2 . The conditional distribution of the variance parameter of speed component ζ_j ($j=0, 1, 2$), is given by

$$\sigma_{j\zeta}^2 | \boldsymbol{\mu}_\zeta, \boldsymbol{\zeta}_j, \mathbf{T} \sim IG \left(\sum_i (\zeta_{ji} - \mu_{\zeta j})^2 / 2 + g_1, \frac{N}{2} + g_2 \right), \quad (4.10)$$

and the three variance parameters define the diagonal of the covariance matrix $\boldsymbol{\Sigma}_\zeta$. The speed components are conditionally normally distributed, and it follows that

$$\zeta_i | \theta_i \sim N(\boldsymbol{\mu}_\zeta + \boldsymbol{\Sigma}_{\zeta\theta} \sigma_\theta^{-2} (\theta_i - \mu_\theta), \boldsymbol{\Sigma}_\zeta - \boldsymbol{\Sigma}_{\zeta\theta} \sigma_\theta^{-2} \boldsymbol{\Sigma}_{\theta\zeta}), \quad (4.11)$$

In this conditional distribution, the covariance $\boldsymbol{\Sigma}_{\zeta\theta}$ in the mean term is considered to be a regression parameter. The conditional distribution of this parameter is normal with variance

$$\boldsymbol{\Omega} = E(\boldsymbol{\Sigma}_{\zeta\theta}) = (\sigma_\theta^2 - \boldsymbol{\Sigma}_{\theta\zeta} \boldsymbol{\Sigma}_\zeta^{-1} \boldsymbol{\Sigma}_{\zeta\theta})^{-1} \left(\left(\boldsymbol{\Sigma}_\zeta^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta) \right)^t \boldsymbol{\Sigma}_\zeta^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta) \right) + \boldsymbol{\Sigma}_0^{-1} \quad (4.12)$$

and mean

$$\boldsymbol{\Sigma}_{\zeta\theta} = \boldsymbol{\Omega}^{-1} \left(\boldsymbol{\Sigma}_\zeta^{-1} (\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)^t (\theta_i - \mu_\theta) (\sigma_\theta^2 - \boldsymbol{\Sigma}_{\theta\zeta} \boldsymbol{\Sigma}_\zeta^{-1} \boldsymbol{\Sigma}_{\zeta\theta})^{-1} \right). \quad (4.13)$$

From the conditional distribution of θ_i given ζ_i , the distribution of the variance parameter can be derived. This variance parameter, $\sigma = \sigma_\theta^2 - \Sigma_{\theta\zeta} \Sigma_\zeta^{-1} \Sigma_{\zeta\theta}$, is inverse-gamma distributed with scale parameter

$$SS = \sum_i \left((\theta_i - \mu_\theta) - \Sigma_{\theta\zeta} \Sigma_\zeta^{-1} (\zeta_i - \mu_\zeta) \right)^2 / 2 + g_1 \quad (4.14)$$

and shape parameter $N/2 + g_2$. Subsequently, from the sampled variance parameter, σ , a sampled value of the variance parameter σ_θ^2 can be obtained using the sampled value for Σ_ζ .

Without the identification restriction on the covariance matrix, Σ_ζ , the values of the complete covariance matrix of the person parameters,

$$\Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \Sigma_{\theta\zeta} \\ \Sigma_{\zeta\theta} & \Sigma_\zeta \end{pmatrix}, \quad (4.15)$$

are sampled from an inverse-Wishart distribution with scale matrix

$$\sum_i \begin{pmatrix} \theta_i - \mu_\theta \\ \zeta_i - \mu_\zeta \end{pmatrix} \begin{pmatrix} \theta_i - \mu_\theta \\ \zeta_i - \mu_\zeta \end{pmatrix}^t + g_1 \quad (4.16)$$

and degrees of freedom $N+Q$. The mean of the speed components, μ_ζ is fixed to zero.

4.3 Simulation study

In this simulation study, attention was focused on the speed process, represented by three different response-time models, using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) to estimate the model parameters. In this first study, different latent growth models for working speed are investigated given the item response times and without considering response data. In the second study, the joint model for responses and response times was used to generate the data, and a modified version of the cirt program of Fox et al. (2007) was used to estimate all model parameters.

The speed trajectories of 1,000 persons were simulated across a set of 20 items. Under each simulation setting, twenty response time sets for the 1,000 persons and 20 items were generated, and the final parameter estimates were the average values over all 20 estimates. It was investigated if the empirical means of the estimates (across replications) closely resembled the generated parameters. The standard error of each empirical mean estimate was computed as the estimated posterior standard deviation divided by the square root of 20. A normal distribution was assumed for the errors, representing the difference between the final

mean value and the estimates from each replicated data set. Subsequently, a 95% (frequentist) confidence interval was constructed with the boundary at two times the estimated standard error. It was investigated if the true values were located in these 95% confidence interval. Then, it was concluded that the generated parameters could be recovered.

4.3.1 Simulation Design

Three different variable speed models were evaluated with respect to the parameter recovery capability. The three speed models that are discussed provide insight in variable speed modeling and illustrate the identification issues when working with variable speed models.

Model 1 represents a variable log-normal speed model with a common linear trend component,

$$\ln T_{ik} = \lambda_{\mu} - (\zeta_{i0} + \gamma X_{ik}) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2), \quad (4.17)$$

where γ is the common linear trend in speed. In this model, the time intensity parameters are equal, denoted as λ_{μ} . The time intensities are restricted to be equal to identify a non-zero linear trend in working speed. Otherwise, the change in response times could be explained by item time intensity differences and by working speed changes over items. The WinBugs code for this model is given in Appendix A. In model 2, the linear trend in speed is modeled as a random effect over persons, and the time intensity are also allowed to vary over items,

$$\ln T_{ik} = \lambda_k - (\zeta_{i0} + \zeta_{i1} X_{ik}) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2) \quad (4.18)$$

where the means of the random effects are fixed to zero. By restricting the mean of the random linear trend effect to zero, the time intensities are defined. In Appendix B, the WinBugs code is given for this model. In model 3, a quadratic random time effect is added, which leads to

$$\ln T_{ik} = \lambda_k - (\zeta_{i0} + \zeta_{i1} X_{ik} + \zeta_{i2} X_{ik}^2) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2). \quad (4.19)$$

The random speed components have a mean of zero to identify the time intensities. The WinBugs code of model 3 is given in Appendix C. The R2WinBugs software was used to estimate the variable speed model parameters. Each MCMC run contained 5,000 iterations, where the first 1,000 iterations were discarded in the estimation of the means and standard deviations of the model parameters.

4.3.2 Simulation Results

The estimation results of model 1, model 2, and model 3 are represented in Table 4.1, 4.2, and 4.3, respectively. The true parameter values used to generate the data are given in each table. In Table 4.1, the

twenty measurement error variance and population parameter estimates for model 1 are displayed. The true value of measurement error variance was set to one and the estimated measurement error variances were around 1.001. The estimated standard deviation is around .046 such that the estimated standard error of the reported variance estimates is around .01. It follows that for item 15, the true value of one is just included in the 95% confidence interval. The confidence intervals of the other parameters contain the true values. The common time-intensity parameter and the linear trend component were correctly estimated. The results show a good recovery of the simulated parameter values.

Table 4.1
Simulated and estimated parameter values of Model 1.

Par.	True	Winbugs		Par.	True	Winbugs	
		Mean	SD			Mean	SD
$\sigma_{\varepsilon 1}^2$	1	1.001	0.047	$\sigma_{\varepsilon 11}^2$	1	0.998	0.046
$\sigma_{\varepsilon 2}^2$	1	1.002	0.047	$\sigma_{\varepsilon 12}^2$	1	1.007	0.047
$\sigma_{\varepsilon 3}^2$	1	1.000	0.046	$\sigma_{\varepsilon 13}^2$	1	0.997	0.047
$\sigma_{\varepsilon 4}^2$	1	0.999	0.047	$\sigma_{\varepsilon 14}^2$	1	0.996	0.046
$\sigma_{\varepsilon 5}^2$	1	1.004	0.047	$\sigma_{\varepsilon 15}^2$	1	0.989	0.046
$\sigma_{\varepsilon 6}^2$	1	1.004	0.047	$\sigma_{\varepsilon 16}^2$	1	1.000	0.047
$\sigma_{\varepsilon 7}^2$	1	0.995	0.046	$\sigma_{\varepsilon 17}^2$	1	1.000	0.047
$\sigma_{\varepsilon 8}^2$	1	1.000	0.047	$\sigma_{\varepsilon 18}^2$	1	1.004	0.047
$\sigma_{\varepsilon 9}^2$	1	0.997	0.047	$\sigma_{\varepsilon 19}^2$	1	1.007	0.048
$\sigma_{\varepsilon 10}^2$	1	1.007	0.048	$\sigma_{\varepsilon 20}^2$	1	1.006	0.047
$\sigma_{0\zeta}^2$	0.5	0.500	0.014				
λ_{μ}	0	-0.001	0.020				
γ	0.25	0.256	0.025				

In Table 4.2, the estimates of the time intensity parameter are shown, and it can be seen that the estimated values are close to the true values. Although not shown, the measurement error variances were also correctly recovered. (see Table J.1 in Appendix J.). The 95% confidence interval of the estimated random intercept variance is [.486–.500] and for the estimated random trend variance [.474–.500]. The confidence intervals just include the true variance. The random effects variances were slightly

underestimated estimated, which might be caused by the limited number of replications. It was also apparent that the variability in the time intensities was slightly underestimated, since the 95% confidence interval [.732–.949] does not contain the true value. The random effect variances are related. When the time intensities are more alike, there is less variability possible in speed values.

Table 4.2
Simulated and estimated parameter values of Model 2.

Par.	True	Winbugs		Par.	True	Winbugs	
		Mean	SD			Mean	SD
λ_1	0.438	0.428	0.036	λ_{11}	0.177	0.178	0.036
λ_2	0.036	0.029	0.036	λ_{12}	-0.067	-0.081	0.036
λ_3	-0.157	-0.154	0.035	λ_{13}	-0.125	-0.127	0.037
λ_4	-0.089	-0.099	0.036	λ_{14}	-0.308	-0.305	0.038
λ_5	-0.162	-0.169	0.036	λ_{15}	0.117	0.127	0.037
λ_6	0.215	0.225	0.037	λ_{16}	-0.279	-0.286	0.038
λ_7	-0.106	-0.116	0.036	λ_{17}	-0.060	-0.061	0.037
λ_8	0.279	0.288	0.036	λ_{18}	-0.316	-0.322	0.038
λ_9	0.327	0.327	0.036	λ_{19}	0.045	0.053	0.038
λ_{10}	0.057	0.052	0.036	λ_{20}	-0.024	-0.022	0.038
σ_λ^2	1	0.841	0.243				
$\sigma_{0\zeta}^2$	0.5	0.493	0.015				
$\sigma_{1\zeta}^2$	0.5	0.487	0.030				

In Table 4.3, the quadratic variable speed model estimates are given. Again the time intensities were correctly estimated. The measurement error variance were also correctly estimated (see Table J.2 in Appendix J). The estimates of the three random speed components are given in Table 4.3. The 95% confidence intervals based on normally distributed errors show a slight underestimation of the random intercept variance ([.430–.484]), and slight overestimation of the random trend variance [.339–.402].

Table 4.3
Simulated and estimated parameter values of Model 3.

Par.	True	Winbugs		Par.	True	Winbugs	
		Mean	SD			Mean	SD
λ_1	-0.114	-0.120	0.035	λ_{11}	0.334	0.330	0.037
λ_2	-0.079	-0.095	0.035	λ_{12}	-0.002	-0.012	0.037
λ_3	-0.172	-0.176	0.036	λ_{13}	-0.092	-0.103	0.037
λ_4	0.022	0.015	0.036	λ_{14}	0.552	0.553	0.037
λ_5	0.001	-0.004	0.036	λ_{15}	0.329	0.322	0.037
λ_6	-0.180	-0.185	0.036	λ_{16}	0.100	0.083	0.038
λ_7	-0.378	-0.377	0.036	λ_{17}	-0.113	-0.114	0.039
λ_8	0.050	0.046	0.037	λ_{18}	0.210	0.211	0.039
λ_9	0.134	0.121	0.036	λ_{19}	-0.229	-0.232	0.039
λ_{10}	0.050	0.040	0.036	λ_{20}	-0.422	-0.421	0.039
μ_λ	0	-0.003	0.234				
$\sigma_{0\zeta}^2$	0.5	0.457	0.062				
$\sigma_{1\zeta}^2$	0.3	0.371	0.071				
$\sigma_{2\zeta}^2$	0.5	0.508	0.016				
σ_λ^2	1	1.076	0.396				

For the three models, the time intensity parameters and the measurement error variances were accurately recovered, given the 20 replicated data sets. In model 2 and 3, some of the random effect variances were slightly over or underestimated, when considering 95% confidence intervals. The limited number of replications and items, may have caused that the simulated variability in speed trajectories was not exactly recovered. It must be noted that most estimates were close to the true values, except the empirical confidence intervals of some estimated random effect variances, assuming independent normally distributed errors, show very small deviations.

