Evaluating Goodness–Of–Fit Tests For The Dichotomous Rasch Model

Diplomarbeit

JULIA MANG

Betreuung:
Prof. Dr. Helmut Küchenhoff
Prof. Dr. Markus Bühner

Ludwig–Maximilians–Universität
Institut für Statistik
Lehrstuhl psychologische Methodenlehre und Diagnostik

Abstract

The Rasch model as the model with the strongest model assumptions from Item Response Theory stands for the meaningful characteristics like unidimensionality of the trait, local independency and parallel item characteristic curves.
To test the fit of the Rasch model to simulated input scenarios like Rasch conform data and data from different forms of Rasch model violations is the main part of this work.
Three selected goodness–of–fit tests are analysed on their performance to these scenarios.
The Andersen likelihood ratio test performs well in connection with Rasch conform data and is able to detect violations from intersecting item characteristic curves, additional guessing and parts of multidimensional traits, but not local dependency.
Whereas the Bootstrap test holds the given α–level for Rasch conform data, the test fails to detect any of the simulated Rasch model violations. These bad results can be interpreted by the chosen χ^2 test statistic for the parametric Bootstrap.
The mixed–Rasch model test is detailed analysed about the methodological test procedure and supposed wrong distribution–assumptions.
Applying results to practical applications the nine subtests from the Intelligence–Structure–Test (IST) 2000 R are investigated to fit the Rasch model. The different outcoming results from the selected goodness–of–fit tests are elaborately discussed.
In the appendix the simulation routine, written with the statistical software R, is displayed and explained.

Preface

This interdisciplinary diploma thesis is written at the Department of Statistics and at the Unit Psychological Methodology and Evaluation of the Ludwig–Maximilians–University in Munich.
First I want to thank Prof. Dr. Helmut Küchenhoff from the Department of Statistics and Prof. Dr. Markus Bühner from the Unit Psychological Methodology and Evaluation of the Department Psychology to have been in charge of my work and to have sponsored me during the thesis operation.
Further thank goes to Dr. Moritz Heene and Dr. Clemens Draxler from the Unit Psychological Methodology and Evaluation of the Department Psychology to have helped me with psychological problem statements.
For the computational and methodological implementation of the mixed–Rasch model a special thank is for Prof. Dr. Friedrich Leisch and Dr. Carolin Strobl from the Department of Statistics.
Due to the friendly assistance for the realisation of running the Bootstrap R–Code on the Linux Cluster Server at the Leibniz Rechenzentrum in Munich I want to thank Dr. Ferdinand Jamitzky from the Leibniz–Rechenzentrum.
Last but not least I thank Dr. Patrick Mair for helping me with computational questions in the R–package *eRm*.
Concluding I would like to thank all these persons for the interdisciplinary collaboration between the two sciences statistics and psychology.

Contents

List of Figures

Chapter 1

Introduction

Psychological testing is one of the main issues in applied psychology. The desired measurements, i.e. individual´s properties like intelligence, arithmetic ability or neuroticism, are mostly not directly observable; thus they are called latent traits. Therefore the individual´s responses to well–chosen items ought to measure the extent to a certain latent trait of the individual. Items are in the context of psychological measurement questions or exercises, which can have outcomes like *wrong/false* or *yes/no* in the dichotomous case. This indirect measurement allows the prediction of an individual´s behavior to other items from the same trait.

The theoretical basis for psychological measurement followed the Classical Test Theory (CTT) over decades. Such a theory assumes that the person´s total test score depends on the sum of the true score of the trait and a corresponding error score.

In the 1950s the first Item Response Theory (IRT) models were implemented. Since some of the assumptions from CTT are not appropriate or revisable for real data, the Item Response Theory has become a new paradigm in psychological testing. IRT has stronger assumptions than CTT and therefore obtains stronger results. It is based on the probability of a person´s answer to a certain item, which links the person´s score to the latent trait. A short review of the history and mainly used IRT models will be given in chapter 2.

For the early application of IRT to psychological testing, nowadays it can be regarded as the psychometric method of choice for testing. Thus it is not only limited to psychological subjects. Alagumalai, Curtis and Hungi (2005) as well as Bezruczko (2005) stated, that IRT can also be adopted to applications in e.g. linguistic, economics or health science.

One of the main features in IRT and one of the main topics of this work is the one parameter logistic IRT model, namely the Rasch model. It was proposed by Rasch in 1960.

The (0/1)–response from a person to a certain item is represented by a logistic function which depends on the parameters for the person´s ability and the item difficulty. The Rasch model has exceeding characteristics like specific objectivity, unidimensionality of the trait or local independency. These properties as well as the main assumptions of the Rasch Model including parameter estimation will be described in chapter 3.
An essential question in IRT and in Rasch modeling is the comparison between the empirical data and the Rasch model. If the model fits the data in an appropriate manner, the data can be expressed by the model and its parameters. Chapter 4 provides an outline about mostly used goodness–of–fit tests for the Rasch model.
Three of these goodness–of–fit tests, namely the Andersen test, the parametric Bootstrap test and the mixed–Rasch model test are analyzed in chapter 5. In a comprehensive simulation study the performance of these three tests is investigated under Rasch model conformity and different forms of Rasch model violations.
To relate to experiences in practical work, the nine subtests of the Intelligence–Structure–Test (IST) 2000 R are tested in chapter 6 for goodness–of–fit to the Rasch model. The IST 2000 R measures the extent to deductive reasoning and is one of the most applied intelligence tests in Germany (Steck, 1997; Schorr, 1995).
Finally, a meaningful summary will be given in chapter 7 and some outlooks for further studies are provided.
The simulation study written with the statistical software R is displayed in the appendix.

Chapter 2

Item Response Theory Models

Item Response Theory is an expansion and an improvement in psychological measurement. In comparison with the former prevalent Classical Test Theory (CTT), this approach improves and devises the features in modeling existing tests, constructing new ones, applying tests to non–standard settings and interpreting the results of measurements (Molenaar, 1995a).
The main idea behind IRT is the probability, that a person answers a certain item correctly. This probability can be expressed by means of the person´s position on the latent trait, i.e. the ability of a person, plus one or two parameters defining the particular item. The probability of a specific answer as a function of the latent trait is given by the Item Characteristic Curve (ICC, van den Wollenberg, 1982) or also called Item Response Function (IRF) and plays a major role in detecting and defining the properties of the IRT model.
A short outline of the initial provided IRT models is given below. For an elaborate overview as for a review of the transition from CTT to IRT refer to Fischer (1974), Hambleton & Swaminathan (1985) and Baker (1992).
In this work it will only be related to dichotomous IRT Models. For information about extensions to the dichotomous Rasch model refer to Fischer & Molenaar (1995).

2.1. The Initial IRT Models

The first IRT models were developed in the 1950s and are nowadays of little use, since more sophisticated models have been developed. Nevertheless these models play an important role in the development of IRT models.
To outline the varieties of the following models, the Item Characteristic Curve, denoted by ICC, will give an outstanding impression of the main model properties. The para-

meter for the latent dimension θ is displayed on the abscissa and the solving probability according to the corresponding latent parameter $P(\theta)$ is indicated on the ordinate, respectively.

The first IRT model was implemented by Guttman (1950) and is called *perfect scale* model. As shown in figure (2.1), the ICC is a deterministic step curve, which is given

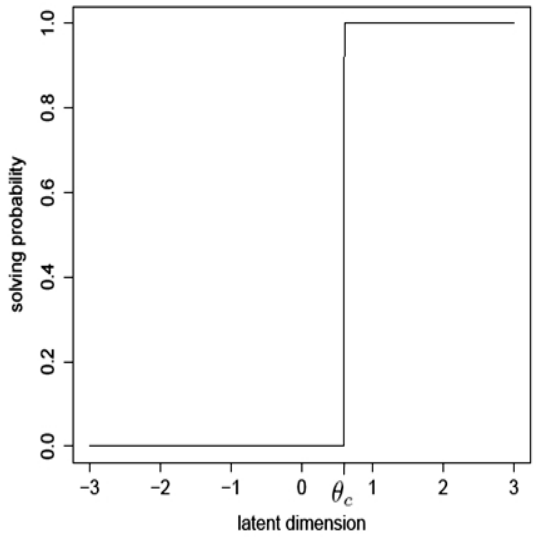

Figure 2.1.: ICC for Guttman´s *perfect scale* model

by the indicator function

$$P(\theta) = \begin{cases} 1 & \theta > \theta_c \\ 0 & \theta \leq \theta_c \end{cases} . \tag{2.1}$$

θ_c is the critical value on the latent scale from which a person answers a certain item correctly.

It is obvious, that this model does not satisfy the user´s pretensions. A meaningful conclusion can only be stated, if a person has θ_c as parameter value, otherwise no meaningful statement about the latent trait position of the person can be made.

Thus a more stochastic model, namely the *latent distance* model, was provided by Lazarsfeld in 1950. This model, displayed in figure (2.2), expands the deterministic

ICC from Guttman to a stochastic ICC. Also here θ_c is the critical ability value for

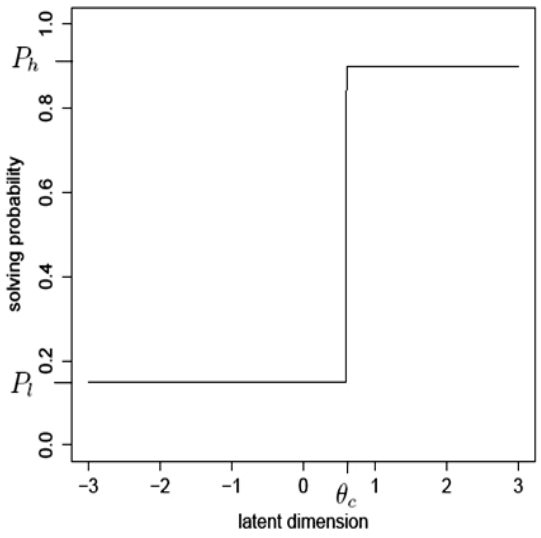

Figure 2.2.: ICC for Lazarsfeld´s *latent distance* model

solving an item. However the possibility to solve a particular item with even low ability parameter values is here defined by a probability P_l. On the other hand persons with high ability parameter values can even fail to solve a certain item. This boundary is given with the upper probability P_h.
Hence the indicator function for the *latent distance* model is denoted by

$$P(\theta) = \begin{cases} P_h & \theta > \theta_c \\ P_l & \theta \leq \theta_c \end{cases} . \tag{2.2}$$

However this easy indicator function is not useful for real data. The connection between the person´s ability parameter and the solving probability is implausible to be described by the *latent distance* model curve.
Due to these findings, Lazarsfeld (1959) stated, that the ICC can be seen as a linear function.
The probability to solve an item is in that case proportional to the person´s position on the latent trait. Therefore the corresponding ICCs are given by

$$P(\theta) = P(\text{Positive Response}|\theta) = c + a\theta \quad . \tag{2.3}$$

As it can be seen in figure (2.3) the resulting probabilities can possibly be beyond the

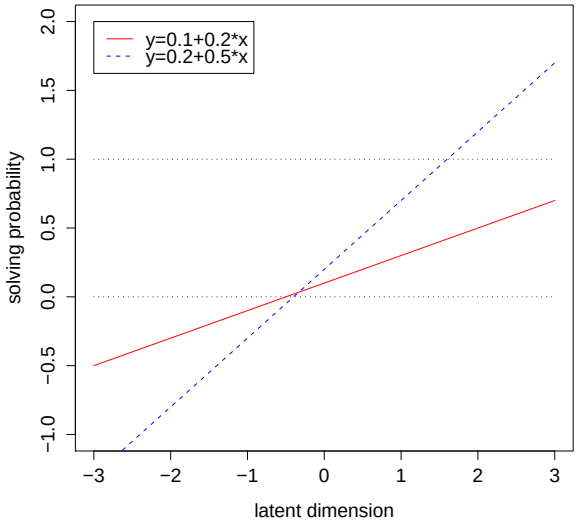

Figure 2.3.: ICCs for Lazarsfeld´s linear model

permitted domain for probabilities, i.e. negative probabilities or probabilities greater than unity. The appearance of these inaccurate probabilities leads to the impracticality of this model.

2.2. The Rasch Model

From a theoretical point of view, as well as from empirical studies (Rost, 2004), the only reasonable curve to link the latent dimension to the solving probability is the logistic function. The solving probability changes only insignificantly when items are very easy or very difficult, whereas it increases almost linearly in conjunction with moderately heavy items. The curve converges on the left hand side to 0 and on the right hand side to 1.
For further comprehension some notations must be announced:

X_{vi}	response of person v to item i
θ_v	position on the latent trait for person v
β_i	item difficulty for item i
α_i	discrimination for item i
γ_i	lower asymptote (guessing) for item i

The simplest model with such a logistic ICC is the 1–parameter logistic model, namely the Rasch model. The probability of solving an item is, as before, not only dependent on the ability parameter θ, but also on the item difficulty parameter β. The dichotomous response to an item is 0 by answering the item falsely and 1 for a correct answer. Thus the basic equation for the Rasch model is given by

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \quad . \tag{2.4}$$

Figure (2.4) shows three typical curves for Rasch model data. The item difficulty pa-

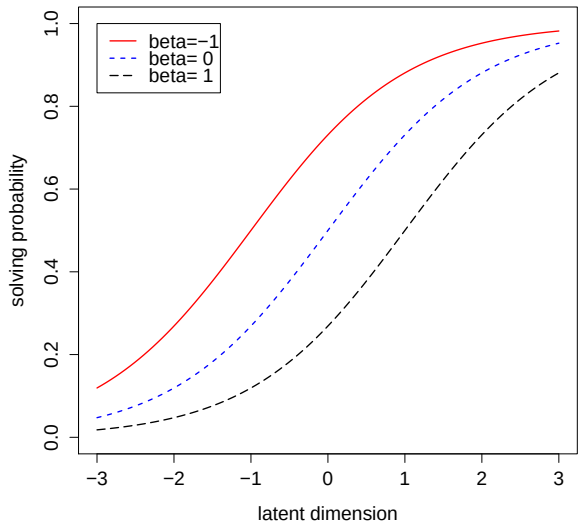

Figure 2.4.: ICCs for the Rasch model

rameter β_i is displayed on the same scale as the person´s ability parameter θ_v, but is not explicitly written down. The scale of θ and β is defined to have an interval scale level. The ICCs can be shifted only horizontally due to differences in item difficulty, i.e. an ICC shifted to the right stands for a heavier item. All items have the same discrimination, which means the items differentiate equal between persons with variable abilities.

A characteristic feature of the Rasch model is its parallel ICCs. Parallelism in this context means non intersecting curves, whereas total parallelism is only reached in the middle slope of the ICC. This feature is related to the Rasch model´s exceeding properties like specific objectivity, local independency or unidimensionality of the trait. Since the Rasch model is one main part of this work, all assumptions as well as model prop-

erties are described in chapter 3, elaborately.

2.3. The 2–Parameter Logistic Model

By taking another parameter α into account the basic mathematical statement for the 2–parameter logistic model (2–PL model), implemented by Birnbaum (1968), is denoted by

$$P(X_{vi} = 1|\theta_v, \beta_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_v - \beta_i))}{1 + \exp(\alpha_i(\theta_v - \beta_i))} \quad , \tag{2.5}$$

where α_i is the discrimination parameter for item differences. It is the slope of the Item Characteristic Curves. As it can be seen from figure (2.5), the slope of the curves increases with rising α_i. That means, if an item discriminates quite good the slope of the ICC is steeper, while a less discriminating item has a ICC which is flatter. One

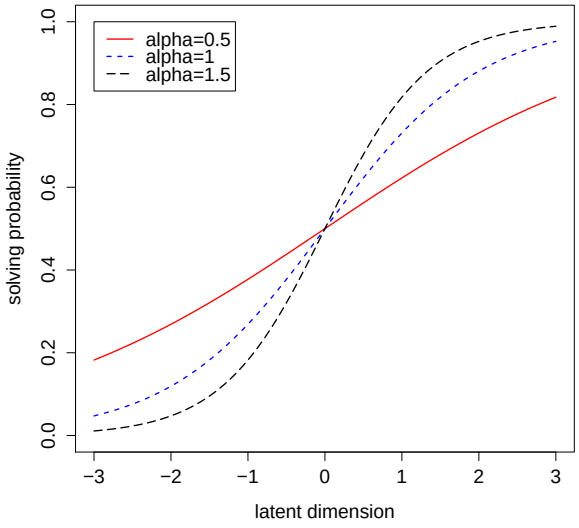

Figure 2.5.: ICCs for the 2–parameter logistic model

of the main properties of the Rasch model, i.e. that the ICCs are parallel, commonly are not valid anymore. This can cause interpretation problems, when a person with a higher ability fails to solve an item, whereas a person with a lower ability solves this item. In this case the ICCs overlap. Hence the 2–PL model violates the assumptions for the Rasch model and is therefore analysed in the simulation study in chapter 5.

2.4. The 3–Parameter Logistic Model

Birnbaum (1968) also implemented the 3–parameter logistic model (3–PL model) where, similar to Lazarsfeld´s *latent distance* model, a probability is given for persons with lower ability to solve an item. This effect is realized by the new defined lower asymptotic parameter γ and is also called the *guessing* parameter. γ moves the logistic curve up to the parameter value. Thus the model equation is given by

$$P(X_{vi} = 1|\theta_v, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp(\alpha_i(\theta_v - \beta_i))}{1 + \exp(\alpha_i(\theta_v - \beta_i))} \quad . \tag{2.6}$$

Three typical ICCs for the 3–PL model are depicted in figure (2.6). Like the 2–PL model

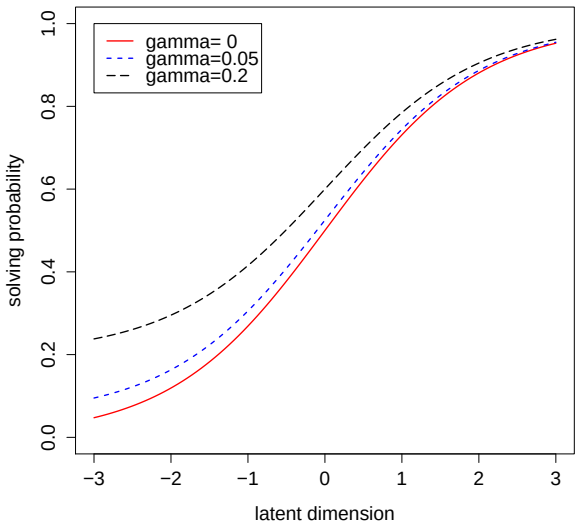

Figure 2.6.: ICCs for the 3–parameter logistic model

also the 3–PL model does not have parallel ICCs. Interpretation problems can arise here, too. Thus the 3–PL model is also analysed in the simulation study in chapter 5 as a Rasch model violation.
The Rasch model, the 2–PL and the 3–PL model are the most popular and most used IRT models in practise. But there are still many possible variations of IRT models. For further study refer e.g. to Embretson & Rice (2000) and Hullin et al. (1983).

Chapter 3

Rasch Models

The Rasch model, as it is introduced in chapter 2, is the simplest Item Response Theory model, but has the strongest assumptions and properties of all the models. The probability X_{vi} for solving an item is dependent on the ability parameter θ_v of a person and the item difficulty parameter β_i. The ability parameter θ_v stands for the position of person v on the latent trait and the difficulty parameter β_i denotes the difficulty of item i. The outcomes X_{vi} of the responses to the items are in the dichotomous case 1 for a correct answer and 0 for a false one.

3.1. The Dichotomous Rasch Model

The connection between the person and the item parameter to the solving probability is given in the dichotomous Rasch model by the logistic function. The basic equation for the Rasch model is, as mentioned before, denoted by

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{exp(\theta_v - \beta_i)}{1 + exp(\theta_v - \beta_i)} \quad , \tag{3.1}$$

or expressed differently by

$$P(X_{vi}|\theta_v, \beta_i) = \frac{exp(x_{vi}(\theta_v - \beta_i))}{1 + exp(\theta_v - \beta_i)} \quad . \tag{3.2}$$

3.1.1. Rasch Model Assumptions

The Rasch model provides the strongest model assumptions of all IRT models, thus it is a very powerful model for measurement. In the following the main model assumptions will be described.

Hambleton, Swaminathan & Rogers (1991) stated that there are three main assumptions for IRT models. Below, these statements will be specialized to the Rasch model.

1. The items on a test are statistically independent of each other after taking a fixed position of the latent trait into account
   $\Rightarrow$ *Local Independency.*
2. A single latent trait is assumed to influence the item performance
   $\Rightarrow$ *Unidimensionality Of The Trait.*
3. A mathematical function exists, that relates the probability of a person´s item–answer to the latent trait measured by the item
   $\Rightarrow$ *(Parallel) ICCs.*

1. *Local Independency*
   One main assumption for the Rasch model is the *local independency*. Statistically this is given by

$$P(X = (X_{v1} = x_{v1}, ..., X_{vk} = x_{vk})|\theta_v) = \prod_{i=1}^{k} P_i(X_{vi} = x_{vi}|\theta_v) \quad , \tag{3.3}$$

   where X is the vector of k responses X_{vi} for a given person v with ability parameter θ_v. Responses from different items are under the restriction of the person´s fixed position on the latent trait independent from each other, thus the *local independency* is also called *conditional independency* for a fixed ability. If there are correlations between responses, only variations in the latent trait will be the cause of this correlation.
   If equation (3.3) can be applied, then the items are indicators for the latent variable.
   It should be noted, that items can correlate in general, but *local independency* implies that items are uncorrelated in subpopulations with similar abilities.

2. *Unidimensionality Of The Trait*
   It is obvious in the context of the Rasch model that selected items measuring more than one latent trait are senseless since a person´s answer to a certain item cannot be linked to one latent trait.
   Formally *unidimensionality of the trait* is stated by

$$P(X_{vi} = x_{vi}|\theta_v, y_1, ..., y_w) = P(X_{vi} = x_{vi}|\theta_v) \quad , \tag{3.4}$$

where $y_1, ..., y_w$ are further variables which are not associated with x_{vi} and θ_v. Equation (3.4) declares that the variables $y_1, ..., y_w$ provide no more information to the probability for person v to solve item i. All knowledge is already given by the dependence on the one latent trait.
Of course, in practise items often refer to more than one latent trait. However it can be demonstrated, that the Rasch model can be applied in the case of sufficient unidimensionality.
This assumption of dimensionality must be provided a priori, which can be achieved e.g. with the help of a factor analysis. If the latter defined *local independency* is given *unidimensionality of the trait* can be directly concluded. This coherence can be applied also vice versa.
Both statements can be proved with the adoption of the Rasch model.

