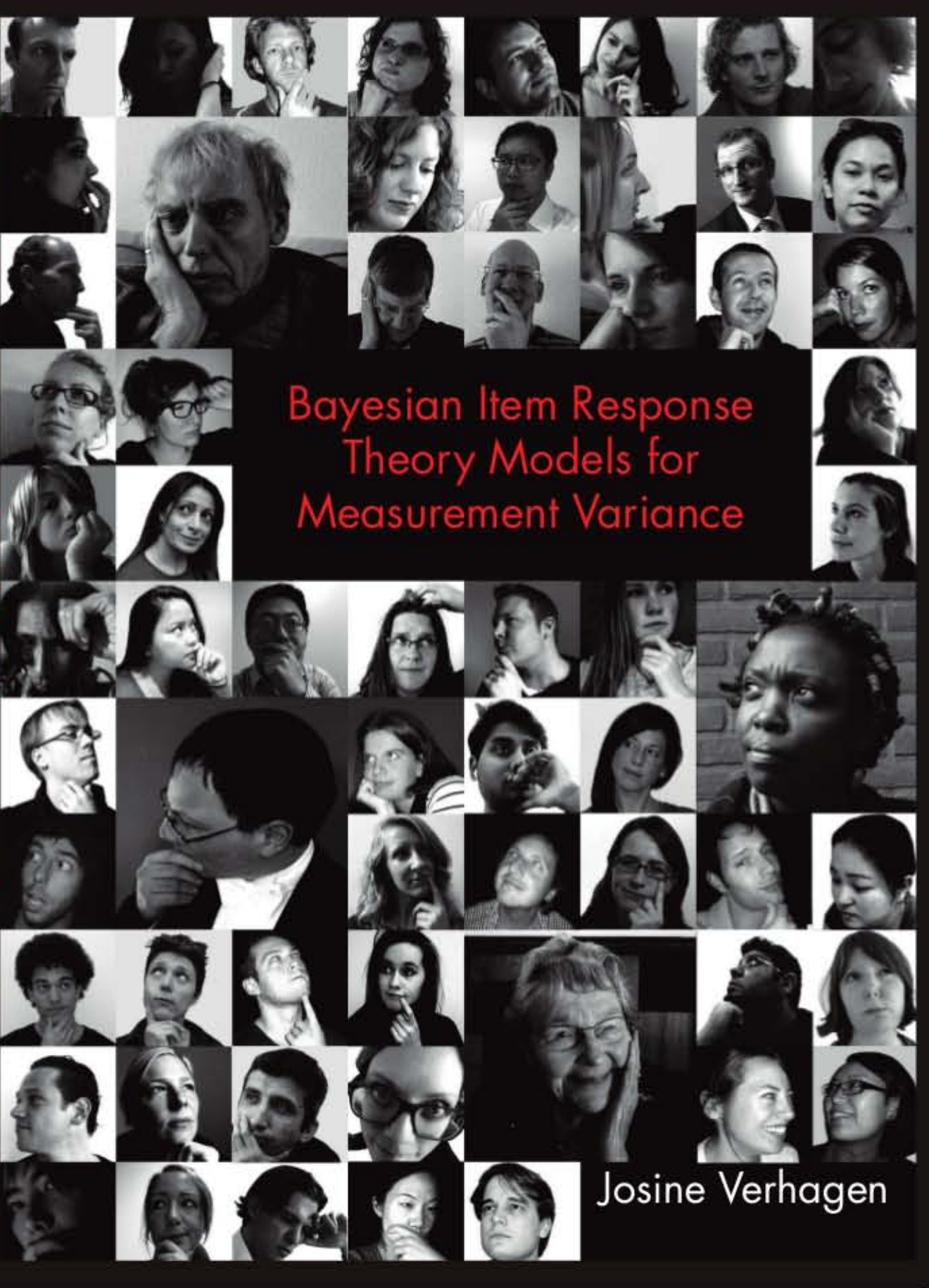

Bayesian Item Response Theory Models for Measurement Variance

Josine Verhagen

Bayesian Item Response Theory Models for Measurement Variance

A.J. Verhagen

November 16, 2012

Graduation Comittee

Chair	Prof. Dr. K. I. van Oudenhoven-van der Zee
Promotor	Prof. Dr. C. A. W. Glas
Assistant promotor	Dr. Ir. G. J. A. Fox
Members	Prof. Dr. Ir. T. J. H. M. Eggen
	Prof. Dr. C. W. A. M. Aarts
	Prof. Dr. J. J. Hox
	Prof. Dr. G. Maris
	Prof. Dr. J. A. M. van der Palen

Verhagen, Anna Jozina
Bayesian Item Response Theory Models for Measurement Variance
Phd Thesis University of Twente, Enschede. - Met samenvatting in het Nederlands.
ISBN: 978-90-365-3469-7
doi: 10.3990/1.9789036534697
printed by: PrintPartners Ipskamp B.V., Enschede
Cover designed by Josine Verhagen with the help of a diverse group of people thinking about a survey question:
Rinke Sophia Rhys Floortje Koos Esmee Bram Chiu Margarita Wim Hinky Alex Amina Christian Shawi Gerhard Annelies Daniël Merel Sylvester Eefje Maartje Mariet Silja Fede Marianna Meen Qil Christina Chen Josien Bas Milou Connie Jesper Ron Marjolein Sabi Aimee Joris Linda Ruud Laura Johan Jory Scott Stephanie Eugene Holly Oma Haitham Danny Dylan Elize Semirhan Lise Lucie Qi Wei Joods Marloes Rick Tiff and Giovane.

BAYESIAN ITEM RESPONSE THEORY MODELS FOR MEASUREMENT VARIANCE

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, November 16, 2012 at 16.45

by

Anna Jozina Verhagen

born November 28, 1982
in Rotterdam, The Netherlands

This dissertation is approved by the following promotores:

Promotor: Prof. Dr. C. A. W. Glas
Assistant promotor: Dr. Ir. G. J. A. Fox

If we cannot end now our
differences, at least we can help
make the world safe for diversity.

John F. Kennedy

Acknowledgements

After a bit more than four years in Enschede, the final product of my work is there. As any PhD project, this has been quite a process, and i will use this section to thank all the people that helped me get to this point.

I really enjoyed working in the department of research methodology, measurement and data analysis (OMD). First of all I would like to thank Jean-Paul Fox, my supervisor, for everything he taught me and his inspirational enthusiasm for mathematical models. Rinke, thank you for your support in the first two years: for your patience in helping me with Fortran, for being a sparring partner and a listening ear, for the "across the hall" conversations, but also for continuing to be there when i needed an external view on things and for your feedback in the last weeks before finishing this thesis, and most of all for being a great friend. Furthermore i would like to thank Marianna and Iris for being wonderful office mates; Connie, Qi Wei, Hanneke, Muirne and Caroline for the many lunch walks we made; Stephanie for the all the advice and inspirational discussions; Sebi for supporting me in my teaching experience and Cees for the support from the sideline at moments in which it really mattered. During the last year of this PhD project I visited Arizona State University, and i would like to thank Roger Millsap en Roy Levy for the discussions we had during that period. They really helped me grow as an independent researcher.

I never regretted joining the board of the PhD Network of the University of Twente (P-NUT). It has been a great experience and I learned a lot about the university, politics and about running and transforming an organization. I would like to thank Anika for teaching me all about P-NUT, and Sergio, Shashank, Juan, Juan Carlos, Bjorn, Silja, Giovane and Rense for the great cooperation and the friendship that grew in this period. I think we can all be proud of where P-NUT is today.

Toastmasters helped me improve my public speaking, listening, improvisation and leadership skills. I would like to thank the toastmasters in all the clubs i visited for giving me feedback, for everyone in Twente toastmasters for becoming such a friendly environment, and in particular Josien van Lanen for being an inspirational mentor.

These 4 years and a bit would not have been the same without the "Enschede crew" to provide the necessary distractions and opportunities to let off steam. I would like to thank Servan for recruiting all the pretty ladies in cubicus for the zeskamp team which was the start of it all. Aimee, thank you for the great friendship and many wine-filled evenings and conversations. A big thank you also

goes to "the ladies" Fede, Lucie and Lise for being such amazing company and for being there for me when needed. Thanks to everyone else of the crew for all the Enschede fun!

Last but not least, thank you mum, dad and Meen for always enquiring about the progress and being supportive when times were rough, and thank you Daniël for the support and for being a welcome distraction during these intense last months.

Contents

Chapter 1

Introduction

In the design and analysis of measurement instruments such as cognitive tests, psychological questionnaires, consumer surveys or attitude questionnaires, a major concern is that the questions should measure the same construct in the same way in all groups the instrument is intended for. Questions should have the same meaning for boys and girls, Chinese and Americans, and elderly and teenagers, at least if we want to make valid comparisons between the total test scores of these groups .

This is not easily achieved. Especially if scores are to be compared between a large number of groups, for example between countries in large international surveys, it is very hard to ascertain that all the questions measure a construct in the same way in all groups. Mathematical problems in an educational test can be more difficult for children with the same ability from countries in which the curriculum does not include similar problems. For males, agreeing with the statement "I had crying spells" indicates a higher level of depression than for females. "I wish I could have more respect for myself" is very important for measuring self-esteem in Americans, but is hardly related to self-esteem in Chinese.

In this thesis, the use of Bayesian Item Response Theory models is investigated for situations in which the measurement instrument does not function in the same way in all groups. On the one hand, tests will be developed to diagnose whether measurement instruments function differently across groups. On the other hand, models will be developed which take these differences into account to enable valid score comparisons and to gain insight into the nature of these differences.

1.1 Measurement invariance

When a measurement instrument measures a construct in the same way in all groups, the instrument exhibits measurement invariance. Measurement invariance is defined as the situation in which persons from each group with the same true value of the measured construct have the same probability for each possible response to all items (e.g. Mellenbergh, 1989; Millsap & Everson, 1993). In an educational test, measurement invariance is present when students from different groups (gender, nationality, ethnic background) with the same ability would have

the same probability of giving the correct answer to all questions.

As an illustration of a situation in which the measurement instrument is not invariant, some items will be used measuring attitudes on the perceived consequences and allowance of immigration from the European Social Survey (ESS, 2004), a large European survey on the attitudes, beliefs and behavior patterns of Europe's diverse populations. Table 1.1, shows that there are large differences in the percentage of respondents in Greece, Norway, Poland and Sweden which agreed with three of the statements.

These differences are partly due to overall differences between countries in the attitude on immigration. The low percentages of Swedes agreeing with each of the statements indicates a relatively positive overall attitude towards immigrants in Sweden, while the high percentages for Greece indicate a relatively negative overall attitude there. But while in Poland and Norway the percentages of respondents agreeing with the first two statements is similar, in Poland a much higher percentage of respondents agrees with the statement about immigrants taking away jobs than in Norway. This indicates that the response to this item is probably differently related to the attitude towards immigration in Poland than in Norway, which would make the item, and therefore the measurement instrument, not measurement invariant.

Table 1.1: Percentage of respondents agreeing with immigration items in the ESS

	Greece	Norway	Poland	Sweden
1. Allow from poor countries*	87%	39%	44%	15%
2. Make worse country*	67%	39%	27%	17%
3. Take away jobs*	78%	19%	52%	12%

*Item 1: To what extent do you think [country] should allow (0) people from the poorer countries outside Europe to come and live here? Item 2: Is [country] made a worse (1) or a better (0) place to live by people coming to live here from other countries? Item 3: Would you say that people who come to live here generally take jobs away (1) from workers in [country], or generally help to create new jobs? (0)

There are two main concerns when an instrument is not measurement invariant. First, there is the issue of validity (e.g. Borsboom, Mellenbergh & van Heerden, 2004). When items function differently over groups, this is an indication that the measurement instrument does not measure the same construct in the same way in each group, which raises questions about what it is the instrument actually measures. Second, the aim is to measure and compare scores between the groups in an accurate way: the score on the item should reflect the status of the respondent on the construct being measured. When the relation between the item response and the underlying construct is different within each group, ignoring this fact leads to inaccurate measurement and can hence lead to invalid conclusions.

Even though the easiest solution would be to simply delete these "inconvenient" questions from the measurement instrument, sometimes they are essential for measuring the construct in some groups, and removing them would mean sig-

nificant loss of information. Therefore, methods have been developed to estimate comparable scores, correcting for the fact that items do not measure the construct in the same way in all groups. These methods are based on measurement models in which the relations between the item responses and the underlying construct are allowed to be different within each group. Examples are multi-group confirmatory factor models for items with continuous answer scales (e.g. Meredith, 1993) and item response theory (IRT) models for items with categorical answer scales (e.g. Thissen, Steinberg & Wainer, 1993). The next section describes how item response theory models can be used for modeling measurement variance.