In Figure 4.3, an illustration is given of the estimated speed trajectories across twenty items of twenty persons who worked faster than the population-average of zero. It followed that some trajectories show an increase and some a decrease in speed over items. Both type of trajectories were accurately recovered from the RT data.

Figure 4.3. Latent speed trajectories over items of those persons with the highest speed realizations. The time axis is defined as the ordered items on a scale from zero to one.

4.4 Modeling Variable Speed in the Amsterdam Chess Test Data

The Amsterdam Chess Test (ACT) data of van der Maas and Wagenmakers (2005) were used to identify variable speed trajectories of 259 subjects, who responded to 40 chess tasks. The chess items were divided over three sections; tactical skill (20 items), positional skill (10 items), and end-game skill (10 items). Each item concerned a chess board situation, and the problem solving task was to select the best possible move. The dichotomous response observations one (correct) and zero (incorrect) as well as the response times were stored. Fox (2010, p. 253) analyzed the data using the RTIRT model to identify items not fitting the data using the joint model of Klein Entink et al. (2009).

The purpose of the present study was to investigate whether subjects worked with variable speed, and what type of speed trajectories could be identified. Furthermore, the complex between-person relationship between ability and random speed components was investigated. In a different approach, Molenaar et al. (2015a) considered a function of ability on speed in their generalized linear model for responses and response times. They found a common curvilinear effect of ability on speed for the end-game items, where high-ability subjects tended to use relatively more time in contrast to low-ability subjects who started to answer faster at the end of the test. In their approach, higher-order interaction terms between ability and speed were used to obtain more insight in the relation between speed

and ability, but the higher-order ability components were assumed to be fixed deterministic components (with a common effect across subjects) and it was assumed that speed did not have an influence on ability.

The RTIRT lognormal (constant) speed model and the lognormal (random quadratic) variable speed model (Equation (4.7)) were used to analyze the data. The MCMC algorithm was ran for 10,000 iterations to estimate all model parameters, where a burn-in period of 1,000 iterations was used. In Figure 4.4, for the variable speed model, trace plots of four (variance and covariance) person parameters are given to show the fast convergence and the stable behavior of the MCMC chains. The MCMC algorithm converged rapidly without specifying informative starting values. The other trace plots showed similar behavior. The R-coda package was used to investigate the chains, and the commonly used convergence diagnostics (e.g., Geweke, Heidelberger and Welch) did not show any issues.

Figure 4.4. Trace plots of the ability and average speed population variance parameters, and the covariance between ability and the slope and quadratic speed components.

The RTIRT with a constant speed factor was fitted to the data. In Figure 4.5, the item parameter estimates are given of the 40-item test. It can be seen that there is sufficient variation in difficulty and item intensity to measure the ability and speed factors accurately at the different levels of the scale. The time-discriminations are higher for the first ten items (they define the tactical skill cluster), which means that responses to those items show more variation between low and fast working subjects. The

time discriminations for the end-game items were less high, indicating less power to discriminate between the working speed of the subjects. Most items discriminate sufficiently between subject's ability levels, where only around five items had a low discriminating value of below 0.5.

Figure 4.5. Item parameter estimates of the 40 chess items. The left plot shows the discrimination (diamond symbol) and difficulty (closed diamond symbol) estimates, where the right plot shows the time discrimination (diamond symbol) and time intensity (closed diamond symbol) estimates.

The covariance estimates of the subject's random factors and the item parameters of the joint model with a constant speed factor are given in Table 4.4. There is substantial between-item variation in difficulty and intensity but less in discrimination and time discrimination. The mean response time residual variance was around .25, and ranged from .15 to .50. The correlation between discrimination and difficulty was around .35, and between time-discrimination and intensity around -.92. This strong negative correlation of -.92 showed that for high time-intensive items the speed factor did not explain much variation in response times, where the speed factor did for low time-intensive items. According to the model, for a time-intensive item an increase in working speed has not much effect on the response time due to the low time-discrimination parameter. The influence on the response time due to a change in speed is much higher for the low time-intensive items which have high time discriminations.

The strong correlation of around .80 between item difficulty and item intensity was also apparent. The difficult items were clearly taking much more time to be solved than the easy items.

Table 4.4
ACT Chess: Covariance components and correlation estimates.

Variance Components		Constant Speed			Variable Speed		
		Mean	SD	Cor.	Mean	SD	Cor.
Person Covariance Matrix							
Ability	σ_{θ}^2	0.317	0.040		0.425	0.139	
	$\rho_{\theta\zeta_0}$	0.105	0.013	0.640	0.115	0.016	0.720
Speed	$\sigma_{\zeta_0}^2$	0.085	0.008		0.06	0.009	
	$\rho_{\theta\zeta_1}$				-0.004	0.014	-0.018
	$\sigma_{\zeta_1}^2$				0.113	0.024	
	$\rho_{\theta\zeta_2}$				-0.014	0.015	-0.086
	$\sigma_{\zeta_2}^2$				0.062	0.015	
Item Covariance Matrix							
Discrimination	Σ_{11}	0.479	0.175		0.533	0.188	
	Σ_{12}	0.252	0.138	0.352	0.289	0.150	0.371
	Σ_{13}	0.030	0.052	0.108	0.019	0.064	0.053
	Σ_{14}	0.067	0.151	0.079	0.147	0.07	0.408
Difficulty	Σ_{22}	1.073	0.259		1.141	0.271	
	Σ_{23}	-0.268	0.085	-0.645	-0.317	0.105	-0.603
	Σ_{24}	1.008	0.272	0.794	0.414	0.111	0.786
Time Discrimination	Σ_{33}	0.161	0.047		0.242	0.076	
	Σ_{34}	-0.450	0.123	-0.915	-0.130	0.047	-0.536
Time Intensity	Σ_{44}	1.503	0.375		0.243	0.058	

For the person parameters there was not much variation in speed levels (around .085), or in ability levels (around .32). Under the constant working speed assumption, the correlation between ability and speed was around .65, which showed that high-ability subjects were also completing the items faster. They were able to identify the solution to the chess problem faster than the low-ability subjects.

This covariance structure holds under the assumption that subjects were working with a constant speed. To investigate variable speed trajectories of subjects, the joint model with the random quadratic variable speed model was also fitted. In Table 4.4, the covariance estimates are given under the label *Variable Speed*. For the covariance between item parameters, it can be seen that the strong correlation between time

discrimination and time intensity diminished to $-.54$. Apparently, the additional speed components explained more variation in the response times, reducing the strong correlation between time discrimination and intensity. The correlation between the item discrimination and time intensity increased to $.41$. The highly discriminating chess items in ability were also the time-intensive items. This relates to the positive correlation between ability and speed. It is likely that the subjects showed different speed behavior in responding to well-discriminating items based on the ability of the student.

The correlation between the average speed-level and ability was around $.72$, which was slightly higher than under the constant speed model. The corresponding 95% highest posterior density (HPD) interval was $[0.637, 0.785]$. A slightly negative correlation of $-.02$ was estimated between ability and the random slope speed component (95% HPD interval equaled $[-0.157, 0.099]$). This means that high-ability subjects were more likely to decrease their speed in a linear way. The correlation between ability and the quadratic speed component was around $-.09$ (95% HPD interval equaled $[-0.266, 0.035]$), which means that high-ability subjects were more likely to show an acceleration in the negative trend in speed. However, both estimated correlation parameters were not significantly different from zero, since zero was included in the 95% HPD intervals.

From the trace plots of the covariance parameters, see Figure 4.4, it can also be seen that the drawn covariance values are not significantly different from zero. Since the correlations with ability were not significantly different from zero, characteristics of the speed trajectories could not be explained by differences in ability.

In Figure 4.6, the estimated person parameter estimates of ability is plotted against the random components of speed. It can be seen that there is positive strong relation between ability and average speed, where the relation between ability and the slope and quadratic speed components is not significantly different from zero. The estimated average speed component was conditionally estimated on the two other random speed components, which accounted for non-constant speed behavior. The positive correlation between the linear and quadratic speed component showed that a more negative (positive) trend in speed was accelerated, leading to an even slower (higher) working speed.

Figure 4.6. Random person parameter estimates; the average speed (ζ_0), slope speed (ζ_1), and the quadratic speed (ζ_2) component plotted against ability (θ). The slope of speed is plotted against the quadratic speed component.

In Figure 4.7, from the total sample of $N=259$ subjects, the fitted item-specific working-speed measurements of twenty high and low ability participants were plotted. The participants started working at different speed, where the high-ability group started to work faster than the low-ability group. Some high-ability participants increased their working speed towards the end of the test, but others decreased their level of working speed around half-way the test. The low-ability participants showed opposite behavior. Most of the low-ability participants increased their working speed half-way the test, where only a few showed a constant decrease in working speed. It is possible that high-ability participants (the 10% highest scoring subjects) were more focused and eager to make all items correct, where low-ability participants (the 10% lowest-scoring subjects) might be less motivated half-way the test, and went more quickly through the other part of the test. Molenaar et al. (2015a) reported about this pattern, based on a higher-order interaction effect between ability and speed. With the quadratic variable speed model, each participant's trajectory of working speed was estimated showing the patterns, while controlling for the correlation with ability. This made it possible to estimate variable working speed behavior of each participant.

Figure 4.7. Fitted latent speed trajectories over items of high and low-ability participants.

4.5 Discussion

Computer-based testing makes it possible to collect item-response time information as well as response information, by simply recording the total time spent on each item and the response. The response times can be used to make more accurate inferences about student's ability (e.g., van der Linden et al., 2010) and item characteristics. Response times can also reveal new information about the test characteristics, student's response behavior and student's ability that would not be identified when only using response information.

The developed latent growth model for working speed can be used to measure variable working speed according to a time scale defined by the order in which the items were responded to. In the present model, random slope of speed and random quadratic speed components were added to model deviations from a constant speed model. The extension to higher-order random effect components can be made. However, this will require a sufficient number of item observations to estimate all model parameters. The higher-order terms can also be included for groups of persons using discrete latent random effects.

This model can be used to measure a more complex relationship between ability and speed, and is a generalization of the constant speed model proposed by van der Linden (2007). When the trajectory of speed includes higher-order components, the relation between working speed and ability can be defined as the weighted correlation between ability and

all the speed components. In that case, the additional random speed components are used to control for non-constant speed to improve the estimation of the relation between speed and ability.

The so-called cross-relation between speed and accuracy was also modeled by Molenaar et al., (2015a), who considered different functions of higher-order ability components on speed. They introduced two person factors, ability and speed. In the proposed model, several random person variables were introduced to better describe this relationship by assuming a variable speed model. The developed MCMC algorithm can handle numerous random effects, since it is a simulation-based estimation procedure.

Several model extensions can be considered to make this model suitable for multiple group or multiple latent group settings, polytomous or nominal response data. The multiple group modeling approach of Azevedo, Andrade and Fox (2012) might be used to extend the joint modeling of responses and responses times to a multiple group setting. Another interesting extension would be to consider also a latent growth model for ability. This would lead to multivariate latent growth modeling framework for ability and speed, to model changes in the factor variables (speed and ability) over time. Then, changes in speed and ability can be jointly modeled to investigate, for example, changes in the accuracy-speed trade off over time.

CHAPTER 5

Evaluation Tools for Joint Models for Speed and Accuracy⁵

5.1 Introduction

In computer-based testing, responses and responses times can be collected. The response times used to respond provide information about the working speed, where responses provide information about the ability of the test taker. A joint modeling framework for ability and speed can be used to measure both latent variables, while accounting for the complex dependencies between item and person characteristics.

Following the general modeling framework of van der Linden (2007), and Klein Entink, Fox, and van der Linden (2009), among others, a log-normal response time model is used to measure speed, and an item response theory model is used to measure ability. Although model comparison criteria have been proposed (e.g., Klein Entink et al., 2009), model-fit statistics for the evaluation of the fit of the joint model to the data have been sparsely developed and evaluated. Bolsinova and Maris (2015) developed a test for conditional independence between responses and response times given an exponential family model for the responses. Some model-fit tests for item response theory model can be used for the joint model but their performances for the joint model have not been fully tested. Tests to evaluate the fit of response time models (e.g., Marianti et al., 2014) have also not been tested for the joint model.

Within a Bayesian modeling approach, Bayesian significance testing is considered to evaluate the fit of the joint model. The extremeness of each test statistic value can be quantified by computing the posterior probability (p-value) that the test value is greater than a certain threshold given the data. Each computed p-value can be directly interpreted, since it is defined on a natural scale.

The common parametric person-fit tests, based on the conditional log-likelihood of a response pattern (Drasgow, Levine, & Williams, 1985) and of a response time pattern (Mariani et al., 2014), are used for the joint

⁵ Marianti, S., & Fox, J.-P. (2015). Evaluation Tools for Joint Models for Speed and Accuracy. Manuscript submitted for publication.

model to identify aberrant response and response time patterns. The joint distribution of ability and speed is included in the computation of each person-fit test to account for the relation between speed and ability in evaluating the extremeness of the response and response time pattern. The performance of the person-fit statistics is evaluated using simulation studies. This Bayesian significance testing approach is also used to evaluate the fit of items.