3. *Parallel ICCs*
   The solving probability can be seen as a function of the item difficulty parameter and the person´s ability parameter. As initiated in chapter 2, this connection is displayed with the Item Characteristic Curve. In the case of the Rasch model this context is provided by the logistic function.
   Referring to figure (2.4), the solving probabilities increase with rising ability. Items, which only differ in difficulty, can be recognized by the horizontal shifted curves. All items have the same discrimination, therefore the ICCs are parallel and do not cross. This graphical advice is a main assumption of the Rasch model.
   Furthermore the point of inflection for every Rasch model ICC is given at the 0.5-probability to solve an item. Hence this point can be treated as a threshold level for item difficulty.

The latter described assumptions can also be seen as model properties, therefore the Rasch model assumptions and the Rasch model properties merge.

3.1.2. Rasch Model Properties

Two further properties will expand the already explained postulates in respect to the Rasch model.

1. *The Raw Score As A Sufficient Statistic*
   The unweighted raw score

$$r_v = \sum_{i=1}^{k} x_{vi} \tag{3.5}$$

   serves as a sufficient statistic for the ability parameter θ_v in the Rasch model. Not the single "0/1" entries need to be regarded but the raw score r_v implies all information about the ability parameter θ_v, thus it is not necessary to know which items were solved from a certain person with ability θ_v. This aspect is the key point in the conditional maximum likelihood parameter estimation, where the ability parameter is eliminated by conditioning due to this sufficiency (refer here to chapter 3.1.3).

2. *Specific Objectivity*
   As mentioned in chapter 2.2, the person´s ability as well as the item difficulty parameter lies on the same scale. This scale is defined to have an interval scale level. Hence parameter values from the ability and difficulty are directly comparable. In addition both, the person parameter values are comparable among each other and the difficulty parameter values are comparable among each other, respectively.
   This leads to the property of the *specific objectivity* (Rasch, 1960). Traub and Wolfe (1981) declared, that the comparison of two items does not depend on which sample of individuals is used to evaluate them, and the comparison of two individuals is not dependent on which item set is used to test them. Statements about items can be made without respecting the taken sample. On the other hand the person´s ability can be expressed without regarding the reviewed items. Due to *specific objectivity*, person and item parameters can also be estimated independently.

The *raw score as a sufficient statistic*–property as well as the *specific objectivity*–property in the present case only refer to the Rasch model and not to any other IRT models. Thus Fischer (1995) stated that the Rasch model plays a singular role within IRT and due to its extraordinary properties is a popular model in applied psychological measurement.

3.1.3. Parameter Estimation

As mentioned in chapter 2.2 and chapter 3 there are two parameters denoting the Rasch model, i.e. the person´s ability parameter θ and the item difficulty parameter β. For

the application of the Rasch model these parameters must be estimated.
Parameter estimation in the Rasch model can be based on two different structures. The first structure depends on the person parameter estimation, namely if they are

- estimated jointly with the item parameters
- eliminated by conditioning due to the raw score as a sufficient statistic
- integrated out by marginalization.

The other structural aspect depending on parameter estimation is the type of estimation algorithm, which is used:

- maximum likelihood
- some other (heuristic) methods.

This work only refers to the conditional maximum likelihood (CML) method where the ability parameter is eliminated by conditioning due to the sufficiency of the raw score. In the following the CML approach is described elaborately. However a short explanation about joint maximum likelihood (JML), marginal maximum likelihood (MML) and other methods for item parameter estimation is provided.

1. *Conditional Maximum Likelihood*
   Parameter estimation by conditional maximum likelihood (CML) was developed by Andersen (1972,1973a). Based on *specific objectivity*, the ability parameter is eliminated by conditioning, because the raw score r_v, explained in section (3.1.2), serves as a sufficient statistic for the ability parameter θ_v.
   According to Molenaar (1995) some substitutions must be made. (Since the parameter ε, which is used in Molenaar (1995), stands in most statistical relations for an error expression, a different substitution parameter ϕ is chosen.)
   With the equations $\xi_v = \exp(\theta_v)$ and $\phi_i = \exp(-\beta_i)$ follows

$$P(X_{vi}|\theta_v, \beta_i) = \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)} \;\Rightarrow\; P(X_{vi}|\xi_v, \phi_i) = \frac{(\xi_v \phi_i)^{x_{vi}}}{1 + \xi_v \phi_i} \quad . \tag{3.6}$$

   Hence the likelihood equation is given by the products over all persons and items

$$L(\xi, \phi) = \prod_{v=1}^{N} \prod_{i=1}^{k} P(X_{vi}) = \prod_{v=1}^{N} \prod_{i=1}^{k} \frac{(\xi_v \phi_i)^{x_{vi}}}{1 + \xi_v \phi_i} \quad . \tag{3.7}$$

With some transformations this equation changes to the conditional likelihood

$$L_c(\phi|r_v) = \left(\prod_{v=1}^{N} \gamma_{r_v}\right)^{-1} \prod_{i=1}^{k} \phi_i^{x_{.i}} \quad , \tag{3.8}$$

where $x_{.i}$ are the item raw scores and γ_{r_v} the elementary symmetric function. The elementary symmetric function γ_{r_v} is related to the sum of all products of the response patterns with raw score r_v

$$\gamma_{r_v}(\phi) = \sum_{x_1,\ldots,x_k|r_v} \prod_{i=1}^{k} \phi_i^{x_i} \quad . \tag{3.9}$$

To illustrate equation (3.9) some results are given by

$$\begin{aligned} \gamma_0 &= 1 \quad , \\ \gamma_1 &= \phi_1 + \phi_2 + \phi_3 + ... + \phi_k \quad , \\ \gamma_2 &= \phi_1\phi_2 + \phi_1\phi_3 + ... + \phi_{k-1}\phi_k \quad , \\ &\vdots \\ \gamma_k &= \phi_1\phi_2...\phi_k \quad . \end{aligned} \tag{3.10}$$

As it can be seen in equation (3.8) the conditional likelihood is no longer dependent on the ability parameter θ or its substitution ξ.
In order to maximize the conditional likelihood L_c, the logarithmic version thereof as a monotonous transformation is used, because derivations are calculated easier, but the results remain the same. By evaluating all derivations with respect to ϕ_i and setting them equal to 0, the maximum likelihood estimator, which maximizes L_c is computed. The derivations of the symmetric function are calculated by

$$\frac{\partial \gamma_{r_v}(\phi)}{\partial \phi_i} = \gamma^{(i)}_{r_{v-1}} \quad , \tag{3.11}$$

where $\gamma_{r_{v-1}}$ is the elementary symmetric function of row sum r_{v-1} with ϕ/ϕ_i.
With some algebraic transformations the following equation emerges from the maximization step

$$x_{.i} - \sum_{v=1}^{N} \frac{\phi_i \gamma^{(i)}_{r_{v-1}}}{\gamma_{r_v}} = 0 \quad , \tag{3.12}$$

for $i = 1, ..., k$. The second order derivation of L_c with $\gamma^{(i,j)}_{r_{v-2}}$ proves the statement about the maximum.
Equation (3.12) must be solved iteratively. Thereby two problems can arise:

a) the evaluation of the elementary symmetric function
b) a slow convergence rate.

In this work the elementary functions are calculated by an algorithm described by Fischer (1974). It is based on the two following equations

$$\gamma_{r_v} = \phi_i \gamma^{(i)}_{r_{v-1}} + \gamma^{(i)}_{r_v} \tag{3.13}$$

$$r_v \gamma_{r_v} = \sum_{i=1}^{k} \phi_i \gamma^{(i)}_{r_{v-1}} \quad . \tag{3.14}$$

The elementary symmetric functions can easily be computed recursively. For other appropriate solving methods refer e.g. to Formann (1986), Fischer and Ponocny (1994) and Liou (1994).
Molenaar (1995) stated that nowadays the second problem with a slow convergence rate appears only with complex data sets like very long tests, complicated patterns with incomplete observations and very uncommon parameter configurations.
It should be noted that extreme response patterns like only 0–answers and only 1–answers do not contain any information for the CML estimation and therefore are excluded.
CML estimates are asymptotically consistent for $n \to \infty$ (Pfanzagl, 1994).

2. *Joint Maximum Likelihood*
The joint maximum likelihood (JML) parameter estimation method refers to the statement, that marginals are sufficient statistics for the model parameters, i.e. the item raw score, denoted as $x_{.i}$, is a sufficient statistic for the item difficulty parameter β_i and the person raw score, denoted as r_v, is a sufficient statistic for the person´s ability parameter θ_v. Under these constraints the likelihood from equation (3.7) changes to

$$L(\theta, \beta) = \frac{\left(\prod_{v=1}^{N} \exp(r_v \theta_v)\right) \left(\prod_{i=1}^{k} \exp(-x_{.i} \beta_i)\right)}{\prod_{v=1}^{N} \prod_{i=1}^{k} (1 + \exp(\theta_v - \beta_i))} \quad . \tag{3.15}$$

With the likelihood formula (3.15) the person parameter θ and the item parameter

β can be estimated concurrently.
The JML estimation procedure provides inconsistent estimates (Andersen, 1971 and 1973a; Haberman, 1977) and are therefore adaptive only for a large number of items.

3. *Marginal Maximum Likelihood*
Under the marginal maximum likelihood (MML) estimation procedure (Glas, 1989) the person parameter θ is integrated out. Thus the likelihood for MML estimation is given by

$$L(G,\beta) = \prod_{v=1}^{N}\left(\int_{-\infty}^{\infty}\prod_{i=1}^{k}\frac{\exp(x_{vi}(\theta_v-\beta_i))}{1+\exp(\theta_v-\beta_i)}dG(\theta_v)\right) \quad , \tag{3.16}$$

where $G(\theta_v)$ is the cumulative distribution function of the person parameter θ_v. Thus the person parameters from the observed individuals are a random sample from $G(\theta_v)$ (Molenaar, 1995). MML requires a postulation of the distribution for the latent trait. If this postulate is wrong, the resulting MML estimates can be inconsistent.
Also with MML persons with perfect score and zero–scores can be included in the estimation procedure. Although these persons do not have any information for the estimation of the item parameters, they contribute to find the right distribution $G(\theta_v)$.

4. *Other Heuristic Methods*
There are some heuristic methods for estimating the Rasch model parameters that do not depend on maximum likelihood. These methods do not provide an asymptotic standard error, but mostly they are much easier to compute.
Methods like *Logistic Regression With Iteratively Reweighed Least Squares* (Verhelst and Molenaar, 1988), *Explicit Method*, *Symmetrizing* and *Minchi* (Fischer, 1974) should be mentioned here.

In the following, a short survey about person parameter estimation is provided. According to Hoijtink and Boomsma (1995) four different types of estimation exist: maximum likelihood estimation (MLE), Bayes modal estimation (BME), weighted likelihood estimation (WLE) and Bayes expected a posteriori estimation (EAP). All methods despite the maximum likelihood estimation are related to the probably unknown and therefore

only assumed distribution of the person parameter θ. Thus just this estimation is explained in the following (for the rest refer to Hoijtink and Boomsma, 1995).
Assuming the item parameter to be known, e.g. estimated by the CML approach, then the maximum likelihood estimator of θ is calculated by

$$\frac{\partial \log\left(\prod_{v=1}^{N}\prod_{i=1}^{k}\frac{\exp(x_{vi}(\theta_v-\beta_i))}{1+\exp(\theta_v-\beta_i)}\right)}{\partial\theta}=0 \quad . \tag{3.17}$$

In equation (3.17) the logarithmised likelihood equation from (3.7) is differentiated with respect to θ and the resulting term is set equal to 0. After some transformations the proximate equation holds

$$r_v=\sum_{i=1}^{k}\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)} \quad . \tag{3.18}$$

With the aid of expression (3.18), the person parameter will be estimated iteratively (Fischer, 1974).
To conclude this section about parameter estimation, the CML method for item parameter estimation is recommended. Since JML provides inconsistent estimates, MML must find the right distribution for θ and other methods do not establish a standard error, the CML estimation yield for asymptotic efficient item parameter estimates in an acceptable computational time.
For person parameter estimation the maximum likelihood method is suggested for the same reasons.

3.2. The Dichotomous Mixed–Rasch Model

The mixed–Rasch model, provided by Rost (1988), is a remarkable possibility of combining Rasch and latent class analysis models. It stated that the Rasch model holds within subgroups but not for the entire population. Statements about differences within subgroups have a quantitative character, whereas statements between classes only have qualitative features.
Testing the fit of the ordinary Rasch model by the usage of the mixed–Rasch model could become one of its main applications. Therefore in the simulation study in chapter 5 the goodness–of–fit for the Rasch model is also tested with the help of the mixed–Rasch model (for theory refer to chapter 4.6). Since for this test the parameter estimations for the mixed–Rasch model parameters are needed, a short overview is given in section

(3.2.2).

3.2.1. Mixed–Rasch Model Properties

In mixture distribution models the probability function for an observed variable X is described by a number of conditional probability distribution functions. The condition, which the functions are referred to, is a mixing variable. In the context of mixed–Rasch models the mixing variable is the latent trait, expressed by the continuous person parameter θ. Hence the probability function of X is denoted by

$$P(X = x) = \sum_{c=1}^{C} \pi_c P(X = x|c) \quad , \tag{3.19}$$

where π_c is the class size parameter and $P(X = x|c)$ is given by

$$P(X = x|c) = \int_{-\infty}^{\infty} \prod_{i=1}^{k} \frac{\exp(x_i(\theta_c + \beta_{ic}))}{1 + \exp(\theta_c + \beta_{ic})} dF_c(\theta_c) \quad . \tag{3.20}$$

Like in the CML procedure (for CML estimation refer to section (3.1.3)), the latent trait parameters θ_c are eliminated by conditioning due to the sufficient statistic r for each class c. With the introduction of the elementary function (explained in section (3.1.3)) γ_{rc} for class c and the conditional score probability $P(r|c) = \pi_{r|c}$, the equation (3.20) changes to

$$P(X = x|c) = \pi_{r|c} \frac{\exp(\sum_{i=1}^{k} x_i \beta_{ic})}{\gamma_{rc}(\exp(\beta_c))} \quad . \tag{3.21}$$

This function is called the mixed–Rasch model with its three parameters, in particular the class size π_c, the latent score probabilities $\pi_{r|c}$ and the item parameters β_{ic} for each class. For parameter normalization conditions refer to Rost and von Davier (1995).
By multiplying over all outcomes of X the likelihood is described as

$$L(\pi_c, \pi_{r|c}, \beta_{ic}) = \prod_{x} \left(\sum_{c} \pi_c \pi_{r|c} \frac{\exp(\sum_{i=1}^{k} x_i \beta_{ic})}{\gamma_{rc}(\exp(\beta_c))} \right)^{n(x)} \quad , \tag{3.22}$$

where $n(x)$ denotes the observed number of response patterns x.
It should be noted that the number of classes is not a model parameter but it has to be provided a priori. By comparing the fit of different class sizes the right number of

classes can be estimated.

3.2.2. Parameter Estimation

The mixed–Rasch model parameters $\pi_c, \pi_{r|c}$ and β_{ic} can not be estimated directly. Dempster, Laird and Rubin (1977) provided an EM–algorithm to estimate such model structures. With this background Rost (1990) implemented an EM–algorithm especially for the mixed–Rasch model.
Within this procedure there are two steps, the *Estimation*–step (E–step) and the *Maximization*–step (M–step). The *Estimation*–step calculates expected pattern frequencies for each latent class c referring to preliminary estimates (in the first iteration starting values) of the model parameters

$$\hat{n}(x,c) = n(x)\frac{\pi_c P(x|c)}{\sum_{c=1}^{C}\pi_c P(x|c)} \quad , \tag{3.23}$$

where $n(x)$ is the number of expected frequencies of vector $X = x$ and $\hat{n}(x,c)$ is an estimate out of it for class c. $P(x|c)$ is given by equation (3.21).
The *Maximization*–step is done by conditional maximum likelihood estimation of the item parameter based on the expected pattern frequencies demonstrated in equation (3.23) within each class. This is done in order to obtain better estimates of $\pi_c, \pi_{r|c}$ and β_{ic}.
The ML–estimates are calculated by maximizing the log–likelihood for each latent class c

$$\ln L_c = \sum_x \hat{n}(x,c)\left[\ln\pi_{r|c} + \sum_{i=1}^{k} x_i\beta_{ic} - \ln[\gamma_r(\exp(\beta_c))]\right] \quad . \tag{3.24}$$

Setting the first derivations of equation (3.24) with respect to β_{ic} to 0, the ML-estimator results for class c

$$\hat{\beta}_{ic} = \ln\frac{n_{ic}}{\sum_{r=0}^{k} m_{rc}\gamma_{r-1}^{(i)}/\gamma_r} \quad . \tag{3.25}$$

n_{ic} stands for preliminary estimates for the number of persons with response 1 on item i in class c. m_{rc} is the estimate for the number of persons with score r in class c. The two parameters were both computed with $\hat{n}(x,c)$, provided by the E–step. The elementary

symmetric function with score $r-1$ and without item i is denoted by $\gamma_{r-1}^{(i)}$.
The score probability $\pi_{r|c}$ and the class size parameter π_c can then be estimated by

$$\hat{\pi}_{r|c} = \frac{m_{rc}}{n_c} \tag{3.26}$$

$$\hat{\pi}_c = \frac{n_c}{N} \quad , \tag{3.27}$$

where n_c is the number of persons in class c, also calculated with $\hat{n}(x,c)$, given by the E–step.
The latent trait parameter for each class θ_c can then, if required, be estimated analogously to the person parameter θ in the Rasch model in section (3.1.3).
As with CML estimation perfect and zero responses are also excluded, since they can not be assigned to a certain class and do not contain any information about the class–specific item parameters.

Chapter 4

Goodness–Of–Fit Tests For The Rasch Model

Since the development of the Rasch model (Rasch, 1960) many goodness–of–fit procedures have been designed to test the fit of the data with the Rasch model.
The null hypothesis in each test is the assumed application of the estimated Rasch model, whereas the alternative hypothesis is the rejection of the Rasch model to a level of significance at 5%, respectively.

$$H_0 : \text{Adoption of the Rasch model} \qquad H_1 : \text{Rejection of the Rasch model} \tag{4.1}$$

It should be noted, that the effect of rejecting the Rasch model is severer than rejecting a classical statistic model (Mair, 2006). More precisely when rejecting a statistical model, a more general model with more parameters can be chosen. By refusing the Rasch model, the principal assumptions of the Rasch model, e.g. specific objectivity, do not hold anymore. Hence the application of the Rasch model is no longer acceptable. Of course, a more parametric model such as the 2–PL Birnbaum model can be applied, but the main properties of the Rasch model are lost. Therefore the Rasch model is still a model with high scientific claims.
There are different approaches to arrange a taxonomy of tests for model fit (Glas & Verhelst, 1995). One taxonomy refers to the assumptions and properties of the model to be tested. Dissimilar forms of model violations, like non–parallel item characteristic curves, no unidimensionality or no local independency can be detected with these tests. There is no possibility to separate these assumptions to detect the corresponding violations. Hence the tests often have power for violations of more than one assumption.
The second taxonomy relates to the mathematical sophistication of the procedure, particularly to the knowledge of the distribution of the test statistic. From this point of view the tests can be divided into two classes. In the first class the distribution of the

test statistic is (asymptotic) known, whereas in the other class the distribution must be approximated.
The third taxonomy relies on the type of statistic which is used. This grouping form relates to the same classification as in discrete statistical models. The main focus here is the type of family, which the test statistic belongs to. This taxonomy is chosen in this work, because it follows the typical statistical partitioning scheme. The main global test procedures for model fit will be presented in the following sections.

4.1. Pearson-Type Tests

All these tests follow the formula

$$\chi^2 = \sum \frac{(o-e)^2}{e}, \tag{4.2}$$

where o represents the observed frequencies and e the expected frequencies. This test statistic is asymptotically χ^2–distributed. Since often the expected frequencies are very low, the asymptotic to a χ^2–distribution can be doubted (for further considerations refer to section (4.5)).

4.1.1. Martin–Löf Test

The Martin–Löf test is sensitive to the violation of strictly monotonically increasing and parallel ICCs.
Let n_{ir} be the observed frequency of correct answers to item i for those persons who have r correct answers. With the background of CML estimation the conditional probability that a person answers item i with raw score r correctly is

$$\frac{\beta_i \gamma_{r-1}^{(i)}}{\gamma_r} \quad , \tag{4.3}$$

where $\beta = (\beta_1, ..., \beta_k)$ is the vector of item parameters. γ_r represents the elementary symmetric function according to the raw score r and $\gamma_{r-1}^{(i)}$ the first derivative of the symmetric function to the raw score r with respect to each item without item i (for more detailed information about the CML estimation refer to chapter 3.1.3). Therefore, if the model fits the data the following equation holds

$$n_{ir} \approx n_r \frac{\beta_i \gamma_{r-1}^{(i)}}{\gamma_r} \quad , \tag{4.4}$$

where n_r denotes the number of persons with r correct answers.
From this approximation a χ^2–sum is generated with the deviations of observed and predicted frequencies.
Let $q_r^T = (n_{1r}, n_{2r}, ..., n_{kr})$ be the corresponding vector of observed frequencies over all k items and $t_r^T = (n_r \frac{\beta_1 \gamma_{r-1}^{(1)}}{\gamma_r}, n_r \frac{\beta_2 \gamma_{r-1}^{(2)}}{\gamma_r}, ..., n_r \frac{\beta_k \gamma_{r-1}^{(k)}}{\gamma_r})$ be the vector of the expected frequencies.
Then the χ^2 test statistic is

$$T = \sum_{r=1}^{k-1} (q_r - t_r)^T V_r^{-1} (q_r - t_r) \quad , \tag{4.5}$$

where V_r is a $k \times k$ variance–covariance matrix with the following elements:

$$n_r \frac{\beta_i \gamma_{r-1}^{(i)}}{\gamma_r}, \quad \text{for} \quad i = j$$
$$n_r \frac{\beta_i \beta_j \gamma_{r-2}^{(ij)}}{\gamma_r}, \quad \text{for} \quad i \neq j \quad .$$

T is asymptotically χ^2–distributed with $(k-1)(k-2)$ degrees of freedom. The null hypothesis, i.e. that the Rasch model holds, has to be rejected to a level of significance for 5% when $T > \chi^2_{1-\alpha}$ with $df = (k-1)(k-2)$. The parameter α stands for the level of significance.
Glas (1981 (not in Bibliography, because in Dutch language), 1988) designed the R_{1c}–test and showed that the Martin–Löf test is equivalent to it (Glas, 1981). For this reason T was transformed to R_{1c}, because the R_{1c}–test fits into the framework of generalized Pearson statistics. This leads to a variety of applications in which the test can be used. In section (4.1.3) this R_{1c}–test will be explained.