1.2 Item Response theory models for measurement variance

There are many ways to model the relation between a person's answers to a test or questionnaire and the score that this person acquires on the test. The traditional and easiest way is to just sum the scores for all the questions. The score then consists of the number of correct answers or of the sum of the ratings on a 5 point scale, for example . However, sum scores do not take into account differences between items, like the difficulty or discriminative power of the items. In this thesis, item response theory (IRT) models will be used to model the relation between test or questionnaire scores and item responses. In IRT models, the probability of a response to an item is a function of the underlying score of a person, also called the person parameter, and the characteristics of the item, called the item parameters. The information about item characteristics provided by IRT models is useful, for instance, to create tests which measure constructs accurately and reliably for respondents from all levels; for comparing different populations or tests; to construct computerized adaptive tests (Van der Linden & Glas, 2000); and for the investigation of measurement invariance. The most basic item response model is the Rasch model (Rasch, 1960), in which the probability of answering an item correctly is a function of the difficulty or threshold of an item and the underlying score or person parameter of a respondent. In an educational test, for example, the probability of a correct answer will be lower for more difficult items, and higher for persons with a higher ability. Other item characteristics can be the discriminative power of the item (Lord & Novick, 1968) or the probability of guessing the answer (See also: Embretson & Reise, 2000).

As an illustration, the three items from the European Social Survey (ESS) described before will be modeled with the two parameter normal ogive (2PNO) IRT model. In this model, the probability of endorsing a statement in the ESS survey for person i can be formulated as a function of the respondent's attitude towards immigration, θ_i, and item parameters for the threshold b_k and discriminative value a_k of the item k:

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k), \tag{1.1}$$

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

Figure 1.1 shows the resulting so-called item characteristic curves (ICC) for the three items presented in Table 1.1, indicating the probability of endorsing an

item as a function of the person parameters, that is, the strength of the attitude θ.

The threshold parameter b_k indicates the attitude value θ at which the probability of agreeing with the statement in the item reaches .5; the higher the threshold parameter, the stronger the attitude of a respondent has to be before he or she agrees with the statement in the item. A higher threshold is represented by the ICC being shifted more to the right. In Figure 1.1, for example, Item 1 has a lower threshold than the other two items.

The discrimination parameter a_k indicates how well the item discriminates between respondents with low and high scores on the attitude, which can be interpreted as how relevant the item is for measuring the attitude. A higher discrimination parameter is represented by a steeper slope of the ICC. In Figure 1.1, for example, Item 3 has a lower discrimination parameter than the other two items.

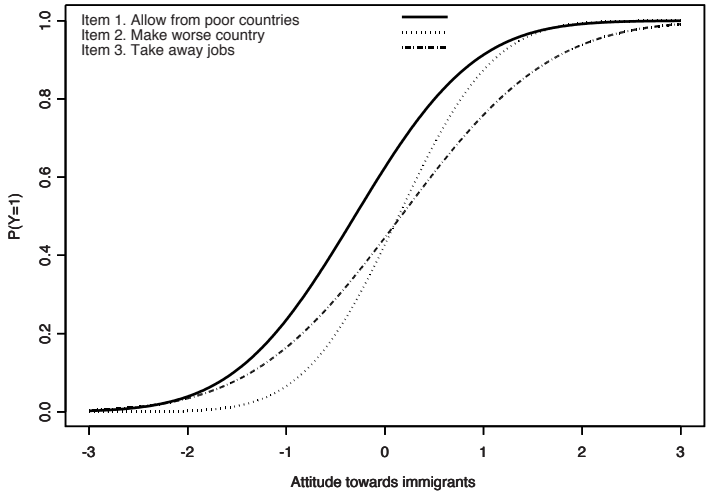

Figure 1.1: Illustration of IRT model: Item characteristic curves for three ESS items.

Measurement invariance for IRT models is defined as the situation in which the item (threshold, discrimination) parameters are equal for all groups. In order to investigate whether this is the case, a model is needed in which the item parameters can be different for each group. Those models will be referred to as multi-group IRT models. Equation 1.1 can be extended with group-specific item parameters for each group j:

$$P(Y_{ijk} = 1 \mid \theta_i, a_{kj}, b_{kj}) = \Phi(a_{kj}\theta_i - b_{kj}). \tag{1.2}$$

Of course, in a situation with multiple groups, one is also interested in overall or mean attitude differences between groups. Therefore, group specific means for the attitude towards immigration θ, μ_{θ_j}, will be specified.

Table 1.2: Country means, country-specific threshold parameters for item 3 and percentage of respondents endorsing item 3*

	Greece	Norway	Poland	Sweden
Country attitude means μ_{θ_j}	1.01	-0.42	-0.36	-1.09
Country thresholds b_{3j}	-0.18	0.76	-0.35	0.67
% endorsing item 3	78%	19%	52%	12%

*Item 3: Would you say that people who come to live here generally take jobs away (1) from workers in [country], or generally help to create new jobs? (0)

Applying this model to the items and countries in Table 1.1, the estimated country means and country threshold parameters for item 3 are given in Table 1.2. As expected, the overall differences in attitude between the countries, as represented by the country attitude means μ_{θ_j}, explain a large portion of the differences in percentages of respondents endorsing the items in Table 1.1. The high mean attitude score of 1.01 in Greece results in a high percentage of Greek respondents endorsing all of the items, while the low mean score of -1.09 in Sweden has the opposite effect. Norway ($-.42$) and Poland ($-.36$), however, have a similar mean on the attitude toward immigrants, but very different percentages of respondents agreeing with the third item. This can be explained by a difference in threshold parameters in both countries b_{3j}, as introduced in the model in Equation 1.2. In Poland this item has a much lower threshold (-.35) than in Norway (.76), indicating that respondents with a relatively low anti-immigrant attitude will agree with this item in Poland, while in Norway the anti-immigrant attitude has to be relatively high for respondents to agree with this statement. The big difference in item parameters between the two countries indicates that measurement invariance does probably not hold.

Many procedures have been developed to test whether differences between item parameters are large enough to conclude that an item is not invariant (see for an overview e.g. Teresi, 2006; Vandenbergh & Lance, 2000). One of the most widely known traditional methods to test for measurement invariance in IRT models is the likelihood ratio test (e.g., Thissen, Steinberg, & Wainer, 1993). Disadvantages of this method include the requirement to indicate some items as invariant beforehand and the large amount of tests necessary in situations with a large number of groups.

The content of the third ESS item covers immigrants taking away jobs. When considering the economic situation of Norway and Poland, one could readily imagine that a concern for jobs might be more pressing in Poland than in Norway. This suspicion is increased by the low item parameter for Greece and the high item parameter for Sweden. This could be investigated by including explanatory information, like the GDP of a country, in the measurement model to explain differences in item parameters between groups. As the traditional multi-group IRT models are predominantly developed for invariance testing, their flexibility to include this type of explanatory information about why item parameters differ across groups is limited.

This thesis explores the use of Bayesian item response theory models to test for differences in item parameters and to model these differences to enable valid score comparisons. Bayesian item response theory models will be developed which allow more flexibility in the measurement model, such as the possibility to include information to explain differences in item parameters. In addition, a different way of testing for measurement invariance is developed.

1.3 Towards Bayesian IRT models and tests for measurement variance

Recently, Bayesian versions of the well-known IRT models have been developed (Albert, 1992; Patz & Junker, 1999a, 1999b). Bayesian estimation methods like Markov Chain Monte Carlo (MCMC) are especially useful for the estimation of complex models. Extending the already complex IRT models with group-specific parameters results in such a complex model, which makes it attractive to use Bayesian estimation methods.

The starting point for this thesis will be the Bayesian multilevel IRT model with random item parameters, as proposed by Fox (2010). In this model, a multilevel or random effects structure is assumed for both the scores on the measured construct, also known as person parameters, and for the item parameters. The multilevel structure on the person parameters (e.g., Fox, 2007; Fox and Glas, 2001) models the person parameters to be normally distributed around their group mean, with different group means and variances for each group. The multilevel structure on the item parameters consists of group-specific threshold (b_{kj}) and discrimination (a_{kj}) parameters, which are normally distributed around higher level general item parameters for each item (a_k and b_k).

The example from the ESS survey can again be used as an illustration. Figure 1.2 illustrates how for the third item, the item characteristic curves for the separate countries vary around one general item characteristic curve for this item, which is represented by the bold line in the center. The differences between the threshold parameters of the countries are represented by the differences in the locations of the curves. The curves for Norway and Sweden are more to the right and the curves for Poland and Greece are more to the left, expressing the relatively high and low threshold parameters for these countries. In addition, the curves for Poland and Greece are steeper, which indicates that this item is more relevant for measuring the attitude towards immigration in these countries, which is represented in higher discrimination parameters.

Bayesian methods (e.g. Gelman et al., 2004) are especially useful for the estimation of this type of hierarchical models, which creates difficulties for estimation in a frequentist framework (see, however Cho & Rabe -Hesketh (2011) for a frequentist estimation method). The hierarchical structure creates standard forms for the conditional posterior distributions of the parameters, which can be used in an MCMC sampling algorithm. To obtain samples from the posterior distribution, iterative draws from these conditional distributions are taken to form a Markov Chain which converges to the posterior distribution. After convergence, these draws are used to obtain estimates of the model parameters based on their

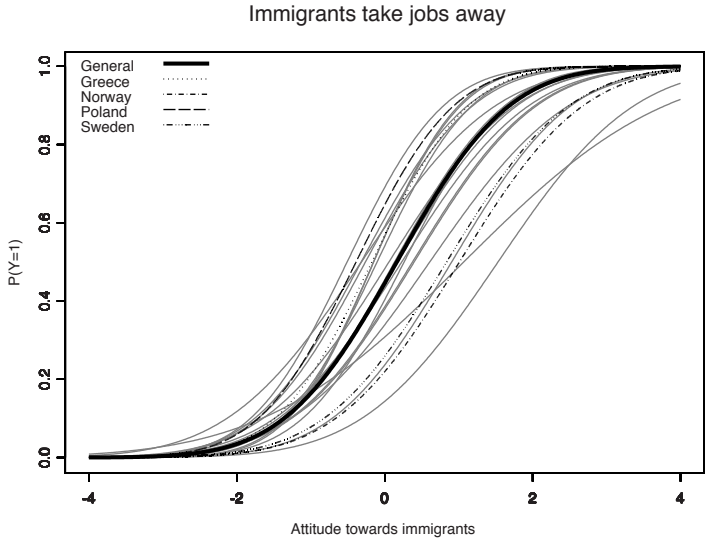

Figure 1.2: Illustration of random item characteristic curves

posterior distribution.

Bayesian IRT models for measurement variance can be extended to make them applicable in a wide range of testing situations. In this thesis, several extensions will be developed. First of all, the specification of multilevel structures allows for the inclusion of person and group level explanatory information on person parameters (chapter 2) and item and group level information on the item parameters (chapter 3). Furthermore, they are easily adapted to account for variance over measurement occasions in longitudinal studies (chapter 4), including time-varying or fixed covariates to explain growth in parameters over occasions. When items have multiple answer categories, a Bayesian version of the Generalized Partial Credit Model (GPCM) (Masters, 1982; Muraki, 1992) can be adapted to account for measurement variance (chapter 4). The multilevel structure on the group-specific parameters represents the assumption that the groups are a random sample from a larger population of groups. This structure only works with a reasonably large number of groups, however. In chapter 5, Bayesian IRT models are described which can be used in case of a smaller number of (fixed) groups. Another extension can be made by including latent class structures to classify, for example, which items are and which are not invariant (chapter 5).

Within the Bayesian framework, an alternative way of testing for measurement invariance can be developed, for example using Bayes factors. Given the data and the prior distributions for the item parameters and item parameter variances, the Bayes factor provides a ratio of the probability of the data under the null hypothesis of invariance (H_0) and the alternative hypothesis of non-invariance (H_1). In this way, support for either hypothesis can be gathered, providing a differentiated view of the evidence for both H_0 and H_1. This in contrast to the

traditional frequentist hypothesis tests, in which the focus is on whether or not the null hypothesis H_0 can be rejected. In chapters 3 and 5, Bayes factor tests will be developed which make it possible to test for measurement invariance within the flexible Bayesian IRT models.

1.4 Outline

In this thesis, the Bayesian IRT model with random item parameters, as proposed by Fox (2010), is explored, extended, and generalized to fit into a framework of Bayesian IRT models for different measurement situations. In addition, a Bayesian framework for testing measurement variance is developed.

Chapter 2 will focus on two aspects of the random item effects multilevel 2PNO model for dichotomous items. First, the flexibility of the model to include a structural multilevel population model, as well as explanatory covariates on the person parameters representing the measured construct, while at the same time accounting for variance in the measurement instrument over groups is investigated. Second, a simulation study will investigate recovery of simulated parameters and convergence, as well as the sensitivity of the parameter estimates to the chosen priors for the variance components. An example based on the Programme for International Student Assessment (PISA) 2003 will illustrate how a measurement variant model with explanatory information on the student level can be estimated and evaluated for fit.