The conditional distribution of the residuals is assumed to be normal in the log-normal response time model. The latent residuals given augmented responses (e.g., Albert and Chib, 1993; Fox, 2010) are also assumed to be normally distributed. The empirical conditional distribution of both residuals is compared to the normal distribution using the Kolmogorov-Smirnov test. Within the MCMC algorithm, in each iteration the conditional empirical distribution is compared to the normal distribution and a Bayesian p-value is computed, which quantifies the extremeness of the violation of the normality assumption. Furthermore, the extremeness of each realized residual in the joint model is evaluated through the posterior probability that the realized residual is greater than a certain threshold value.

The tools for the joint model evaluation are also applied to more general joint models for ability and speed. In the first extension, the joint model with a three-parameter normal ogive model for speed is considered to account for guessing behavior. In the second extension, the log-normal response time model is extended to account for variable working speed. The fit statistics and computation of residuals are also used for the evaluation of the extended joint models. For all joint models, MCMC will be used for parameter estimation, which closely connects to the algorithms of Klein Entink et al. (2009) and Fox (2010), which will be extended to handle guessing behavior and variable speed.

After introducing the joint model for ability and speed, person-fit statistics are defined under the log-normal RT (response time) model and IRT (item response theory) model. It will be shown that given all information, response and RT patterns can be identified as aberrant with a specific posterior probability, according to the Bayesian significance test procedure. In a simulation study, the power to detect the aberrancies is investigated by simulating various types of aberrant responses and response times. A real data study is used to illustrate the use of fit statistics for different joint models. Several directions for future research are presented.

5.2 The Joint Modeling Framework

Responses and RTs can be jointly modeled using a hierarchical latent variable model. The working speed and ability are assumed to underlie the RTs and responses, respectively. Fox, Klein Entink, and van der Linden (2007), Klein Entink, Fox et al. (2009), and van der Linden (2007) have developed a Bayesian modeling framework where a lognormal RT model and an IRT model are used to model the level-1 observations. At a higher level, population distributions are defined for the latent variables speed and ability, and for the item parameters in the RT model and the IRT model.

Another joint modeling approach considered the generalized linear model for responses and RTs (Molenaar, Tuerlinckx, & van der Maas, 2015a). This generalized linear approach restricts items to have equal discriminations. However, it is more realistic to assume that the effect of an increase in ability (working speed) can have a different effect over items on the responses (response times). A joint IRT modeling approach for categorical response times has also been considered (e.g., DeBoeck & Partchev, 2012; Partchev & DeBoeck, 2012; Ranger & Kuhn, 2012a). Treating continuous time observations as categorical RTs will always reduce the available amount of information, where the measurement precision can be increased by modeling the continuous RTs.

In the present joint modeling approach, the IRT model for binary responses can be the two-parameter or three-parameter IRT model and the log-normal model includes time-discriminations and item-specific error variances. For the RT model, let RT_{ik} denote the RT of person i ($i = 1, \dots, N$) on item k ($k = 1, \dots, K$). A lognormal RT distribution is considered to account for the positively skewed characteristic of RT distributions. The lognormal distribution for the RT is given by

$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2), \quad (5.1)$$

where the time intensity parameter λ_k represents the average time needed to complete the item (on a logarithmic scale), the speed parameter, ζ_i , represents the working speed of test taker i , and the time discrimination parameter ϕ_k the item-specific effect of working speed on the RT. Fox et al. (2007) and Klein Entink et al. (2009) introduced the time-discrimination parameter as a slope parameter for speed, which models the sensitivity of the item for different speed-levels of the test takers.

This item time discrimination parameter differs from the time discrimination parameter defined by van der Linden (2009a). In his approach, the reciprocal of the standard deviation of the measurement error is defined to be the time discrimination. This also allows for item-specific variances. However, the time discriminations in Equation (5.1) also model co-variances between RTs. Furthermore, the additional error term can model variations in RTs due to stochastic behavior of the test taker. When test takers operate with different speed values, take small pauses during the test, or change their time management, the RTs might show more systematic variation than explained by the structural mean term. The item-specific error component might accommodate for these differences and avoid bias in the parameter estimates.

Besides observing the RT for item k , let Y_{ik} denote the response of person i on item k . For the binary responses, a three-parameter IRT model is considered to describe the responses. The probability of a correct response is given by

$$P(Y_{ik} = 1 | \theta_i, b_k, a_k) = c_k + (1 - c_k) \Phi(a_k \theta_i - b_k) \quad (5.2)$$

where Φ denotes the cumulative normal distribution function, a_k the discrimination parameter, b_k the difficulty parameter, and c_k the guessing parameter. When using the underlying latent response formulation, two latent response variables are defined. An auxiliary variable S_{ik} is defined which is equal to one, when the test taker knows the correct response and equal to zero otherwise (Beguin and Glas, 2001). The latent response variable Z_{ik} is normally distributed with mean $a_k \theta_i - b_k$ and variance 1, and truncated from below by zero when the test taker knows the response ($S_{ik} = 1$), or truncated above by zero when the test taker doesn't know the response ($S_{ik} = 0$). For the two-parameter model, the response Y_{ik} is the indicator of the latent response variable Z_{ik} being positively truncated. The joint model for the latent continuous responses and RTs is given by,

$$\begin{aligned} Z_{ik} &= a_k \theta_i - b_k + e_{ik}, \quad e_{ik} \sim N(0, 1) \\ \ln RT_{ik} &= \lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2). \end{aligned} \quad (5.3)$$

The test takers are assumed to be randomly selected from a population and the ability and speed variable are assumed to have a multivariate normal population distribution

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\zeta} \\ \rho_{\theta\zeta} & \sigma_\zeta^2 \end{pmatrix} \right), \quad (5.4)$$

where the population means μ_θ and μ_ζ represent the population average level of ability and working speed, respectively. The population variances σ_θ^2 and σ_ζ^2 represent the variance in ability and working speed in the population, where $\rho_{\theta\zeta}$ represents the common covariance between ability and speed.

The population distribution of the item characteristics is a multivariate normal, which is given by,

$$\begin{pmatrix} a_k \\ b_k \\ \phi_k \\ \lambda_k \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_a \\ \mu_b \\ \mu_\phi \\ \mu_\lambda \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho_{ab} & \rho_{a\phi} & \rho_{a\lambda} \\ \rho_{ab} & \sigma_b^2 & \rho_{b\phi} & \rho_{b\lambda} \\ \rho_{a\phi} & \rho_{b\phi} & \sigma_\phi^2 & \rho_{\phi\lambda} \\ \rho_{a\lambda} & \rho_{b\lambda} & \rho_{\phi\lambda} & \sigma_\lambda^2 \end{pmatrix} \right), \quad (5.5)$$

where item characteristic means specified and the variances in item characteristics and their covariances. The guessing parameter will be treated as a mixture model parameter and specifies the probability, c_k , that the response is correctly guessed. Then, the $1-c_k$ represents the probability that the response is not guessed correctly and follows an IRT model.

5.3 Person Fit for Speed and Accuracy

When a joint model is used to measure ability and speed, person-fit statistics are required to distinguish test takers with aberrant item response patterns and/or RT patterns from test takers with normal item response patterns and/or RT patterns. Scores on a test can be spuriously high or low due to cheating, guessing or random responding (e.g., Meijer, 1996), which can be directly detected from the item response pattern. However, it is also possible that the aberrant response behavior is manifested in the RT pattern. Subsequently, an inaccurate working speed measurement can lead to a peculiar relationship with the ability measurement. It is also possible that both the item response pattern and the RT pattern indicate aberrant response behavior. In that case the RT simply contributes to the evidence to identify aberrant-responding test takers.

Meijer and Sijtsma (2001) and Karabatsos (2003) give an overview of the large number of person-fit statistics to detect aberrant response patterns. The parametric tests are based on the principle that the fit of responses to a set of items are evaluated under the IRT model. The

popular person-fit statistic of Drasgow et al. (1985) and the standardized version of Levine and Rubin (1979) is based on the log-likelihood of the response pattern. This person-fit statistic is often used in educational measurement and is shown to have power to identify aberrant response patterns (Karabatsos, 2003; Dimitrov and Smith, 2006).

The log-likelihood of the responses is used to evaluate the fit of a response pattern. The conditional person-fit statistic, denoted as l^y , given the model parameters, is based on the two-parameter IRT model for the responses; that is,

$$\begin{aligned} l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) &= \log p(\mathbf{y}_i | \theta_i, \mathbf{a}, \mathbf{b}) = \sum_{k=1}^K \log p(y_{ik} | \theta_i, a_k, b_k) \\ &= \sum_{k=1}^K y_{ik} \log p(y_{ik} | \theta_i, a_k, b_k) + (1 - y_{ik}) \log(1 - p(y_{ik} | \theta_i, a_k, b_k)). \end{aligned} \quad (5.6)$$

The expected value and variance are given by

$$\begin{aligned} E(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i)) &= \sum_{k=1}^K p(y_{ik} = 1 | \theta_i, a_k, b_k) \log p(y_{ik} | \theta_i, a_k, b_k) + \\ &\quad (1 - p(y_{ik} | \theta_i, a_k, b_k)) \log(1 - p(y_{ik} | \theta_i, a_k, b_k)), \end{aligned} \quad (5.7)$$

and

$$Var(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i)) = \sum_{k=1}^K p(y_{ik} = 1 | \theta_i, a_k, b_k) (1 - p(y_{ik} | \theta_i, a_k, b_k)) \log \left(\frac{p(y_{ik} | \theta_i, a_k, b_k)}{1 - p(y_{ik} | \theta_i, a_k, b_k)} \right)^2, \quad (5.8)$$

respectively. The standardized version of this person-fit statistic is given by

$$l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) = \frac{l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) - E(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i))}{\left(Var(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i)) \right)^{\frac{1}{2}}} \quad (5.9)$$

which is assumed to be approximately standard normally distributed. For the joint model, the computation of the person-fit statistic is more complex due to relationships between model parameters. The ability, item discrimination and difficulty parameters are related to the speed, time discrimination and time intensity parameters. This correlation needs to be taken into account in the computation of the person-fit statistic. The MCMC algorithm developed for the joint model supports the computation of expected a posteriori estimates of the model parameters taking into account the full covariance structure between model parameters.

A Bayesian significance test can be defined to compute the extremeness of each response pattern. The person-fit statistic is an increasing function of the likelihood of the responses. Therefore, increasing values of the negative person-fit statistic correspond to misfit.

The negative person-fit statistic is considered in the computation of a critical region. For increasing l_s^y values the posterior probability of misfit is increasing. Then, the posterior probability that the computed person-fit statistic is greater than a certain threshold value is given by,

$$P\left(l_s^y\left(\hat{\theta}_i, \hat{\mathbf{a}}, \hat{\mathbf{b}}; \mathbf{y}_i\right) > C\right) = \Phi\left(l_s^y\left(\hat{\theta}_i, \hat{\mathbf{a}}, \hat{\mathbf{b}}; \mathbf{y}_i\right) > C\right) = p_{ly}. \quad (5.10)$$

A dichotomous classification variable F_i^y can be defined, which equals 1 when the pattern is marked aberrant and zero otherwise, given the population and item parameters,

$$F_i^y = \begin{cases} 1 & \text{if } P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) > C\right) \\ 0 & \text{if } P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) \leq C\right). \end{cases} \quad (5.11)$$

The status of F_i^y can be computed in each MCMC iteration, and the average over MCMC iterations is used as an estimate of the posterior probability of an aberrant response pattern. Therefore, it is possible to identify response patterns for which the person-fit statistic value is greater than the threshold value, with a specific posterior probability. The extremeness of each response pattern can be quantified with a posterior probability given the threshold C .

When the underlying model is the three-parameter model, the person-fit statistic can be adapted to include the probability of a correct response according to Equation (5.2) (e.g., Glas and Meijer, 2003). This would evaluate the fit of an RT pattern, which include guessed responses. However, it is also possible to evaluate the response pattern conditional on the auxiliary data $\mathbf{s}$. Then, the person-fit statistic evaluates the fit of the responses for which the candidate knows the response ($S=1$) or doesn't know the response ($S=0$). The underlying model is again the two-parameter IRT model, and the person-fit statistic can be stated as,

$$p_{ly} = \int \Phi\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{s}_i\right) > C\right) p\left(\mathbf{s}_i | \mathbf{y}, \theta_i, \mathbf{a}, \mathbf{b}\right) d\mathbf{s} \quad (5.12)$$

where the auxiliary data is generated using the MCMC algorithm. So, for the identified two-component three-parameter response model, one component describes the correctly guessed responses and the other component describes the remaining responses. Then, this person-fit statistic can be used to identify aberrant response behavior for these remaining responses, given the correctly guessed responses.