4.1.2. Q_1–Test

The Q_1–test was developed by van den Wollenberg (1979). It is also sensitive to violations of strictly monotonically increasing and parallel ICCs.
For this test, as well as for the three following tests, a stochastic variable M_{1gi} is defined with its realization m_{1gi}. This realization is the count of persons belonging to score g and giving a right response to item i. Given k items, there are $k-1$ score groups.
Let $\mathbb{E}(M_{1gi}|\hat{\omega}, \hat{\beta})$ be the CML expected value of this variable. This is the expected value given the frequency distribution of the persons sum scores $\hat{\omega}$ and the CML estimates of the item parameters $\hat{\beta}$.

The test is based on the first–order frequencies, namely the deviation of the observed and expected frequencies.

$$d^*_{1gi} = m_{1gi} - \mathbb{E}(M_{1gi}|\hat{\omega}, \hat{\beta}) \tag{4.6}$$

To obtain the test statistic of the Q_1–test these differences are divided by their estimated standard deviation and become a standardized binomial variable z_{1gi}. Therefore the test statistic is given by

$$Q_1 = \frac{k-1}{k} \sum_{i=1}^{k} \sum_{g=1}^{G} z^2_{1gi} \quad , \tag{4.7}$$

for item i. This statistic is asymptotically χ^2–distributed with $(G-1)(k-1)$ degrees of freedom. The null hypothesis will be rejected if $Q_1 > \chi^2_{1-\alpha}$ with $df = (G-1)(k-1)$.

4.1.3. R_1–Test

The R_1–test is also based on first–order frequencies, but additionally takes the covariance of the deviations into account. As with the latter tests, this test is sensitive to non-parallel ICCs (Glas, 1988).
Depending on the type of parameter estimation two tests have been proposed. The R_{1c}–test relates to CML estimation and the R_{1m}–test is based on MML estimation. Because this work only relies on CML estimation refer to Glas and Verhelst (1995) for full details of the R_{1m}–test.
If the single deviations in (4.6) are dependent over all items i, the χ^2–distribution can be doubted. Therefore Glas (1988) developed this test and included the covariance structure into the test statistic.
Let d_{1g} be the vector of elements $d_{1gi} = d^*_{1gi}/\sqrt{n}$, where n equals the sample size. Further let W_{1g} be the corresponding estimated variance–covariance matrix. Then the test statistic is denoted by

$$R_{1c} = \sum_{g=1}^{G} d^{\mathrm{T}}_{1g} W^{-}_{1g} d_{1g} \quad . \tag{4.8}$$

The regularity of W is not always guaranteed, so it is proposed to use the generalized inverse W^- (Glas and Verhelst, 1995).
The R_{1c}–test statistic is asymptotically χ^2–distributed with $(G-1)(k-1)$ degrees of

freedom and again, the null hypothesis is rejected if $R_{1c} > \chi^2_{1-\alpha}$ with $df = (G-1)(k-1)$.

4.1.4. Item-Oriented Tests U_i, S_i

The tests presented here are based on the same differences as the tests before, but explicitly focus on specific items. As mentioned in section (4.1.2) the variable z_{1gi} is a standardized binomial variable. These scaled differences can be used as a diagnostic tool for analysing variations in item discrimination.
Molenaar (1983) developed a statistic where the outcome signals whether the discrimination of the item is too high or too low. Let $c1$ and $c2$ denote cut-off points which divide the score into a low, middle and high area. These boundaries are usually chosen in such a way, that each summation over them includes 25% of the persons sampled, i.e. the low as well as the high area contains 25 % of the individuals.
Then the test statistic is given by

$$U_i = \frac{\sum_{g=1}^{c_1} z_{1gi} - \sum_{g=c_2}^{k-1} z_{1gi}}{(c_1 + k - c_2)^{1/2}}. \tag{4.9}$$

U_i is approximately standard normal distributed and the null hypothesis, namely that the item is Rasch conform, is rejected if $U_i > z_{1-\alpha}$. U_i is only approximately normal distributed because the values of z_{1gi} rely on parameter estimates and are therefore not independent.
This test has also been transformed into the framework of generalized Pearson tests (Verhelst and Eggen, 1989 (not in Bibliography because of Dutch language); Verhelst, Glas and Verstralen, 1994). The differences d_{1gi} are squared and scaled by a variance–covariance matrix. This results in a statistic with a χ^2–distribution with 1 degree of freedom.
The last test statistic based on a specific item is S_i (Verhelst and Eggen, 1989; Verhelst, Glas and Vestralen, 1994, Glas and Verhelst, 1995). It relies on the same deviations between observed and expected responses in homogeneous score groups as the R_{1c}-test, but operates on item level. The test has like the U_i statistic power for differences in item discrimination.
Denote d_i as the vector of the elements d_{1gi}. Thus the test statistic is given by partitioning every score level into equivalent classes and computing the difference between the observed and expected number of correct responses to an item i.

$$S_i = d_i^{\mathrm{T}} W_i^{-} d_i \tag{4.10}$$

S_i is asymptotically χ^2–distributed with $(G-1)$ degrees of freedom. Again the null hypothesis is rejected if $S_i > \chi^2_{1-\alpha}$ with $df = (G-1)$.

4.1.5. Q_2–Test

The Q_2–test as well as the previously described R_2–test (section (4.1.6)) are sensitive to violations of unidimensionality and stochastic independence assumption. Under unidimensionality and fix conditions of one person´s position on one latent trait the association between the items must disappear. If this is not the case a second dimension can play a role in the questionnaire. Therefore tests for unidimensionality must rely on the association between items. The Q_2–test as well as the R_2–test are therefore based on second–order statistics.
Again a stochastic variable M_{2gij} is defined. Its realization m_{2gij} is the number of correct answers to items i and j in the subsample g. The CML expected value is denoted by $\mathbb{E}(M_{2gij}|\hat{\omega}, \hat{\beta})$. Thus

$$d_{2gij} = m_{2gij} - \mathbb{E}(M_{2gij}|\hat{\omega}, \hat{\beta}) \quad . \tag{4.11}$$

The standardized variable z_{2gij} is computed by dividing the latter differences by their standard deviations. The test statistic is denoted by

$$Q_2 = \frac{k-3}{k-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{g=1}^{G} z^2_{2gij} \quad . \tag{4.12}$$

Estimating the item parameter in each subgroup yields to a good approximation of a χ^2–distribution. In this cases van den Wollenberg (1982) showed, that Q_2 approximates a χ^2–distribution with $df = Gk(k-3)/2$ degrees of freedom. If $Q_2 > \chi^2_{1-\alpha}$ with $df = Gk(k-3)/2$, the null hypothesis must be rejected.

4.1.6. R_2–Test

As mentioned before the R_2–test is also sensitive for violations of unidimensionality and stochastic independence.
Also the R_2–test can be constructed under different forms of parameter estimation. The R_{2c}–test belonging to CML parameter estimation will be explained in the following. For the R_{2m}–test related to MML estimation refer again to Glas and Verhelst (1995).
Based on second-order statistics this test belongs not to a partition of the sample into

subsamples but to the covariance between the following differences

$$d^*_{2ij} = m_{2ij} - E(M_{2ij}|\hat{\omega}, \hat{\beta}) \quad . \tag{4.13}$$

m_{2ij} denotes the number of persons who answered both items i and j correct. Further let d_2 be the vector of elements $d_{2ij} = d^*_{2ij}/\sqrt{n}$ and W_2^{-1} be the inverse of the estimated covariance matrix from the differences of all pairs of items. Then the test statistic is given by

$$R_{2c} = d_2^{\mathrm{T}} W_2^{-1} d_2 + d_1^{\mathrm{T}} W_1^{-1} d_1 \quad . \tag{4.14}$$

To simplify the derivation of the distribution of this statistic the quadratic form of persons getting only one item correct, namely $d_1^{\mathrm{T}} W_1^{-1} d_1$, is also taken into account.
R_{2c} is asymptotically χ^2–distributed with $k(k-1)/2$ degrees of freedom. The null hypothesis will be rejected if $R_{2c} > \chi^2_{1-\alpha}$ with $df = k(k-1)/2$.

4.2. Likelihood Ratio Tests

The principle of the likelihood ratio test, denoted by LR, is to analyse the ratio between a full model and a reduced model.

$$LR = \frac{L(\hat{\beta}_0; y)}{L(\hat{\beta}_1; y)} \tag{4.15}$$

$\hat{\beta}_0$ implies the ML estimator of the full model, whereas $\hat{\beta}_1$ is the ML estimator of the reduced model. The corresponding values of the likelihood function are represented by $L(\hat{\beta}_0; y)$ and $L(\hat{\beta}_1; y)$, respectively.
As in chapter 3.8 with the CML estimation also here, the logarithmic version as a monotonous transformation of LR is used normally. Then the ratio becomes a deviation and the asymptotic distribution of the LR statistic is a χ^2–distribution, while all other properties of the LR term remain the same.
In the following, the Andersen LR test and the Martin Löf LR test will be explained. Both tests belong to the CML context.

4.2.1. Andersen LR Test

The Andersen likelihood ratio test is based on the main property of the Rasch model, the specific objectivity. Statements about persons and items can be made independently from the drawn sample. This test is the most commonly used test and is available in most Rasch softwares.
The data is divided into subgroups. The splitting criterion can be formed on the basis of score levels or on the basis of external criteria.
The conditional likelihood for each subgroup and for the whole data is generated. According to specific objectivity, the ML estimator for the subgroups must not differ (apart from some random deviances) from the ML estimator for the whole data, i.e. the estimated item parameter $\hat{\beta}$ for the whole sample X remains constant in the estimated item parameter $\hat{\beta}_g$ for the subsamples X_g

$$\hat{\beta} = \hat{\beta}_1 = \hat{\beta}_2 = ... = \hat{\beta}_G \quad . \tag{4.16}$$

Andersen (1973b) proposed a splitting criterion according to raw scores into two subgroups. Thus the one subgroup contains persons with higher scores and the other subgroup consists of lower capable individuals. But on that account different splitting criterions like gender or the division after the mixed–Rasch model (section (4.6)) can be used as well.
Thus the test statistic is given by

$$LR = 2\left(\sum_{g=1}^{G} \ln L_c(\hat{\beta}_g; X_g) - \ln L_c(\hat{\beta}; X)\right) \quad , \tag{4.17}$$

LR is asymptotically χ^2–distributed. The degrees of freedom are equal to the number of parameters estimated in the subgroups minus the number of parameters estimated in the Rasch model, i.e. $df = G(k-1)-(k-1) = (G-1)(k-1)$. Thus the null hypothesis will be rejected if $LR > \chi^2_{1-\alpha}$–distribution with $df = (G-1)(k-1)$.
Andersen (1973b) stated that this test is sensitive to violations of the property from parameter invariance and of the property from parallel ICCs. A simulation study from van den Wollenberg (1979) confirms this statement. Mead (1976) and Gustaffson (1977) analysed the sample size, besides Gustaffson (1980) showed that the LR-test has power for violations based on the 2–PL and the 3–PL models.
With the simulation study in chapter 5 of this work, the Andersen LR test will be

analysed in detail.

4.2.2. Martin Löf LR Test

Although Martin Löf´s LR test (1973, not in Bibliography, because in Swedish language) was constructed to test whether two sets of items form a Rasch scale, it can also be seen as a test for the assumption of unidimensionality.
The items are split into two subsets of k_1 and k_2 items, where $k_1 + k_2 = k$.
$r = (r_1, r_2)'$ with $r_1 = 0, ..., k_1$ and $r_2 = 0, ..., k_2$ denotes the score patterns of the two subsets and n_r the number of persons obtaining score pattern r. r stands for the total sum score of a person´s response pattern, thus $r = r_1 + r_2$. In comparison to the other already mentioned χ^2 tests, the division into subgroups under the alternative hypothesis does not follow only one but two latent dimensions, i.e. one persons total score r represents a subscore r_1 on the one latent variable and a subscore r_2 on the other latent variable. Hence, this test is also sensitive to the violation of unidimensionality but only as far as it is dependent on the chosen item grouping.
The test statistic for the Martin Löf LR test is given by

$$LR = 2\left(\sum_r n_r \ln\left(\frac{n_r}{N}\right) - \sum_r n_r \ln\left(\frac{n_r}{N}\right) - \ln L_c + \ln L_c^{(1)} + \ln L_c^{(2)}\right). \tag{4.18}$$

L_c, $L_c^{(1)}$ and $L_c^{(2)}$ are the likelihood functions calculated by using CML estimates in the complete case, the first subgroup, and the second subgroup, accordingly. Equation (4.18) is under the null hypothesis, i.e. the items construct a Rasch scale, asymptotically χ^2–distributed with $k_1 k_2 - 1$ degrees of freedom.
The first two sums in equation (4.18) are needed because L_c is conditioned on the frequency distribution of scores on the complete test, while $L_c^{(1)}$ and $L_c^{(2)}$ are only conditioned on the frequency distribution of the scores on the subtests.
The null hypothesis will be rejected if $LR > \chi^2_{1-\alpha}$ with $df = k_1 k_2 - 1$.

4.3. Wald–Type Tests

Wald–type tests have a lot in common with the Likelihood Ratio tests explained in section (4.2). With defining a general model it is tested whether a certain restriction holds. This means in the context of Rasch models that parameter estimates of two meaningful sample subgroups are compared. This section will only refer to two subgroups, but the

generalization to more subgroups is straightforward.
Like the Likelihood Ratio tests also Wald–type tests are sensitive to violations of the property of invariance of parameters and the property of parallel ICCs.
Let $\beta_g = (\beta_{g1}, ..., \beta_{gm})^T, g = 1, 2$, be the model parameters for the g–th subgroup and m the number of free parameters. In the context of CML estimation $m = k - 1$. Note that the parameter for the k–th item is set to zero and the parameters $\beta_{g1}, ..., \beta_{g,k-1}$ are the difficulty parameters of the items 1 to $(k - 1)$ in subgroup g, respectively.
Let $\beta^T = (\beta_1^T, \beta_2^T)$. Thus the null hypothesis, i.e. that the Rasch Model holds, can be stated as

$$h_j(\beta) = \beta_{1j} - \beta_{2j} = 0, \quad j = 1, ...q \quad . \tag{4.19}$$

Therefore, the whole restriction vector is denoted by $h(\beta)^T = (h_1(\beta), ..., h_q(\beta))$. Let $\Sigma_g, g = 1, 2$, be the corresponding variance–covariance matrix of the ML estimator of β_g. If the responses between the two subgroups are independent, the variance–covariance matrix will be given by

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \quad . \tag{4.20}$$

Let $T(\beta)$ be the $2m \times q$ matrix $[t_{gi}]$ of the partial derivations of $h_j(\beta)$ regarding β_g. More precisely t_{gi} is given by

$$t_{gi} = \frac{\partial h_j(\beta)}{\partial \beta_g} \quad . \tag{4.21}$$

Then the Wald test statistic is denoted by the quadratic form

$$W = h^T(\hat{\beta})[T^T(\hat{\beta})\Sigma T(\hat{\beta})]^{-1}h(\hat{\beta}) \quad . \tag{4.22}$$

The test statistic is asymptotically χ^2–distributed with q degrees of freedom. In the Rasch context with CML estimation and two subgroups the degrees of freedom are $df = k - 1$.
Glas and Verhelst (1995) stated that one should be careful with interpreting the differences between groups of item parameter estimates. The origin and unit which are chosen for the scale in the two subgroups must be equal, otherwise an interpretation is senseless. Hence it is necessary to implement some restrictions, which are, unlike in equation

(4.19) independent from the chosen normalization in both subgroups. One possibility to achieve this is, that the restrictions are functions of the differences between a particular item parameter estimate and the estimates of the other item parameters, e.g.

$$h_{1i} = \sum_{j \neq i} \left(\hat{\beta}_{1i} - \hat{\beta}_{1j} \right) - \sum_{j \neq i} \left(\hat{\beta}_{2i} - \hat{\beta}_{2j} \right) = 0 \ , \quad (i = 1, ..., k) \quad . \tag{4.23}$$

Because these sums can be both positive and negative and therefore some terms might be canceled, a further restriction is needed:

$$h_{2i} = \sum_{j \neq i} \left(\hat{\beta}_{1i} - \hat{\beta}_{1j} \right)^2 - \sum_{j \neq i} \left(\hat{\beta}_{2i} - \hat{\beta}_{2j} \right)^2 = 0 \ , \quad (i = 1, ..., k) \quad . \tag{4.24}$$

If the scale is chosen that the sum of the parameters is zero, the terms (4.23) and (4.24) will be reduced to

$$h_{1i} = 2(\hat{\beta}_{1i} - \hat{\beta}_{2i}) = 0 \quad , \tag{4.25}$$

and

$$h_{2i} = \left(k\hat{\beta}_{1i}^2 + \sum_j \hat{\beta}_{1j}^2 \right) - \left(k\hat{\beta}_{2i}^2 + \sum_j \hat{\beta}_{2j}^2 \right) = 0 \quad , \tag{4.26}$$

for $i = 1, ..., k$, respectively. Relating to these results, one can see that the test will be sensitive to differences in the variance to item parameter estimates as well as to the square thereof. To adjust equation (4.25) in the general framework of Wald type tests, the test statistic W_{1i} is denoted by

$$W_{1i} = \frac{(\hat{\beta}_{1i} - \hat{\beta}_{2i})^2}{\sigma_{1i}^2 + \sigma_{2i}^2} \quad , \tag{4.27}$$

where σ_{1i} and σ_{2i} represent the i–th diagonal element of the variance–covariance matrix of a solution normalized to a zero sum in both groups. W_{1i} is asymptotically χ^2–distributed with 1 degree of freedom.
According to equation (4.27) also the second term, stated in (4.26), can be transformed into the general expression for the Wald statistic. Thus W_{2i} is then given by

$$W_{2i} = \frac{h_{2i}^2}{t_1^T \Sigma_1 t_1 + t_2^T \Sigma_2 t_2} \quad , \tag{4.28}$$

where t_g, g=1,2 is the k–dimensional vector with

$$t_{gi} = 2(1 + \delta_{ij}k)\beta_{gj} \quad , \tag{4.29}$$

and δ_{ij} stands for the Kronecker symbol.

4.4. Nonparametric Tests

Almost all of the available goodness–of–fit tests are based on parameter estimation and asymptotic distribution assumptions, mostly the χ^2–distribution. These tests need large calibration samples for the parameter estimation. To avoid this, some exact nonparametric tests were implemented which do not refer to any asymptotics or parameter estimation.

The first nonparametric approach was presented by Rasch (1960). It is based on the same statement as the CML estimation, namely that the marginal sums from a data matrix X, i.e. the row and column sums, are sufficient statistics of X. This means the probability of X does not depend on the single 0/1 elements from X, but only on their given marginals.

If the Rasch model holds, all data matrices X with the same marginals as the observed one are equally likely. Formally, the conditional distribution of X given the marginal sums is uniform:

$$P(X = x | X_{i+} = x_{i+}, X_{+j} = x_{+j} \quad \text{for} \quad i, j = 1, ..., g) = \frac{1}{N((x_{i+}), (x_{+j}))}, \tag{4.30}$$

where x_{i+} are the row sums, x_{+j} are the column sums and $N((x_{i+}), (x_{+j}))$ is the number of all data matrices X with the same marginals. Sampling from this conditional distribution one can approximate the distribution of the null hypothesis of any unknown test statistic from the data matrix X and thus construct a nonparametric test of the Rasch model (Besag & Clifford, 1989).

Unfortunately, in respect to computational efficiency no algorithm has been found so far to create all possible data matrices with the same marginals. However some procedures for sampling from the sample space in a nonuniform way have been implemented and algorithms for correcting the distribution from the deviance to uniformity have been presented.

These procedures can be divided into two classes. The one class samples matrices independent from a nonuniform distribution with the help of "importance sampling"

(Snijders, 1991; Chen et al., 2005; Chen & Small, 2005).
Snijders (1991) suggested to generate data matrices from a proposal distribution differing from the desired uniform distribution and then load these matrices with an importance weight. Any test statistic under the uniform distribution can then be generated with this weighted average. Since it is not easy to find an appropriate proposal distribution, especially for high-dimensional problems such as sampling 0/1 matrices with given marginals, Snijder (1991) and Chen & Small (2005) use "sequential importance sampling", in which the generation of the data matrices are made column by column. Proposal distributions for every column can then be found without emerging problems.
Chen & Small (2005) improved this algorithm to make it faster and to achieve the Monte Carlo standard error in such a way that the user can choose the number of simulations.
The other class is based on Markov Chain Monte Carlo (MCMC) applications, in which the data matrices in the sample space are regarded as states. The transition probabilities from one state to another are given by the sampling scheme. Under specific conditions the distribution of various states converges to a stationary distribution. To get close to this stationary distribution a burn–in phase is needed, i.e. a series of sampled matrices, which are dropped before the point when the stationary distribution is reached. However the sampled matrices are not independent.
These methods have been studied by Connor and Simberloff (1979), Besag and Clifford (1989), Roberts and Stone (1990), Rao et al. (1996), Ponocny (2001) and recently by Verhelst (2008).
Ponocny (2001) uses a variation of the MCMC algorithm. According to the Markov Chain he generated 0/1 matrices, but the stationary distribution is not uniform. By taking a weighted average of the samples from the Markov chain he achieves a consistent estimate of the p–value of any test statistic. But the generation of these weights takes a long time.
Recently, Verhelst (2008) proposed an MCMC algorithm, which is much faster than the existing ones and can be used for larger matrices. Also in this case the stationary distribution is not uniform. Like in Ponocny´s method (2001) the importance sampling algorithm is used, where the stationary distribution gets nearly uniform. In addition a Metropolis–Hastings algorithm is used to get the stationary distribution uniform.

4.5. The Parametric Bootstrap

Most of the goodness–of–fit test statistics for the Rasch model are based on a asymptotic χ^2–distribution. But some assumptions must suite to confirm the use of this distribution.
The number of items as well as the number of response categories, in the dichotomous Rasch model i.e. two categories, define the number of all possible response patterns in a model for categorical data. Even if there are large sample sizes, the number of possible response patterns will easily exceed the number of observed response patterns. Thus most of the observed frequencies in the contingency table will be zero. The table is denoted to be sparse (Agresti & Yang, 1986).
To support this statement the rule of thumb of a minimum expected frequency is denoted by $MIN_x(E(x)) \geq 5$. Even simplified practice rules of thumb (Read & Cressie, 1988) like $MIN_x(E(x)) \geq 1$ are far too high for the sparse data. Since many of the possible response patterns are not observed and therefore also the expected frequencies are very small, the asymptotic for a χ^2–distribution is not given, thus the distribution of the test statistic under the null hypothesis is unknown.