Chapter 3 will investigate Bayesian tests for measurement invariance. A Bayes factor test for the invariance of individual item parameters, a deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) for comparing models with and without invariant item parameters, and a highest posterior density region test (Box and Tiao, 1973) (Appendix C) to assess differences between item parameters will be evaluated. In the second part of this chapter, a model for including explanatory information about differences in item parameters between groups is introduced, to explore why the items are not invariant. After a simulation study showing the performance of the tests, both the tests and the explanatory item information will be illustrated with an application to the European Social Survey attitude towards immigrants questionnaire.

Chapter 4 focuses on item parameter variance in a longitudinal context. First,the random item parameter model and the Bayes Factor tests for invariance are extended to a Generalized Partial Credit Model (Muraki, 1992) for items with more than two answer categories. This involves truncated and correlated multivariate random effects. A joint longitudinal growth structure on both item and person parameters is then implemented. A simulation study will investigate parameter recovery. The model will be illustrated with data from a randomized clinical trial concerning the treatment of depression by increasing psychological acceptance.

Chapter 5 will place the previously described models in a broader framework of Bayesian IRT models for measurement variance. A distinction will be made between Bayesian IRT models for fixed and random groups, and extensions to latent or multiple groups will be discussed. Variations on the Bayes Factor test described in chapter 3 will be presented for implementation in WinBUGS and evaluated in a simulation study comparing them to a traditional likelihood ratio

test for evaluating invariance. The different models will be illustrated for two real test situations.

Chapter 6 concludes the thesis with a discussion of the main conclusions which can be drawn from these chapters, and some suggestions for future directions.

Chapter 2

Random Item Effects Modeling for Cross-National Survey Data

Abstract

The analysis of response data from large-scale surveys is complex due to measurement invariance issues, cross-national dependency structures, and complicated sampling designs. It will be shown that the item response theory model with (cross-national) random item effects is particularly useful for the analysis of cross-national survey data. In this study, the properties of the model and a powerful estimation method are discussed. Model extensions for the purpose of explaining cross-national variation in test characteristics are discussed. An illustration is given of a real-data application using PISA 2003 response data.

2.1 Introduction

Item response theory (IRT) methods are standard tools for the analysis of large-scale assessments of student's performance. In educational survey research, the National Assessment of Educational Progress (NAEP) is primarily focused on scaling the performances of a sample of students in a subject area (e.g., mathematics, reading, science) on a single common scale, and measuring change in educational performance over time. The Organization for Economic Cooperation and Development (OECD) organizes the Program for International Student Assessment (PISA). The programme is focused on measuring and comparing abilities in reading, mathematics and science of 15-year-old pupils over 30 member-countries and various partner countries every three years and started in 2000. Another example is the Trends in International Mathematics and Science Study (TIMSS) conducted

Adapted from: Fox, J.-P. & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 467–488). London: Routledge Academic.

by the International Association for the Evaluation of Educational Achievement (IEA) to measure trends in students' mathematics and science performance.

Large-scale (educational) survey studies can be characterized by (1) the ordinal character of the observations, (2) the complex sampling designs with individuals responding to different sets (booklets) of questions, (3) booklet effects are present (the performance on an item depends on an underlying latent variable but also on the responses to other items in the booklet), and (4) presence of missing data. The presence of booklet effects and missing data complicates an IRT analysis of the survey data. The analysis of large-scale survey data for comparative research is further complicated by several measurement invariance issues (e.g., Meredith & Milsapp, 1992; Steenkamp & Baumgartner, 1998), as assessing comparability of the test scores across countries, cultures and different educational systems is a well-known complex problem. The main issue is that the measurement instrument has to exhibit adequate cross-national equivalence. This means that the calibrations of the measurement instrument remain invariant across populations (e.g., nations, countries) of examinees.

It will be shown that a random item effects model is particularly useful for the analysis of cross-national survey data. The random item effects parameters vary over countries, which leads to non-invariant item characteristics. Thus, cross-national variation in item characteristics is allowed and it is not necessary to establish measurement invariance. The random item effects approach supports the use of country-specific item characteristics and a common measurement scale. Further, the identification of the random item effects model does not depend on marker or anchor items. In current approaches to measurement invariance, at least two invariant marker items are needed to establish a common scale across countries. In theory only one invariant item is needed to fix the scale, but an additional invariant item is needed to be able to test the invariance of this item. Further, a poorly identified scale based on one marker item can easily jeopardize the statistical inferences. Establishing a common scale by marker items is very difficult when there are only a few test items and/or when there are many countries in the sample.

The focus of the current study is on exploring the properties and the possibilities of the random item effects model for the analysis of cross-national survey data. After introducing the model, a short description of the estimation method will be given. Then, in a simulation study, attention is focused on the performance and global convergence property of the estimation method by re-estimating the model parameters given simulated data. Subsequently, an illustration is given of a real-data application using PISA 2003 data.

2.2 Random Item Effects Modeling

IRT methods provide a set of techniques for estimating individual ability (e.g., attitude, behavior, performance) levels and item characteristics from observed discrete multivariate response data. The ability levels cannot be observed directly but are measured via a questionnaire or test. The effects of the persons and the items on the response data are modeled by separate sets of parameters. The person parameters are usually referred to as the latent variables, and the item parameters

are usually labeled item difficulties and item discrimination parameters.

Assume a normal ogive IRT model for binary response data for $k = 1, \ldots, K$ items and $i = 1, \ldots, n$ respondents. The overall item characteristics are denoted as $\boldsymbol{\xi}_k = (a_k, b_k)^t$ representing item difficulty and item discrimination parameters, respectively. The individual ability level is denoted as θ_i. The probit version of the two-parameter IRT model also known as the normal ogive model is defined via a cumulative normal distribution,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k) = \int_{-\infty}^{a_k\theta_i - b_k} \phi(z)dz, \tag{2.1}$$

where $\Phi(.)$ and $\phi(.)$ are the cumulative normal distribution function and the normal density function, respectively. The a_k is referred to as the discrimination parameter and the b_k as the item difficulty parameter.

In Equation 2.1, the item parameters apply to each country and can be regarded as the international item parameters. Without a country-specific index, cross-national variation in item characteristics is not allowed. Following the modeling approach of De Jong, Steenkamp, and Fox (2007), country-specific item characteristics are defined. Let $\tilde{a}_{kj}$ and $\tilde{b}_{kj}$ denote the discrimination and difficulty parameters of item k in country j $(j = 1, \ldots, J)$. As a result, the success probability depends on country-specific item characteristics, that is,

$$P\left(Y_{ijk} = 1 \mid \theta_i, \tilde{a}_{kj}, \tilde{b}_{kj}\right) = \Phi(\tilde{a}_{kj}\theta_i - \tilde{b}_{kj}). \tag{2.2}$$

The country-specific or nation-specific item parameters are based on the corresponding response data from that country. When the sample size per country is small and response bias (e.g. extreme response style, nonrepresentative samples) is present, the country-specific item parameter estimates have high standard errors and they are probably biased. This estimation problem can be averted by a random item effects modeling framework in which the country specific item parameters are considered random deviations from the overall item parameters. The main advantage of this hierarchical modeling approach is that information can be borrowed from the other country-specific item parameters. Therefore, a common population distribution is defined at a higher level for the country-specific item parameters. As a result, a so-called shrinkage estimate comprises the likelihood information at the data level and the information from the common assumed distribution. Typically, the shrinkage estimate of country-specific item parameters has a smaller standard error and gives a more robust estimate in case of response bias.

For each item k, assume an exchangeable prior distribution for the country-specific item parameters. This means that the joint distribution of the country-specific item parameters is invariant under any transformation of the indices. A priori there is no information about an order of the country-specific item characteristics. That is, for each k, for $j = 1, \ldots, J$ holds that:

$$\tilde{\boldsymbol{\xi}}_{kj} = \left(\tilde{a}_{kj}, \tilde{b}_{kj}\right)^t \quad \sim \quad \mathcal{N}\left((a_k, b_k)^t, \boldsymbol{\Sigma}_{\tilde{\xi}}\right), \tag{2.3}$$

where (a_k, b_k) are the international item parameter characteristics of item k and $\boldsymbol{\Sigma}_{\tilde{\xi}}$ the cross-national covariance structure of country-specific characteristics. This

covariance structure is allowed to vary across items. Here, a conditionally independent random item structure is defined with $\mathbf{\Sigma}_{\tilde{\xi}}$ a diagonal matrix with elements $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$.

In most cases there is not much information about the values of the international item parameters. Without a priori knowledge to distinguish the item parameters it is reasonable to assume a common distribution for them. A multivariate normal distributed prior is assumed for the item parameters. It follows that,

$$\boldsymbol{\xi}_k = (a_k, b_k)^t \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi, \mathbf{\Sigma}_\xi\right) \tag{2.4}$$

where the prior parameters are distributed as

$$\begin{aligned} \mathbf{\Sigma}_\xi &\sim \mathcal{IW}(\nu, \mathbf{\Sigma}_0) && (2.5)\\ \boldsymbol{\mu}_\xi \mid \mathbf{\Sigma}_\xi &\sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_\xi/K_0), && (2.6) \end{aligned}$$

for $k = 1, \ldots, K$. The multivariate Normal distribution in Equation (2.4) is the exchangeable prior for the set of K item parameters $\xi_{\mathbf{k}}$. The joint prior distribution for $(\mu_\xi, \mathbf{\Sigma}_\xi)$ is a Normal inverse Wishart distribution, denoted as $\mathcal{IW}$, with parameters $(\mu_{\mathbf{0}}, \mathbf{\Sigma_0}/\mathbf{K_0}; \nu, \mathbf{\Sigma_0})$ where K_0 denotes the number of prior measurements, and ν and $\mathbf{\Sigma_0}$ describe the degrees of freedom and scale matrix of the inverse-Wishart distribution. These parameters are usually fixed at specified values. A proper vague prior is specified with $\mu_{\mathbf{0}} = \mathbf{0}$, $\nu = 2$, a diagonal scale matrix $\mathbf{\Sigma_0}$ with elements 100 and K_0 a small number.

To summarize, the random item effects model can be specified as a normal ogive IRT model with country-specific item parameters, in Equation 2.2. The country-specific item parameters are assumed to have a common population distribution with the mean specified by the international item parameters (Equation 2.4). At a higher level, conjugated proper priors are specified for the international item prior parameters.

In different ways and for different purposes IRT models with item parameters defined as random effects have been proposed. Albers et al. (1989) defined a Rasch model with random item difficulty parameters for an application where items are obtained from an item bank. De Boeck (2008) also considered the Rasch model with random item difficulty parameters. Janssen et al. (2000) defined an IRT model where item parameters (discrimination and difficulty) are allowed to vary across criterions in the context of criterion-referenced testing. Glas et al. (2003) and Glas, Van der Linden and Geerlings (2010) considered the application of item cloning. In this procedure, items are generated by a computer algorithm given a parent item (e.g., item shell or item template). De Jong et al. (2008) used cross-national varying item parameters (discrimination and difficulty) for measuring extreme response style. style.

2.3 Modeling Respondent Heterogeneity

In large-scale survey research, the sampled respondents are often nested in groups (e.g., countries, schools). Subsequently, inferences are to be made at different levels of analysis. At the level of respondents, comparisons can be made between

individual performances. At the group level, mean individual performances can be compared. To facilitate comparisons at different hierarchical levels, a hierarchical population distribution is designed for the respondents.

Common IRT models assume a priori independence between individual abilities. Dependence of results of individuals within the same school/country is to be expected, however, since they share common experiences. A hierarchical population distribution for the ability of the respondents can be specified that accounts for the fact that respondents are nested within clusters. The observations at level-1 are nested within respondents. The respondents at level-2 are nested within groups (level-3) and indexed $i = 1, \ldots, n_j$ for $j = 1, \ldots, J$ groups. Let level-2 respondent-specific covariates (e.g. gender, SES) be denoted by $\mathbf{x}_{ij}$ and level-3 covariates (e.g. school size, mean country SES, type of school system) by $\mathbf{w}_{qj}$ for $q = 0, \ldots, Q$.

A hierarchical population model for the ability of the respondents consists of two stages: the level-2 prior distribution for the ability parameter $\theta_i j$, specified as

$$\theta_{ij} \mid \boldsymbol{\beta}_j \quad \sim \quad \mathcal{N}\left(\mathbf{x}_{ij}^t \boldsymbol{\beta}_j, \sigma_\theta^2\right), \tag{2.7}$$

and the level-3 prior, specified as

$$\boldsymbol{\beta}_j \quad \sim \quad \mathcal{N}\left(\mathbf{w}_j \boldsymbol{\gamma}, \mathbf{T}\right), \tag{2.8}$$

An inverse-gamma prior distribution and an inverse-Wishart prior distribution are specified for the variance components σ_θ^2 and $\mathbf{T}$ respectively. The extension to more levels is easily made.