In Marianti et al. (2014), a person-fit statistic is defined for the RT pattern. Instead of considering the likelihood of the response pattern, the likelihood of the RT pattern is considered. Let $RT_{ik}^* = \ln(RT_{ik})$ denote the

logarithm of the RT of test taker i on item k . Given the model specification in Equation (5.1), the likelihood of a response pattern is represented by the product of RTs,

$$\begin{aligned} -2\log p(\mathbf{rt}_i^* | \zeta_i, \lambda, \phi, \sigma^2) &= -2\sum_{k=1}^K \log p(r_{ik}^* | \zeta_i, \lambda_k, \phi_k, \sigma_k^2) \\ &= \sum_{k=1}^K \left(\left(\frac{r_{ik}^* - \mu_{ik}}{\sigma_k} \right)^2 + \log(2\pi\sigma_k^2) \right) \\ &= \sum_{k=1}^K (Z_{ik}^2 + \log(2\pi\sigma_k^2)), \end{aligned} \quad (5.13)$$

where Z_{ik} is standard normally distributed, since it represents the standardized error of the normally distributed logarithm of RT. The sum of standardized errors is an increasing function of the negative log-likelihood of RTs. This error function is used as the likelihood-based person-fit statistic for RTs,

$$l_i^t(\zeta_i, \lambda, \phi, \sigma^2; \mathbf{rt}_i^*) = \sum_{k=1}^K Z_{ik}^2 \quad (5.14)$$

where unusually large statistic values indicate a misfit. The statistic represents a departure of the RTs from expected RTs under the model. The posterior distribution of the statistic can be used to examine whether a pattern of observed RTs is extreme under the model. The distribution of the test statistic is chi-squared with K degrees of freedom given the model parameters. When the critical region is computed in each MCMC iteration, the chi-square distribution of the test statistic can be used to evaluate the extremeness of the RT pattern. Let threshold C define the boundary of a critical region. This critical region contains the set of values for the observed statistic value for which the null hypothesis is rejected. This critical value C can be determined from the chi-squared distribution,

$$P(l_i^t(\mathbf{rt}_i^*) > C) = P(\chi_K^2 > C) = p_{t'}. \quad (5.15)$$

The $p_{t'}$ is the posterior probability that the observed statistic value is larger than C given the RT pattern. Again a dichotomous classification variable can be defined, F_i^t , which equals 1 when the RT pattern is marked aberrant and zero otherwise, given the population and item parameters,

$$F_i^t = \begin{cases} 1 & \text{if } P(l_i^t(\zeta_i, \phi, \lambda; \mathbf{rt}_i) > C) \\ 0 & \text{if } P(l_i^t(\zeta_i, \phi, \lambda; \mathbf{rt}_i) \leq C). \end{cases} \quad (5.16)$$

In the same way, the status of F_i^t can be computed in each MCMC iteration, and the average over MCMC iterations is used as an estimate of

the posterior probability of an aberrant RT pattern. Test takers with aberrant RT and response patterns can be identified with a third classification variable, which equals one when $F_i^t = 1$ and $F_i^y = 1$. In that case,

$$F_i^{t,y} = \begin{cases} 1 & \text{if } P(l'_s(\zeta_i, \phi, \lambda; \mathbf{rt}_i) > C, l'_s(\theta_i, \alpha, \beta; \mathbf{y}_i) > C) \\ 0 & \text{if } 1 - P(l'_s(\zeta_i, \phi, \lambda; \mathbf{rt}_i) > C, l'_s(\theta_i, \alpha, \beta; \mathbf{y}_i) > C). \end{cases} \quad (5.17)$$

In each MCMC iteration, the status of classification variable is evaluated and the average over MCMC iterations is an estimate of the marginal posterior probability of an aberrant test-taker concerning the response and RT pattern. Note that the posterior significance probabilities and the classification probabilities are computed using MCMC, and therefore, accounting for the relation between parameters from the IRT and lognormal model.

5.4 Residual Analysis

Albert and Chib (1995), Johnson and Albert (1999), and Fox (2010) considered Bayesian residuals to evaluate the fit of the model. For the joint model in Equation (5.3), two types of residuals can be considered. The latent residuals e_{ik} defined as the difference between the latent continuous response and the mean can be estimated to identify extreme outliers and the total percentage of extreme outliers per item and per test taker. Following the procedure of Albert and Chib (1995), the expressions of the conditional expected latent residuals is given by

$$E(e_{ik} | Y_{ik} = 0, a_k, b_k, \theta_i) = \frac{-\varphi(b_k - a_k \theta_i)}{\Phi(b_k - a_k \theta_i)} \quad (5.18)$$

and

$$E(e_{ik} | Y_{ik} = 1, a_k, b_k, \theta_i) = \frac{\varphi(b_k - a_k \theta_i)}{\Phi(a_k \theta_i - b_k)}. \quad (5.19)$$

The extremeness of a latent residual is computed by the posterior probability that the latent residual is greater than a specific threshold r . It follows that

$$P(|e_{ik}| > r | Y_{ik} = 0, a_k, b_k, \theta_i) = \frac{\Phi(-r)}{1 - \Phi(a_k \theta_i - b_k)} \quad (5.20)$$

and

$$P(|e_{ik}| > r | Y_{ik} = 1, a_k, b_k, \theta_i) = \frac{\Phi(-r)}{\Phi(a_k \theta_i - b_k)}. \quad (5.21)$$

The log-RT residuals, ε_{ik} , are defined as the difference between the RT and the mean. The extremeness of the RT residual can be expressed as the posterior probability that the residual is greater than a threshold q . In the same way, it follows that,

$$P(|\varepsilon_{ik}| > q | \zeta_i, \lambda_k, \phi_k, r_{ik}^*) = \Phi\left(-q - \frac{\varepsilon_{ik}}{\sigma_k}\right) + 1 - \Phi\left(q - \frac{\varepsilon_{ik}}{\sigma_k}\right), \quad (5.22)$$

where $\varepsilon_{ik} = r_{ik}^* - (\lambda_k - \phi_k \zeta_i)$. The extremeness of the latent residuals and RT residuals are computed as by-products of the MCMC algorithm for the joint model, thereby accounting for relationships between the joint model parameters.

5.5 Evaluating Distributional Assumptions

The fit of the distribution of the RT residuals can be evaluated using the Kolmogorov-Smirnov test (KS test). This is a non-parametric test that is used to compare the empirical distribution of the residuals to the assumed normal distribution. The distance between the distribution of the realized residuals and the normal cumulative distribution function is evaluated in each MCMC iteration. Under the null hypothesis, the RT residuals are assumed to be normally distributed, and a significant statistic value shows evidence for non-normally distributed RT residuals.

The KS test is used as a Bayesian goodness-of-fit-test by computing the marginal posterior probability that each set of item residuals is non-normally distributed. For the computed residuals of item k of n test-takers, the empirical distribution is given by

$$F_n(\varepsilon) = \frac{1}{n} \sum_{i=1}^n I(\varepsilon_{ik} < \varepsilon)(\varepsilon_{ik}) \quad (5.23)$$

where $\varepsilon_{ik} = r_{ik}^* - (\lambda_k - \phi_k \zeta_i)$ and $I(\cdot)$, the indicator function, equals one when $\varepsilon_{ik} < \varepsilon$ and zero otherwise. The KS test statistic is given by,

$$D_n = \sup_{\varepsilon} |F_n(\varepsilon) - \Phi(\varepsilon)|. \quad (5.24)$$

The distribution of D_n is the Kolmogorov distribution and the D_n converges to zero when the residuals are normally distributed. Subsequently, the posterior significance probability is computed in each MCMC iteration,

$$p_{ks}(\lambda_k, \phi_k, \zeta) = P(D_n > c | \mathbf{rt}_k^*, \lambda_k, \phi_k, \zeta), \quad (5.25)$$

where the average significance probability over MCMC iterations is used as an estimate of the marginal posterior probability that the residuals of item k are non-normally distributed.

The KS test could also be applied to evaluate the normality assumption of the latent residuals. However, this application of the KS test does not have

power, since the latent residuals already depend on simulated normally distributed latent continuous responses.

5.6 MCMC Estimation

A Markov chain Monte Carlo (MCMC) algorithm was implemented to estimate the parameters of the joint model. The MCMC procedure is based on the algorithm developed by Fox et al., (2007) and Klein Entink et al., (2009). The MCMC algorithm was extended to include the estimation of the three-parameter normal-ogive model parameters. This extension was based on the simulation of auxiliary data $\mathbf{S}$, and sampling of guessing parameters c_k , where a Beta prior distribution was used for the guessing parameter. The other priors were similar to those of Fox et al., (2007) and Klein Entink et al., (2009). The R-code of the estimation method is translated into an R-program referred to as LNIRT.

5.7 Simulated Data Analysis

5.7.1 Parameter recovery of the joint model with guessing

The performance of the MCMC algorithm to estimate the joint model with a three-parameter IRT model for the responses was evaluated. Therefore, data were simulated according to the joint model, and the MCMC algorithm was used to recover the simulated true parameter values.

Figure 5.1. For each item, the differences between simulated and estimated values of the item parameters are plotted.

Figure 5.1 shows the discrepancy between estimated and true value of each item parameter. The discrepancies are small, generally ranging from -0.077 to 0.045, for all item parameters. It was concluded that there exists a close agreement between estimated values and true values. The estimated posterior standard deviation ranged from 0.022 to 0.127. It shows that the MCMC method, for estimating the joint model with a three-parameter IRT model, was able to recover the true item parameter estimates.

5.7.2 Evaluate performance person-fit tests

Two types of aberrance response patterns were generated. One pattern represented cheating, related to response patterns, and the other random response behavior, related to RT patterns.

Aberrant responses due to cheating were simulated by letting low-ability test takers correctly answer difficult items (Meijer, 1996; St-Onge et al., 2011). In order to determine the performance of the person-fit test in detecting these aberrant patterns, 10 to 20 percent of the low-ability test takers were selected and they had a probability of .75 to answer 10 difficult items correctly in a 20-item test.

RT patterns with random response times were generated to simulate RTs from test takers who do not solve the items but provide copied answers or use a cheat sheet. They might know the average time to complete the test, but don't know the average time to complete each item. Aberrant RTs were generated from a log-normal distribution with the mean equal to the average RTs, and the standard deviation equal to three times the average standard deviation of RTs. The average test time for the aberrant RT patterns was similar to the average test time of non-aberrant RT patterns.

In Table 5.1, the detection rates, averaged over 50 replicated data sets, are given for both person-fit statistics. Aberrant RT (random response behavior) and response (cheating) patterns were generated for 100 test takers (10%) and for 200 test takers (20%) for 10 items. A significance level of .05 was used. The Type-I errors are given under the heading Model Fit: 4.2% and 3.1% of the test takers were detected as aberrant according to I_t and I_s^y , respectively. The Type-I errors were slightly underestimated but will improve when increasing the number of test takers.

In the first condition (*Random Response Behavior and Cheating*), both types of aberrances were simulated. The results show that with 10% aberrant test takers the I_t and I_s^y statistics were able to detect 9.4% and 10% of them, respectively. When 20% of the test takers showed aberrant

responses, the l_t and l_s^y were able to detect 15.6% and 20% of them, respectively.

Table 5.1

Detection rates of person fit tests l_t and l_s^y , in identifying aberrant response and RT patterns for $N=1,000$ and $K=20$.

		l_t statistic		l_s^y statistic	
		Aberrant	Population	Aberrant	Population
Model Fit		-	4.2	-	3.1
<i>Significance level 0.05</i>					
10%	<i>Condition 1</i>	9.4	9.7	10	13.4
	<i>Condition 2</i>	9.4	9.7	-	3.2
	<i>Condition 3</i>	-	3.9	10	13.1
20%	<i>Condition 1</i>	15.6	15.8	20	23.3
	<i>Condition 2</i>	15.6	15.8	-	3.2
	<i>Condition 3</i>	-	4.1	19.9	23.2

Note: Condition 1, random response behavior and cheating; condition 2, random response behavior; condition 3, cheating.

The performance of each person fit statistic was also evaluated under each condition separately. Both the l_t and l_s^y showed less performance when increasing the percentage of aberrant test takers from 10% to 20%. This corresponds to the finding of Karabatsos (2003), who reported that generally, detection rates decreases as the percentage of aberrant patterns increases. Both statistics show good detection rates in identifying aberrant patterns and the performance is similar to the results of the condition with both violations. In the condition with only aberrant patterns due to cheating, the l_t statistic showed detection rates close to the detection rates obtained under the Model Fit condition. In the condition with only aberrant patterns due to random response behavior, the l_s^y statistic showed detection rates close to the Type-I error.

5.7.3 Evaluate performance KS test

The simulated RTs under the joint model were manipulated to simulate positively (right-) skewed RT error distributions. Therefore, for items 1 to 5 of a 20-item test, Gamma distributed noise (with different shape and rate parameters) was added to the simulated log-normally distributed RTs under the model. The positively distributed errors led to positively skewed RT distributions. A total of 50% noise was added to the RTs of item 1, 33% to item 2, 25% to item 3, 20% to item 4, and 17% to item 5 for the 1,000 test takers. The posterior probability of non-lognormally distributed

residuals was computed for each item and each specification of the distribution of added noise. The joint model with time discriminations equal to one was used to simulate the response and RT data. This joint model was also estimated given the responses and RTs with positively skewed distributed noise.

Using Equation (5.25) for the computation of posterior significance probability of non-normally distributed errors, the detection rates of non-normally distributed item residuals was computed for 100 replicated data sets using a significance level of .05, which are given in Table 5.2. The Gamma distributed added noise had a shape parameter of 2, and rate parameters of 1, 1.25, and 1.5. When increasing the rate parameter the average value of the noise reduced given a constant shape parameter.