4.5.1. The Bootstrap Draft

A good solution to this problem is to use the parametric Bootstrap (Efron, 1979 & 1982) for goodness–of–fit testing (Bollen & Stine 1993; Langeheine et al., 1996; von Davier, 1997; Tollenaar & Mooijaart, 2003). This method provides an empirical distribution that represents the distribution of the test statistic under the null hypothesis.
Von Davier (1997) stated, that parametric IRT models as well as Latent Class Analysis can be tested with the parametric Bootstrap goodness–of–fit, as they provide a probability for each response in the observed data matrix. Thus the dichotomous Rasch model can be tested by the use of the parametric Bootstrap goodness–of–fit.
The parametric Bootstrap test can be based on any goodness–of–fit statistic. Mostly applied are members of the so called power–divergence family $CR(\lambda)$ (Cressie & Read, 1984), which is denoted by

$$CR(\lambda) = \sum_{i=1}^{m^k} \frac{1}{\lambda(\lambda+1)} O(x_i) \left[\left(\frac{O(x_i)^\lambda}{E(x_i)} \right) - 1 \right] \quad , \tag{4.31}$$

where $O(x_i)$ is the observed frequency of response pattern x_i and $E(x_i)$ the related expected frequency. Also the Pearson χ^2 statistic, as well as the likelihood ratio statistic are members of this family (critical results from simulation studies to these statistics are given in chapter 5.2.7.
The procedure for the parametric Bootstrap is given by:

1. The observed test statistic, denoted by T_{obs}, is computed with the given data.
2. Item as well as person parameters from the observed data are estimated.
3. With these parameters a new data set is simulated (refer here to Section (4.5.2)).
4. For these simulated data the test statistic T is calculated.
5. Steps (3) and (4) are repeated B times to define the empirical distribution for the test statistic under the null hypothesis.

If the observed test statistic is larger than the $(1-\alpha)B$–th percentile of the ordered Bootstrap statistics, then the model must be rejected to a level of significance for 5%. Thus the p–value of the observed test statistic is then estimated by

$$\frac{1+\sum_{i=1}^{B} I(T_i > T_{obs})}{(B+1)} \quad , \tag{4.32}$$

where $I(T_i > T_{obs})$ is the indicator function for the number of $T_i > T_{obs}$.
The number B of Bootstrap replications must be chosen with respect to the aim of the Bootstrap (Efron & Tibshirani, 1993). If the Bootstrap is used to estimate the distribution of a test statistic or to estimate confidence intervals, B has to be very high. If the aim of the Bootstrap is to estimate a standard error, the number of the replications can be quite small. Most of the simulation studies for testing the goodness–of–fit in the Rasch model uses a B of 1000 (e.g. von Davier, 1997; Tollenaar & Mooijaart, 2003). Also in this work B is set to 1000 replications.
The duration of a Bootstrap goodness–of–fit test is very long and highly computer intensive. Not only the simulation of new data sets, but mainly the estimation of parameters for calculating the test statistic in each set boosts the computational time (Langeheine et al., 1996; von Davier, 1997).

4.5.2. Simulation Data Matrices With Given Marginals

In mostly used Bootstrap goodness–of–fit tests for the Rasch model the simulation of new data sets with estimated parameters is done without respecting the given marginals of

the data. But in the context of CML estimation the main issue on parameter estimation refers to the person scores as a sufficient statistic for the single 0/1 entries of the data matrix. Clemens Draxler came up with the idea to include this aspect in the Boostrap routine. Thus a new simulation algorithm is implemented in this work, which generates the new data matrices with respect to the given marginals, i.e. the person scores.
As pointed out in section (3.1) the solving probability for a certain item i and person v within the Rasch context is

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{exp(\theta_v - \beta_i)}{1 + exp(\theta_v - \beta_i)} \quad , \tag{4.33}$$

which below is stated as p_i for one person v.
The following equations refer only to one person from the sample, i.e. one row in the data matrix. According to that, the whole data matrix can be built up row-wise.
The solving probabilities given the marginals can therefore be seen as a function of the person score and the unconditional solving probabilities from equation (4.33).
The unconditional probability for the person´s score (here denoted as c) is given by

$$P(S = c) = \sum_{(x_1,...,x_k)|\sum x_i = c} \prod_{i=1}^{k} p_i^{x_i}(1 - p_i)^{(1-x_i)} \quad , \tag{4.34}$$

where p_i is the probability from equation (4.33) and x_i is out of (0,1). $P(S = c)$ will be denoted as SFK in the following:

$$P(S = c) = SFK(p_1, ..., p_k, c) \quad . \tag{4.35}$$

More precisely, imagine a response pattern like "1100". The corresponding SFK will then be calculated by

$$\begin{aligned} SFK(p_1, p_2, p_3, p_4, 2) =& p_1p_2(1 - p_3)(1 - p_4) + p_1(1 - p_2)p_3(1 - p_4) + \\ & p_1(1 - p_2)(1 - p_3)p_4 + (1 - p_1)p_2p_3(1 - p_4) + \\ & ... + (1 - p_1)(1 - p_2)p_3p_4 \quad . \end{aligned} \tag{4.36}$$

With the common probability to solve e.g. the first item and the given person score, the conditional probability to solve this item conditional on the person score follows

$$P(X_1 = 1|S = c) = \frac{P(X_1 = 1 \wedge S = c)}{P(S = c)} = \frac{p_1 \cdot SFK(p_2, ..., p_k, (c - 1))}{SFK(p_1, ..., p_k, c)} \quad . \tag{4.37}$$

Using these probabilities the algorithm for sampling with given marginals can be constructed as follows:

- Sample X_1 from $P(X_1 = 1|S = c)$ and then successively
- X_2 from $P(X_2 = 1|S = c, X_1 = x_1)$
- X_3 from $P(X_3 = 1|S = c, X_1 = x_1, X_2 = x_2)$
- ...
- X_i from $P(X_i = 1|S = c, X_1 = x_1, X_2 = x_2, ..., X_{i-1} = x_{i-1})$.

The general recursive algorithm is thus given by

$$P\left(X_i = 1|S = c, \sum_{j=1}^{i-1} X_j = d\right) = \frac{p_i \cdot SFK(p_{i+1}, ..., p_k, (c-d-1))}{SFK(p_i, ..., p_k, (c-d))} \quad . \tag{4.38}$$

Adopting this algorithm to every person of the sample, i.e. to every row of the data matrix, a simulated data matrix with the same person scores as the observed matrix results. The 0/1 entries of the response matrix are generated by the decision rule

$$X_{vi} = \begin{cases} 1 & p_{vi} = P\left(X_{vi} = 1|S_v = c, \sum_{j=1}^{i-1} X_{vj} = d\right) \geq p^*_{vi} \\ 0 & p_{vi} = P\left(X_{vi} = 1|S_v = c, \sum_{j=1}^{i-1} X_{vj} = d\right) < p^*_{vi} \end{cases} , \tag{4.39}$$

where p^*_{vi} are the entries of a $n \times k$ matrix P^* generated by random deviates of the standard uniform distribution.
In this work the parametric Bootstrap goodness–of–fit test is used in the simulation study in chapter 5 with the latter defined method to reproduce data matrices with given marginals. As test statistic the Pearson χ^2 statistic is used based on the observed pattern frequencies $O(x)$ and the expected pattern frequencies $E(x)$

$$\chi^2 = \sum_{i=1}^{m^k} \frac{(O(x_i) - E(x_i))^2}{E(x_i)} \quad . \tag{4.40}$$

4.6. Mixed–Rasch Model Test

In this work the concept of the mixed–Rasch model is used to determine the fit of data to the Rasch model.
Analogically to the likelihood–ratio test in section (4.2.1), the mixed–Rasch model test is also based on the main assumption of the Rasch model, namely specific objectivity. Within the subsamples the item parameter estimates must not differ, except from a sampling error, over these groups.
The main challenge here is to find the best fitting partition with given heterogeneity of the individuals. This division can be found with the help of the mixed-Rasch model. The latent classes of individuals represent those latent subpopulations that maximize the person´s homogeneity within classes, and heterogeneity between classes. As a result, classes of individuals are obtained where the item parameter estimates maximally differ between the classes (Rost & von Davier, 1995).
Hence the mixed–Rasch model test in the present case, which compares the one class with the two class solution, must be at least as powerful as the likelihood ratio test by Andersen with any other obvious two–class division of the sample.
Thus the test statistic is given by

$$LRMR = 2\left(\sum_{g=1}^{G} lnL_c(\hat{\beta}_g; X_g) - lnL_c(\hat{\beta}; X)\right) \quad , \tag{4.41}$$

where the splitting criterion is not due to raw scores, but to the class size parameters provided by the mixed–Rasch model.
This test statistic is assumed not to be asymptotically χ^2–distributed. Since one class size parameter must be set to 0 to obtain the Rasch model as a nested model of the mixed–Rasch model, the parameter space is not open and therefore the asymptotic to a χ^2–distribution is, unlike in the Andersen LR–test, not given anymore. Thus the distribution of the test statistic under the null hypothesis should be seen as unknown.
In the simulation study in chapter 5 of this work this mixed–Rasch model test will be taken up again.

Chapter 5

Simulation Studies

Three different goodness–of–fit tests will be investigated for performance in this chapter, namely the Andersen likelihood ratio test (for theory information see chapter 4.2.1), the Bootstrap–test (for theory information see chapter 4.5) and the mixed–Rasch model test (for theory information see chapter 4.6).
The simulation studies are based on the Monte Carlo simulation technique. Results are obtained by repeated random sampling from different inputs.
Here results from the Monte Carlo simulation are the type–one error rates for the Rasch conform data and the power for the Rasch violated data. Rasch conform and Rasch violated data matrices serve as inputs for the simulation. Variations in sample sizes and number of items as well as variations within different forms of model violations generate divers scenarios, in order to analyse the performance of the chosen tests. For each test scenario 100 replications, i.e. 100 varying simulated data matrices, are constructed to assign the corresponding type–one error rate or the corresponding power, respectively. The following section provides an informative overview about the simulation design.

5.1. Simulation Design

The simulation design is based on two different aspects. The first aspect is to investigate the performance of the tests under the constraint of simulated Rasch conform data. Thus the type–one error rate is computed. The second aspect is analysed under the constraint of Rasch violated data of different violation strength. Hence the power of the test can be calculated.
For each aspect the null hypothesis of the certain tests is that the Rasch model holds. The corresponding model of the violated data is interpreted as the alternative hypothesis

for the certain test. Therefore the test hypotheses are given by

$$H_0 : \text{Rasch model} \qquad H_1 : \text{model of violation} \tag{5.1}$$

For each scenario, i.e. Rasch conform data and Rasch violated data, the solving probability $p_{vi} = P(X_{vi} = 1)$ of a person v to an item i is generated due to the scenario´s model. The resulting $n \times k$ solving matrix is denoted by P. The scenario´s models, the constitution of the model´s parameters as well as the calculation of the solving matrix P will be described in section (5.1.1) for Rasch conform data and in section (5.1.2) for Rasch violated data.
With these solving probabilities the final dichotomous response matrix can then be computed with the following decision rule

$$X_{vi} = \begin{cases} 1 & p_{vi} \geq p^*_{vi} \\ 0 & p_{vi} < p^*_{vi} \end{cases}, \tag{5.2}$$

where p^*_{vi} are the entries of a $n \times k$ matrix P^* which contains random deviates of the standard uniform distribution.
With the resulting response matrix the CML estimates for the Rasch model parameters $\hat{\theta}$ and $\hat{\beta}$ and subsequent the specific test statistic can be calculated. The outcome of the corresponding p–value indicates the rejection or adoption of the model assumption to a level of significance at 5%. If the p–value is less than an α–level of 0.05 the null hypothesis of the researched test, i.e. that the Rasch model holds, must be rejected. On the other hand if the p–value of the test is larger or equal to this α–level the null hypothesis can be accepted.
After 100 replications of this computation the type–one error rate for Rasch data or the power for Rasch violated data can be generated based on counting rejections of the null hypothesis.

5.1.1. Rasch Data

For the simulation of a single Rasch conform data matrix, the person´s ability parameter θ and the item difficulty parameter β are drawn form a standard normal distribution. Mair (2006) stated that a $N(0, 1)$–distribution is appropriate because most of the persons are at an average ability and only a few persons do have high or low ability values. The same holds for the consideration concerning items.

With the formal equation of the Rasch model

$$p_{vi} = P(X_{vi} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \quad , \tag{5.3}$$

the $n \times k$ solving probability matrix P with its entries $[p_{vi}]$ can then be provided. The resulting matrix is thus given by

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nk} \end{pmatrix} \quad . \tag{5.4}$$

In order to obtain the final dichotomous 0/1–matrix X, the decision rule from equation (5.2) is applied.
Different scenarios are computed due to variations in sample sizes and in the number of items. Here the sample sizes are 100, 250, 500 and 1000. For each chosen sample size the number of items is set to 10, 20 or 30, respectively.
For each scenario 100 replications are obtained. By estimating the Rasch parameters for each replication and applying the selected test the type–one error rate can be generated. Knowing from the simulation assumptions that the data are Rasch data, because they are constructed from the Rasch model theorems, the type–one error rate is then be enumerated by counting the rejections of the null hypothesis (i.e. that the Rasch model holds) from the data with the help of obtained p–values.

5.1.2. Rasch Violated Data

Four types of Rasch model violations are simulated. The corresponding models of violations are interpreted as the alternative hypothesis in testing situations. The four types are

1. *Non–Parallel Item Characteristic Curves*
2. *Guessing*
3. *No Local Independency*
4. *No Unidimensionality.*

The person´s ability parameter θ and the item difficulty parameter β are like with the Rasch conform data drawn from a $N(0, 1)$–distribution. Also variations in sample sizes and variations in the number of items form different test scenarios. Samples sizes are in comparison to the Rasch data 100, 250, 500 and 1000, while the numbers of items vary

from 10, 20 to 30.
However each violation has specific additional parameters to itemize model properties or rather in this study–context to define model violation characteristics. According to the latter defined variations in sample sizes and item numbers also the degree of violation is varied in order to obtain different test scenarios.

1. *Non–parallel Item Characteristic Curves*
   For this violation the 2–parameter logistic model by Birnbaum (1968), introduced in chapter 2.3, is simulated. In addition to the parameters θ and β a discrimination parameter α for the item differences is provided. This parameter is decisive for intersecting ICCs. If an item discriminates quite good the ICC must be steeper, while a less discriminating item has a ICC which is flatter. Additionally, the discrimination parameter must not become negative. Based on these assumptions the parameter α is drawn from a logarithmic normal distribution with $\log N(0, \sigma^2)$. The variations of the corresponding standard deviation σ for the test scenarios are

   - $\sigma = 0.12$ for a weak violation,
   - $\sigma = 0.25$ for a medium violation and
   - $\sigma = 0.50$ for a strong violation of the Rasch model.

   With the latter drawn parameters the solving probabilities for the 2–PL model are simulated through

   $$p_{vi} = P(X_{vi} = 1|\theta_v, \beta_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_v - \beta_i))}{1 + \exp(\alpha_i(\theta_v - \beta_i))} \quad . \tag{5.5}$$

   Applying the decision rule from equation (5.2) the dichotomous response matrix X results.
   For each combination of persons, items and discriminations, 100 replications of generated response matrices are made. With estimating the Rasch parameters and applying the chosen test for each replicated data matrix the power of the specific test is computed.
   The test hypotheses are provided by

   $$H_0 : \text{Rasch model} \qquad H_1 : \text{2–PL model} \quad . \tag{5.6}$$

   The power is calculated by counting the number of rejections of the Rasch model. From the simulation design it is a fact that the data are not Rasch conform, hence the power increases if the test rejects the Rasch model.

2. *Guessing*

   An additional guessing parameter to the 2–PL model is introduced by Birnbaum (1968) which leads to his 3–parameter logistic model. This model is already presented in chapter 2.4. The Person parameter θ and the item parameter β are drawn again from a $N(0,1)$–distribution. The discrimination parameter α is analog to the 2–PL model drawn from the $\log N(0,\sigma^2)$–distribution. The standard deviation of the log normal–distribution is also set to 0.12, 0.25 and 0.50 for a weak, medium and strong violation, respectively.

   The value of the guessing parameter is the lower asymptote of the characteristic item curve. From a logical point of view it can be stated, that the probabilities for solving an item due to guessing lies in the range of $[0;0.3]$, e.g. 1 out of 5 items can be solved realistically due to guessing.

   Such a asymmetric probability domain can be simulated with the help of the $Beta$–distribution. Glas and Meijer (2003) use a $Beta(5,17)$–distribution to generate simulated guessing values, because the mean value of this distribution is 0.2, which is often used as an guideline for guessing. In this work the guessing parameter is drawn from a $Beta(2, \text{number of items})$–distribution to take the changing number of items, i.e. 10, 20 and 30 items, into account.

   Hence the solving probability matrix P with its entries p_{vi} can be simulated by

$$p_{vi} = P(X_{vi} = 1|\theta_v, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1-\gamma_i)\frac{\exp(\alpha_i(\theta_v - \beta_i))}{1+\exp(\alpha_i(\theta_v - \beta_i))} \quad . \tag{5.7}$$

   The response matrix X results from the decision rule (5.2).

   Again for each combination of parameter values 100 replications are obtained. With the Rasch parameter estimates for each replication the power of the goodness–of–fit tests can be calculated.

   For this violation the test hypotheses will be given by

$$H_0 : \text{Rasch model} \qquad H_1 : \text{3–PL model} \quad . \tag{5.8}$$

   The power is, as in the 2–PL model, computed by counting the number of rejections of the Rasch model.

3. *No Local Independency*

   One of the main assumptions of the Rasch model is the *local independency*. Fixing the ability parameter of a person, the responses to certain items must not correlate. To violate this characteristic, 5 items are chosen to be dependent in each simulation scenario. The additional parameter δ_{ij} indicates the degree of dependence between item i and item j.

   This violation, i.e. the correlation of the two items, is set in this work to

   - 0.25 for a weak violation,
   - 0.50 for a medium violation and
   - 0.75 for a strong violation of the Rasch model.

   Kelderman (1984), Jannarone (1986) and Kelderman & Rijkes (1994) proposed the following model equation for the solving probabilities given a dependency between certain items

$$P(X_{vi} = 1|X_{vj}; \theta_v, \beta_i, \delta_{ij}) = \frac{\exp(\theta_v - \beta_i + x_{vj}\delta_{ij})}{1 + \exp(\theta_v - \beta_i + x_{vj}\delta_{ij})} \quad , \tag{5.9}$$

   where X_{vi} is the response of person v to item i and X_{vj} the response of person v to item j.

   The resulting response matrix X is again calculated by applying the decision rule from equation (5.2). For each scenario 100 replications are made and the CML estimates for the Rasch model are provided.

   The power of the considered tests is generated by applying the specific test to each replicated data matrix and its corresponding estimates. The alternative hypothesis is a local dependence model, thus the test hypotheses are

$$H_0 : \text{Rasch model} \qquad H_1 : \text{local dependent model} \quad . \tag{5.10}$$

   Counting the number of rejections serves as an indicator for the power.

4. *No Unidimensionality*

   The violation of *unidimensionality* can be achieved by the multidimensional Rasch model from Glas (1989, 1992). The model relates in its original form to polytomous items (multiple categories), but dichotomous items can be seen as a special case. Applying this special case, the multidimensional dichotomous Rasch model

is denoted by

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{\exp\left(\sum_{d=1}^{D}(\theta_{vd} - \beta_{id})\right)}{1 + \exp\left(\sum_{d=1}^{D}(\theta_{vd} - \beta_{id})\right)} \quad , \tag{5.11}$$

where $d = 1, ..., D$ is the dimension of the model.
The multidimensional model in this work is applied to 2 dimensions, i.e. 2 latent traits, but it can be disposed to even more dimensions in general.
The parameters for this model are, as applied from Suárez-Falcón and Glas (2003), the correlation parameter $r_{\theta_1\theta_2}$ for latent trait θ_1 and latent trait θ_2 and again the person and item parameter. $r_{\theta_1\theta_2}$ indicates the correlation between the 2 dimensions. To vary the strength of violation $r_{\theta_1\theta_2}$ is set to

- 0.75 for a weak violation,
- 0.50 for a medium violation and
- 0.25 for a strong violation of the Rasch model.

Unlike the other explained data generation methods the person parameter θ is drawn from a multivariate normal distribution $N(0, \Sigma)$ where Σ is given by

$$\Sigma = \begin{pmatrix} 1 & r_{\theta_1\theta_2} \\ r_{\theta_1\theta_2} & 1 \end{pmatrix} \quad . \tag{5.12}$$

The item parameter still is drawn form the standard normal distribution.
Also the number of items which reference to the 2 dimensions is varied. They are either constructed so that the loading–proportion is 50 : 50, which means that 50% of the items load on dimension θ_1 and 50% load on dimension θ_2, or so that the ratio is set to 80 : 20.
With the decision rule from equation (5.2) the corresponding response matrix X can be generated. With 100 replications for each test scenario the estimated Rasch parameters and the resultant researched test statistic serves for the power of the tests computed by the counts of rejections to the Rasch model.
The test hypotheses are

$$H_0 : \text{Rasch model} \qquad H_1 : \text{2–dimensional Rasch model} \quad . \tag{5.13}$$

Except for the 3–PL data, all these data simulation methods are realised in the programming routine by the R package *eRm*. The code as well as informative explanations

can be found in the appendix.

5.2. Results

All results will be demonstrated graphically.
The first type of figures is related to type–one error rates. The number of items are outlined on the abscissa and the corresponding error rates on the ordinate. For each sample size there is a sub–graph given in the figure.
The second type of figures displays the power of the certain test on the ordinate and the degree of violation on the abscissa. Each combination of number of items and sample sizes are provided in a sub–graph of the image.
It should be mentioned that the time factor for the Bootstrap test is immense. On the authors laptop (CPU: Intel Pentium Dual–Core Processor, 1,73 GHz, main memory: 2GB/Go DDR2 SDRAM) each of the different test scenarios would last about 1 to 20 days. Thus acceleration methods must have been found in order to obtain results in time.
First the R–Code for the Bootstrap–routine was parallelised from the usually used 1 CPU to 32 CPUs. Then the whole Bootstrap program was run on the Linux Cluster Server at the Leibniz–Rechenzentrum in Munich. These two acts served as a huge time improvement, thus the results were obtained within a reasonable time scale.

5.2.1. The Mixed–Rasch Model Test

As mentioned in chapter 4.6, the distribution of the mixed–Rasch model test under the null hypothesis should be taken as unknown. This is the key point of the simulation study from the mixed–Rasch model test.
The first approach of this part of the work is to assume a χ^2–distribution with $(G(k-1)+(G-1)+G(k-2)+2)-(k-1)$ degrees of freedom for the test statistic. From the principles of the LQ–test it is known, that the degrees of freedom are calculated from the difference between the number of free parameters in the restricted model and the number of free parameters in the full model. The free parameters in the mixed–Rasch model are given by $(G(k-1)+(G-1)+G(k-2)+2)$. The first summand relates to the item parameters within each class g, the second summand to the class size parameters and the third summand to the class–specific score probabilities. Since extreme responses can not be allocated to a certain class, the last summand takes these two facets into account.