This structural hierarchical population model is also known as a multilevel model (e.g., Aitkin & Longford, 1986; Bryk & Raudenbush, 1993; de Leeuw & Kreft, 1986; Goldstein, 1995; Snijders & Bosker, 1999).

2.4 Identification and Estimation

The common IRT model (assuming invariant item parameters) with a multilevel population model for the ability parameters is called a multilevel IRT model (MLIRT) (e.g., Fox, 2007; Fox and Glas, 2001). In empirical multilevel studies, estimated ability parameters are often considered to be measured without an error and treated as an observed outcome variable. Ignoring the uncertainty regarding the estimated abilities may lead to biased parameter estimates and the statistical inference may be misleading.

Several comparable approaches are known in the literature. Zwinderman (1991) defined a generalized linear regression model for the observed responses with known item parameters at the lowest level of hierarchy. Adams, Wilson and Wu (1997), Raudenbush and Sampson (1999), and Kamata (2001), defined a generalized linear regression model for the observed responses with item difficulty parameters at the lowest level. This model consists of a Rasch model for the observed responses and a multilevel regression model for the underlying latent variable. Note that a two-parameter IRT model extended with a multilevel model for the latent variable leads to a more complex nonlinear multilevel model since the conditional density

of the responses given the model parameters is not a member of the exponential family which seriously complicates the simultaneous estimation of the model parameters (Skrondal & Rabe-Hesketh, 2004).

In the MLIRT modeling framework the multilevel population model parameters are estimated from the item response data without having to condition on estimated ability parameters. In addition, this modeling framework allows the incorporation of explanatory variables at different levels of hierarchy. The inclusion of explanatory information can be important in various situations, this can for example lead to more accurate item parameter estimates. Another related advantage of the model is that it can handle incomplete data in a very flexible way.

Here, the MLIRT model is extended with a random item effects measurement model. In fact, this is the MLIRT model with non-invariant item parameters as the item parameters are allowed to vary across countries. This MLIRT model with random item effects is not identified since the scale of the latent variable is not defined. When the item parameters are invariant, the model is identified by fixing the mean and variance of the latent scale. In case of non-invariant item parameters, in each country, there is indeterminacy between the latent country-mean (parameterized by a random intercept) and the location of the country-specific item difficulties (parameterized by random difficulty parameters). This indeterminacy is solved by restricting the sum of country-specific difficulties to be zero in each country. The variance of the latent scale can be defined by restricting the product of international item discrimination parameters to be one, or by imposing a restriction on the variance of the latent variable.

The model parameters are estimated simultaneously using an MCMC algorithm that was implemented in Fortran which will be made available in the MLIRT R-package of Fox (2007). The MCMC algorithm consists of drawing iteratively from the full conditional posterior distributions. The chain of sequential draws will converge such that, after a burn-in period, draws are obtained from the joint posterior distribution. These draws are used to make inferences concerning the posterior means, variances, and highest posterior density intervals of parameters of interest.

2.5 Simulation study

The estimation method for the MLIRT model with random item effects is evaluated by investigating convergence properties and by comparing true and estimated parameters for a simulated data set. Different priors for the cross-national discrimination parameter variances are used to investigate the prior influence on the estimation results.

2.5.1 Data Simulation

A data set was simulated with 10,000 cases, 15 items and 20 groups of 500 students. The ability parameters were generated in two steps. First, the mean group ability parameters β_j were generated from a normal $\mathcal{N}\left(0, \tau^2\right)$ distribution, with τ^2 from an inverse gamma $\mathcal{IG}(1,1)$ distribution. The individual ability parameters θ_{ij} were subsequently generated from a normal $\mathcal{N}(\beta_j, \sigma_\theta^2)$ distribution, with σ_θ^2 equal to 1.

International item parameters a_k and b_k were sampled independently from a lognormal distribution with mean $\mu_a = 1$ and standard deviation $\sigma_a = .15$, and a normal distribution with mean $\mu_b = 0$ and standard deviation $\sigma_b = .30$, respectively. Subsequently group specific parameters a_{kj} and b_{kj} were sampled independently from a lognormal distribution with mean a_k and between group standard deviation $\sigma_{a_k} = .20$, and a normal distribution with mean b_k and between group standard deviation $\sigma_{b_k} = .40$, respectively. As a result the group specific discrimination parameters ranged from .32 to 1.79 and the group specific difficulty parameters from -1.16 to 1.32.

Responses were generated by applying the random effects normal ogive IRT model to acquire the success probabilities, comparing this probability with a random number r from a uniform distribution on $(0, 1)$ and assigning a value one when $P\left(Y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_{kj}\right) < r$ and a value zero otherwise.

2.5.2 Procedure

The model was estimated using an MCMC algorithm implemented in Fortran that will be made available in the MLIRT Package (Fox, 2007). To be able to use an MCMC algorithm, prior distributions and initial values for the estimated parameters need to be specified. The initial values were generated from a standard normal distribution for the individual ability parameters and set to zero for the group-specific ability parameters. International and country-specific difficulty parameters were set to zero and the discrimination parameters were set to one. All initial values for the variances were set to one. A number of $20,000$ iterations was run, of which the first $1,000$ iterations were discarded as burn-in period. As an indication of the accuracy of the estimation, correlations between true and estimated parameters, the mean absolute difference between the true and the estimated parameters and the root mean of the squared differences between the true and estimated parameters were computed, all over items and countries.

2.5.3 Investigating Cross-National Prior Variance Dependence

The non-informative priors for the variance components should have as little impact as possible on the final parameter estimates. It is not desirable that cross-national differences in item characteristics are implied by the prior settings. In this section the sensitivity of the prior for the cross-national item discrimination variances is investigated. Analyses showed that prior settings were highly influencing the results.

To examine the prior sensitivity of the cross-national variance of the discrimination parameters $\sigma^2_{a_k}$, several inverse gamma (IG) priors with different scale and shape parameters (1, 1; .1, .1; .01, .01 ;1, .1 ; 1, .01) were investigated for this parameter. The similar correlations between the true and the estimated parameters ($\rho_a = .89 - .91$, $\rho_b = .95$), the similar root mean squared differences ($RMSD_a = .11 - .13$, $RMSD_b = .17$) and the mean absolute differences ($MAD_a = .09$, $MAD_b = .13$) across different priors show that the choice of prior does not affect the difficulty parameter estimates at all and the discrimination

Table 2.1: True and Estimated Cross-National Discrimination Variances for Different Priors.

	True	IG(1 , 1)		IG(1 , 0.1)		IG(1 , 0.01)	
Item	$\sigma^2_{a_k}$	Mean	Sd	Mean	Sd	Mean	Sd
1	0.03	0.15	0.05	0.05	0.05	0.03	0.01
2	0.04	0.16	0.06	0.05	0.05	0.04	0.02
3	0.04	0.14	0.05	0.04	0.04	0.02	0.01
4	0.03	0.13	0.04	0.03	0.03	0.02	0.01
5	0.02	0.13	0.04	0.03	0.03	0.02	0.01
6	0.04	0.15	0.05	0.05	0.05	0.04	0.01
7	0.06	0.15	0.05	0.05	0.05	0.04	0.01
8	0.05	0.16	0.06	0.06	0.06	0.05	0.02
9	0.03	0.17	0.06	0.06	0.06	0.04	0.02
10	0.06	0.16	0.06	0.06	0.06	0.05	0.02
11	0.05	0.16	0.05	0.05	0.05	0.04	0.02
12	0.03	0.14	0.05	0.04	0.04	0.03	0.01
13	0.07	0.17	0.06	0.07	0.07	0.06	0.02
14	0.05	0.15	0.05	0.05	0.05	0.03	0.01
15	0.04	0.16	0.05	0.06	0.06	0.04	0.02

Table 2.2: True and Estimated Cross-National Difficulty Variances for Different Gamma Priors.

	True	IG(1 , 1)		IG(1 , 0.1)		IG(1 , 0.01)	
Item	$\sigma^2_{b_k}$	Mean	Sd	Mean	Sd	Mean	Sd
1	0.19	0.24	0.08	0.14	0.14	0.13	0.05
2	0.17	0.26	0.09	0.17	0.17	0.16	0.06
3	0.13	0.23	0.08	0.13	0.13	0.12	0.04
4	0.12	0.23	0.08	0.13	0.13	0.12	0.04
5	0.11	0.20	0.07	0.10	0.10	0.09	0.03
6	0.12	0.22	0.07	0.12	0.12	0.11	0.04
7	0.13	0.22	0.07	0.12	0.12	0.11	0.04
8	0.12	0.20	0.07	0.10	0.10	0.09	0.03
9	0.17	0.29	0.10	0.19	0.19	0.18	0.06
10	0.15	0.19	0.07	0.09	0.09	0.08	0.03
11	0.15	0.25	0.08	0.15	0.15	0.14	0.05
12	0.11	0.20	0.07	0.10	0.10	0.09	0.03
13	0.13	0.22	0.07	0.12	0.12	0.11	0.04
14	0.19	0.26	0.09	0.17	0.17	0.16	0.05
15	0.13	0.23	0.08	0.13	0.13	0.12	0.04

parameter estimates only slightly. The cross-national item parameter variance estimates are influenced, however.

Table 2.1 and Table 2.2 show that an IG(1, 1) prior resulted in estimates of the cross-national item parameter variances that were consistently too high, and the IG(1, .01) prior resulted in estimates that were consistently slightly lower than the original variances, but within the range of the 95% highest posterior density (HPD) interval. The 95% HPD interval is the interval over which the integral of the posterior density is .95 and the height of the posterior density for every point in the interval is higher than the posterior density for every point outside the interval. Because the posterior density is the distribution of the estimated parameter, the interpretation of this interval is that given the observed data this interval contains the parameter with 95% probability. The other priors performed almost equally well in this respect. With exception of the IG(1,1) prior, all IG prior settings gave almost equal results, so unless a too informative prior is taken the results are not dependent on the choice of prior.

2.5.4 Convergence and parameter recovery

To check whether the MCMC chains have converged, convergence diagnostics and trace plots are inspected for both the cross-national item parameter variances and the international item parameters. The Geweke Z convergence diagnostic is computed by taking the difference between the mean of (a function of) the first n_A iterations and the mean of (a function of) the last n_B iterations, divided by the asymptotic standard error of this difference which is computed from spectral density estimates for the two parts of the chain (Cowles & Carlin, 1996). The result is approximately standard normally distributed. A large Z means that there is a relatively big difference between the values in the two parts of the chain, which indicates the chain is not yet stationary. The autocorrelation is the correlation between values in the chain with a certain lag between them.

The traceplots show a homogeneous band around a mean that after a burn in period stays more or less the same, without trends or large scale fluctuations. The international difficulty parameters and the cross-national variances of the difficulty parameters showed good convergence, with an autocorrelation below .15 and Geweke Z values under 3. This was similar for most discrimination parameters except for the discrimination parameter for item 9, which had an autocorrelation of .31.

In Figure 2.1, examining the traceplot of this parameter some trending is observed, but not in an extreme way. The high discrimination parameter corresponds with high information in a small region of latent scores. As the latent scores in some groups will fall predominantly outside this area, parameter estimates for the item are difficult to make for these groups. In similar situations higher autocorrelations have been found (e.g. Wollack et al., 2002). In general, estimation is better when the highest item information is matched to the latent trait distribution in the sample.

The true item parameters that were used to simulate the dataset were recovered well, as is illustrated in Figure 2.2. The correlations between the true and estimated country-specific and international item parameters were all larger than

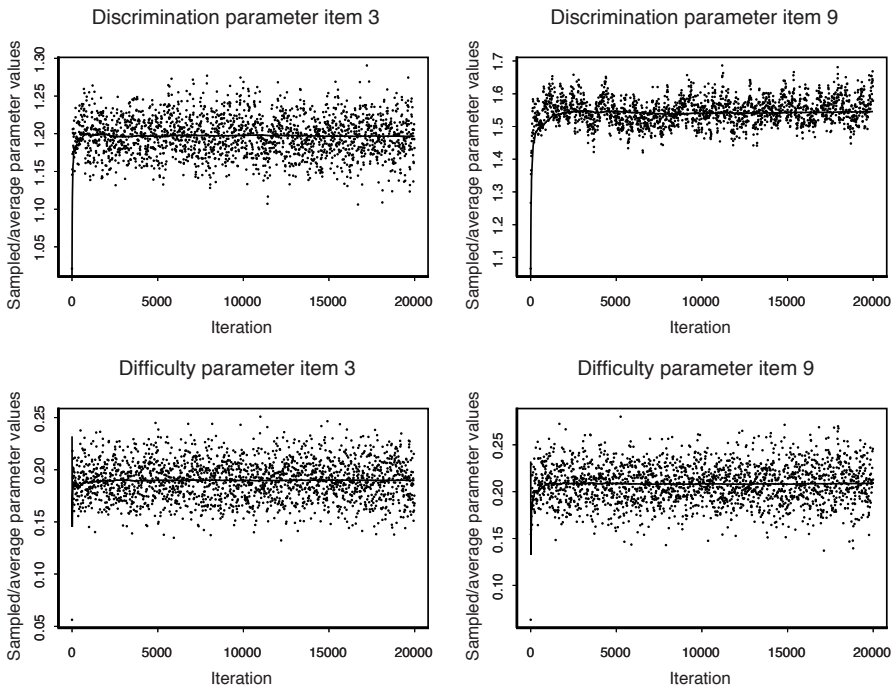

Figure 2.1: Traceplots and moving averages for the item parameters of item 3 and 9.