The detection rates were around 1 for all items when the added noise was distributed with a rate parameter of 1. The added Gamma distributed noise was larger than the normally distributed errors under the model leading to a violation of normality. When adding Gamma distributed noise with an increasing rate parameter the detection rates are decreasing. When the percentage of added noise is smaller than 33%, the detection rates decreased for decreasing percentages of noise added. The item residuals of the remaining items 6 to 20 did not show significant violations of normality, since for these items the logarithm of RTs were generated from normal distributions.

Table 5.2

Performance of the KS test in identifying non-normally distributed item residuals. The reported detection rates are based on one hundred replications using a significance level of .05.

Item	%	Gamma (α, β)		
		$\alpha = 2, \beta = 1$	$\alpha = 2, \beta = 1.25$	$\alpha = 2, \beta = 1.5$
1	50	1	0.91	0.68
2	33.33	1	0.89	0.68
3	25	0.98	0.88	0.58
4	20	0.99	0.76	0.49
5	16.67	0.94	0.64	0.45

5.8 Real Data Analysis

A real data set consisting of responses and RTs of 723 test takers responding to 170 test items were analyzed using the joint model. The LNIRT program was used to estimate all model parameters and to compute the several goodness-of-fit statistics. The joint model was identified by restricting the population means of ability and speed to zero

and by restricting the product of time discriminations and discriminations to one.

The MCMC convergence diagnostics were used to evaluate the convergence of the chains. According to the diagnostics, a burn-in period of 1,000 iterations and a total of 5,000 MCMC iterations were made to estimate the model parameters.

In Figure 5.2, the estimated item parameters are shown of both measurement models. The plotted item difficulties and time intensities in the left subplot were rescaled to have a mean of zero. The estimated mean of the item difficulties was $-.56$ and the item difficulties ranged from -1.71 to $.79$. The estimated mean of the time intensities was around 4.00 and the time intensities ranged from 2.89 to 4.85 . So, the average RT to complete each item ranged from $\exp(2.89) \approx 18s$ to $\exp(4.85) \approx 128s$. It can be seen that the variety in item difficulty is relatively large, which make it possible to estimate accurately test-takers' level of ability on the entire range of the scale. Some of the item discriminations were relatively small (slightly above $.30$). The highly discriminating items were also the most difficult items. The variety in time discriminations is not very high, which ranges from $.4$ to 1.6 on a logarithmic scale. The average population level of speed was fixed to zero.

Figure 5.2. The estimated item parameters; difficulty and time intensity (left subplot) rescaled to have a mean of zero, and item discrimination and time-discrimination (right subplot) .

The person fit statistics to detect aberrant response behavior were computed. In Figure 5.3, the estimated person-fit statistic values, l' , given the RTs are plotted against the posterior probability of significance. The statistic values are chi-square distributed with 170 degrees of freedom, under the joint model. Subsequently, the critical statistic value is 201.4, when the level of significance equals $.05$. Estimated statistic values higher

than 201.4 were located in the critical region. Given this significance level, the estimated number of aberrant patterns was 124, which was 17.15% of the test takers.

Figure 5.3. Estimated person fit statistics l_t with respect to the RT patterns plotted against the corresponding posterior significance probability.

In Figure 5.4, the estimated person fit statistic values l_s^y are plotted against the posterior significance probability (P-value). The critical area is above the statistic value of 1.645, when considering a significance level of .05. Test takers who had a statistic value higher than 1.645, were located in the critical region. In this study, 16 persons (2.213%) were identified in the critical region and hence, were detected as persons with aberrant response patterns.

Figure 5.4. Estimated person fit statistic values l_s^y with respect to response patterns plotted against the corresponding posterior significance probability.

Figure 5.5. Person fit statistic l_t (related to RTs) plotted against l_s^y (related to responses).

In Figure 5.5, the person fit statistic l_s^y (x-axis) is plotted against the person-fit statistic l_t (y-axis). For both statistics the threshold value of the significant area is marked with a dotted line. It can be seen that with respect to aberrant RT patterns, l_t , a serious number of test takers are marked as aberrant since their statistic value is greater than 201.4. A few test takers were marked as aberrant with respect to their response pattern and had a statistic value above 1.645. Test takers who were marked as aberrant with respect to their response and RT pattern are marked with a triangle. Only 6 test takers were marked as aberrant for both patterns. The plotted statistic scores concerning RT and response patterns do not seem to be related.

The KS test was used to check the normality assumption of the RT item residuals. It followed that for 18 of the 170 items (around 11%) the normality assumption was rejected with a significance level of .05, where the p-values ranged from .000 to .042. The extremeness of the residuals was investigated using Equation (5.22), where $q=2$. Around 12.3% of the standardized residuals were considered extreme, which partly explains the non-normality residual distribution of the 18 items. For most of the identified extreme residuals, the observed RT was larger than 200s and ranging to 612s, where 128s was the highest population-average response time. It was concluded that the log-normal distribution failed to

include this relatively high percentage of extreme RTs since the tails of the log-normal distribution were too flat.

The extremeness of the latent residuals were also investigated, using Equation (5.20) and Equation (5.21), where $r=2$. Around 1.9% of the latent response residuals were marked as extreme. These residuals corresponded to incorrect responses of high-ability test takers (with an ability level around 0.5) to easy items (with a difficulty level below -1).

The speed-accuracy trade-off in the population can be investigated by plotting the estimated abilities against speed for each test taker. In Figure 5.6, the relationship between speed and ability for the identified non-aberrant and aberrant test takers is plotted. The aberrant group of 124 test takers was identified according to the l_i statistic with a significance level of .05. It can be seen that both groups show a comparable positive correlation between speed and ability. Three regression lines of ability on speed are represented in Figure 5.6, a dotted line for the aberrant test takers, a dashed line for the non-aberrant test takers, and a straight line for all test takers. It can be seen that the correlation between speed and ability is just slightly smaller for the aberrant test takers. The aberrant RT patterns did not strongly influence the relationship between speed and ability.

Figure 5.6. Estimated level of ability plotted against speed for the identified non-aberrant and aberrant test takers given the statistic values l_i .

In Table 5.3, the covariance and correlation estimates are given of the population parameters of the joint model under the label LNIRT. For all test takers, it can be seen that the estimated correlation between ability and speed, when speed is constant, is around 0.486. The positive correlation indicates that the high-ability test takers worked faster than the low-ability test takers. The variance in speed values over test takers is around .022, which is rather small. The test takers' speed values range from -0.6 to 0.56. Most of the variation between RTs is explained by the differences in time intensities.

There exists a high correlation between item discrimination and time discrimination, and item difficulty and time intensity, around 0.501 and 0.464, respectively. This means that the discriminating items with respect to ability also discriminate well with respect to speed. The positive relation between the item difficulty and time intensity means that the time-intensive items are the more difficult items.

Table 5.3
Covariance and correlation estimates of person and item population parameters of the joint model (LNIRT) and the joint model with variable speed.

Variance Components		LNIRT			Variable Speed		
		Mean	SD	Cor.	Mean	SD	Cor.
Person Covariance Matrix							
Ability	σ_{θ}^2	0.093	0.006		0.119	0.008	
	$\rho_{\theta\zeta_0}$	0.022	0.002	0.486	0.024	0.002	0.505
Speed	$\sigma_{\zeta_0}^2$	0.022	0.001		0.019	0.001	
	$\sigma_{\zeta_1}^2$				0.01	0.001	
	$\sigma_{\zeta_2}^2$				0.01	0.001	
Item Covariance Matrix							
Discrimination	$\sum_{11}$	0.287	0.041		0.264	0.044	
	$\sum_{12}$	-0.062	0.022	-0.236	0.043	0.022	0.176
	$\sum_{13}$	0.091	0.018	0.501	0.091	0.019	0.497
	$\sum_{14}$	0.013	0.014	0.076	-0.007	0.014	-0.042
Difficulty	$\sum_{22}$	0.24	0.027		0.227	0.026	
	$\sum_{23}$	-0.07	0.015	-0.421	-0.026	0.014	-0.153
	$\sum_{24}$	0.073	0.013	0.464	0.071	0.013	0.464
Time Discrimination	$\sum_{33}$	0.115	0.016		0.127	0.018	
	$\sum_{34}$	-0.035	0.01	-0.322	-0.034	0.01	-0.297
Time Intensity	$\sum_{44}$	0.103	0.011		0.103	0.011	

To further investigate the relatively high percentage of aberrant test takers with respect to RT patterns, a variable speed model was fitted. The variable speed model can accommodate for non-constant speed behavior of test takers. This might explain the identified aberrant RT patterns under the joint model assuming a constant speed level.

Let a variable X_{ik} denote the (time) scale representing the order in which the items were made for each test taker. Then, a growth model is defined to model variable working speed of a test taker using a random intercept, a random trend and a random quadratic time component. The joint model with a variable speed component is represented by

$$\begin{aligned} Z_{ik} &= a_k \theta_i - b_k + \omega_{ik}, \quad \omega_{ik} \sim N(0,1) \\ \ln RT_{ik} &= \lambda_k - \phi_k (\zeta_{i0} + \zeta_{i1} X_{ik} + \zeta_{i2} X_{ik}^2) + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_\varepsilon^2). \end{aligned} \quad (5.26)$$

The random person parameters in the joint model are assumed to be multivariate normally distributed,

$$\begin{pmatrix} \theta_i \\ \zeta_{i0} \\ \zeta_{i1} \\ \zeta_{i2} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_\theta \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\zeta_0} & 0 & 0 \\ \rho_{\theta\zeta_0} & \sigma_{\zeta_0}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\zeta_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\zeta_2}^2 \end{pmatrix} \right) \quad (5.27)$$

where a covariance is modelled between ability and the random intercept of speed. The data did not provide support for a covariance between ability and the random linear (trend) and quadratic speed components. Therefore, the remaining random speed components were assumed to be independently distributed.

For each test taker, a variable speed measurement per item can be defined; that is, $\zeta_{ik} = \zeta_{i0} + \zeta_{i1} X_{ik} + \zeta_{i2} X_{ik}^2$ for each i and k . This item-specific measurement of speed can be used to compute the test statistic l' to identify aberrant test takers while accounting for variable working speed.

In Table 5.3, under the heading Variable speed, the covariance and correlation estimates are given of the person and item population parameters. The estimates are slightly different from the constant speed model. It can be seen that the estimated variance of the linear and quadratic speed components are almost zero. Since the mean and variance of the random trend and quadratic component is zero and approximately zero, respectively, it can be concluded that test takers hardly changed their working speed during the test. Most of the variation between RTs is explained by the item differences comparable to the constant working speed model. This also explained that the data did not

support non-zero covariance estimates between ability and the linear and quadratic speed components.

Under the variable speed model, around 136 test takers were identified to be aberrant, where 122 of them were also identified to be aberrant under the constant speed model. It can be concluded that the aberrant RT patterns were not typically from test takers who worked with a variable working speed, since they were also identified to be aberrant under the variable speed model. A closer inspection of the aberrant RT patterns showed that for 121 aberrant test takers on average 2 items (ranging from 1 to 7) had observed RTs of 200s or more, which is much higher than the highest time intensity. So, most of the aberrant RT patterns had a corresponding person-fit statistic value which was considered extreme due to a few high RT observations.

Finally, the KS test was used to evaluate whether the RT item residuals, under the lognormal model with variable speed, were normally distributed. For eleven items the assumption of normality was rejected using a significance level of 0.05. So, the variable speed model partly improved the fit of 7 items, which were rejected to have normally distributed residuals under a constant speed model but not under the variable speed model. The percentage of RT residuals considered to be extreme was still around 12.5%.

5.9 Discussion

The joint modeling of responses and RTs can be used to make inferences about ability and speed given educational test data. The relationship between speed and ability provides information about the test taker and items. It increasingly receives attention due to the increase in computer-based testing. To make correct inferences from the joint model for speed and ability, statistical tests have been developed to evaluate the fit of the joint model.

Bayesian significance testing is used to evaluate person fit for observed RT and response patterns. The developed person fit statistics can be used to identify aberrant test takers with respect to their RT pattern or their response pattern or both patterns. It was shown in a simulation study that under different conditions aberrant test takers were correctly identified. In the present study only two different types of aberrant behavior was simulated. A more comprehensive simulation study is needed to fully investigate the performance of the person-fit tests for joint models.

The residual statistics concerning both patterns can be used to evaluate the fit of the model. The residuals and residual statistics are

easily computed as by-products of the MCMC algorithm. The KS test, based on the realized residuals, showed good performance to identify non-normally distributed RT item residuals. The KS test was used as a Bayesian significance test, where the posterior probability of an extreme KS test statistic was computed. In future research, other residual statistics can be considered to further develop Bayesian significance testing for the joint model. Finally, the developed tests can be easily extended to more general joint models for responses and RTs. It was shown that the fit statistics and computation of residuals can also be used for the evaluation of the joint model with variable speed.

CHAPTER 6

Summary and Discussion

6.1 Summary

In this dissertation, response time modeling, in the context of computerized based testing, has been discussed. Response times reveal information about the working speed of test takers. Taking response times and responses into account results in more information that can be used to improve insights in test taker's behavior during the test. Therefore, several chapters in this dissertation have been dedicated to response times as well as responses.