It is observed in the corresponding simulation study, that in almost all results of test scenarios the Rasch model is rejected to a given α–level of 5%.
Analyzing these results it comes out, that the Rasch model as a nested model of the mixed–Rasch model is only obtained by setting one class size parameter to zero. But this violates one main assumption of the approximation of a χ^2–distribution, namely that the parameter space must be open. This means the class size parameter must be arranged in the open space between]0;1[. By setting one class size parameter to zero, a χ^2–distribution of the test statistic under the null hypothesis can not be assumed any more, therefore the distribution should be seen as unknown.
Due to time limits in the diploma thesis the approach to generate the distribution under the null hypothesis with the help of Bootstrapping–methods or to compare the Rasch model and the mixed–Rasch model with the help of information criterions like AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion), has not been realised. These procedures are only stated in the outlook for further studies (refer to chapter 7).
Because the achieved results from the mixed–Rasch model test with approximated χ^2–distribution are assumed to be incorrect, they are not demonstrated in this work.

5.2.2. Type–One Error Rates

For the calculation of the type–one error rates the rejections of the Rasch model are counted due to the restriction, that Rasch conform data have been simulated (this simulation design is explained in section (5.1.1)). A nominal α–level of 0.05 is assumed. The number of rejections of the specific test must be small in order to hold this level.
Figure (5.1) demonstrates the type–one error rates for the Andersen test. The x-coordinates are given by the number of items and the y-coordinates denote the corresponding type–one error rate. For each sample size a sub–graph is displayed. The nominal α–level is marked by the red dotted line.
Neither variations in the number of items nor variations in sample sizes strongly affect the outcome of the type–one error rate. Each scenario holds approximately the nominal α–level.
The simulation study from Suárez-Falcón and Glas (2003) confirms this statement. Indeed they vary the scenarios differently. The number of items are 15, 50 and 75, whereas sample sizes change from 100, 250, 500, 1000 to 4000.
The type–one error rates for the Bootstrap test are displayed in figure (5.2). Except for two type–one error rates of 0.06 all rates are equal or less than the nominal α–level.

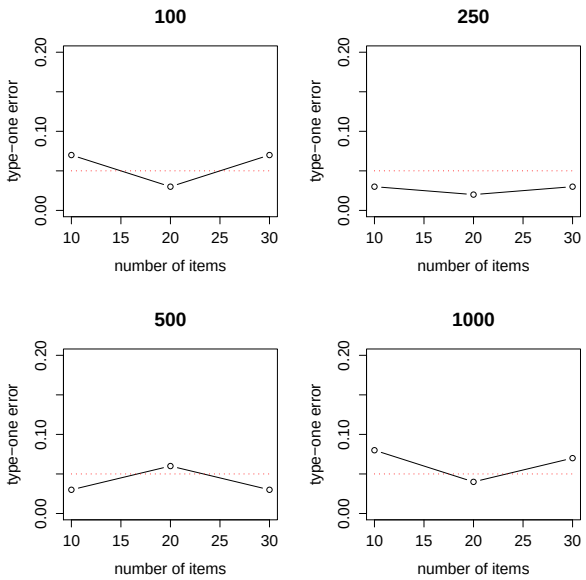

Figure 5.1.: Type–one error rates of the Andersen test

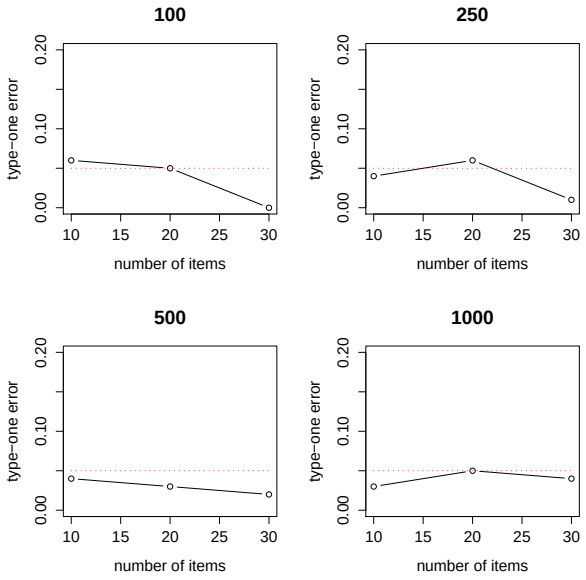

Figure 5.2.: Type–one error rates of the Bootstrap test

Thus this test holds the α–level quite well.
The sample size do not affect the outcome of the simulation. However a tendency of lower error rates with rising numbers of items can be stated.
Comparing the Andersen test and the parametric Bootstrap test, the parametric Bootstrap test performs better than the Andersen test. Considering only type–one errors the

Bootstrap test must be recommended.
However really checking a test on performance, also the power under different forms of Rasch model violations must be regarded. Such scenarios will be investigated in the following sections.

5.2.3. Non–Parallel Item Characteristic Curves

As introduced in chapter 5.1.2, different forms of model violations form different test scenarios. With these scenarios the functionality of the certain tests will be analysed. Counting the number of rejections serves as an indicator for the power of the test, since the simulated input data are Rasch model violated. As with the type–one error rates the given α–level is assumed to be 0.05.
The first scenarios are referred to the *non–parallel ICCs* violation.
In figure (5.3) the degree of violation is displayed on the abscissa and the corresponding

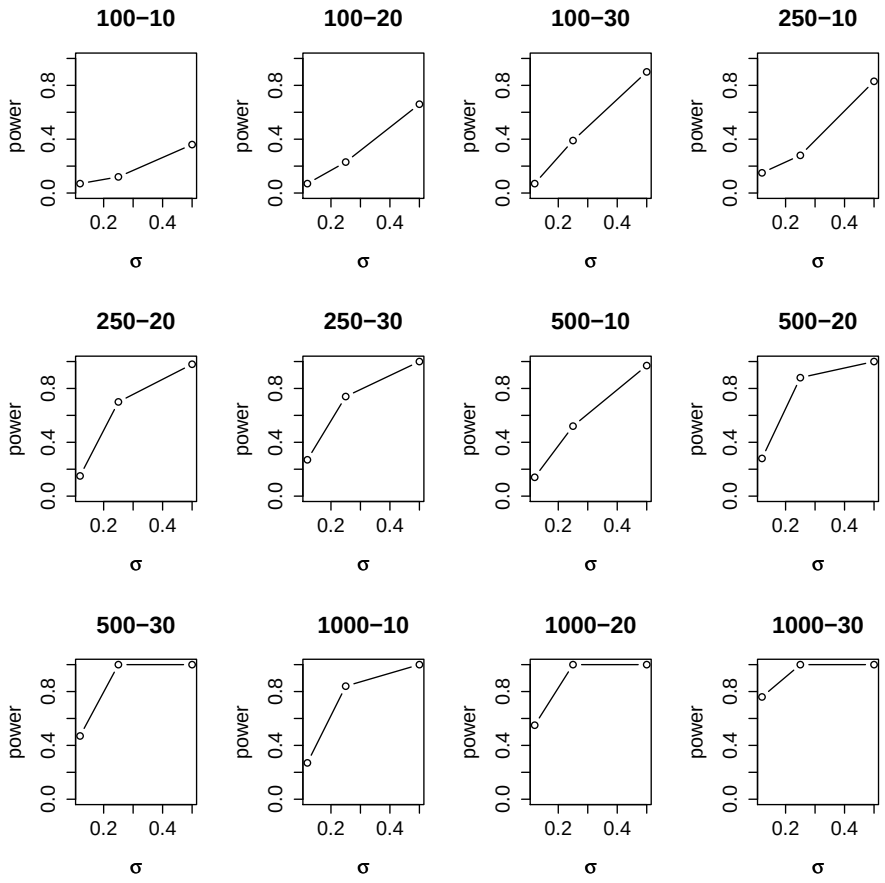

Figure 5.3.: Power of the Andersen test – non–parallel ICCs

power on the ordinate. The parameter σ on the abscissa denotes the standard deviation of the $log(N(0, \sigma^2))$–distribution, from which the discrimination parameter α is drawn in the 2–PL model (for the corresponding simulation design refer to section (5.1.2)). For each combination of the number of items and the sample sizes a sub–graph is shown.
As Gustaffson (1980) stated the Andersen test has power for violations based on the 2–PL model. Both, a rising number of items as well as an increasing sample size improve the power of the tests. Also the strength of the violations is detected right. Besides power values for light violation almost all remaining powers lie above 50%. Summing up it can be stated that the Andersen test performs well in connection with the violation by using the 2–PL model.
Power values for the Bootstrap test are displayed in figure (5.4).

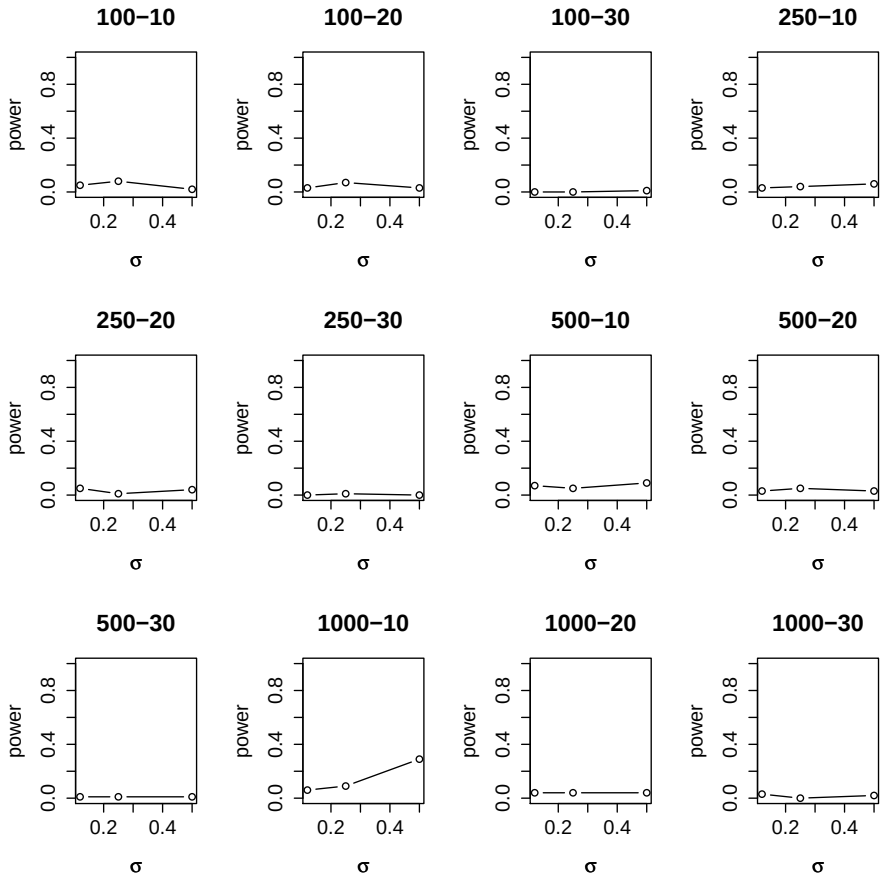

Figure 5.4.: Power of the Bootstrap test – non–parallel ICCs

Besides the combination of 1000 persons, 10 items and a strong violation, all powers are at an unacceptable low level. Neither the degree of violation, nor the number of items or

the size of the samples influence the extent of the power. Except from one power value all powers are less or equal to the value of 0.1. Therefore this parametric Bootstrap test fails to perform admissible for detecting violations in parallelism.
For an elaborate analysis of bad power results refer to section (5.2.7).
In comparison the Andersen test performs by far better than the Bootstrap test and is therefore preferable.

5.2.4. Guessing

The second violation scenario results in the 3–PL model, where additional to the 2–PL model a lower asymptote is introduced to constitute guessing (refer to section (5.1.2)). Figure (5.5) provides the power values for the Andersen test. Analogously to the 2–PL

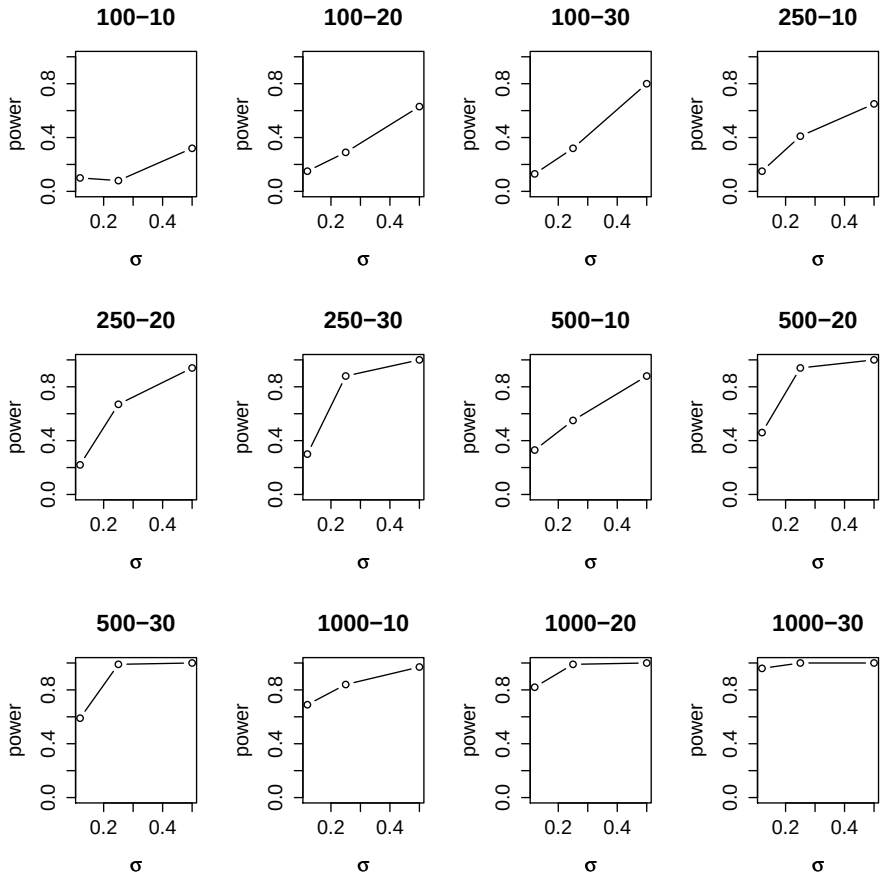

Figure 5.5.: Power of the Andersen test – Guessing

model the 3–PL model is well distinguished by the Andersen test (Gustaffson, 1980). With rising number of items and persons the counts of rejections increases and therefore

the power gets better. Only for weak violations ($\sigma = 0.12$) is the performance of the test poor. Hence the Andersen test performs for a noticeable violation quite good.
The power values for the Bootstrap test according to violations due to the 3–PL model

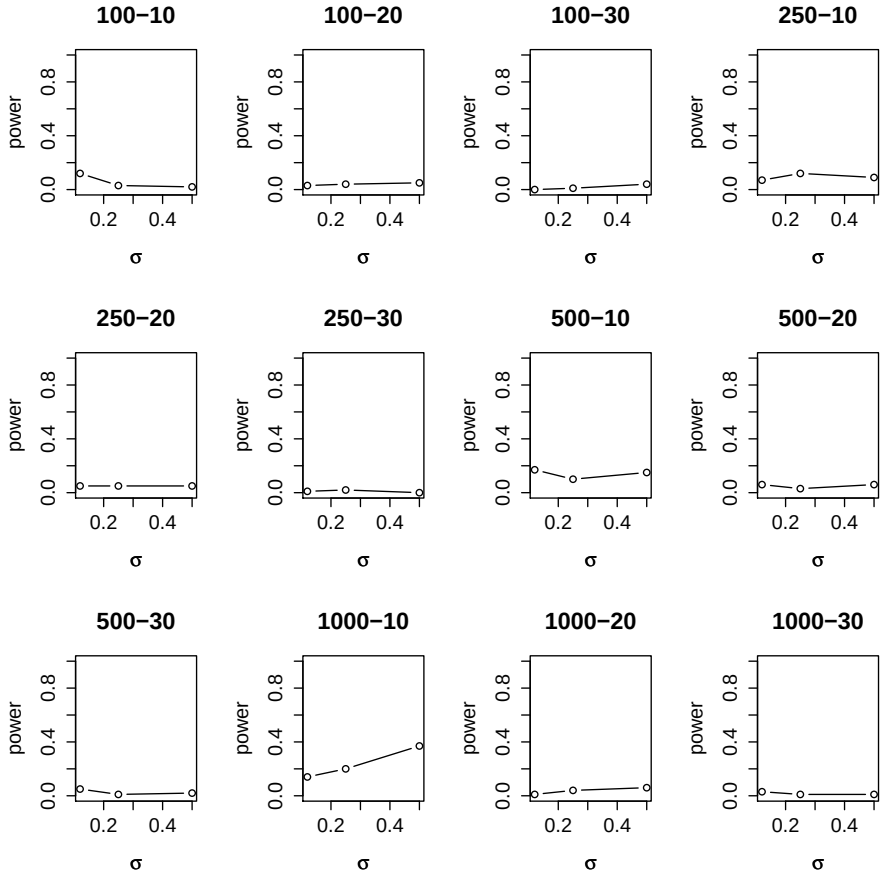

Figure 5.6.: Power of the Bootstrap test – Guessing

are demonstrated in figure (5.6).
As with the 2–PL model only the combination of 1000 persons and 10 items leads to a almost acceptable result. Again neither the degree of violation, nor the number of items, nor the sample sizes influence the outcome of the power. Also with violations coming from the 3–PL model the Bootstrap test fails to perform well.
Because of these unexpected bad results an extensive analysis is given in section (5.2.7). By comparing the Andersen test and the Bootstrap test due to intersecting ICCs and a lower asymptote, the Andersen test is recommended.

5.2.5. No Local Independency

For the third violation 5 items in different violation strengths are correlated and thus the assumptions *local independency* of the Rasch model is violated (explanations to this simulation design are provided in section (5.1.2)).
The power values for the Andersen test are given in figure (5.7). It can be concluded

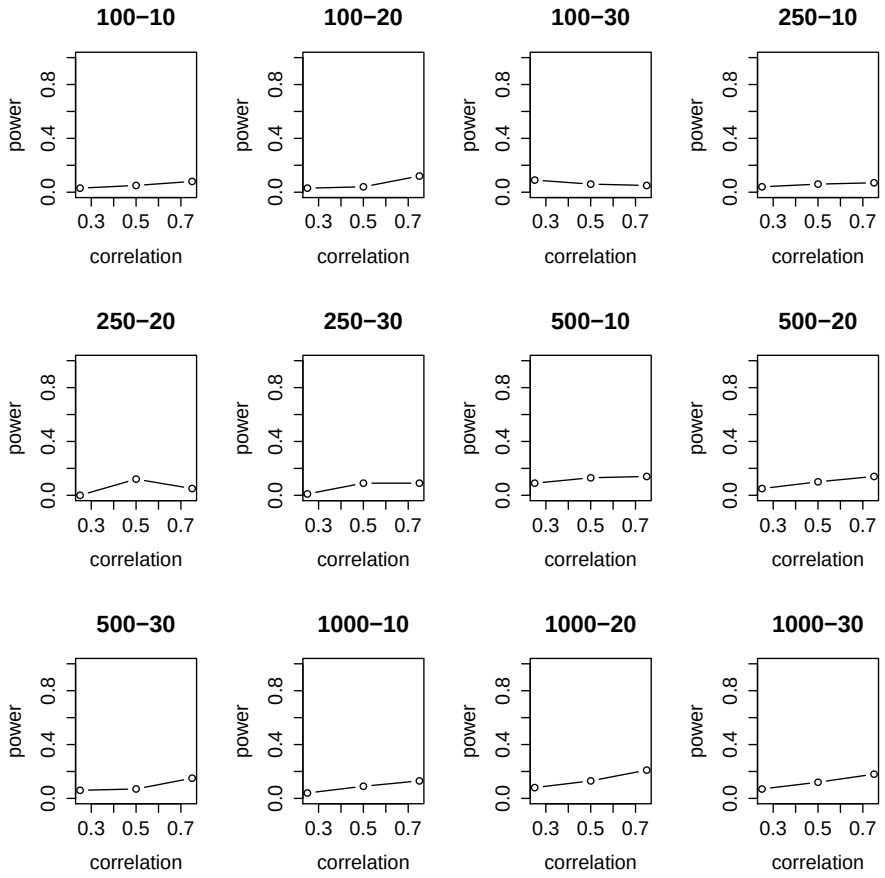

Figure 5.7.: Power of the Andersen test – No local independency

that the Andersen test fails to detect correlations between certain items. None of the power values achieve even the 25% border. Although no influence from the number of items and the sample size is observable, the power raises slightly with the rising degree of violation. Summing up the Andersen test does not perform well in distinguishing local dependency.
For explanation, the Andersen LR–test is based on first–oder frequencies. From the for-

mal representation in equation (4.17) only comparisons between subgroups can be made. The test is not constructed to detect violations referring to second–order frequencies between two items and therefore has not the ability to detect such violations.
In comparison to the Andersen test the Bootstrap test performs just as bad as the

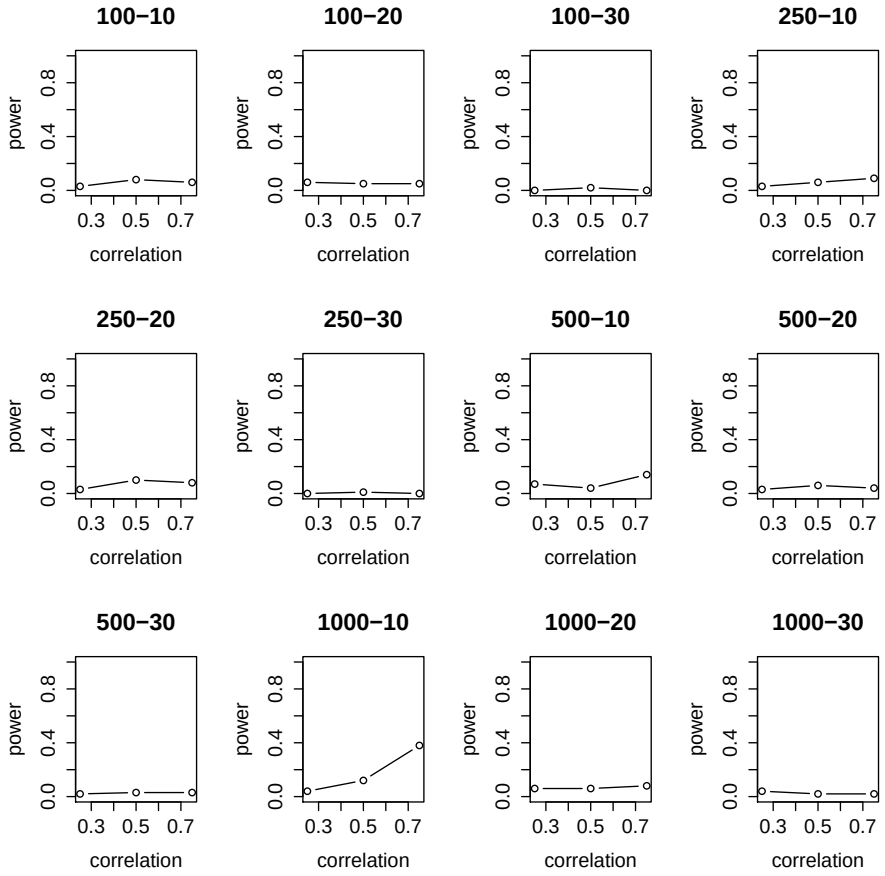

Figure 5.8.: Power of the Bootstrap test – No local independency

Andersen test. In figure (5.8), the power values for this test can be regarded. Again only in the case of 1000 persons and 10 items a comprehensible mechanism is observable despite the power of the test in this case is nevertheless worse. In this scenario the test identifies more of these violations with rising degree of violation.
Also the explanation of the Andersen test could be the reason for bad results. In the Bootstrap routine the used Pearson χ^2–statistic is also related to first–oder observations (see equation (4.40)). On the other hand the reasons explained in section (5.2.7) may serve for bad power rates as well.
For detecting correlations between items, neither the Andersen test nor the Bootstrap

test is recommended. For such detections a test based on second–order frequencies like the Q_2–test or the R_2–test should be preferred.