.91. All true values fall into the 95% HPD intervals, and all estimated parameters were in the right direction. The cross-national item parameter variances and the group means of the ability parameters were also very accurately estimated.

2.6 PISA 2003: Mathematics Data

In this section the random item effects (MLIRT) model will be applied to a data set collected by the Programme for International Student Assessment (PISA) in 2003. PISA is an initiative of the Organization for Economic Cooperation and Development (OECD). Every three years PISA measures the literacy in reading, mathematics and science of 15-year old students across countries, where literacy refers to 'the capacity to apply knowledge and skills and to analyze, reason and communicate effectively as problems are posed, solved and interpreted in a variety of situations' (OECD, 2004). In each data collection round one subject area is emphasized. In 2003 this was mathematic literacy, which resulted in four subdomains for mathematic performance. In addition to subject-specific knowledge, cross-curricular competencies as motivation to learn, self-beliefs, learning strategies and familiarity with computers were measured. Furthermore the students

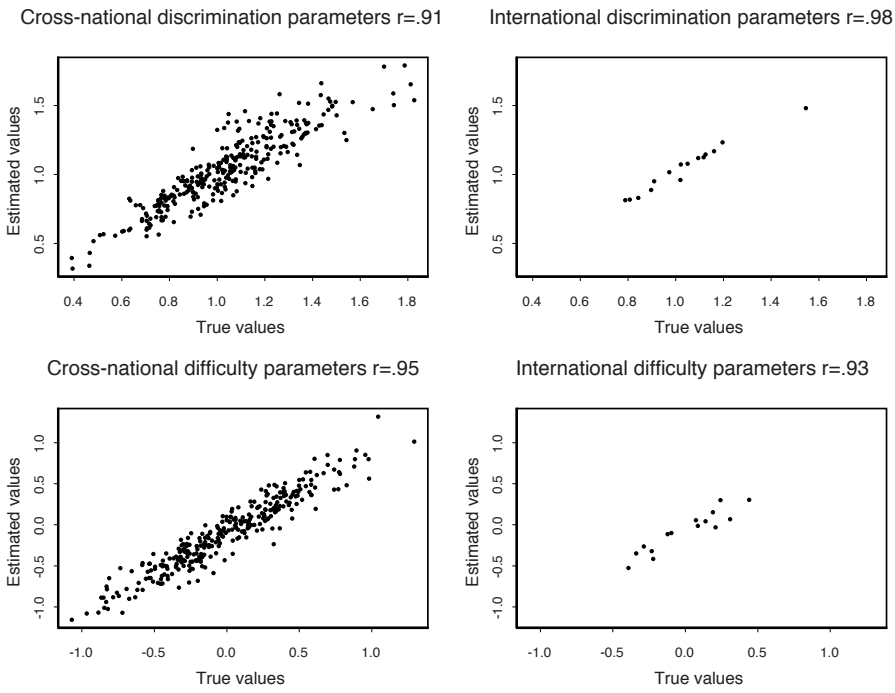

Figure 2.2: Plots of true and estimated international and cross-national item parameters.

answered questions about their background and their perception of the learning environment, while school principals provided school demographics and an assessment of the quality of the learning environment. The current practice in PISA for items that show signs of differential item functioning between countries is to delete them in all or in some countries, or to treat them as different items across countries. Item by country interaction is used as an indication for DIF, based on whether the national scaling parameter estimates, the item fit, and the point biserial discrimination coefficients differ significantly from the international scaling values (OECD, 2005). The (international) item parameters are then calibrated for all countries simultaneously, in order to create a common measurement scale. In practice, cross-national differences in response patterns are present, which makes the assumption of invariant item parameters unlikely. Goldstein (2004) argued that the Rasch measurement model used for the PISA data is too simplistic for such cross-national survey data as the multilevel nature of the data and country-specific response differences are not acknowledged. The proposed random item effects model deals with these problems by simultaneously including a multilevel structure for ability and allowing item parameters to differ over countries while at the same time a common measurement scale is retained. In addition covariates can be included in the model to explain within and between country variance in

ability and item parameters. We hypothesize that a random item effects model will acknowledge the real data structure more and therefore will fit the data better than the Rasch model. We chose to use items from the domain that measured skills in quantitative mathematics, which consists essentially of arithmetic or number-related skills applied to real life problems (e.g. exchange rates, computing the price of assembled skateboard parts). PISA works with a large item pool, from which students receive only limited clusters of items. In this way testing time is reduced, while at the same time the full range of topics is covered. Fourteen booklets with different combinations of item clusters were used, equally distributed over countries and schools. Due to this (linked) incomplete design the test scores can later be related to the same scale of estimated ability using IRT. To avoid booklet effects and simultaneously keep all countries well represented, we chose to use the first booklet, in which eight quantitative mathematics items were present. Due to a lack of students, the data from Liechtenstein were removed. This resulted in test data from 9769 students across 40 countries on eight quantitative mathematics items.

As covariates we used gender, index of economic, social and cultural status, minutes spend on math homework, mathematical self-concept and school student behavior. Gender differences and social economic status are generally known to be predictors of mathematical performance. The index of economic, social and cultural status was a combined measure of parental education, parental occupational status and access to home educational and cultural resources. A student questionnaire measured engagement in mathematics, self-beliefs concerning mathematics and learning strategies in mathematics. As all the latter measures correlated strongly with self-beliefs in mathematics, we chose to include self-concept in mathematics (belief in own mathematical competence). A school questionnaire was given to the school principals to asses aspects of the school environment. From these questions student behavior (absenteeism, class disruption, bullying, respect for teachers, alcohol/drug use) was the best predictor for mathematical performance. In addition, from the time spent on total instruction, math instruction and math homework, minutes spent on math homework was the best predictor of mathematical performance. Missing values in the covariates (ranging from $5-22$ percent) were imputed by the SPSS MISSING VALUE ANALYSIS REGRESSION procedure based on 20 variables. This procedure imputes the expected values from a linear regression equation based on the complete cases plus a residual component chosen randomly from the residual components of the complete cases.

2.6.1 PISA 2003: Results

Three random item effects models were estimated with the MLIRT package. The most general model, denoted as $M3$, allows for random item effects and random intercepts and covariates on the ability parameters. The other two models are nested in this model. Model $M3$ is presented as:

$$\begin{aligned} P\left(Y_{ijk}=1 \mid \theta_{ij}, \tilde{a}_{kj}, \tilde{b}_{kj}\right) &= \Phi(\tilde{a}_{kj}\theta_{ij} - \tilde{b}_{kj}) \\ \left(\tilde{a}_{kj}, \tilde{b}_{kj}\right)^t &= (a_k, b_k)^t + (\epsilon_{a_k}, \epsilon_{b_k})^t \end{aligned} \tag{2.9}$$

where the residual cross-national discrimination and difficulty effects are normally distributed with variance $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$, respectively, and

$$\begin{aligned}\theta_{ij} &= \gamma_{00} + \beta_1 HOMEWORK_{ij} + \beta_2 BEHAVIOR_{ij} + \\ &\quad \beta_3 SELFCONCEPT_{ij} + \beta_4 ESCS_{ij} + \beta_5 FEMALE_{ij} + u_{0j} + e_{ij},\end{aligned}$$

where $e_{ij} \sim \mathcal{N}\left(0, \sigma^2_\theta\right)$ and $u_{0j} \sim \mathcal{N}\left(0, \tau^2\right)$. The restricted model $M1$ only allows for random intercepts on the ability parameters and restricted model $M2$ allows for country-specific item parameters in addition to $M1$. Model $M1$ is identified by restricting the mean and variance of the latent ability scale to zero and one, respectively. Model $M2$ and $M3$ are identified by restricting the variance of the latent ability scale to one and by restricting the sum of country-specific item difficulties to zero in each country. There are no restrictions specified for the discrimination parameters since the models assume factor variance invariance.

The first $1,000$ iterations were discarded, the remaining $19,000$ iterations were used for the estimation of the model parameters. The program took approximately 2.5 hours to complete the estimation. To check whether the chains reached a state of convergence, trace plots, and convergence diagnostics were examined. The diagnostics and trace plots did not indicate convergence problems, except for a somewhat high autocorrelation for the discrimination parameter of item 2 in both random item effects models, model $M2$ and $M3$. The high autocorrelation results from the fact that this item has both a high discrimination and a high difficulty parameter. Since the item information function for this item is very steep and centered around the difficulty parameter value the parameters of this item will be very hard to estimate, especially in countries where the ability level is low. In Brazil, for example, only 13 out of the 250 selected students had an estimated ability that was higher than the difficulty level of the item, which indicates that there was very little information to base the estimated parameters on in this country. For the three models, the estimated international item parameter estimates are given in Table 2.3. For model $M2$ and $M3$, the estimated cross-national discrimination and difficulty standard deviations are also given.

Cross-national variance

The estimated international discrimination parameters of M1 and M2 are very similar. The estimated international difficulty parameters of model M2 are higher, because the identification rules for the two models differ. However, the estimated difficulty parameters of model M1 can be transformed to the scale of model M2. For item 1, the transformed estimated item difficulty of $M1$ resembles the estimated item difficulty of $M2$ ($.73 \cdot .82 - .59 \approx .01$). Note that the estimated variances of both ability scales are approximately equal.

In Table 2.3, the estimated -2 log-likelihood of the IRT part and the structural multilevel part are given. Both terms are used to estimate an DIC that also contains a penalty function for the number of model parameters. When comparing model $M1$ with $M2$, the log-likelihood of the IRT part is improved and the log-likelihood of the multilevel part is almost equal. The DIC also shows a clear improvement in fit due to the inclusion of random item effects. This supports the hypothesis of non-invariant item parameters.

Table 2.3: Parameter estimates of the MLIRT model and two random item effects models.

	Model M1		Model M2				Model M3			
item	$\hat{a}_k$	$\hat{b}_k$	$\hat{a}_k$	$\hat{\sigma}_{a_k}$	$\hat{b}_k$	$\hat{\sigma}_{b_k}$	$\hat{a}_k$	$\hat{\sigma}_{a_k}$	$\hat{b}_k$	$\hat{\sigma}_{b_k}$
1	.81	-.59	.82	.09	.00	.14	.78	.08	-.02	.14
2	1.06	.19	1.10	.24	.99	.12	1.16	.20	1.03	.15
3	.73	-.04	.72	.07	.48	.11	.69	.06	.46	.11
4	.69	-.36	.70	.12	.14	.11	.70	.10	.14	.11
5	.56	-.02	.58	.12	.40	.08	.61	.13	.41	.09
6	.37	-1.51	.40	.16	-1.26	.10	.38	.16	-1.26	.10
7	.69	-.78	.69	.10	-.29	.10	.66	.09	-.30	.10
8	.66	-.94	.69	.12	-.46	.08	.67	.11	-.46	.08
	Mean	HPD	Mean	HPD			Mean	HPD		
γ_{00}	.01	[-0.14,0.15]	0.73	[0.58,0.88]			1.01	[0.88,1.14]		
σ^2	0.79	[0.77,0.82]	0.79	[0.77,0.82]			0.57	[0.54, 0.59]		
τ^2	0.22	[0.13,0.33]	0.22	[0.13,0.33]			0.14	[0.08, 0.21]		
β_1(Homework)							-0.37	[-0.44,-0.30]		
β_2(Behavior)							0.07	[0.05, 0.09]		
β_3(Self-concept)							0.28	[0.25, 0.30]		
β_4(ESCS)							0.33	[0.31, 0.36]		
β_5(Female)							-0.07	[-0.11,-0.03]		
-2 LL IRT		-36,129.56		-35,642.12				-35,813.18		
-2 LL ML		-12,897.95		-12,901.53				-11,261.54		
DIC MLIRT		105,431.03		104,481.66				101,973.38		

The estimated cross-national variance in item discriminations and item difficulties supports the hypothesis of cross-national item parameter variance. Item six does not discriminate well between students with lower and higher ability in math, probably because the item is too easy. The estimated country-specific discrimination parameters of model M2 show that in some countries (e.g. Japan: .614), the item discriminates better, while in other countries (e.g. Switzerland: .149 and Belgium: .192) the item hardly discriminates at all. Item two is the most discriminating item, the estimated country specific discriminations range from .634 (Indonesia) and .751 (Tunisia) to 1.415 (Hungary) and 1.602 (Japan). The estimated difficulty parameters for this item range from .784 (USA) to 1.127 (Ireland). Figure 2.3 shows the item characteristic curves (ICCs) for item eight. The relatively low discrimination parameters for Denmark, Indonesia and the Netherlands make the curves for those countries relatively flat, while their difficulty parameters separate their curves in horizontal direction. The relatively high discrimination parameters for Thailand and Japan make their curves very steep.