In Chapter 2, three person-fit statistics referred to as the l'_z , l'_s , and l' are discussed. These test statistics have approximately normal sampling distributions and an exact chi-squared distribution, respectively. Through simulation studies, the performance of the person-fit statistics for RT patterns was evaluated. Different types of aberrant response behavior were considered and the best results were obtained for random response behavior using l' or l'_s , with a detection rate close to one under different conditions.

In Chapter 3, a dynamic factor model for modeling working speed is presented, which describe the transition of changes in working speed over blocks of items. Test takers who work under a stationary or a nonstationary process can be identified by a mixture modeling approach. Results of the simulation study showed that in order to obtain accurate parameter estimates of the dynamic speed model, around 10 blocks of items, each containing at least two items were required.

In Chapter 4, a latent growth modeling approach for working speed is presented, which can be used to measure variable working speed according to a time scale defined by the order in which the items were solved. According to the simulation study, satisfied estimation results were obtained using MCMC. A random slope and a random quadratic speed component were added to model deviations from a constant speed trajectory. More higher-order terms can be included to describe more flexible latent trajectories of speed.

The fit of the distribution of the response time residuals was evaluated using a Kolmogorov-Smirnov test (KS test). Person-fit statistics were proposed to detect aberrant patterns of responses and/or response times. In Chapter 5, these methods have been illustrated for both simulated and

real data. Based on a simulation study, it was shown that the performance of the person-fit statistics worked well in identifying aberrant patterns. The model evaluation tools can be used for a joint model, which also can include a component for guessing behavior and a component for variable working speed behavior.

The log-normal and IRT models in this dissertation have been applied to real data. The MCMC methods for estimating the model parameters were implemented in R and are referred to as: Log-Normal Response Times (LNRT), Log-Normal Response Times Moving Average (LNRTMA), and Log Normal Item Response Theory (LNIRT). The programs are discussed in appendix (G, H).

6.2 Discussion

Computers are often used to administer standardized test. This makes it possible to record test taker's responses and response times. One of the advantages of collecting response and/or response times is the possibility to identify aberrant behavior by evaluating differences between expected and observed data. The two different sets of data, response times and responses, which are often observed on continuous and ordinal scales respectively, are a potential source of information which can reveal information about the item characteristics as well.

Aberrant patterns can be identified by person fit statistics, given result from the simulation studies. However, in realistic testing situations, the test statistics don't reveal any information about the type of aberrance detected.

In a time limited test, response times can be used for identification of strategies adopted by test takers during the test. For example, when running out of time, test takers might change their current strategy to work faster (Schnipke & Scrams, 2002). Another scenario can be, that a test taker changes his working speed due to fatigue.

Several models of working speed have been discussed in this dissertation in order to describe how test takers complete an item with respect to time. These models include the constant working speed model, the stationary and non-stationary speed over blocks of items and latent growth model for speed.

The complete data of responses and response times have also been analyzed with a joint model to measure ability and working speed, while accounting for the relationships between item and person characteristics.

When response times are available, they can be used for diagnostic purposes to improve the accuracy of inferences. Response times can also reveal new information about the test characteristics, and test taker's

characteristics. Observed availability of response time information can lead to a better understanding of item characteristics, and test takers response behavior.

In practice, labelling a test taker as aberrant is a serious judgement that can affect further decisions. The statistical outcomes with respect to response and/or response times should not be the only source of evidence. Additional information, such as observations made during the test, is required in order to support the judgments made about the test takers' behavior.

6.3 Future Research

The developed methods presented in this dissertation can be applied in educational testing. They can be used as tools in order to create improved inferences about competencies of test takers. However, several model extensions can be considered. For example, variable speed modeling can be applied to non-cognitive test data. Consideration of both of the observations, responses and response times, in non-cognitive tests has been done by Meng et al. (2014), who proposed a general model for responses and response times in personality tests while assuming the speed constant for the entire test. It might be interesting if speed is a constant factor included in the model for personality tests.

When an item's time intensity differs across groups, it can indicate differential item functioning (DIF). Detecting DIF of time intensity parameters can be of interest to improve the measurement of speed.

The lognormal model can be used to describe response times and demonstrates in general a good model fit (Schnipke and Scrams, 1997; Van Zandt, 2000; van der Linden, 2006). However, different modeling approaches, i.e. Gamma and Weibull distributions (Van Zandt, 2000), for describing the response times can also be considered to improve the model fit when the lognormal distribution fails to give a good description of the data.

In this dissertation, it has already been discussed that the working speed might change over items. It can also be interesting extension to consider a latent growth model for ability. This would lead to multivariate latent growth modeling framework for ability and speed to model changes in speed and ability over items.

References

- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology*, 15(2), 163–181.
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422), 669-679.
- Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82(4), 747-769.
- Assink, N., Van der Lubbe, R. H., & Fox, J.-P. (2015). Does time pressure induce tunnel vision? An examination with the eriksen flanker task by applying the hierarchical drift diffusion model. In Y. Wang, P. Borne, & I. Rudas (Eds.), *New Developments in Computational Intelligence and Computer Science* (pp. 30-40). Vienna, Austria.
- Azevedo, C. L., Andrade, D. F., & Fox, J.-P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational statistics & data analysis*, 56(12), 4399-4412.
- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4), 541-561.
- Bolsinova, M. & Maris, G. (2015). A test for conditional independence between response time and accuracy. *British Journal for Mathematical and Statistical Psychology*, doi: 10.1111/bmsp.12059.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39, 331–348.
- Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4), 383–430.
- Box, G. E., Hunter, J. S., & Hunter, W. G. (1978). *Statistics for experimenters*. New York: Wiley.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48(1), 1-28.

- Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied measurement*, 7(2), 170.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67-86.
- Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, 98, 555–563.
- Evans, F., & Reilly, R. (1973). A study of speededness as a source of test bias. *Journal of Educational Measurement*, 9, 123–131.
- Ferrando, P. J., & Lorenzo-Seva, U. (2007a). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, 31(6), 525-543.
- Ferrando, P. J., & Lorenzo-Seva, U. (2007b). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research*, 42(4), 675-706.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.
- Fox, J.-P., Klein Entink, R., & Linden, W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, 20(7), 1–14.
- Geerlings, H., van der Linden, W. J., & Glas, C. A. W. (2013). *Applied Psychological Measurement*, 37, 140–161.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 8, 881–889.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27(3), 217-233.
- Johnson, V. and Albert, J. (1999). *Ordinal Data Modeling*. New York: Springer Verlag.

- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298.
- Klein Entink, R. H., & Herman, R. (2009). *Statistical models for responses and response times*. University of Twente.
- Klein Entink, R. H., Fox, J.-P., & Van Der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74(1), 21-48.
- Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14(1), 54.
- Lawrence, I. (1993). The effect of test speededness on subgroup performance. Research Report 93-49, Princeton, NJ: Educational Testing Service.
- Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychol Test Assess Model*, 53(3), 359-379.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, 4(4), 269-290.
- Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization* (Vol. 8). Oxford University Press.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325--337.
- Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, 58(3), 445-469.
- Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of educational and behavioral statistics*, 39(6), 426-451.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3-8.

- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8(3), 261–272.
- Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing* (Vol. 03, No. 03). Law School Admission Council.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135.
- Meng, X. B., Tao, J., & Shi, N. Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, 84(1), 1-21.
- Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, 58, 525–543.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55(1), 75–106.
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015a). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, 68(2), 197-219.
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015b). A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, 50(1), 56-74.
- Parshall, C. G. (2002). *Practical considerations in computer-based testing*. Springer.
- Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated?. *Intelligence*, 40(1), 23-32.
- Ranger, J., & Kuhn, J. T. (2012a). A flexible latent trait model for response times in tests. *Psychometrika*, 77(1), 31-47.
- Ranger, J., & Kuhn, J. T. (2012b). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, 54 , 128-148.

- Ranger, J., & Kuhn, J. T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, 67, 388-407. doi: 10.1111/bmsp.12025
- Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 4, 59–74.
- Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Toward a taxonomic model. *Learning and Individual Differences*, 11, 1–120.
- Roskam, E. E. (1997). Models for speed and time-limit tests. In *Handbook of modern item response theory* (pp. 187-208). Springer New York.
- Rubin, D. B. (1984). Comment: assessing the fit of logistic regressions using the implied discriminant analysis. *Journal of the American Statistical Association*, 79(385), 79–80.
- Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response models. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theories and applications* (pp. 205–240). Cambridge, England: Cambridge University Press.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, 34(3), 213-232.
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. *Computer-based testing: Building the foundation for future assessments*, 237-266.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331–342.
- Stan Development Team. (2014). Stan modeling language users guide and reference manual, version 2.5.0 [Computer software manual]. Retrieved from <http://mc-stan.org/>
- St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, 0146621610391777.

- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203) New York: Academic Press.
- van der Linden, W. J. (2005). Model-based innovations in computer-based testing. *Computer-Based Testing and the Internet*, 39.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308.
- van der Linden, W. J. (2009a). A bivariate log-normal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, 34, 378-394.
- van der Linden, W. J. (2009b). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, 46(3), 247-272.
- van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, 48, 44–60.
- van der Linden, W. J. (2015). Lognormal response-time model. In W.J. van der Linden (Ed.), *Handbook of item response theory: Vol 1. Models*. Boca Raton, FL: Chapman & Hall/CRC.
- van der Linden, W. J., & Fox, J.-P. (2015). Joint hierarchical modeling of responses and response times. In W. van der Linden (Ed.), *Handbook of item response theory: Vol 1. Models*. Boca Raton: FL: Chapman & Hall/CRC.
- van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, 75(1), 120-139.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365-384.
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 195-210.

- van der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American journal of psychology*, 29-60.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16(1), 44.
- van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7(3), 424-465.
- Wang, C., Chang, H., & Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66 , 148-168.
- Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time *Applied Psychological Measurement*, 29(5), 323-339.
- West, M. (1993). Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Computing Science and Statistics*, 24, 325–333.
- West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models* (2nd ed.) New York: Springer.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163-183.
- Wise, S.L., Pastor, D.A., Kong, X.J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, 22, 185-201.

Samenvatting

In dit proefschrift is het modelleren van responsietijden besproken, binnen de context van computer-ondersteund toetsen. De responsietijden geven informatie over de snelheid van werken van respondenten. Er kan meer inzicht verkregen worden in het gedrag van respondenten wanneer zowel responsietijden als responsie meegenomen worden in de analyse. Verschillende hoofdstukken in dit proefschrift beschouwen analyses op basis van zowel responsietijden als responsie.

In hoofdstuk 2, drie statistieken zijn besproken, en benoemd als l_z^t , l_s^t , en l^t . De statistieken evalueren de modelpassing van de responsietijden van respondenten. De drie statistieken hebben ieder hun eigen (asymptotische) verdeling onder het model. De l^t heeft een chi-kwadraat verdeling, en de andere twee zijn bij benadering normaal verdeeld. Doormiddel van simulatiestudies zijn de eigenschappen van de statistieken geëvalueerd. Verschillend afwijkend gedrag van personen was gesimuleerd en de statistieken l^t en l_s^t , lieten de beste resultaten zien om dit afwijkende gedrag te detecteren. De detectiegraad was dicht bij 1, waardoor bijna alle afwijkende responsietijd patronen gedetecteerd werden.

In hoofdstuk 3 werd een dynamisch factor model voor de werksnelheid van personen geïntroduceerd. Het dynamisch factor model beschrijft de veranderingen in werksnelheid van de respondenten gedurende de test. De respondenten die met een gemiddelde constante snelheid werken en respondenten die met een tijdsvariante gemiddelde snelheid werken konden gedetecteerd worden. Hiervoor werd een model gebruikt met beide componenten waardoor respondenten kon worden toegewezen aan een van beide klassen. Simulatiestudie resultaten lieten zien dat redelijk accurate schattingen van de model parameters verkregen werden wanneer minimaal tien blokken van (minimaal) twee items werden gebruikt om de verandering van snelheid over blokken te modelleren.

In hoofdstuk 4 werd er een groei-model gebruikt om de trajectorie van snelheid van werken te modelleren. De onderliggende tijd as werd gedefinieerd op basis van de opeenvolging van items die zijn beantwoord. De ordening van items bepaalde de tijd as van iedere trajectorie. Alle model parameters konden accuraat geschat worden gebruikmakend van simulatietechnieken. Het groei-model kon afwijkingen modelleren van een constante werksnelheid door een trend en een kwadratische

tijdscomponent met persoon-specifieke parameters. Meer flexibele trajectoriën konden geschat worden met hogere orde tijdscomponenten.

In hoofdstuk 5 werd er aandacht gegeven aan de geschiktheid van de verdeling van de responsetijden doormiddel van de Kolmogorov-Smirnov toets. Hierbij werd er getoetst op de normaliteitsassumptie van de residuele afwijkingen. De statistieken om de passing van personen te evalueren werd uitgebreid om gezamenlijk de passing voor zowel responsietijden alsmede responsie patronen te evalueren. Op basis van een simulatiestudie werd geconcludeerd dat ook de statistieken voor de modelpassing van responsie en responsietijden goed functioneren.