5.2.6. No Unidimensionality

The analysis of *unidimensionality* is based on two aspects. The first aspect belongs to the correlation between the considered dimensions. In this work two dimensions are chosen. The second aspect refers to the loading proportion of the items to the dimensions (this simulation design is also described in detail in chapter 5.1.2).
First the Andersen results will be discussed. In Figure (5.9) the power values can be regarded.
The loading proportion of the items is here 50:50, that means 50% of the items load on

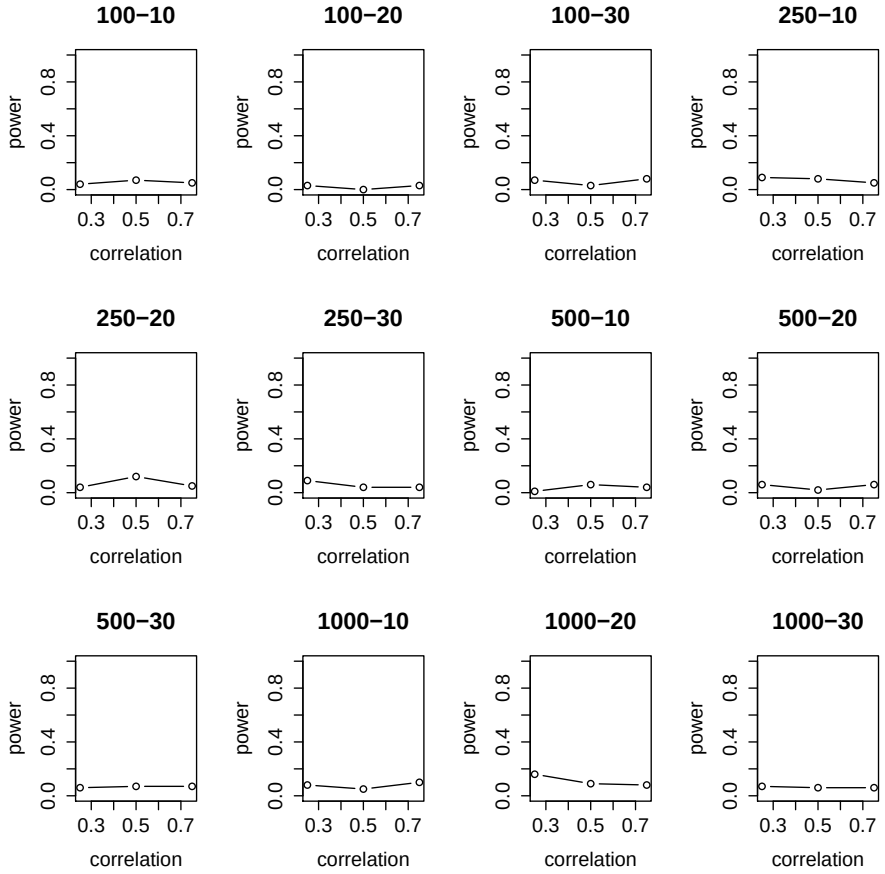

Figure 5.9.: Power of the Andersen test – No unidimensionality – Items loading 50:50

the first dimension θ_1 and 50% on the second dimension θ_2.
From the bad results one can assume that the Andersen test is not able to detect viola-

tions in the question of *unidimensionality*. The number of items, the sample sizes and the rising form of correlation between dimensions do not affect the power of the test.
Surprisingly by regarding the figure (5.10) the Andersen test performs well in detecting violations due to a two–dimensional model, where the loading proportion is set to 80:20. Note that a correlation of $r_{\theta_1\theta_2} = 0.25$ indicates a strong model violation and $r_{\theta_1\theta_2} = 0.75$

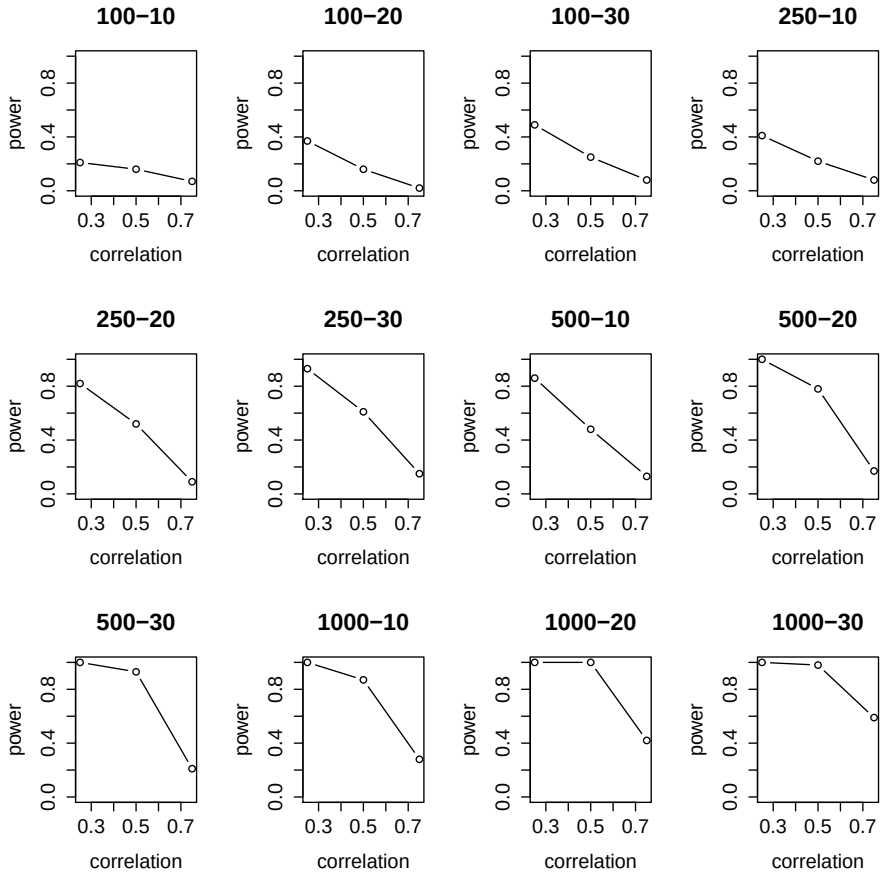

Figure 5.10.: Power of the Andersen test – No unidimensionality – Items loading 80:20

a light violation, because with $r_{\theta_1\theta_2} = 0.75$ the dimensions are strongly correlated and are basically the same. Therefore the typical curve, i.e. the power gets better with rising number of items, sample sizes and model violations, is here upside down.
The only suggestive reason for the phenomena of different results due to item loadings can be based on the subgroups which are generated in the context of the Andersen LR–test. Maybe in the case of 80:20 loadings the subgroup consisting of more capable persons refers more to one of the dimensions, that means these persons answer e.g. the items measuring the extent to the first dimension θ_1 more often correctly than the items

based on the second dimension θ_2. Thereby the test becomes responsive for the violation of *unidimensionality*.

Otherwise when 50% of the items load on dimension θ_1 and the remaining items refer to dimension θ_2, the subgroup with the more capable persons will answer mostly items correctly on both dimensions. Thus the test will not be able to detect multidimensionality.

Hence it can be stated that the LR–test from Andersen is able to detect multidimensionality only in the case of unequal partitioning of item loading.

According to Andersen´s LR–test Bootstrap test results for the violation of *unidimensionality* are displayed in figure (5.11) and figure (5.12). Figure (5.11) relates to a loading relation of 50:50 and figure (5.12) is based on a proportion of 80:20.

Both results only detect model violation in the case of 1000 persons, 10 items and a small

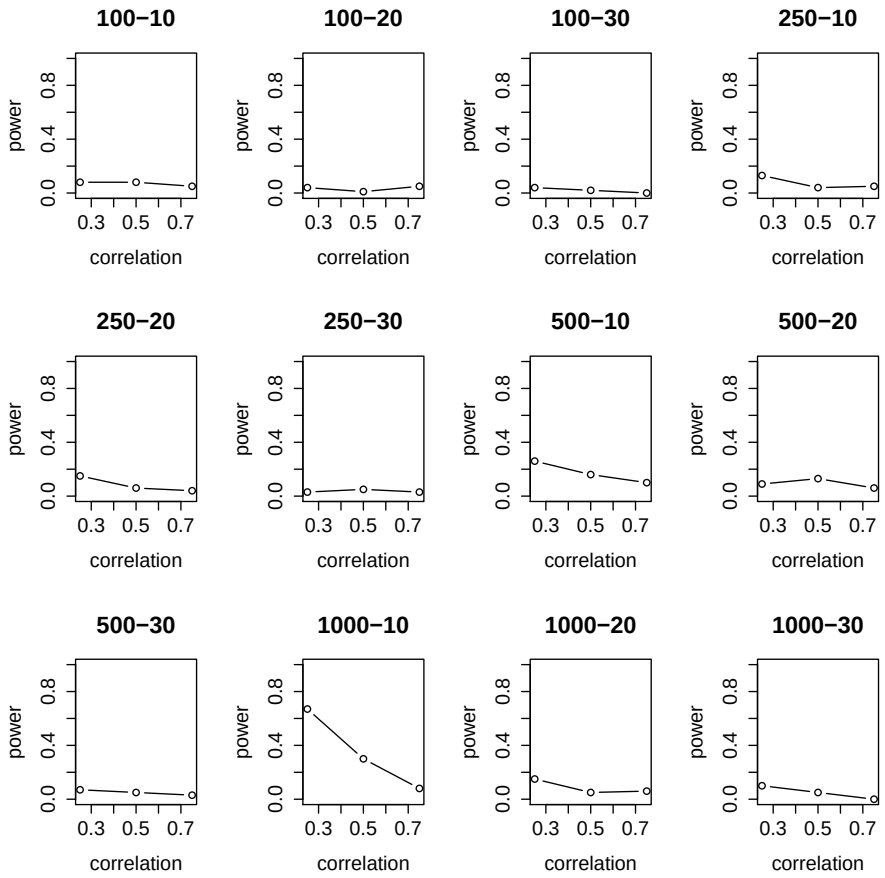

Figure 5.11.: Power of the Bootstrap test – No unidimensionality – Items loading 50:50

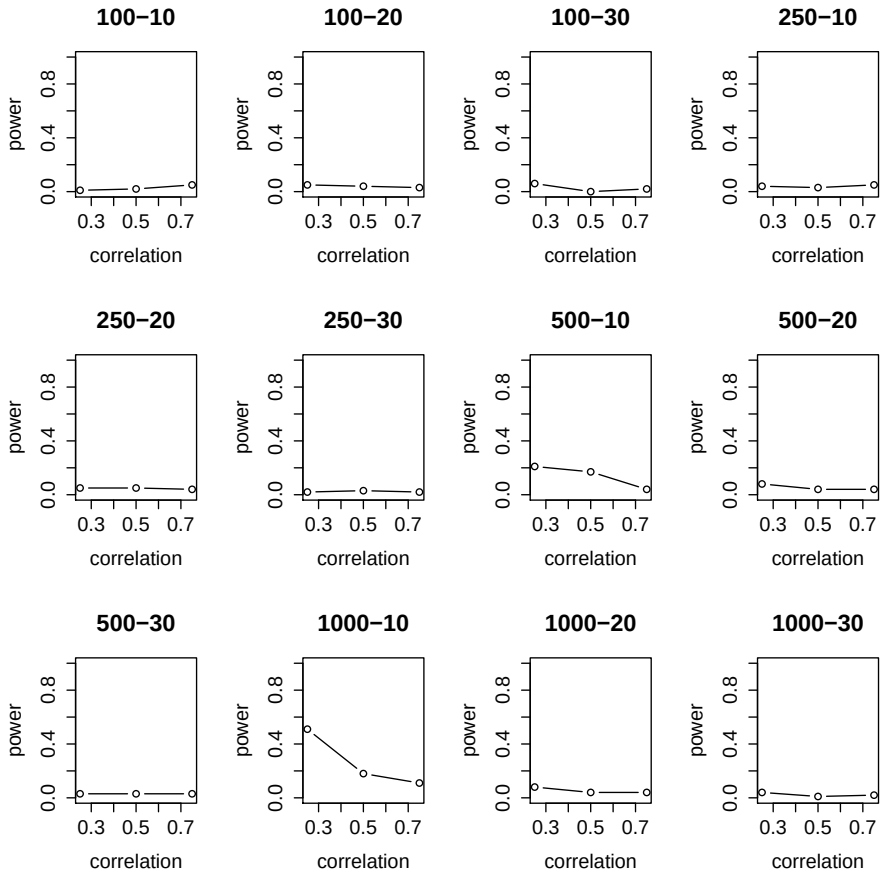

Figure 5.12.: Power of the Bootstrap test – No unidimensionality – Items loading 80:20

correlation between the dimensions. In all other compositions the findings are as bad as the latter obtained results from the parametric Bootstrap test. The power response to none of the variable inputs, i.e. violation strength, the number of items and the sample sizes. A comprehensive analysis of these bad results will be investigated in the following section.

5.2.7. Analysis Of Bad Bootstrap Power

Because of bad Bootstrap test results a comprehensive analysis is performed since parametric Bootstrap techniques usually serve for acceptable test results.
First the impact of the chosen test statistic is investigated. Since the Pearson χ^2 statistic refers to all possible response patterns, the number of such frequencies becomes huge for already 10 items. According to this finding the values of required expected frequencies become tiny. Statements relating to such tiny outcomes must be taken with care and

regarded doubtfully. Since these expected frequencies are based on the null hypothesis, i.e. the Rasch model assumption, they can barely be distinguished from expected frequencies based on the alternative hypothesis. Hence the Rasch model is accepted in almost all cases.

Thus the Bootstrap test is performed with only five items in order to check the performance of the test with a small number of items. Here the number of possible response patterns constrained to convenient $2^5 = 32$ cases. Results for the type–one error rates and the violation of parallel item characteristic curves are demonstrated in figure (5.13) and figure (5.14), respectively.

It can be seen in figure (5.13) that, expectedly with only five items the Bootstrap test

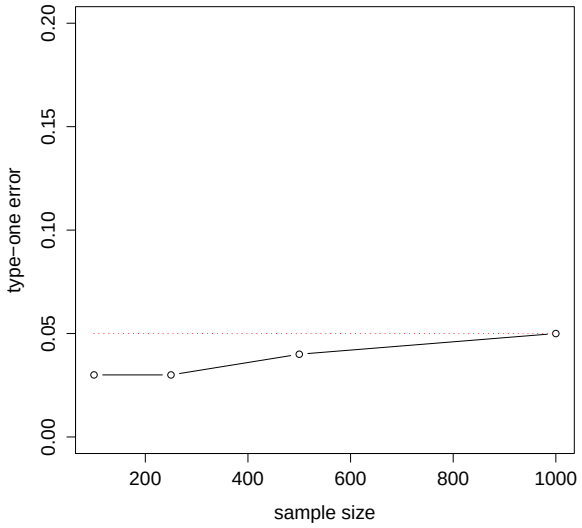

Figure 5.13.: Type–one error rates of the Bootstrap test with five items

also holds the nominal α–level of 5%.

The power of the test for the violation of intersecting item characteristic curves becomes better due to the enormous reduction of all possible response patterns, although the power achieves only acceptable results in the cases of 500 and 1000 persons and a strong model violation (see figure (5.14)).

Another approach to explain the bad Bootstrap results is created by choosing another test statistic which does not relate to all possible frequencies but only to the observed single responses. Such a test statistic is R_ϕ. This statistic is based on the maximum range of inter–item correlation.

Also for this scenario, results of type–one error rates and the violation of parallel item

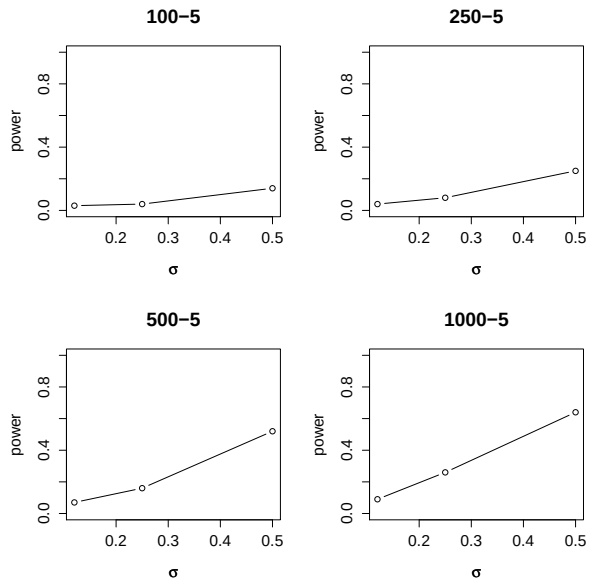

Figure 5.14.: Power of the Bootstrap test with five items – non–parallel ICCs

characteristic curves are calculated and thus displayed in figure (5.15) and (5.16). Except for one outlier in the case 500 persons and 30 items, all test scenarios for the

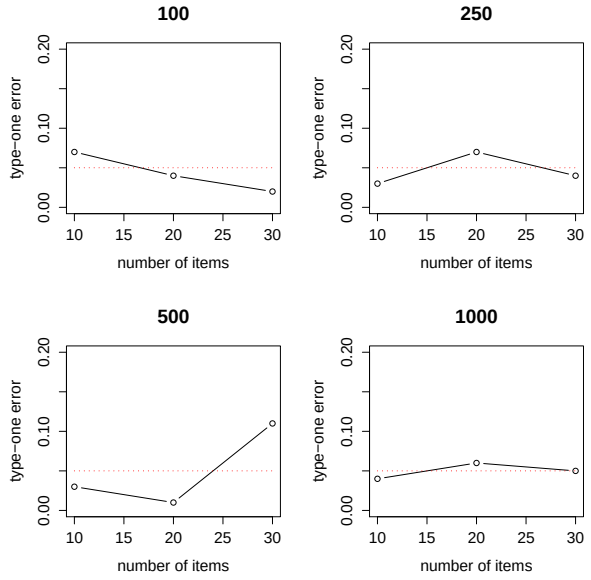

Figure 5.15.: Type–one error rates of the Bootstrap test with test statistic R_ϕ

Bootstrap test with the R_ϕ test statistic hold approximately the nominal given α–level of 0.05 (see figure (5.15)).

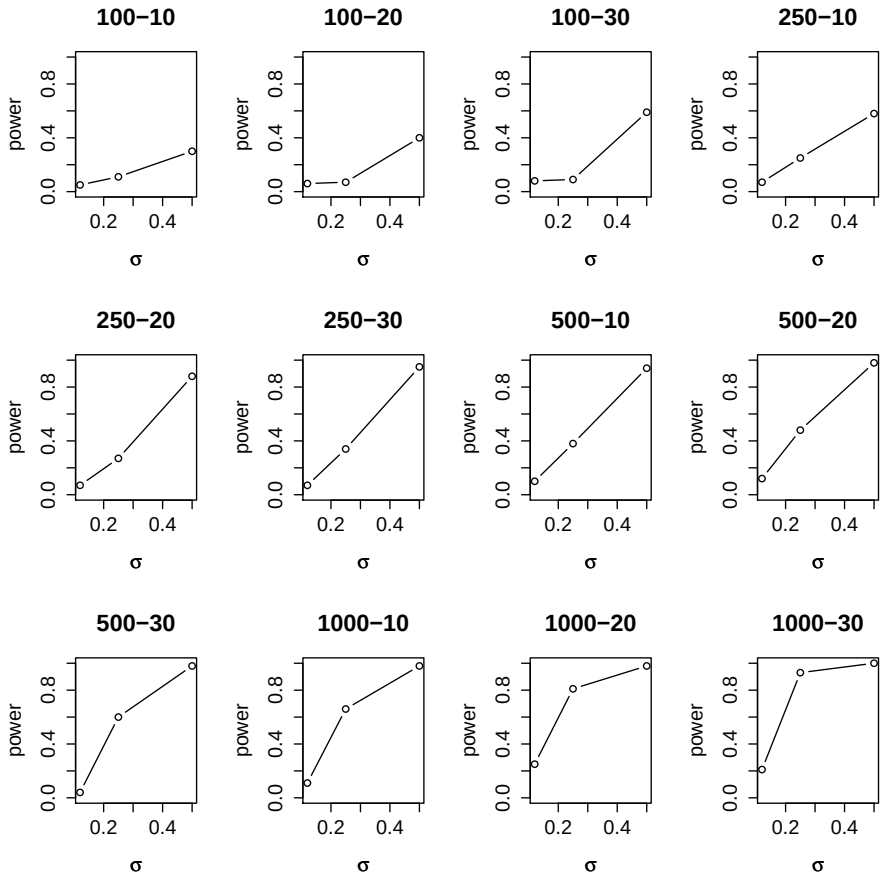

Figure 5.16.: Power of the Bootstrap test with test statistic R_ϕ – non–parallel ICCs

Demonstrated in figure (5.16), the power values of the different simulated scenarios are as desired. With rising number of items, sample sizes and violation strength, the power increases to an admissible level. Hence it can be stated that the Bootstrap test performs well in conjunction with test statistics based on the single responses.

A third analysis is the inspection of the distribution of the obtained p–values. From statistical theory it is known that the distribution of the p–values under the null hypothesis should follow a uniform distribution (see Lehmann & Romano, 2005). This aspect can be seen in figure (5.17) on the left hand side. The red line denotes the reference uniform distribution function.