The data supports the grouping of respondents in countries. The estimated intraclass correlation coefficient shows that 21% of the total variance in latent ability is explained by mean ability differences across countries.

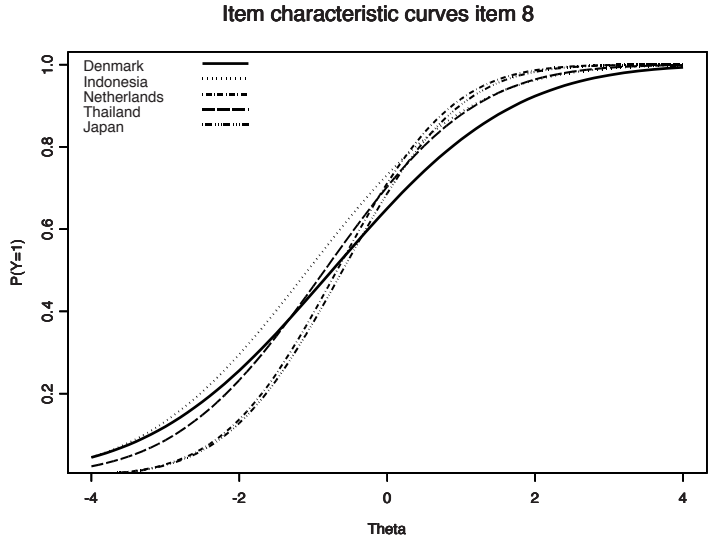

Figure 2.3: Cross-national item specific curves for item 8

Covariates on the ability parameters

Model $M2$ is extended with explanatory information at the individual-level and this leads to model $M3$. It is to be expected that the estimated international item parameters and the estimated cross-national item variances of model $M2$ and $M3$ are equal since individual-based explanatory information is incorporated. From Table 2.3, it can be seen that the estimated international item parameters and estimated cross-national item variances are approximately the same. Thus, the covariates do not explain cross-national item variance.

The log-likelihood of the IRT part did not change a lot but the log-likelihood of the multilevel part shows a clear improvement of model fit. The DIC shows that model $M3$ fits the data better than the other two models, indicating that the inclusion of explaining covariates on the individual level is an improvement of the model.

The covariates explain around 28% of the level-2 variance in ability between students and around 36% of the level-3 variance in ability between countries. The explained variance is within as well as between group variance, the conditional intraclass correlation stays almost the same at .20. The parameter γ_{00} is no longer the general latent mean, but the intercept in a regression equation that predicts the latent scores for the individuals conditional on the covariate effects.

The effects of all five covariates were strong, as can be seen from the estimated HPD intervals. Time spent on math homework and being female were predictive of a lower ability and a higher self-concept in mathematics, a higher economic, social and cultural status and absence of negative student behavior at the school of a student had a positive effect on math ability. The negative effect of time spent

on math homework can be explained by the fact that weak math students spend more time on their homework. This is in line with results found in PISA (OECD, 2004) and in other studies about predictors for math performance (e.g. Chiu & Klassen, 2009).

2.7 Concluding Remarks

A random item effects model was introduced for cross-national item response data. The model supports the use of country-specific item parameters. Further, a structural multilevel population model for the respondents was specified where cross-national differences in mean abilities are allowed. As a result, cross-national differences in item characteristics and respondents' abilities can be modeled and measurement invariant items are not needed. The corresponding identifying restrictions make the often difficult search for common invariant items unnecessary. The object is to explain the variation using covariate information at the item, individual, and group level. The model discussed here gives the opportunity to include covariates to predict ability, which gives the opportunity to estimate explanatory effects simultaneously with the IRT parameters, taking into account the measurement error in the ability parameters. The simulation study showed accurate recovery of the simulated parameters, although the estimates were a bit unstable for high discrimination parameters. The convergence criteria showed that MCMC draws are obtained from the joint posterior (target) distribution and they can be used to make posterior inferences. The prior choices can be evaluated via a prior sensitivity analysis to evaluate whether the prior choices substantially affect inferences. Here, the inverse gamma prior of the cross-national discrimination variances was investigated. It was shown that an informative prior (inverse gamma prior with a shape and scale parameter of one) can lead to significant cross-national variation in item discriminations. The study also showed that stable results were obtained for smaller prior parameter values. The PISA data supported the hypothesis that large-scale educational response data are more accurately analyzed with a random item effects model as this model fitted the data better. Substantial cross-national item parameter variance was detected. In PISA 2003, each item was selected and tested to be measurement invariant, however, the present analysis showed that the item characteristics were not invariant across countries. To avoid this complex issue of establishing partial or full measurement invariance, the proposed methodology for handling complex cross-national response data takes a different approach and provides a flexible framework for making meaningful cross-national comparisons.

Chapter 3

Bayesian Tests of Measurement Invariance

Abstract

Random item effects models provide a natural framework for the exploration of violations of measurement invariance without the need for anchor items. Within the random item effects modeling framework, Bayesian tests (Bayes factor, deviance information criterion) are proposed, which enable testing multiple marginal invariance hypotheses simultaneously. The performance of the tests is evaluated with a simulation study, which shows the tests have high power and a low type I error rate. Data from the European Social Survey are used to test measurement invariance of the attitude towards immigrant items and to show that background information can be used to explain cross-national variation in item functioning.

3.1 Introduction

Large cross-national surveys are increasingly used in education (PISA, TIMMS), social survey (SHARE, ESS, WVS, ISSP) and consumer research. Using a large number of countries, test scores or attitudes are compared to evaluate, for example, differences in policy. A well-known and widely investigated problem with this kind of data is that, to be able to make those comparisons, the measurement instrument should be at least partially invariant. Invariance of the measurement instrument implies that persons from each group with the same value of the underlying construct have the same probability of endorsing the items. Measurement non-invariance can arise through for instance translation errors, differential or multiple meanings of words in some countries, unequal familiarity with the item content or format, or dissimilar response tendencies over countries (see for an overview Sireci, Patsula, & Hambleton, 2005; Van de Vijver & Poortinga, 2005; Van de Vijver & Tanzer, 1998).

Adapted from: Verhagen, A. J. , & Fox, J.-P (2012). Bayesian tests of measurement invariance. *The British Journal of Mathematical and Statistical Psychology.*

Recently, Bayesian modeling techniques have made it straightforward to estimate non-invariant country-specific item parameters as well as country-specific latent distribution parameters on a common scale, without specifying anchor items in advance.

For this purpose, random item effects models have been proposed (e.g., De Jong & Steenkamp, 2010; De Jong, Steenkamp & Fox, 2007; Fox, 2010; Fox & Verhagen, 2010) as well as mixture IRT models in which some items are classified as anchor items in the estimation process (Soares, Goncalves & Gamerman, 2009). These Bayesian models make it possible to estimate one latent underlying scale for the person parameters across countries, taking variations in country-specific item parameters into account. This enables the comparison of individuals and countries on a common latent scale. The proposed method solves invariance problems regarding item bias in the measurement instrument. Variance in the definition of the underlying construct over countries or other forms of method bias (Van de Vijver & Tanzer, 1998) will remain possible, however.

In many applications the object is to apply invariance assumptions where they are valid; that is, to make cross-national comparisons using a model which adequately describes the observed data but is not unnecessarily complicated. In addition, cross-national comparisons on the estimated latent scale might be based on incorrect non-invariance assumptions, which might lead to less sharp distinctions since a more complex model can accommodate a larger set of potential observations. Furthermore, testing multiple measurement invariance hypotheses can be a goal in itself, as it provides information about detected cross-national differences in item parameters as well as in the distributions of the latent person parameters. Hence, although the random item effects model enables the estimation of non-invariant item parameters, it is also important to assess the significance of violations of invariance.

The best known traditional methods to detect items that are not invariant are based on nonparametric analysis, linear regression, factor analysis, and item response theory (IRT) (see for an overview e.g. Teresi, 2006; Vandenbergh & Lance, 2000). One of the most widely used parametric methods is the likelihood ratio test, which compares IRT models with all items constrained to invariance to models with only some invariant (anchor) items and non-invariant item parameters for the other items (e.g., Thissen, Steinberg, & Wainer, 1993). A similar procedure can also be carried out in a confirmatory factor analysis model (e.g. Meredith, 1993) using Mplus (Muthén & Muthén, 1998-2006) or LISREL (Jöreskog & Sörbom, 1996).

The traditional parametric stepwise procedures have some disadvantages. First, when group-specific item and latent distribution parameters (means and variances) are estimated, additional restrictions are needed to identify the model (e.g. Reise, Widaman & Pugh, 1993). Most estimation procedures need at least one anchor item to identify the underlying latent scale, while a larger set of anchor items is preferred to make reliable inferences. However, anchor items are extremely difficult to obtain in large-scale cross-national surveys (e.g. May, 2006; Rensvold & Cheung, 2001). Second, most of these procedures are based on comparing the fit of several models, which means that for each hypothesis a separate model needs to be fitted. This is a time-consuming procedure, and comparing a large number

of models on the same data set also makes it sensitive to data snooping (White, 2000).

In this paper, a Bayesian testing procedure is proposed for the simultaneous evaluation of multiple invariance hypotheses. The random item effects model eliminates the need for anchor items. Measurement invariance will be tested through direct evaluation of the variance components of the random item effects by a Bayes factor. This Bayesian test can handle a set of multiple individual marginal invariance hypotheses concerning the invariance of each item parameter. The posterior probability that each null hypothesis is true is computed. Consequently, null hypotheses can be rejected based on their posterior probabilities (Efron et al., 2001; Storey, 2003). In addition, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) can be used to compare the most general model with restricted models to test the measurement invariance hypotheses. Once the proposed tests have determined whether the variance in item parameters is significant, two additional steps can be taken. First, the model needs to be adjusted by restricting the parameters which were flagged as invariant. When modeling an invariant item parameter as a random effect, a normality assumption is violated and a worse model fit will be obtained. Second, when variance in item parameters over groups has been detected, this paper presents a way of explaining this variance with group-level covariates.

The remainder of this article is organized as follows. The random item effects Multilevel IRT (MLIRT) model is reviewed with a focus on the different variance components. A model extension is presented to explain cross-national variance in item and person parameters at the country, individual, and item level using covariates. Then, Bayesian tests are proposed for evaluating multiple invariance hypotheses. A simulation and a real data study will be used to explore the performance of the tests in detecting invariance.

3.2 Random Item Effects MLIRT Model

In cross-national surveys, observed (binary) data come from respondents nested in clusters. This clustering of respondents leads to additional dependencies between the observations. The multilevel IRT (MLIRT) model of Fox (2007) and Fox and Glas (2001), and closely related models (e.g. Adams et al., 1997; Kamata, 2001; Maier, 2001; Rabe-Hesketh et al., 2004; Raudenbush & Sampson, 1999), account for these additional dependencies, as they introduced a multilevel population model for the latent variable of person parameters. This latent variable is measured using an IRT model and all items are assumed to be invariant across the clusters of respondents. However, in cross-national surveys the items are often not invariant and show differential item functioning. Response models with random item effects parameters have been developed to account for the random variation in item functioning across countries. This random item effects modeling approach has the advantages that all items are allowed to be non-invariant and that anchor items are not needed to identify the scale of the latent variable, while accounting for group-specific differences in the person parameters. The random item effects MLIRT model (De Jong, Steenkamp, & Fox, 2007, Fox & Verhagen, 2010) is an extension of the MLIRT model (Fox, 2007; Fox & Glas, 2001) with ran-

dom item parameters to account for differential item functioning across countries. This flexible modeling framework will be used to develop tests for measurement invariance, to detect anchor items, and to explain and/or control for item-covariate information.

For each person, the K binary item responses are assumed to be conditionally independent given the random person effects parameter. This is the common local independence assumption in item response modeling. In each nation, it is also assumed that the n_j responses to each item are conditionally independent given the random item effects parameters (item difficulty $\tilde{b}_{kj}$ and discrimination $\tilde{a}_{kj}$).