De verschillende methoden zijn geïllustreerd op data verkregen uit de praktijk. De statistieken voor modelpassingen kunnen ook toegepast worden op een model met een gokparameter en een variabele werksnelheid, waardoor er rekening gehouden kon worden met eventueel gokgedrag van respondenten.

Het lognormale model en de item responsie modellen in dit proefschrift zijn toegepast op data verkregen uit de praktijk. De simulatietechnieken die gebruikt zijn voor het schatten van de parameters zijn geïmplementeerd in de statistische software R. Ze worden in de Appendix benoemd als de Log-Normal Response Times (LNRT), Log-Normal Response Times Moving Average (LNRTMA), en de Log Normal Item Response Theory (LNIRT). De programma's worden verder beschreven in de appendix (G, H).

Appendix A

The Expected Statistic Value and Its Variance as a Function of Response Times

The marginal distribution of the RT data is used to evaluate the fit of an RT pattern. This l_0 statistic as defined in Equation (2.5) can be standardized to derive the null distribution. The standardized version is denoted as l'_z , which requires the computation of the expected value and the variance.

The l_0 follows from the independently normally distributed logarithm of RTs as stated in Equation (2.6). Then the expected statistic value as a function of the RT is given by,

$$\begin{aligned} E(l_0) &= E\left(\sum_i \frac{(T_{pi}^* - \mu_{pi})^2}{\sigma_i^2} + \log(2\pi\sigma_i^2)\right) = \sum_i \left(E\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2 + \log(2\pi\sigma_i^2)\right) \\ &= \sum_i \left(E(Z_{pi}^2) + \log(2\pi\sigma_i^2)\right) = \sum_i (1 + \log(2\pi\sigma_i^2)) = I + \sum_i \log(2\pi\sigma_i^2) \end{aligned}$$

since the Z_{pi} is standard normally distributed and the expected value of a squared standard normally distributed variable equals one ($E(Z_{pi}^2) = \text{Var}(Z_{pi}) = 1$).

The variance of the statistic value as a function of the RT is given by

$$\begin{aligned} \text{Var}(l_0) &= \sum_i \text{Var}\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right) \\ &= \sum_i E\left(\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)^2 - \left(E\left(\frac{T_{pi}^* - \mu_{pi}}{\sigma_i}\right)^2\right)^2 \\ &= \sum_i E(Z_{pi}^4) - (E(Z_{pi}^2))^2 = \sum_i (3 - 1) = 2I. \end{aligned}$$

The expected value of the fourth power of a standard normally distributed variable follows from a variable transformation. Let $y = Z_{pi}^2$. Then, $E(Z_{pi}^4) = E(y^2)$, which can be expressed as a Gamma distribution with shape parameter 5/2 and scale parameter 2. The value three follows from the fact that the Gamma density integrates to one over the range of positive numbers.

Appendix B

Stationary Speed Process

A stationary speed process is the condition where the expected response times and covariances are only influenced by the average working speed and the time intensities. Under the measurement model in Equation (3.4) and the evolution equation defined in Equation (3.5), the sequence of response times follows a stationary speed process under certain conditions. It is shown that the stochastic process can be characterized as speed covariance stationary by showing that the mean and all covariances are not influenced by block-specific speed parameters when the start of the sequence changes. It follows that

$$\begin{aligned} E(T_{ik_c} | \zeta_i, \lambda_{k_c}) &= E\left(\lambda_{k_c} - \zeta_i - \sum_{j=1}^{c-1} \rho_j r_{ij} - r_{ic} + e_{ik_c}\right) \\ &= \lambda_{k_c} - \zeta_i - E\left(\sum_{j=1}^{c-1} \rho_j r_{ij} - r_{ic} + e_{ik_c}\right), \\ &= \lambda_{k_c} - \zeta_i \end{aligned}$$

since the error terms are independently distributed with a mean of zero. Thus, the expected response time only changes over items due to different item intensities such that

$$E(T_{ik_c} | \zeta_i, \lambda_{k_c}) - \lambda_{k_c} = E(T_{ik_{c-j}} | \zeta_i, \lambda_{k_{c-j}}) - \lambda_{k_{c-j}}.$$

The conditional variance of the response time is given by

$$\begin{aligned} Var(T_{ik_c} | \zeta_i, \lambda_{k_c}) &= Var\left(\lambda_{k_c} - \zeta_i - \sum_{j=1}^{c-1} \rho_j r_{ij} - r_{ic} + e_{ik_c}\right) \\ &= Var\left(\sum_{j=1}^{c-1} \rho_j r_{ij} + r_{ic} + e_{ik_c}\right) \\ &= \sigma_\rho^2 \left(1 + \sum_j \rho_j^2\right) + \sigma_k^2. \end{aligned}$$

The first term on the right-hand side follows from the moving average process for modeling the working speed; the second term represents the measurement error variance for item k . The variance term from the speed process is invariant across items, which is a necessary condition for a

stationary speed process. Furthermore, as long as $\sum_j \rho_j^2$ is finite, it follows that the variance of the response time is finite.

Finally, the conditional covariance between subject's i response time in block c and block $c-s$ is given by

$$\begin{aligned} Var(T_{ik_c}, T_{ik_{c-s}} | \zeta_i) &= Var\left(\zeta_i + \sum_{j=1}^c \rho_j r_{ic-j}, \zeta_i + \sum_{j=1}^c \rho_j r_{ic-s-j}\right) \\ &= \sigma_\zeta^2 + \sigma_\rho^2 (\rho_j + \rho_1 \rho_{s+1} + \rho_2 \rho_{s+2} + \dots) \end{aligned}$$

It does not depend on any block parameters, but only on the number of blocks separating the response times.

Appendix C

WinBUGS Algorithm:

Two-component Mixture Model for Speed

```

model{
  for(i in 1:N) {
    eta[i,1] <- speed[i] + beta[i]* r.0[i] + r[i,1]
    for(k in 1 : (K/C) ) {#item blocks
      T[i,k] ~ dnorm(mu [ i , k ] , sigmam [ k ] )
      mu[i,k] <- bsp[k] - eta[i,1]
    }

    for(c in 2:C) { # speed in blocks 2—C: define mixture of MA(1)
      and AR(1) eta[i,c] <- speed[i] + (1 - betaF[class1[i]]) *
        beta[i]*r[i,c - 1] + r[i,c] + betaF[class1[i]]*eta[i,c - 1]
      for(k in ((K/C)*c - ((K/C) - 1)) : ((K/C)* c) ) {
        T [i,k] ~ dnorm (mu [i,k],sigmam[k])
        mu[i,k] <- bsp[k] - eta[i,c]
      } }

    for(c in 1:C) {
      r[i,c] ~ dnorm(0,sigmaspt) # item block error variance
    }
    r.0[ i ] <- 0
    dummy[i] ~ dbeta(a,b)
    beta[i] <- 2*dummy[i] - 1
    # transition parameter with identification restriction
    speed[i] ~ dnorm(mus1, sigmasp1) # average speed
    class[i] ~ dbern(q) # mixture component distribution
    class1[i] <- class[i] + 1
  }

  for(k in 1:K) {
    bsp[k] ~ dnorm(mub,sigmasb) # time intensities
    sigmam[k] ~ dgamma (.1,. 1)
  }

  #hyper prior specification
  q~dbeta(1,1)
  betaF[1] <- 0

```

```
betaF[2] <- 1  
mus1 <- 0  
a ~ dunif(0,100)  
b ~ dunif(0,100)  
mub ~ dnorm(0,1.0E-1)  
sigmab ~ dgamma (.1,.1)  
sigmasp1 ~ dgamma (.1,.1)  
sigmasp2 ~ dgamma (.1,.1)  
sigmaspt ~ dgamma (.1,.01)  
}
```

Appendix D

Winbugs Code for Model 1 (Equation 4.17)

```
model
{
  for (i in 1:N) {
    for (k in 1:K) {
      T[i,k] ~ dnorm(mu[i,k],sigmam[k])
      mu[i,k] <- mub + speed[i,k]
      speed[i,k] <- b1*X[i,k] + b0[i]
    }
  }
  for (i in 1:N) {
    b0[i] ~ dnorm(0,sigmasp)
  }
  for (k in 1:K) {
    sigmam[k] ~ dgamma(.1,.1)
    sigmamn[k] <- 1/sigmam[k]
  }
  b1 ~ dnorm(0,1.0E-2)
  mub ~ dnorm(0,1)
  sigmasp ~ dgamma(1,.1)
  sigmaspn <- 1/sigmasp
}
```


Appendix E

Winbugs Code for Model 2 (Equation 4.18)

```

model
{
  for (i in 1:N) {
    for (k in 1:K) {
      T[i,k] ~ dnorm(mu[i,k],sigmam[k])
      mu[i,k] <- bsp[k] - speed[i,k]
      speed[i,k] <- b1[i]*X[i,k] + b0[i]
    }
  }
  for (k in 1:K) {
    bsp[k] ~ dnorm(mub,sigtab);
    sigmam[k] ~ dgamma(.1,.1)
    sigmamn[k] <- 1/sigmam[k]
  }
  for (i in 1:N) {
    b0[i] ~ dnorm(0,sigmasp)
    b1[i] ~ dnorm(mub1,sigtab1)
  }
  mub1 <-0
  mub ~ dnorm(0,1)
  sigtab1 ~ dgamma(1,.1)
  sigtab1n <- 1/sigtab1
  sigtab ~ dgamma(1,.1)
  sigtabn <- 1/sigtab
  sigmasp ~ dgamma(1,.1)
  sigmaspn <- 1/sigmasp
}

```


Appendix F

Winbugs Code for Model 3 (Equation 4.19)

```

model
{
  for (i in 1:N) {
    for (k in 1:K) {
      T[i,k] ~ dnorm(mu[i,k],sigmam[k])
      mu[i,k] <- bsp[k] - speed[i,k]
      speed[i,k] <- b1[i]*X[i,k]+b2[i]*pow(X[i,k],2)+b0[i]
    }
  }
  for (k in 1:K) {
    bsp[k] ~ dnorm(mub,sigmasp);
    sigmam[k] ~ dgamma(1,.1)
    sigmamn[k] <- 1/sigmam[k]
  }
  for (i in 1:N) {
    b0[i] ~ dnorm(0,sigmasp)
    b1[i] ~ dnorm(mub1,sigmasp)
    b2[i] ~ dnorm(mub2,sigmasp)
  }
  mub1 <- 0
  mub2 <- 0
  mub ~ dnorm(0,1)
  sigmasp ~ dgamma(1,.1)
  sigmaspn <- 1/sigmasp
  sigmasp1 ~ dgamma(1,.1)
  sigmasp1n <- 1/sigmasp1
  sigmasp2 ~ dgamma(1,.1)
  sigmasp2n <- 1/sigmasp2
  sigmasp ~ dgamma(1,1)
  sigmaspn <- 1/sigmasp
  sigmasp1 ~ dgamma(1,.1)
  sigmasp1n <- 1/sigmasp1
  sigmasp2 ~ dgamma(1,.1)
  sigmasp2n <- 1/sigmasp2
  sigmasp ~ dgamma(1,1)
  sigmaspn <- 1/sigmasp
}

```


Appendix G

R Package: LNIRT

Log Normal Item Response Theory (LNIRT) is a program written in R (modified version of the *cirt* R-package of Fox et al., 2007) for estimating the model parameters of response times under log normal model and/or responses under Item response theory model.

Examples:

```
data <- simLNIRT(N=500,K=10,rho=0)
```

N : Person
K : Item
Y : Response matrix of dim (N, K)
RT : log-response time matrix (time spent on solving an item)
XG : Number of iterations for the MCMC algorithm
Rho : Covariance between person parameters

Response Time:

```
out<-LNIRT(RT=data$RT,XG=20000)
```

Response Time and Response:

```
out<-LNIRT(RT=data$RT,Y=data$Y,XG=20000)
```

R Code for Data Simulation:

```
simLNIRT <- function(N,K,rho){
  library(MASS)
  mutheta <- rep(0,2)
  covtheta <- diag(2)
  covtheta[1,2] <- covtheta[2,1] <- rho
  theta <- mvrnorm(N, mutheta, covtheta, empirical = TRUE)
  covitem <- diag(4)
  for(ii in 1:4) {
    covitem[ii,] <- covitem[ii,]*rep(c(.03,1),2)
  }
  muiteM <- rep(c(1,0),2)
  ab <- mvrnorm(K, muiteM, covitem)
  ab[,c(2,4)] <- ab[,c(2,4)] - t(matrix(colMeans(ab[,c(2,4)]),2,K))
  ab[,1] <- abs(ab[,1])
  ab[,1] <- ab[,1]/(prod(ab[,1])^(1/K))
  ab[,3] <- abs(ab[,3])
  ab[,3] <- ab[,3]/(prod(ab[,3])^(1/K))
  par <- theta[,1] %*% matrix(ab[,1],nrow=1,ncol=K) -
    t( matrix(ab[,2],nrow=K,ncol=N) )
  probs <- matrix(pnorm(par),ncol = K,nrow = N)
  Y <- matrix(runif(N*K),nrow = N, ncol = K)
  Y <- ifelse(Y < probs,1,0)
  time <- matrix(rnorm(N * K,sd=1), nrow = N, ncol = K)
  time[1:N, ] <- time[1:N, ] - theta[,2]
  for (ii in 1:K) {
    time[1:N, ii] <- ab[ii,3]*time[1:N, ii] + ab[ii,4]
  }
  RT <- time
  return(list(Y=Y,RT=RT,theta=theta,ab=ab))
}
```

Appendix H

R Package: LNIRTQ

Log Normal Item Response Theory Quadratic (LNIRTQ) is a program written in R (modified version of the *cirt* R-package of Fox et al., 2007) for estimating response and response time which includes variable speed in the model. The code of LNIRT was extended by adding random person effects, referred as Q, which exhibits the number of speed components.