Interestingly by regarding the distribution of the p–values under the alternative hypothesis, which means under the hypothesis of Rasch violated data, an approximate uniform distribution function follows (refer here to figure (5.17) right hand side). Normally a different, mostly unknown, distribution function must appear.

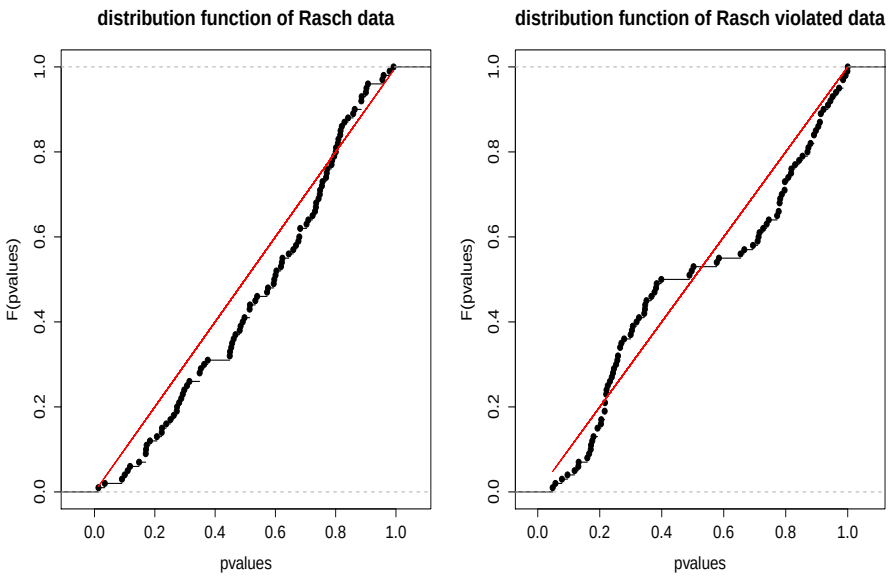

Figure 5.17.: Comparison of distribution functions of obtained p–values

This concludes that the distribution function of the null hypothesis and the distribution function of the alternative hypothesis are almost the same and therefore nearly no difference can be detected. Out of it the Rasch model will not be rejected.
This third analysis also confirms the assumption of doubted usage of test statistics which relate to all possible response patterns for a number of items equal or larger than 10 cases. Therefore the parametric Bootstrap test should be recommended to be used in conjunction with test statistics based on the single response outcomes by exercising with equal or more than 10 items.

Chapter 6

Practical Analysis: I–S–T 2000 R

The Intelligence–Structure–Test (IST) 2000 R is a test based on the structural model of intelligence. It is one of the most applied intelligence tests in Germany (Steck, 1997; Schorr, 1995). There exist three different forms of the representation of the I–S–T 2000 R:

1. Basic module short–form which contains nine subtests for deductive reasoning
2. Basic module short–form with additional two retentivity tests
3. Extension module with nine knowledge tests.

In addition to this enumeration there are two different sequences of items. These are denoted by form A and B.
This work refers to the already taken sample from Bühner, Ziegler, Krumm and Schmidt-Atzert (2006), which refers to the Basic module short–form A with nine subtests. These tests measure the extend to verbal, numerical and figural intelligence, respectively. Thus these subtests are given by

- Sentence Completion
- Verbal Analogies
- Similarities
- Numerical Calculations
- Number Series
- Numerical Signs
- Figure Selection
- Cubes
- Matrices.

For these nine subtests it is interesting to test the fit of the Rasch model. Due to the fact that the tests are time limited the question arises, if the subtests measure only the

latter stated traits and not additional speed.
The sample of 273 individuals was taken from students mostly of the following fields of study: psychology, business administration, medicine, pedagogy and law. There were 176 female and 97 male participants. The students were from 18 to 39 years old (average: 22,9 and standard deviation: 3,3) and have studied between 1 and 21 semesters (average: 4,5 and standard deviation: 3,9).
Bühner, Ziegler, Krumm and Schmidt-Atzert (2006) tested the fit to the Rasch model with the common Bootstrap test applied in the software *WINMIRA* (see von Davier, 1997). There the parametric Bootstrap test refers to the Pearson χ^2 test statistic and to the Cressie–Read test statistic, but Bootstrap replications are not sampled from matrices with constant row sums.
To compare the results obtained from Bühner, Ziegler, Krumm and Schmidt-Atzert (2006), these values will be taken up again next to the results from the here discussed Andersen LR–test and the Bootstrap test with given marginals.
All calculated p-values are displayed in table (6.1),

	B (χ^2)	B (CR)	B wgm.(χ^2)	B wgm.(R_ϕ)	LR
Sentence Completion	0.26	0.31	0.44	0.12	0.00
Verbal Analogies	0.21	0.32	0.79	0.21	0.00
Similarities	0.04	0.06	0.06	0.03	0.01
Numerical Calculations	0.32	0.38	0.90	0.00	0.00
Number Series	0.13	0.15	0.95	0.00	0.00
Numerical Signs	0.02	0.03	0.21	0.00	0.00
Figure Selection	0.00	0.01	0.01	0.00	0.00
Cubes	0.00	0.00	0.01	0.00	0.00
Matrices	0.40	0.63	0.59	0.01	0.59

Table 6.1.: Results for Rasch model fit of I–S–T 2000 R

where B denotes the abbreviation of Bootstrap, *B wgm.* the abbreviation for Bootstrap with given marginals and LR stands for likelihood ratio. If the p–value is less than the nominal stated α–level of 0.05, the null hypothesis of the researched test, i.e. that the Rasch model holds, must be rejected. On the other hand if the p–value of the test is larger or equal to this α–level, the null hypothesis can be accepted.
It can be seen that results from the Bootstrap versions used in Bühner, Ziegler, Krumm and Schmidt-Atzert (2006) (table (6.1) columns: B (χ^2), B (CR)) and results obtained from the Bootstrap version used in this work based on constant row sums (table (6.1)

column: B wgm.(χ^2)) achieve almost similar results. *Sentence Completion*, *Verbal Analogies*, *Numerical Calculations*, *Number Series* and *Matrices* can be assumed to be unidimensional with the Bootstrap test versions. For the subtests of *Similarities* and *Numerical Signs* different results emerge. The fit to the Rasch model for the subtests *Figure Selection* and *Cubes* must be rejected to a level of significance for 5%.

According to findings from simulation study in chapter 5.2.7 the Bootstrap test based on test statistics referring to expected frequencies of response patterns might not be useful. Since all possible response patterns must be taken into account already up from 10 items the number of possible response patterns becomes huge and therefore the extent of the expected frequencies tiny, so statements based on these outcomes must be doubted.

The Andersen LR–test (table (6.1) column: LR) rejects the null hypothesis for all but one subtest. Bühner, Ziegler, Krumm and Schmidt-Atzert (2006) stated that the asymptotic χ^2–distribution for the Andersen LR–test must be doubted because the rule to approximate this distribution is normally violated. This rule contains that expected frequencies of each possible response pattern must be equal or larger than 1. Since the number of possible response patterns for each subtest is $2^{20} \sim 1 * 10^6$ the expected frequencies are in most cases by far less than 1. From this point of view the usefulness of the LR–test results might be doubtful.

However relating to obtained results from simulation studies in chapter 5, the Andersen LR–test provides a good evaluation for data fit to the Rasch model. From this perspective, results from the Andersen LR–test should also be considered.

Out of interest, the Bootstrap test with constant row sums is yet analysed with another test statistic (see also chapter 5.2.7). The statistic R_ϕ is based on the maximal inter–item correlation, thus only on observed values and not on all possible response patterns. Results from this routine can be found in the next to last column in table 6.1 (B wgm.(R_ϕ)). This version of a parametric Bootstrap test rejects the null hypothesis more often than the other three Bootstrap test versions. Only the subtests *Sentence Completion* and *Verbal Analogies* can be assumed to fit the Rasch model and therefore to be unidimensional.

Which of these results users should believe in, remains unsolved. Thus further studies must be taken to solve this problem (refer also to chapter 7).

Chapter 7

Summary and Outlook

Item Response Theory (IRT) has become a main fragment in psychological testing. Unlike the previous preferred Classical Test Theory, the IRT is based on solving probabilities which link the person´s ability to a latent trait like intelligence.
The Rasch model stands out of the range of IRT models. With its exceeding properties like unidimensionality of the trait, local independency and parallel item characteristic curves the Rasch model is a widely used and popular model for testing.
The main attention of this work is based on testing the fit of the Rasch model. Therefore three different tests are selected from the variety of Rasch model tests to analyse their performance, namely the Andersen test, the Bootstrap test and the mixed–Rasch model test.
In a diverse simulation study the type–one error rates for simulated Rasch conform data and the power for different forms of simulated Rasch model violations are investigated. Results from the Andersen test showed that the test is sensitive for distinguishing intersecting item characteristic curves and an additional lower asymptote. The test fails to detect violations of local independency. In the case of multidimensionality the test performs well only if the loading proportion of the items to the dimensions is unequal. This means a great deal of the items relate to one dimension and only a small part of the items refer to the other dimension. Besides from power results the type–one error rates retain approximately the nominal α–level of 0.05.
Also the Bootstrap test holds this nominal α–level. However the test fails to detect any of the chosen violations, i.e. intersecting item characteristic curves, a lower asymptote, no local independency and multidimensionality. Despite very few exceptions all power values do not exceed the 10% boundary.
The reason for the bad Bootstrap results can be found in the used Pearson χ^2 test statistic. For its calculation all possible response patterns must be taken into account, thus

the number of all possible response patterns for even 10 items becomes huge and therefore the outcomes of expected frequencies becomes tiny. Hence statements about such tiny expected frequencies can be doubtful. Further studies with a different test statistic R_ϕ based on the maximal range of inter–item correlation affirm this assumption. Also studies with only five items confirm the underlying statements of bad Bootstrap power. Results from the mixed–Rasch model test are not demonstrated in this work since they are based on adopted wrong distribution assumptions of the null hypothesis. Analysing the obtained results, it came out that the distribution of the null hypothesis must be assumed to be unknown.
Applying results to practical applications the nine subtests from the Intelligence–Structure–Test (IST) 2000 R are investigated to fit the Rasch model. Whereas the Andersen LR–test (table (6.1) column: LR) rejects the application to the Rasch model for all except one subtest, the Bootstrap test (table (6.1) column: B wgm.(χ^2)) rejects none except two subtests. Reasons for these different outcomes are thought to be due to the doubtful χ^2–distribution assumption from the LR–test and the huge number of small expected response patterns. Obtained results are compared to already taken studies from Bühner, Ziegler, Krumm and Schmidt-Atzert (2006).
The simulation study is written with the statistical software R. The structure of the study as well as the R–code is demonstrated in the appendix.

Since the Andersen LR–test has already been studied by a range of scientists, further studies for the Bootstrap and the mixed–Rasch model test could be considered.
As the Bootstrap–test performs very poorly in association with the Pearson χ^2 test statistic, it would be interesting to compare the performance of the Bootstrap test with different test statistics. These test statistics should contain statistics based on expected frequencies, which means based on all possible frequencies, and statistics referring to only observed values, i.e. single response values. Thus the impact from the huge number of possible response patterns can be researched.
To achieve a distribution of the null hypothesis for the mixed–Rasch model test a Bootstrap routine can be applied and thus type–one error rates and power values of the mixed–Rasch model test can be analysed.

Appendix A

R Code

In the following the main parts of the programming code from the simulation study are demonstrated. After each mapped code a short summary about this part of the code is provided.
Explanatory notes within the code are characterised by the sign "#".

A.1. Simulation Rasch Data

```
sim_rasch <- function(replication=100,persons=100,items=10,
seed=NULL,cutpoint="randomized")
{
  #predefining list
  X_list <- list()

  for (i in 1:replication)
  {
    #simulating rasch data for each element of list
    X_list[[i]] <- sim.rasch(persons=persons,items=items,seed=seed,
    cutpoint=cutpoint)

    #removing extreme values in persons and items
    sum_r <- apply(X_list[[i]],1,sum)
    ind0_r <- (1:length(sum_r))[sum_r==0]
    ind1_r <- (1:length(sum_r))[sum_r==dim(X_list[[i]])[2]]
    ind_r <- c(ind0_r, ind1_r)

    sum_c <- apply(X_list[[i]],2,sum)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_list[[i]])[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_list[[i]]<-X_list[[i]][,-ind_c]
    if (length(ind_r)>0)
    X_list[[i]]<-X_list[[i]][-ind_r,]

  }#end for i

return(X_list)
}#end function
```

The function *sim_rasch* generates *replication*–times Rasch data sets. The actual data simulation is performed with the help of the R–package *eRm*. There the function *sim.rasch* creates a dichotomous Rasch model data matrix due to instruction from the sample size (*persons*), the number of items (*items*), the possibility to chose a seed for random number generation (*seed*) and the performance of the transformation from the probability matrix into the resulting 0/1 matrix (*cutpoint*).
Because of used CML estimation all extreme response patterns, i.e. response patterns with only 0 or only 1 entries, will be removed from calculations.

A.2. Simulation Rasch Violated Data

```
###################
#Non-parallel ICCs#
###################

sim_violation_non_parallel_ICC <- function(replication=100,persons=100,
items=10,discrim=0.25,seed=NULL,cutpoint="randomized")
{
  #predefining list
  X_list <- list()

  for (i in 1:replication)
  {
    #simulating rasch violated data for each element of list
    X_list[[i]] <- sim.2pl(persons=persons,items=items,discrim=discrim,
    seed=seed,cutpoint=cutpoint)

    #removing extreme values in persons and items
    sum_r <- apply(X_list[[i]],1,sum)
    ind0_r <- (1:length(sum_r))[sum_r==0]
    ind1_r <- (1:length(sum_r))[sum_r==dim(X_list[[i]])[2]]
    ind_r <- c(ind0_r, ind1_r)

    sum_c <- apply(X_list[[i]],2,sum)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_list[[i]])[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_list[[i]]<-X_list[[i]][,-ind_c]

    if (length(ind_r)>0)
    X_list[[i]]<-X_list[[i]][-ind_r,]

  }#end for i

return(X_list)
}#end function
```

The violation *non parallel ICCs* is implemented with the simulation of the 2–PL model. The function *sim_violation_non_parallel_ICC* generates 2–PL model data matrices as

often as the given instruction in *replication*. With the help of the function *sim.2pl* from the R–package *eRm* such matrices can be provided. The instructions from the variables *persons, items, seed* and *cutpoint* are the same as for the simulation of Rasch conform data in section (A.1). Additionally, the standard deviation of the log normal–distribution is to specify (*discrim*). Hence from this $logN(0, \sigma^2)$–distribution the discrimination parameter α of the 2–PL model is drawn.
Again extreme response patterns are removed.

```
######################
#No local independency#
######################

sim_violation_no_local_indep <- function(replication=100,persons=100,
items=10,it.cor,seed=NULL,cutpoint="randomized")
{
  #predefining list
  X_list <- list()

  for (i in 1:replication)
  {
    #simulating rasch violated data for each element of list
    X_list[[i]] <- sim.locdep(persons=persons,items=items,it.cor=it.cor,
    seed=seed,cutpoint=cutpoint)

    #removing extreme values in persons and items
    sum_r <- apply(X_list[[i]],1,sum)
    ind0_r <- (1:length(sum_r))[sum_r==0]
    ind1_r <- (1:length(sum_r))[sum_r==dim(X_list[[i]])[2]]
    ind_r <- c(ind0_r, ind1_r)

    sum_c <- apply(X_list[[i]],2,sum)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_list[[i]])[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_list[[i]]<-X_list[[i]][,-ind_c]

    if (length(ind_r)>0)
    X_list[[i]]<-X_list[[i]][-ind_r,]

  }#end for i

return(X_list)
}#end function
```

The correlation between 5 arbitrarily selected items serves for the violation of *local independency*. The function *sim_violation_no_local_indep* constitutes data matrices with correlations between items according to the specification in *replication*. The variables

persons, items, seed and *cutpoint* indicate the same statements as in the latter defined data simulations (e.g. in section (A.1)). *it.cor* denotes the correlation matrix, where entries for 5 items are filled with the corresponding violation degree.
Again extreme outcomes of individuals and items are taken out of the matrices.

```
######################
#No unidimensionality#
######################

sim_violation_no_unidim <- function(replication=100,persons=100,
items=10,Sigma,weightmat,seed=NULL,cutpoint="randomized")
{
  #predefining list
  X_list <- list()

  for (i in 1:replication)
  {
    #simulating rasch violated data for each element of list
    X_list[[i]] <- sim.xdim(persons=persons,items=items,Sigma,weightmat,
    seed=seed,cutpoint=cutpoint)

    #removing extreme values in persons and items
    sum_r <- apply(X_list[[i]],1,sum)
    ind0_r <- (1:length(sum_r))[sum_r==0]
    ind1_r <- (1:length(sum_r))[sum_r==dim(X_list[[i]])[2]]
    ind_r <- c(ind0_r, ind1_r)

    sum_c <- apply(X_list[[i]],2,sum)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_list[[i]])[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_list[[i]]<-X_list[[i]][,-ind_c]

    if (length(ind_r)>0)
    X_list[[i]]<-X_list[[i]][-ind_r,]

  }#end for i

return(X_list)
}#end function
```

Apart from the already known expressions for *replication, persons, items, seed* and *cutpoint* the simulation of the *unidimensionality* violation is based on two further statements. *Sigma* denotes the correlation matrix of the two dimensions, whereas the loading proportion of the items to the dimensions is given in the substitution parameter *weightmat*.
Since extreme response patterns are unutilised in CML estimation, they are removed.

```
#################################################################
#3-PL model - Model with discrimination and guessing parameter#
#################################################################
sim_violation_3PL <- function(replication=100,persons=100,
items=10,discrim=0.25,seed=NULL,cutpoint="randomized")
{
  #predefining list
  X_list <- list()
  for (i in 1:replication)
  {
    #predefining within loop
    psolve <- matrix(0,persons,items)

    #drawing parameters
    if (!is.null(seed)) set.seed(seed)
    capable <- rnorm(persons)

    if (!is.null(seed)) set.seed(seed)
    difficult <- rnorm(items)

    if (!is.null(seed)) set.seed(seed)
    alpha <- rlnorm(items,0,sdlog=discrim)

    if (!is.null(seed)) set.seed(seed)
    gamma <- rbeta(items,2,items)

    #Generating probabilities
    for (j in 1:persons)
    {
           for (k in 1:items)
      {
        psolve[j,k]<-gamma[k]+(1-gamma[k])*(exp(alpha[k]*
        (capable[j]-difficult[k]))/(1+exp(alpha[k]*(capable[j]-difficult[k]))))
      }#end for k
    }#end for j

    #Generating (0,1) matrix with latter probabilities
    if (cutpoint == "randomized")
    {
      if (!is.null(seed)) set.seed(seed)
      X_list[[i]] <-(matrix(runif(items*persons),persons,items) < psolve)*1
    }#end if
    else
    {
      X_list[[i]] <- (cutpoint < psolve)*1
    }#end else

    #removing extreme values in persons and items
    sum_r <- apply(X_list[[i]],1,sum)
    ind0_r <- (1:length(sum_r))[sum_r==0]
    ind1_r <- (1:length(sum_r))[sum_r==dim(X_list[[i]])[2]]
    ind_r <- c(ind0_r, ind1_r)

    sum_c <- apply(X_list[[i]],2,sum)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_list[[i]])[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_list[[i]]<-X_list[[i]][,-ind_c]
    if (length(ind_r)>0)
    X_list[[i]]<-X_list[[i]][-ind_r,]
  }#end for i

return(X_list)
}#end function
```

Besides the already explained simulated data matrices, the simulation of the 3–PL model is not generated with the help of the R-package *eRm*, but build up analogously to their structure.
The replacement character *replication, persons, items, seed* and *cutpoint* have the same features as described in section (A.1). More precisely, the person and the item parameter are drawn from a standard normal distribution and the discrimination parameter from a log normal distribution with zero mean and a standard deviation specified in the appeal of the function. The guessing parameter is drawn from a *Beta(2, items)* distribution, where *items* indicates the number of items.
The solving probabilities are generated according to the 3–PL model equation (refer to chapter 2.4). The transformation to the resulting dichotomous response matrix is realised with random standard uniform deviates.
Also here extreme response patterns are excluded.

A.3. Andersen Test

```
source("myandersen.r")

#########################################################
#Function type I error rates for Rasch homogeneous data#
#########################################################

andersen_typeI_error_rate <- function(X_list, replication=100)
{
  #inits
  reject <- 0

  #creating n pvalues
  pvalues <- list()
  pvalues <- lapply(X_list,myandersen)
  pvalues <- unlist(pvalues)

  #Count type I error
  reject <- sum(pvalues < 0.05)

  type_I_error <- reject/replication

  #return vector of pvalues and type-I error rate
  return(list(andersen_pvalues=pvalues, andersen_type_I_error=type_I_error))
}#end function
```

```
#######################################
#Function power for Rasch violated data#
#######################################

andersen_power <- function(X_list, replication=100)
{
  #inits
  failure <- 0

  #creating n pvalues
  pvalues <- list()
  pvalues <- lapply(X_list,myandersen)
  pvalues <- unlist(pvalues)

  #Type II error rates
  failure <- sum(pvalues > 0.05)

  #calculating power
  power <- 1-(failure/replication)

  #return vector of pvalues and power
  return(list(andersen_pvalues=pvalues, andersen_power=power))
}#end function
```

The Andersen test routine contains two functions. The function *andersen_typeI_error_rate* calculates the type–one error rates for the Rasch conform data sets, generated in section (A.1). And the other function *andersen_power* calculates the power for the Rasch violated data (section (A.2)). *X_list* denotes the outcome of the data simulation routines, i.e. *replication*–times data matrices.
The real Andersen LR–test is executed by the function *myandersen*, explained in the following.
The number of rejections for Rasch data and Rasch violated data indicates the type–one error rate and the power, respectively.

```
myandersen <- function(z)
{
  #estimating Rasch data and performing LRtest
  rm <- RM(z)
  lr_mean <- LRtest(rm, splitcr="mean")
  pvalue <- lr_mean$pvalue

  return(pvalue)

}#end function
```

With the help of the R–package *eRm* the LR–test from Andersen is realised. First the parameter estimates of the item parameter for the whole data are provided through the function *RM* from *eRm*. With these estimates the LR–test with two subgroups based on raw score splitting is performed. The resulting *pvalue* serves for an indicator of model rejection or model adoption for a level of significance for 5%.
Operating the function *andersen_typeI_error_rate* for all variations in Rasch conform

data matrices and the function *andersen_power* for all variations in Rasch violations follow the resulting figures from chapter 5.