The binary item responses are conditionally independently Bernoulli distributed at the observation level (level 1):

$$\begin{aligned} Y_{ijk} \mid \theta_{ij}, \tilde{a}_{kj}, \tilde{b}_{kj} &\sim \mathcal{B}\left(P\left(Y_{ijk}=1 \mid \theta_{ij}, \tilde{a}_{kj}, \tilde{b}_{kj}\right)\right) \\ P\left(Y_{ijk}=1 \mid \theta_{ij}, \tilde{a}_{kj}, \tilde{b}_{kj}\right) &= \Phi\left(\tilde{a}_{kj}\theta_{ij} - \tilde{b}_{kj}\right), \end{aligned} \tag{3.1}$$

where θ_{ij} is the person parameter for person i in country j, and $\tilde{a}_{kj}$ and $\tilde{b}_{kj}$ are the item parameters of item k for country j. The discrimination and threshold parameters are stored in a vector $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{b}_{kj})^t$.

3.2.1 Unconditional Modeling: Exploring Variance Components

Normally distributed latent item responses Z_{ijk} can be defined when the success probabilities are modeled with a probit model (Albert, 1992). The level-1 measurement model for the latent continuous item responses is defined as:

$$Z_{ijk} \mid Y_{ijk}, \tilde{a}_{kj}, \theta_{ij}, \tilde{b}_{kj} \sim \mathcal{N}\left(\tilde{a}_{kj}\theta_{ij} - \tilde{b}_{kj}, 1\right), \tag{3.2}$$

where Y_{ijk} is the indicator that Z_{ijk} is positive. The mean term consists of three random effects parameters to model the complex data structure in which the level 1 (latent) observations Z_{ijk} are not strictly hierarchically structured, but cross-classified by two level 2 units (country-specific items and persons).

The random effects parameters at the second level of the model are assumed to be normally distributed; that is,

$$\theta_{ij} \mid \mu_j, \sigma^2_{\theta_j} \sim \mathcal{N}\left(\mu_j, \sigma^2_{\theta_j}\right), \tag{3.3}$$

$$\tilde{a}_{kj} \mid a_k, \sigma^2_{a_k} \sim \mathcal{N}\left(a_k, \sigma^2_{a_k}\right), \tag{3.4}$$

$$\tilde{b}_{kj} \mid b_k, \sigma^2_{b_k} \sim \mathcal{N}\left(b_k, \sigma^2_{b_k}\right), \tag{3.5}$$

where the random person effects θ_{ij} are normally distributed with a country-specific mean μ_j and variance $\sigma^2_{\theta_j}$, and the (country-specific) random item effects $\tilde{a}_{kj}$ and $\tilde{b}_{kj}$ are normally distributed with an item-specific mean (international item parameter) and variance. Note that the third level is country-specific for the person parameters and item-specific for the item parameters.

On the third level, the international item parameters a_k and b_k and the latent group means μ_j can also be modeled as normally distributed random effects. The

international item parameters $\boldsymbol{\xi}_k = (a_k, b_k)^t$, are modeled on the third level as random item-specific deviations from an overall item parameter mean $\boldsymbol{\mu}_\xi$:

$$p\left(\boldsymbol{\xi}_k\right) = \mathcal{N}\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right), \tag{3.6}$$

where $\boldsymbol{\mu}_\xi = (a_0, b_0)^t$ and $\boldsymbol{\Sigma}_\xi$ has diagonal elements σ_a^2 and σ_b^2.

3.2.2 Conditional Modeling: Explaining Variance

The multilevel IRT model (Fox, 2007; Fox & Glas, 2001) defines a second and third level for the nesting of respondents in countries on the latent variable. Subsequently, individual ($\boldsymbol{x}_{ij}$) and country ($\mathbf{w}_j$) information can be introduced at these levels to account for background differences in the person parameters. For example, the effects of country-specific differences in predictors related to student achievement on students' test scores can be evaluated. The background variables can be used to explain variance, as main effects or cross-level interactions, on the individual (level 2) and group (level 3) level:

$$\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma^2_{\theta_j} \sim \mathcal{N}\left(\mathbf{x}_{ij}^t \boldsymbol{\beta}_j, \sigma^2_{\theta_j}\right) \tag{3.7}$$

$$\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T} \sim \mathcal{N}\left(\mathbf{w}_j \boldsymbol{\gamma}, \mathbf{T}\right), \tag{3.8}$$

where $\boldsymbol{\beta}_j$ are the (random) effects of individual covariates and $\boldsymbol{\gamma}$ are the effects of the country covariates on $\boldsymbol{\beta}_j$.

In the same way, explanatory information can be included that might explain or control for parts of the variance in country-specific item parameters. When differences in item parameters are detected, country covariates $\mathbf{v}_{kj}$ can be incorporated to account for these cross-national differences in item parameters

$$\tilde{a}_{kj} \mid \mathbf{v}_{kj}, \boldsymbol{\delta}_{a_k}, \sigma^2_{a_k} \sim \mathcal{N}\left(a_k + \mathbf{v}_{kj}^t \boldsymbol{\delta}_{a_k}, \sigma^2_{a_k}\right) \tag{3.9}$$

$$\tilde{b}_{kj} \mid \mathbf{v}_{kj}, \boldsymbol{\delta}_{b_k}, \sigma^2_{b_k} \sim \mathcal{N}\left(b_k + \mathbf{v}_{kj}^t \boldsymbol{\delta}_{b_k}, \sigma^2_{b_k}\right), \tag{3.10}$$

where the regression coefficients, $\boldsymbol{\delta}$, can be assumed fixed or random across items. Furthermore, the explanatory information can be continuous or categorical. In this way, cross-national differences in item parameters might be partly explained by a categorical covariate that reflects, for example, different cultural subgroups and this effect might vary randomly across items.

A further conditional modeling extension is to use explanatory information $\mathbf{U}_\xi$ about the items at level 3 to specify cross-level interactions $\boldsymbol{\nu}$, with characteristics of the items predicting the magnitude of the effects ($\boldsymbol{\delta}_{a_k}$ and $\boldsymbol{\delta}_{b_k}$) of country-level covariates on the country-specific item parameters at level 2:

$$\left(\boldsymbol{\delta}_{a_k}, \boldsymbol{\delta}_{b_k}\right)^t \mid \mathbf{U}_\xi, \boldsymbol{\nu}, \boldsymbol{\Sigma}_\delta \sim \mathcal{N}\left(\mathbf{U}_\xi \boldsymbol{\nu}, \boldsymbol{\Sigma}_\delta\right). \tag{3.11}$$

Note that the international item parameters $\boldsymbol{\xi}_k = (a_k, b_k)^t$ and the explanatory international item effects $\boldsymbol{\delta}_{ak}$ and $\boldsymbol{\delta}_{b_k}$ can be modeled simultaneously using a multivariate normal prior.

The idea of modeling item parameters was introduced in the linear logistic test model (LLTM) (Fischer, 1973), which assumed a perfect linear decomposition of

the threshold parameter. Recently, random item parameters to model groupings of items or item characteristics with an error term have been used in several studies (Albers et al., 1989; De Boeck, 2008; Janssen et al., 2000; Glas et al., 2003; Glas, Van der Linden & Geerlings, 2010). Here, the conditional modeling of random item characteristics is extended with covariates explaining item parameter differences between items and between countries simultaneously using some combination of fixed and random explanatory item effects.

The modeling possibilities are enormous, since effects of explanatory variables can be modeled to be fixed or to fluctuate across countries or items and to depend on higher-level information. Furthermore, the cross-classified modeling structure makes it possible to include explanatory information at the second and third level of both hierarchies.

3.3 Model Identification and Estimation

In a full Bayesian approach, the parameters of the prior distributions are modeled using hyperpriors. Inverse gamma priors are defined for the variance parameters $\sigma^2_{\theta_j}$, τ^2, $\sigma^2_{a_k}$, and $\sigma^2_{b_k}$, with shape parameter one and scale parameter, .1, .2, .05, and .05, respectively. A normal inverse Wishart prior is defined for the prior parameters of the multivariate normal prior in Equation (3.6), where the mean vector is $(1, 0)$, the scale matrix of rank two has diagonal elements .25 and non-diagonal elements .20, and two degrees of freedom. The inverse gamma priors are chosen in such a way that realistic values of the variances for the different parameters have a non-zero density. Prior sensitivity for this model is low as long as the priors are not chosen too wide (for a discussion of gamma prior sensitivity in this model, see Fox & Verhagen, 2010). Using a normal distribution for the prior corresponds with the idea that the items which constitute the test are assumed to be a sample from a larger pool of items. Normal inverse Wishart priors are also defined for the parameters of the multivariate normal priors in Equation (3.8), and (3.11).

All model parameters are estimated simultaneously using a Gibbs sampler that is based on augmenting the responses with latent continuous response data $\mathbf{Z}$ (Albert, 1992). The Markov Chain Monte Carlo (MCMC) Gibbs sampler for random item effects has been described by Fox (2010). This routine with the extension for tests of invariance was implemented in Fortran, and can be called from Splus. An Splus routine will be available from the authors.

In the uni-dimensional item response model, the mean and variance of the person parameters on the latent scale are directly related to and exchangeable with the threshold and discrimination parameters, respectively. To identify the (two-parameter) item response model, either the latent scale is fixed to have a mean of zero and a variance of one, or the sum of the threshold parameters and the product of the discrimination parameters are restricted to zero and one, respectively. When either group-specific item or person parameters are specified, the model can still be identified in that way.

When modeling group differences in both the item parameters and the person parameters simultaneously the identification problem multiplies (e.g. Reise, Widaman & Pugh, 1993). That is, for each country a common shift of the thresh-

old parameters is exchangeable with a shift in the country mean and a common shift of the discrimination parameters with a shift in the country variance.

There are multiple ways to solve this identification problem. Most traditional methods are based on at least one anchor item and a reference group. The latent scale in the reference group is identified by restricting the latent mean and variance for this group. In addition, at least one invariant item is assumed, which has common item characteristics across countries, to express results on a common scale. This also identifies the location and variance of the scales of the other groups. The reference group and anchor item identify the latent scale, given the configural invariance (Steenkamp & Baumgartner, 1998) assumption that the items load on the same single factor.

In the present approach the group-specific item parameters and the group-specific means of the person parameters are modeled as random effects. A common shift of all the group-specific threshold parameters of group j still results in the same expected probability of a correct answer as a comparable shift, in the opposite direction, of the group-specific latent mean μ_j. For example, when group j scores better on a test this can be the result of a higher group-mean ability, μ_j, or the result of all items being easier for this group.

Instead of using anchor items, the present model will be restricted in such a way that the sum of the threshold parameters within each group equals zero. This corresponds with the intuitive idea that whenever a group j performs better on all items, this will be represented in a higher group-specific mean ability μ_j. This restriction links the scales by specifying the overall threshold to be equal in each group, and restricts the locations of the scales in all groups. The random item effects strengthen the linkage between the scales by shrinking the group-specific item parameters towards a common item-specific mean, which combines information from all groups involved.

In a similar way, a common shift of the discrimination parameters is exchangeable with a comparable shift in the group-specific variance, σ_{θ_j}. This is solved by restricting the product of the discrimination parameters within each group to be equal to 1.

In this way, the group-specific item parameters are prevented from shifting simultaneously in the same direction without the need for anchor items or for interpreting group effects relative to a reference group. Once invariant items are known, either beforehand or after they have been detected by an invariance test, they are fixed. Then, common overall latent scale restrictions, such as restricting the general latent mean and variance, are sufficient to identify the model.

3.4 Testing Assumptions of Invariance

In cross-national survey research, attention is often focused on testing various levels of invariance. It is shown above that the results from the random item effects MLIRT model can be used to compare groups on the latent variable. Restricted versions of the model correspond with interesting tests of metric and scalar invariance (Steenkamp & Baumgartner, 1998). Besides measurement invariance, other invariance issues like factor mean and variance invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998) play a role when comparing groups, which can

also be evaluated under the proposed model. These invariance issues will not be investigated in this paper.

Two cross-national random item effects variances are tested to investigate the assumption of measurement invariance. The hypotheses of interest are whether the cross-national item parameter variances, $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$, differ from zero. Testing the variance of a random effect is complicated, since the point of zero variance lies on the boundary of the parameter space. It is known that the classical procedures can break down asymptotically (e.g., likelihood ratio test) or require modified asymptotic null distributions (e.q., Wald tests, generalized likelihood ratio tests). These test statistics have complicated distributions and are difficult to apply (e.g., Molenberghs & Verbeke, 2007; Pauler, Wakefield & Kass, 1999).