Examples:

Simulation:

```
data <- simvaryspeed(N=1000,K=20,rho=0)
```

N : Person
K : Item
Y : Response matrix of dim (N, K)
RT : log-response time matrix (time spent on solving an item)
XG : Number of iterations for the MCMC algorithm
X : Time scale
Rho : Covariance between person parameters

Estimation:

```
out<-LNIRTQ(Y=data$Y, RT=data$RT, X=data$X, XG=20000)
```


R Code for Data Simulation of Model 1 (Equation 4.17):

```
simvaryspeed <- function(N,K){
  library(MASS)
  b0 <- rnorm(N,mean=0,sd=.5)
  b1 <- 0.25
  A <- (1:K)
  A <- rep(A,N)
  X <- matrix(A,nrow=N,ncol=K,byrow=TRUE)
  X <- (X - 1)/K

  speed <- matrix(0,ncol=K,nrow=N)
  theta <- matrix(rnorm(N,mean=0,sd=1), nrow = N, ncol = 1) #ability
  for(ii in 1:N){
    speed[ii,] <- b0[ii] + b1*X[ii,]
  }
  covitem <- diag(4)
  for(ii in 1:4) {
    covitem[ii,] <- covitem[ii,]*rep(c(.03,1),2)
  }
  muiitem <- rep(c(1,0),2)
  ab <- mvrnorm(K, muiitem, covitem)
  ab[,c(2,4)] <- ab[,c(2,4)] - t(matrix(colMeans(ab[,c(2,4)]),2,K))
  ab[,1] <- abs(ab[,1])
  ab[,1] <- ab[,1]/(prod(ab[,1])^(1/K))
  ab[,3] <- abs(ab[,3])
  ab[,3] <- ab[,3]/(prod(ab[,3])^(1/K))
  par <- theta %*% matrix(ab[,1],nrow=1,ncol=K) -
    t( matrix(ab[,2],nrow=K,ncol=N) )
  probs <- matrix(pnorm(par),ncol = K,nrow = N)
  Y <- matrix(runif(N*K),nrow = N, ncol = K)
  Y <- ifelse(Y < probs, 1,0)

  time <- matrix(0,ncol=K,nrow=N)
  for (kk in 1:K) {
    time[1:N, kk] <- speed[1:N, kk] + rnorm(N,mean=0,sd=1)
  }
  T <- time
  return(list(Y=Y,T=T,X=X,theta=theta,speed = speed,ab=ab))
}
```

R Code for Data Simulation of Model 2 (Equation 4.18):

```

simvaryspeed <- function(N,K){
  library(MASS)
  b0 <- rnorm(N,mean=0,sd=.5)
  b1 <- rnorm(N,0,.5) # sd= 0.5

  A <- (1:K)
  A <- rep(A,N)
  X <- matrix(A,nrow=N,ncol=K,byrow=TRUE)
  X <- (X - 1)/K

  speed <- matrix(0,ncol=K,nrow=N)
  theta <- matrix(rnorm(N,mean=0,sd=1), nrow = N, ncol = 1) #ability
  for(ii in 1:N){
    speed[ii,] <- b0[ii] + b1[ii]*X[ii,] #speed
  }

  covitem <- diag(4)
  for(ii in 1:4) {
    covitem[ii,] <- covitem[ii,]*rep(c(.03,1),2)
  }
  muiitem <- rep(c(1,0),2)

  ab <- mvrnorm(K, muiitem, covitem)
  ab[,c(2,4)] <- ab[,c(2,4)] - t(matrix(colMeans(ab[,c(2,4)]),2,K))
  ab[,1] <- abs(ab[,1])
  ab[,1] <- ab[,1]/(prod(ab[,1])^(1/K))
  ab[,3] <- abs(ab[,3])
  ab[,3] <- ab[,3]/(prod(ab[,3])^(1/K))

  # itemcorrect
  par <- theta %*% matrix(ab[,1],nrow=1,ncol=K) -
    t( matrix(ab[,2],nrow=K,ncol=N) )
  probs <- matrix(pnorm(par),ncol = K,nrow = N)
  Y <- matrix(runif(N*K),nrow = N, ncol = K)
  Y <- ifelse(Y < probs, 1,0)

  time <- matrix(0,ncol=K,nrow=N)
  # response time
  for (kk in 1:K) {
    time[1:N, kk] <- ab[kk,4]-speed[1:N, kk] + rnorm(N,mean=0,sd=1)
  }
  T <- time
  return(list(Y=Y,T=T,X=X,theta=theta,speed=speed,ab=ab,b1=b1))
}

```

R Code for Data Simulation of Model 3 (Equation 4.19):

```

simvaryspeed <- function(N,K){
  library(MASS)
  b0 <- rnorm(N, 0, .5)
  b1 <- rnorm(N, 0, .5) # sd= 0.5
  b2 <- rnorm(N, 0, .3)
  A <- (1:K)
  A <- rep(A,N)
  X <- matrix(A,nrow=N,ncol=K,byrow=TRUE)
  X <- (X - 1)/K
  speed <- matrix(0,ncol=K,nrow=N)

  theta <- matrix(rnorm(N,mean=0,sd=1), nrow = N, ncol = 1) #ability
  for(ii in 1:N){
    speed[ii,] <- b0[ii]+b1[ii]*X[ii,]+b2[ii]*(X[ii,]^2) #speed
  }
  covitem <- diag(4)
  for(ii in 1:4) {
    covitem[ii,] <- covitem[ii,]*rep(c(.03,1),2)
  }
  muiitem <- rep(c(1,0),2)
  ab <- mvrnorm(K, muiitem, covitem)
  ab[,c(2,4)] <- ab[,c(2,4)] - t(matrix(colMeans(ab[,c(2,4)]),2,K))
  ab[,1] <- abs(ab[,1])
  ab[,1] <- ab[,1]/(prod(ab[,1])^(1/K))
  ab[,3] <- abs(ab[,3])
  ab[,3] <- ab[,3]/(prod(ab[,3])^(1/K))

  # itemcorrect
  par <- theta %*% matrix(ab[,1],nrow=1,ncol=K) -
    t( matrix(ab[,2],nrow=K,ncol=N) )
  probs <- matrix(pnorm(par),ncol = K,nrow = N)
  Y <- matrix(runif(N*K),nrow = N, ncol = K)
  Y <- ifelse(Y < probs, 1,0)

  time <- matrix(0,ncol=K,nrow=N)
  # response time
  for (kk in 1:K) {
    #time[1:N, kk] <- ab[kk,3]*speed[1:N, kk] + ab[kk,4] + rnorm(N,mean=0,sd=1)
    time[1:N, kk] <- ab[kk,4]-speed[1:N, kk] + rnorm(N,mean=0,sd=1)
  }
  T <- time
  return(list(Y=Y,T=T,X=X,theta=theta,speed=speed,ab=ab,b1=b1,b2=b2))
}

```

Appendix I

RSTAN Algorithm

```

data{
  int<lower=0> N;
  int<lower=0> K;
  int<lower=0> C;
  int<lower=0> B[K];    //Block of items
  real RT[N,K];        //Response times
}
parameters{
  real zeta[N];
  real eta[N,C];
  real lambda[K];
  real rho[N];
  real<lower=0> sigma2; //inverse variance for inverse gamma prior.
  real<lower=0> sigma2K;
  real<lower=0> tau2;
  real<lower=0> sigmaR2; //block errors on eta
  real<lower=0> sigrho2;
  real muz;
  real murho;
}

transformed parameters {
  real<lower=0> sigma;
  real<lower=0> sigmaK;
  real<lower=0> tau;
  real<lower=0> sigmaR;
  real<lower=0> sigrho;

  tau <- sqrt(1/tau2);
  sigma <- sqrt(1/sigma2);
  sigmaK <- sqrt(1/sigma2K);
  sigmaR <- sqrt(1/sigmaR2);
  sigrho <- sqrt(1/sigrho2);
}

model{
  real etahat[N,C];    // prediction for eta
  real r[N,C];         // error terms
  for(ii in 1:N){
    for(kk in 1:K){
      RT[ii,kk] ~ normal(lambda[kk] - eta[ii,B[kk]],sigma);
    }

    etahat[ii,1] <- zeta[ii];    // assume zeroth error = 0
    r[ii,1] <- eta[ii,1] - etahat[ii,1];s

    for(cc in 2:C){
      etahat[ii,cc] <- zeta[ii] + rho*r[ii,cc-1]; //MA(1) component
      r[ii,cc] <- eta[ii,cc] - etahat[ii,cc];
    }
  }
}

```

Appendix I
138

```
for(cc in 1:C){
  r[ii,cc] ~ normal(0,sigmaR);
}
}
for(ii in 1:N){
  zeta[ii] ~ normal(muz,tau);
  rho[ii] ~ normal(murho,sigrho);
}
for(kk in 1:K){
  lambda[kk] ~ normal(0,sigmaK);
}
// priors
murho ~ normal(0,2);
muz ~ normal(0,10);
tau2 ~ gamma(.01, .01);
sigma2 ~ gamma(.01, .01);
sigma2K ~ gamma(.01, .01);
sigmaR2 ~ gamma(.01, .01);
sigrho2 ~ gamma(.01, .01);
}
```

Appendix J

Simulated and Estimated Measurement Error Variance Parameters

Table J.1. *Simulated and estimated measurement error variance parameters of Model 2.*

Par.	True	Winbugs		Par.	True	Winbugs	
		Mean	SD			Mean	SD
$\sigma_{\varepsilon 1}^2$	1	1.016	0.049	$\sigma_{\varepsilon 11}^2$	1	0.999	0.047
$\sigma_{\varepsilon 2}^2$	1	1.003	0.048	$\sigma_{\varepsilon 12}^2$	1	1.004	0.047
$\sigma_{\varepsilon 3}^2$	1	0.997	0.047	$\sigma_{\varepsilon 13}^2$	1	1.010	0.048
$\sigma_{\varepsilon 4}^2$	1	1.007	0.048	$\sigma_{\varepsilon 14}^2$	1	0.996	0.046
$\sigma_{\varepsilon 5}^2$	1	0.999	0.046	$\sigma_{\varepsilon 15}^2$	1	1.002	0.048
$\sigma_{\varepsilon 6}^2$	1	1.009	0.048	$\sigma_{\varepsilon 16}^2$	1	1.007	0.049
$\sigma_{\varepsilon 7}^2$	1	1.012	0.048	$\sigma_{\varepsilon 17}^2$	1	0.997	0.047
$\sigma_{\varepsilon 8}^2$	1	0.998	0.047	$\sigma_{\varepsilon 18}^2$	1	0.995	0.048
$\sigma_{\varepsilon 9}^2$	1	0.995	0.047	$\sigma_{\varepsilon 19}^2$	1	1.000	0.049
$\sigma_{\varepsilon 10}^2$	1	0.996	0.046	$\sigma_{\varepsilon 20}^2$	1	0.993	0.049

Table J.2 *Simulated and estimated measurement error variance parameters of Model 3.*

Par.	True	Winbugs		Par.	True	Winbugs	
		Mean	SD			Mean	SD
$\sigma_{\varepsilon 1}^2$	1	0.997	0.048	$\sigma_{\varepsilon 11}^2$	1	0.999	0.047
$\sigma_{\varepsilon 2}^2$	1	0.998	0.048	$\sigma_{\varepsilon 12}^2$	1	1.011	0.048
$\sigma_{\varepsilon 3}^2$	1	0.998	0.048	$\sigma_{\varepsilon 13}^2$	1	0.998	0.047
$\sigma_{\varepsilon 4}^2$	1	0.996	0.048	$\sigma_{\varepsilon 14}^2$	1	0.999	0.048
$\sigma_{\varepsilon 5}^2$	1	1.008	0.049	$\sigma_{\varepsilon 15}^2$	1	0.993	0.047
$\sigma_{\varepsilon 6}^2$	1	1.005	0.048	$\sigma_{\varepsilon 16}^2$	1	0.992	0.047
$\sigma_{\varepsilon 7}^2$	1	0.996	0.046	$\sigma_{\varepsilon 17}^2$	1	1.004	0.048
$\sigma_{\varepsilon 8}^2$	1	1.000	0.047	$\sigma_{\varepsilon 18}^2$	1	0.987	0.048
$\sigma_{\varepsilon 9}^2$	1	1.005	0.048	$\sigma_{\varepsilon 19}^2$	1	0.995	0.049
$\sigma_{\varepsilon 10}^2$	1	0.993	0.048	$\sigma_{\varepsilon 20}^2$	1	0.996	0.051