A.4. Bootstrap Test

The R–routine for the Bootstrap test is based on a network of functions. Which function relates to which routine is demonstrated with the command *source*, that means functions that are called within a procedure must be included in this routine.

```
source("bootstrap_function.r")

#########################################################
#Function Type I error rates for Rasch homogeneous data#
#########################################################

bootstrap_typeI_error_rates <- function(X_list, replication=100, B=1000)
{
  #inits
  pvalues <- rep(0,replication)
  reject <- 0
  #creating n pvalues
  for(i in 1:replication)
  {
              #Bootstrapping with constant row sum sampling
              pvalues[i] <- bootstrap_suff(B=B,X=X_list[[i]])

              #Count rejections
             if(pvalues[i] < 0.05)
             {
               reject <- reject+1
             }#end if
  }#end for i

  #Count type I error
  type_I_error <- reject/replication
  #return vector of pvalues and type I error rates
  return(list(bootstrap_pvalues=pvalues, bootstrap_type_I_error=type_I_error))
}#end function
```

```
#######################################
#Function power for Rasch violated data#
#######################################

bootstrap_power <- function(X_list, replication=100, B=1000)
{
  #inits
  pvalues <- rep(0,replication)
  failure <-0
  #creating n pvalues
  for(i in 1:replication)
  {
             #Bootstrap with constant row sum sampling
             pvalues[i] <- bootstrap_suff(B=B,X=X_list[[i]])

             #Count failures
            if(pvalues[i] > 0.05)
            {
              failure <- failure+1
            }#end if
  }#end for i

  #calculating power
  power <- 1-(failure/replication)
  #return vector of pvalues and power
  return(list(bootstrap_pvalues=pvalues, bootstrap_power=power))
}#end function
```

The main functions for the Bootstrap test are the *bootstrap_typeI_error_rates* function for Rasch conform data and the *bootstrap_power* function for Rasch violated data. These operations return the type–one error rate and the power value according to the achieved p–values from the test, respectively. The parameter B in the header of the function indicates the number of Bootstrap replications, which must no be confused with the replications for generating error rates and power values.
The function *bootstrap_suff* is included in the routine and will be explained next.

```
library(multicore)

source("const_rowsum_function.r")
source("sampling_function.r")
source("pearson_chi.r")
source("myb.r")

bootstrap_suff <- function(B=1000,X)
{
  #estimating data
  rm <- RM(X)
  sigma <- -(rm$betapar)
  p.rm <- person.parameter(rm)
  prob_X <- pmat(p.rm)

  #predefinings
  p_indicator <- 0
  pvalue <- 0

  #Test statistic on observed values
  Tobs <- pearson.chi(X,sigma)

  #probability matrix with constant row sum
  SFK <- const_rowsum(X=X, prob_X=prob_X)

  #Bootstrap-repetitions
  Tstat <- list()
  Tstat <- mclapply(1:B,function (i){
                 b <- sampling(X,prob_X,SFK)
                 myb(b)},mc.set.seed=TRUE,mc.cores=32)

  #counting Tstat > T_obs
  p_indicator <- sum(Tstat > Tobs)

  #calculating the p-value with the formula [1+sum_{i_1}^{B}I(T_i>T_obs)/(B+1)]
  pvalue <- (1+p_indicator)/(B+1)

  return(pvalue)
}#end function
```

This function can be seen as the core of the whole Bootstrap routine. All main actions occur here.

First the estimates of the item parameters as well as the estimates of the person parameters are computed and the resulting solving probability matrix is generated. With the subsequent explained function *pearson.chi*, the Pearson χ^2 test statistic for the observed data is calculated.

To sample matrices with given marginals the function *const_rowsum* returns a list with solving probabilities under the constraint of given row sums. With these probabilities response matrices, containing the same row sums as the original data matrices, are provided and for each of the sampled matrix the pearson χ^2 test statistic is calculated.

With the upper defined formula, the p–values can be computed and can be included to the function *bootstrap_typeI_error_rates* and *bootstrap_power*, respectively.

The command *library(multicore)* includes further instructions which are not already implemented in the R–basis source and expand the range of commands. The appeal

mclapply parallelises the R program to the indicated number of CPUs, i.e. here 32 CPU´s.
In the following, included functions from the routine *bootstrap_suff* will be explained.

```
const_rowsum <- function(X, prob_X)
{
  #inits before loops
  prob_X_const <- matrix(nrow=nrow(X), ncol=ncol(X))
  SFK <- list()
  #generating start of recursive algorithm
  #attend: position [1] is sum 0!
  SFK[[1]] <- matrix(0,nrow=nrow(prob_X),ncol=ncol(prob_X)+1)
  SFK[[1]][,1] <- 1-prob_X[,ncol(prob_X)]
  SFK[[1]][,2] <- prob_X[,ncol(prob_X)]

  #for different starting points
  for (k in 2:ncol(prob_X))
  {
    #inits
    SFK[[k]] <- matrix(0,nrow=nrow(prob_X),ncol=ncol(prob_X)+1)
    # for each row of data matrix
    for (i in 1:nrow(prob_X))
    {
      #if sum=0
      SFK[[k]][i,1] <- (1-prob_X[i,ncol(prob_X)-(k-1)])*SFK[[k-1]][i,1]

      #if 0<sum<number of items
      for (j in 2:k)
      {
        SFK[[k]][i,j] <- prob_X[i,ncol(prob_X)-(k-1)]*
        SFK[[k-1]][i,j-1]+(1-prob_X[i,ncol(prob_X)-(k-1)])*SFK[[k-1]][i,j]
      }# end for j

      #if sum= number of items
      SFK[[k]][i,k+1] <- prob_X[i,ncol(prob_X)-(k-1)]*SFK[[k-1]][i,k]
    }#end for i
  }#end for k
#-------------------------------------------------------------------------------
  #for better using of SFK´s: removing of unuseful columns
  for (k in 1:ncol(prob_X))
  {
    if (k < ncol(prob_X))
    {
      weg <- ((ncol(prob_X)+1)-((ncol(prob_X)+1)-(k+1))):ncol(prob_X)+1
      SFK[[k]] <- SFK[[k]][,-weg]
    }#end if
  }#end for
  return(SFK)
}#end function
```

The function *const_rowsum* generates, as already mentioned, a list containing the solving probabilities under the constraint of given marginals, i.e. given row sums. The algorithm used in this routine can be elaborately found in chapter 4.5.2.

```
sampling <- function(X,prob_X,SFK)
{
  #inits before loop
  X_const_sample <- matrix(0,nrow=nrow(prob_X), ncol=ncol(prob_X))
  prob_X_const <- matrix(0,nrow=nrow(prob_X), ncol=ncol(prob_X))
  #iterative sampling algorithm for each element of matrix
  for (i in 1: nrow(prob_X))
  {
      #getting row sum of matrix to be sampled
      summe_ganz <- sum(X[i,])

      for (j in 1: ncol(prob_X))
      {
        #getting row sum of each position in row i
        summe <- sum(X_const_sample[i,0:(j-1)])
        #difference
        diff <- summe_ganz-summe
        #extra assignment for j=number of items -> in else arm
        if (j != ncol(prob_X))
        {
          #extra assignment for difference=0
          if(diff==0)
          {
            prob_X_const[i,j] <- 0
          }#end if
          else
          {
            #attend: position [summe+1], because position [1] is sum=0
            prob_X_const[i,j] <- prob_X[i,j]*
            SFK[[ncol(prob_X)-j]][i,(diff)]/SFK[[ncol(prob_X)-j+1]][i,diff+1]
          }#end else
        }#end if
        else
        {
            #special assignment for last element of row
            if(diff==1)
            prob_X_const[i,j] <- 1
            else
            prob_X_const[i,j] <- 0
        }#end else

        #sampling each element of matrix with latter probabilities
        X_const_sample[i,j] <- (runif(1)<prob_X_const[i,j])*1

      }#end for j
  }#end for i
  return(X_const_sample)
}#end function
```

With the algorithm described in chapter 4.5.2 the function *sampling* samples data matrices with the latter provided solving probabilities with given marginals and returns them into the function *bootstrap_suff* for further calculations.

```
myb <- function(z) {

    Tstat <- 0

    X_const_sample <- z

    #removing extreme values in items
    #there can not be extreme values in persons because of constant row sums
    sum_c <- colSums(X_const_sample)
    ind0_c <- (1:length(sum_c))[sum_c==0]
    ind1_c <- (1:length(sum_c))[sum_c==dim(X_const_sample)[1]]
    ind_c <- c(ind0_c, ind1_c)

    if(length(ind_c)>0)
    X_const_sample<-X_const_sample[,-ind_c]

    #estimating data
    rm_sample <- RM(X_const_sample)
    sigma_sample <- -(rm_sample$betapar)

    #calculating test statistic with constant row sum sampled matrix
    Tstat <- pearson.chi(X_const_sample, sigma_sample)

    return(Tstat)

}#end function
```

Function *myb* removes extreme values in items from the committed data matrix of *bootstrap_suff*. Extreme response patterns for persons can not occur here due to retained row sums from sampling algorithm. Subsequently, item parameter estimates for the sampled matrices are computed and the corresponding χ^2 test statistic is calculated. These values can then be returned to function *bootstrap_suff* for the generation of the p–value.

```
source("symmetric_function.r")

pearson.chi <- function(X, sigma) {

  #preparing data
  X_count <- apply(X, 1, paste, collapse = "/")

  #inits
  Obs <- rep(0,length(X_count))
  Exp <- rep(0,length(X_count))
  SF <- rep(0,ncol(X))
  p_pattern <-rep(0,length(X_count))

  #generating symmetric functions
  SF <- symmetric_function(X,sigma)

  #generating count of unique persons scores
  person.scores <- rowSums(X)
  Freq_r <- as.vector(table(factor(person.scores, levels = 1:ncol(X))))
  #adding score 0
  Freq_r <- c(nrow(X)-sum(Freq_r)[1],Freq_r)

  #observed frequencies
  match <- match(X_count, X_count)
  match <- factor(match, levels = unique(match))
  Obs = as.vector(table(match))

  #expected frequencies
  positions <- rowSums(unique(X))
  summe <- as.matrix(unique(X))\%*\%sigma
  zw <- as.vector(summe[,1])
  p_pattern <- exp(-zw)/SF[positions+1]
  Exp <- p_pattern*Freq_r[positions+1]

  #eliminating complete 0 or 1 response patterns
  Obs <- Obs[-(c(1,nrow(X)))]
  Exp <- Exp[-(c(1,nrow(X)))]

  #respecting Exp==0
  if (any(ind <- Exp == 0))
  Exp[ind] <- 0.001

  #generating Test statistic
  Tstatistic <- sum(((Obs-Exp)^2)/Exp)+sum(Obs)-sum(Exp)

  return(Tstatistic)
}#end function
```

Finally the R–code from the function *pearson.chi* is shown.
The elementary symmetric functions must be generated in order to obtain expected frequencies. Additionally, with the counts of unique person scores the expected frequencies can be calculated.
Observed frequencies are obtained from single enumeration of given frequencies.
After some computational transformations, the Pearson χ^2 test statistic is computed for each sampled matrix with given row sum.
The function *symmetric_function* is illustrated in the following.

```
symmetric_function <- function(X, sigma){

  #inits before loops
  SF <- rep(0,ncol(X)+1)
  g <- matrix(nrow=ncol(X)+1,ncol=ncol(X))
  gsumme <-0

  #vector with element wise exp(-sigma)
  term <-  exp(-sigma)

  #restriction
  g[1,] <- 1

  #symmetric function with score sum =0
  SF[1] <- 1

  #generating symmetric function with recursive algorithm from Fischer (1974)
  for (i in 1:ncol(X))
  {
      gsumme <- sum(g[i,]*term)
      gsumme <- gsumme/i
      SF[i+1] <- gsumme

      for (j in 1:ncol(X))
      {
        g[i+1,j] <- gsumme-term[j]*g[i,j]
      }#end forj
  }#end fori

  return(SF)
}#end function
```

As stated before, the function *symmetric_function* generates the elementary symmetric function of a data matrix and their item parameter estimates. The algorithm used here is from Fischer (1974) and is explained in chapter 3.1.3.

By applying the Bootstrap–routine to each combination of scenarios based on the number of items, sample sizes and different forms of model violations or Rasch model conformity, the results displayed and explained in chapter 5 emerge.

Bibliography

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

Alagumalai, S. & Curtis, D.D. & Hungi, N. (2005). *Applied Rasch Measurement: A book of exemplars*. Dordrecht: Springer-Kluwer.

Agresti, A. & Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis, 5*, 9-21.

Andersen, E.B. (1971). *Conditional inference for multiple choice questionnaires*, Report No. 8, Copenhagen: Copenhagen School for Economics and Business Administration.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42-54.

Andersen, E.B. (1973a). *Conditional inference and models for measuring.* Mentalhygiejnisk Forskningsinstitut: Copenhagen.

Andersen, E.B. (1973b). A goodness of fit test for the Rasch Model. *Psychometrika, 38*, 123-140.

Baker, F.B. (1992). *Item Response Theory. Parameter estimation techniques*. New York: Marcel Dekker.

Besag, J. & Clifford, P. (1989). Generalized Monte-Carlo significance tests. *Biometrika, 76*, 633-642.

Bezruczko, N. (2005). *Rasch Measurement in Health Sciences*. Maple Grove, MN: JAM Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee´s ability. In Lord, F.M. & Novick, M.R., *Statistical Theories of Mental Test Scores*, 395-479. Reading, MA: Addison-Wesley.

Bollen, K.A. & Stine, R.A. (1993). Bootstrapping goodness–of–fit measures in structural equation models. In Bollen, K. & Long, J., *Testing Structural Equation Models*, 111-135. Sage Focus Edition: Newbury Park.

Bühner, M. (2006). *Einführung in die Fragebogenkonstruktion*. München: Pearson Studium.

Bühner, M. & Ziegler, M. & Krumm, S. & Schmidt–Atzert, L. (2006). Ist der I–S–T 2000 R Rasch–skalierbar? [Is the I–S–T 2000 R Rasch scaleable?]. *Diagnostica, 52*, 119-130.

Chen, Y. & Diaconics, P. & Holmes, S. & Liu, J. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association, 100*, 109-120.

Chen, Y. & Small, D. (2005). Exact tests for the Rasch Model via sequential importance sampling. *Psychometrika, 70*, 11-30.

Connor, E.F. & Simberloff, D. (1979). The assembly of species communities: chance or competition? *Ecology, 60*, 1132-1140.

Cressie, N. & Read, T. (1984). Multinomial goodness–of–fit tests. *Journal of the Royal Statistical Society, Series B, 46*, 440-464.

Davier v., M. (1997). Bootstrapping goodness of fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online, 2*, 29-48.

Davier v., M. (1998). WINMIRA: A Windows program for mixed Rasch Models. Kiel: IPN.

Davier v., M. & Yamamoto, K. (2007). Mixture–distribution and hybrid Rasch Models. In Davier v., M. & Carstensen, C.H., *Multivariate and Mixture Distribution Rasch Models. Extensions and Applications*, 99-115. New York: Springer.

Dempster, A.P. & Laird, N. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Draxler, C. (2007). *Sequentielle Tests für das Rasch Modell* [Sequential tests for the Rasch Model]. Published Doctoral Dissertation, Berlin: Logos Verlag.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Ann. Statistics, 7*, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall: New York.

Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch Model. *Psychometrika, 46*, 59-77.

Fischer, G.H. (1995). Derivations of the Rasch Model. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations, Recent Developments, and Applications*, 15-38. New York: Springer.

Fischer, G.H. & Scheiblechner, H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch [Algorithms and programs for Raschs probabilistic test Model]. *Psychologische Beiträge, 12*, 23-51.

Fischer, G.H. & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59*, 177-192.

Fischer, G.H. & Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.

Formann, A.K. (1986). A note on the computation of the second order derivative of the elementary symmetric functions in the Rasch Model. *Psychometrika, 51*, 335-339.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch Model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch Models*. Doctoral Thesis. Enschede: University of Twente.

Glas, C.A.W. (1992). A Rasch Model with a multivariate distribution of ability. In Wilson M., *Objective Measurement: Foundations, Recent Developments, and Applications*, 236-258. Norwood, NJ: Ablex.

Glas, C.A.W. & Verhelst, N. (1995). Testing the Rasch Model. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations, Recent Developments, and Applications*, 69-96. New York: Springer.

Glas, C.A.W. & Meijer, R.R. (2003). A Bayesian approach to person-fit analysis in item response theory models. *Applied Psychological Measurement, 27*, 217-233.

Gustaffson, J.E. (1977). *The Rasch Model of Dichotomous Items: Theory, Applications and a Computer Program*. Göteborg: Institute of Education, University of Göteborg.

Gustaffson, J.E. (1980). Testing and obtaining the fit of data to the Rasch Model. *British Journal of Mathematical and Statistical Psychology, 33*, 205-233.

Guttman, L. (1950). Chapter 2, 3, 6, 8, and 9. In Stouffer, S.A., *Measurement and Prediction*. Princeton, NJ: Princeton University Press.

Haberman, J.S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5*, 815-841.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hambleton, R.K. & Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications.

Hoijtink, H. & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch Model. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations, Recent Developments, and Applications*, 53-68. New York: Springer.

Holland, P.W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hullin, C.L. & Drasgow, F. & Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

Kelderman, H. (1984). Loglinear Rasch Model tests. *Psychometrika, 49*, 223-245.

Kelderman, H. & Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Klauer, K.C. (1995). The assessment of person fit. In Fischer, G.H. & Molenaar I.W., *Rasch Models: Foundations, Recent Developments, and Applications*, 97-110. New York: Springer.

Langeheine, R. & Pannekoek, J. & van de Pol, F. (1996). Bootstrapping Goodness-of-Fit measures in categorical data analysis. *Sociological Methods and Research, 24*, 492-516.

Lazarsfeld, P.F. (1950). The logical and mathematical foundations of latent structure analysis. In Stouffer, S.A. & Guttmann, L. & Suchman, E.A. & Lazarsfeld, P.F. & Star, S.A. & Clausen, J.A., *Studies in World War II, Vol. IV: Measurement and Prediction*, 362-412. Princeton, NJ: Princeton University Press.

Lazarsfeld, P.F. (1959). Latent structure analysis. In Koch, S., *Psychology: A study of a science*, 476-542. New York: McGraw-Hill.

Lehmann, E.L. & Romano, J.P. (2005). *Testing statistical hypotheses*. New York: Springer.

Leisch F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11*, 1-18.

Liou, M. (1994). More on the computation of higher-order derivates of the elementary symmetric functions in the Rasch Model. *Applied Psychological Measurement, 18*, 53-62.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer Verlag.

Mair, P. (2006). *Simulation studies for goodness-of-fit statistics in item response theory*. University of Vienna: Unpublished Master Thesis.

Mair, P. & Hatzinger, R. (2007a). CML based estimation of extended Rasch Models with the eRm package in R. *Psychology Science, 49*, 26-43.

Mair, P. & Hatzinger, R. (2007b). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software, 20*, 1-20.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3-18.

Mead, R. (1976). *Assessment of fit of data to the Rasch Model through analysis of residuals*. Doctoral Dissertation, University of Chicago.

Meiser, T. & Stern, E. & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch Models for the analysis of repeated observations. *Methods of Psychological Research Online, 3*, 75-93.

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

Molennar, I.W. (1983). Some improved diagnostics for failure in the Rasch Model. *Psychometrika, 48*, 49-72.

Molenaar, I.W. (1995a). Some background for item response theory and the Rasch Model. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations,Recent Developments, and Applications*, 3-14. New York: Springer.

Molenaar, I.W. (1995b). Estimation of item parameters. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations,Recent Developments, and Applications*, 39-51. New York: Springer.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In Fischer, G.H. & Laming, D., *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, 249-263. New York: Springer-Verlag.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch Model. *Psychometrika, 66*, 437-460.

R Development Core Team. (2009). R: A language and environment for statistical computing. Munich, Germany. (ISBN 3-900051-12-7), URL http://www.R-project.org/.

Rao, A.R. & Jana, R. & Bandyopadhyay, S. (1996). A Markov Chain Monte Carlo method for generating random (0,1)–matrices with given marginals. *The Indian Journal of Statistics, 58*, 225-242.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Read, T. & Cressie, N. (1988). *Goodness–of–Fit Statistics for Discrete Multivariate Data.* Springer Series in Statistics. Springer: New York.

Roberts, A. & Stone, L. (1990). Island–sharing by archipelago species. *Oecologia, 83*, 560-567.

Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrica, 53*, 327-348.

Rost, J. (1990). Rasch Models in latent classes: an integration of two approaches to item analysis. *Journal of Applied Psychological Measurement, 14*, 271-282.

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook test theory, test design]. Bern: Huber.

Rost, J. & Davier v., M. (1995). Mixture distribution Rasch Models. In Fischer, G.H. & Molenaar, I.W., *Rasch Models: Foundations,Recent Developments, and Applications*, 257-268. New York: Springer.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen [Status and perspectives of applied diagnostic techniques: Results of a representative survey of West German psychologists]. *Diagnostica, 41*, 3-20.

Snijders, T.A.B. (1991). Enumeration and simulation methods for 0–1 matrices with given marginals. *Psychometrika,56*, 397-417.

Steck, P. (1997). Aus der Arbeit des Testkuratoriums: Psychologische Testverfahren in der Praxis. Ergebnisse einer Umfrage unter Testanwendern in der Praxis [The use of psychological tests - A survey of professional psychologists]. *Diagnostica, 43*, 267-284.

Suárez-Falcón, J.C. & Glas, C.A.W. (2003). Evaluation of global testing procedures for item fit to the Rasch Model. *British Journal of Mathematical and Statistical Society, 56*, 127-143.

Tollenaar, N. & Mooijaart, A. (2003). Type I errors and power of the parametric Bootstrap Goodness–of–Fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology, 56*, 271-288.

Traub, R.E. & Wolfe, R.G. (1981). Latent Trait Theories and the assessment of educational achievement. *Review of Research in Education, 9*, 377-435.

Verhelst, N. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika, 73*, 705-728.

Verhelst, N. & Molenaar, I.W. (1988). Logit based parameter estimation in the Rasch Model. *Statistica Neerlandica, 42*, 273-295.

Verhelst, N. & Glas, C.A.W. & Verstalen, H. & Eggen, T. (1994). *OPLM: Computer program and manual*. Arnhem: CITO.

Verhelst, N. & Hatzinger, R. & Mair, P. (2007). The Rasch Sampler. *Journal of Statistical Software, 20*, 1-14.

Wainer, H. & Dorans, N.J. & Flaugher, R. & Green, B.F. & Mislevy, R.J. & Steinberg, L. & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Wollenberg van den, A.L. (1979). *The Rasch Model and time-limit tests*. Doctoral Dissertation. University of Nijmegen: Netherlands.

Wollenberg van den, A.L. (1982). Two new test statistics for the Rasch Model. *Psychometrika, 47*, 123-139.