In the Bayesian framework, the invariance hypotheses are ideally tested by Bayes factors comparing the marginal likelihoods of the models with and without invariance constraints. It will be shown that the MCMC algorithm for estimating the most general model can be used to evaluate various restricted versions of the model. In this way, (1) specific invariance assumptions can be tested without having to rely on invariant items, (2) the invariance assumptions can be tested simultaneously, and (3) only the most general model needs to be estimated. The DIC (Spiegelhalter et al., 2000) is used to compare the fit of models with and without invariance constraints. A deviance term is defined to investigate the difference in deviance, which is induced by a restriction on the model parameters.

Once the tests determined whether the variance in item parameters is significant, the model can be adjusted by restricting the parameters for which no significant variance was found to invariance.

3.4.1 The Bayes Factor

The Bayes factor is expressed as the ratio of the marginal likelihood of model (or hypothesis) M_0 to the marginal likelihood of M_1. The Bayes factor is generally computationally demanding since it requires the evaluation of high-dimensional integrals for both models. However, the computation of the Bayes factor simplifies when the models are nested or have a common conditional distribution of observed data. In both cases, the Bayes factor reduces to a ratio that can be evaluated under the most general model.

Encompassing Prior

Consider the null hypothesis H_0: $\sigma^2_{b_k} = 0$, which states that item k's threshold parameter is invariant across countries. The posterior probability of $\sigma^2_{b_k} = 0$ under the unrestricted model, $p\left(\sigma^2_{b_k} = 0 \mid \mathbf{y}\right)$, can be expressed as

$$p\left(\sigma^2_{b_k} = 0 \mid \mathbf{y}\right) \quad = \quad \frac{p\left(\mathbf{y} \mid \sigma^2_{b_k} = 0\right) p\left(\sigma^2_{b_k} = 0\right)}{\int p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) p\left(\sigma^2_{b_k}\right) d\sigma^2_{b_k}}. \tag{3.12}$$

It follows that the Bayes factor can be expressed as the ratio of the density at the null hypothesis of the prior and posterior distribution under the unrestricted

model (see Dickey, 1971; Verdinelli & Wasserman,1995),

$$\text{BF} = \frac{p\left(\sigma^2_{b_k} = 0 \mid \mathbf{y}\right)}{p\left(\sigma^2_{b_k} = 0\right)} = \frac{p\left(\mathbf{y} \mid H_0\right)}{p\left(\mathbf{y} \mid H_1\right)}. \tag{3.13}$$

However, the specified inverse gamma prior only assigns positive density values to positive variance parameter values. That is, the point $\sigma^2_{b_k} = 0$ has zero prior and posterior probability under the general model. Hoijtink (2011), Klugkist and Hoijtink(2007), and Klugkist (2008) defined an encompassing prior approach, where the prior for the constrained model is obtained by restricting the encompassing prior to a specific area. It follows that the constrained prior is nested within the encompassing prior. The encompassing prior for the invariance tests will be defined as an unconstrained inverse gamma prior with most of its mass in the expected range of the parameter. The problem of zero probability under the null hypothesis is now avoided by defining the invariance hypothesis as $\sigma^2_{b_k} < \delta$, which corresponds closely with the original hypothesis when a very small value is chosen for δ. Furthermore, the Bayes factor is easily evaluated for different values of δ.

Common Likelihood

A slightly different approach is based on the restriction that both models share the same conditional distribution of observed data (Geweke, 2005). Furthermore, the parameter space associated with the prior under the null hypothesis, denoted as Θ_0, is a subset of the parameter space associated with the prior under the alternative hypothesis, Θ_1. The prior densities are allowed to be entirely different and they are not restricted to be nested within each other.

Let $p\left(\sigma^2_{b_k} \mid H_0\right)$, $\sigma^2_{b_k} \in \Theta_0$ and $p\left(\sigma^2_{b_k} \mid H_1\right)$, $\sigma^2_{b_k} \in \Theta_1$, where $\Theta_0 \subseteq \Theta_1$, denote the prior under the null hypothesis and the prior under the alternative hypothesis, respectively. The Bayes factor in favor of the null hypothesis can be expressed as

$$\begin{aligned}
\text{BF} &= \frac{\int_{\Theta_0} p\left(\sigma^2_{b_k} \mid H_0\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}}{\int_{\Theta_1} p\left(\sigma^2_{b_k} \mid H_1\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}} = \frac{\int_{\Theta_0} p\left(\sigma^2_{b_k} \mid H_0\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}}{p\left(\mathbf{y} \mid H_1\right)} \\
&= \int_{\Theta_1} \left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)}\right] \frac{p\left(\sigma^2_{b_k} \mid H_1\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right)}{p\left(\mathbf{y} \mid H_1\right)} d\sigma^2_{b_k} \\
&= \int_{\Theta_1} \left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)}\right] p\left(\mathbf{y} \mid \sigma^2_{b_k}, H_1\right) d\sigma^2_{b_k} \\
&= E\left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)} \mid \mathbf{y}, \right].
\end{aligned} \tag{3.14}$$

where the ratio of prior densities is evaluated using the posterior density of $\sigma^2_{b_k}$ under the alternative hypothesis. As a result, the Bayes factor can be evaluated as the posterior expectation of the ratio of prior densities using the posterior draws from $p\left(\sigma^2_{b_k} \mid \mathbf{y}, H_1\right)$ given that the ratio is bounded on Θ_1. In general, any prior that restricts the parameter space to a subset of the original parameter space can be compared with the more general prior via the Bayes factor using MCMC output from the general model (See Appendix B).

The common likelihood approach is more general, as any form of prior can be used and different priors can be tested against each other. Both approaches to compute the Bayes factor give identical results when nested priors are used. Bayes factors with nested priors will be used to test the following invariance hypotheses: measurement invariant discrimination parameters, $\sigma^2_{a_k} < \delta$, measurement invariant threshold parameters, $\sigma^2_{b_k} < \delta$, invariant latent means, $\tau^2_j < \delta$, and invariant within-country latent variances, $\left|\sigma^2_{\theta_j} - \overline{\sigma}^2_{\theta}\right| < \delta$.

3.4.2 DIC: Comparing Constrained and Unconstrained Models

The fit of models with and without invariance restrictions can be compared using information criteria like the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Deviance information criterion (DIC). The deviance-based criteria are defined as the posterior mean of the deviance corrected for the number of parameters in the model. However, for complex hierarchical models, the number of model parameters is hard to determine. The DIC solves this problem by computing the effective number of parameters. The effective number of parameters, p_D, can be thought of as the ratio of the information in the likelihood about the parameters as a fraction of the total information in the likelihood and the prior (Spiegelhalter, Best, & Carlin, 1998). The p_D is computed by subtracting the deviance at the posterior means from the posterior mean of the deviance.

Let $\mathbf{\Lambda}$ denote the parameters of interest. The deviance function is defined as $D(\mathbf{\Lambda}) = -2\log p(\mathbf{\Lambda})$, which will be a function of the likelihood, and the DIC is given by

$$\begin{aligned} \text{DIC} &= \overline{D(\mathbf{\Lambda})} + \overline{D(\mathbf{\Lambda})} - D\big(\hat{\mathbf{\Lambda}}\big) \\ &= \overline{D(\mathbf{\Lambda})} + p_D, \end{aligned}$$

where $\overline{D(\mathbf{\Lambda})}$ is the posterior mean deviance and $D\big(\hat{\mathbf{\Lambda}}\big)$ the estimated deviance at the posterior estimate of $\hat{\Lambda}$.

Here, the assumption of measurement invariance is tested by comparing the estimated DIC of the general model with the estimated DIC of the model assuming invariant item parameters. The assumption holds when the latter model is associated with the smallest DIC. It follows that, contrary to the previously discussed parameter tests, several models have to be estimated to test the invariance assumption.

3.5 Simulation Study

A simulation study will be performed to explore the performance of the tests via a power study, and to evaluate the recovery of simulated parameters. Therefore, data sets were simulated with 20 items and 40 groups of 200 cases each, to represent response data from a large-scale international survey. Data were simulated using parameters drawn from the prior distributions. The latent group means μ_j were

sampled from a normal distribution with mean zero and between-group variance .50. Individual latent variable values were drawn from a normal distribution with mean μ_j and a variance generated from an inverse gamma distribution such that the within-group variances ranged from .40 to 2.00. International threshold and discrimination parameters were generated from normal distributions with mean zero and one and variance .30 and .10, respectively. Group-specific item parameters were generated from normal distributions with the international parameters as the mean values. The cross-national item parameter variances varied from zero to .13 for the discrimination parameters and from zero to .30 for the threshold parameters across five groups of four items. As a result, the group-specific discrimination parameters ranged between .50 and 1.50, and the threshold parameters ranged between -2.00 and 2.00.

3.5.1 Testing Full and Partial Measurement Invariance

The model-based tests given the unconditional random item effects MLIRT model will be used to test multiple invariance assumptions simultaneously. The measurement invariance assumptions of invariant item discrimination and threshold characteristics were tested for each item nested in five item groups. The items in group one were simulated to be measurement invariant and the items in the other groups to have increasingly varying item parameters. For each item the assumption of (partial) invariance was evaluated by testing whether the item-specific variance parameters $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$ equal zero. The assumption of full item parameter invariance was evaluated by comparing the model assuming full measurement invariance with the model assuming measurement non-invariance.

In Table 3.1, the test results of the measurement invariance hypotheses evaluated by the Bayes factor are presented per item group. The item-level results are averaged for each item group. According to the Bayesian tradition of accumulating support for the null hypothesis, the null hypothesis is accepted at a Bayes factor larger than three, indicating substantial support in favor of the null hypothesis (i.e. Jeffreys, 1961). This is an arbitrary value, higher values of the Bayes factor indicate more support for the null model versus the alternative model, but in case of invariance testing it is often desirable to draw a line somewhere and indicate an item as invariant. The results are shown for three increasing values of the about equality constraint δ. In the first columns, the percentages of items across the 50 data sets for which the null hypothesis of invariance was accepted are shown. The second columns present the posterior probability of the null hypothesis given the data. Here, the prior probability of the null hypothesis is defined as $\pi_0 = P(H_0)$, and the marginal posterior probability is expressed as

$$\begin{aligned} P\left(H_0 \mid \mathbf{y}\right) &= \frac{\pi_0 p\left(\mathbf{y} \mid \sigma^2_{b_k} = 0\right)}{\pi_0 p\left(\mathbf{y} \mid \sigma^2_{b_k} = 0\right) + (1-\pi_0)\int_{\sigma^2_{b_k}\neq 0} p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) p\left(\sigma^2_{b_k}\right) d\sigma^2_{b_k}} \\ &= \left(1 + \frac{1-\pi_0}{\pi_0}\frac{p\left(\mathbf{y} \mid H_1\right)}{p\left(\mathbf{y} \mid H_0\right)}\right)^{-1} \\ &= \left(1 + \frac{1-\pi_0}{\pi_0} BF^{-1}\right)^{-1}. \end{aligned} \tag{3.15}$$

Table 3.1: Simulation study results. Percentage of invariant parameters detected (BF > 3) and average posterior probability of the null hypothesis over 50 replicated data sets.

H_0	BF>3 $P(H_0 \mid \mathbf{y})$		BF>3 $P(H_0 \mid \mathbf{y})$		BF>3 $P(H_0 \mid \mathbf{y})$	
	$\delta < .0016$		$\delta < .0025$		$\delta < .0036$	
$\sigma^2_{a_k} = 0$, invariant item discrimination						
.00	0.92	0.92	0.99	0.97	0.99	0.97
.04	0.05	0.08	0.07	0.15	0.04	0.18
.07	0.01	0.02	0.01	0.08	0.01	0.11
.10	0.01	0.03	0.02	0.07	0.00	0.09
.13	0.01	0.03	0.02	0.07	0.00	0.09
$\sigma^2_{b_k} = 0$, invariant item difficulty						
.00	0.99	0.99	1.00	0.99	1.00	0.99
.05	0.00	0.00	0.00	0.00	0.00	0.00
.10	0.00	0.00	0.00	0.00	0.00	0.00
.20	0.00	0.00	0.00	0.00	0.00	0.00
.30	0.00	0.00	0.00	0.00	0.00	0.00

	DIC	Dhat	p_D
Full Measurement Invariance	181244	173257	7987
Measurement Non-Invariance	179961	170325	9636

The estimated posterior probabilities of the null hypothesis are computed with prior probability $1/2$ for H_0 and H_1.

The Bayes factor, evaluating the item-specific discrimination variances, showed high invariance detection rates: 92- 99% for the three levels of the about equality constraint δ. The average posterior probability of the null hypothesis of invariance for the invariant items was .92-.97, indicating that on average there was substantially more support for the null than for the alternative hypothesis. Invariant threshold parameters were correctly detected for almost all of the items. The $P(H_0 \mid \mathbf{y})$ of .99 for those items indicated substantially more support for the null hypothesis than for the alternative hypothesis.

For the groups of non-invariant items, the rate of incorrect detections was very small ($< .02\%$) for the three groups with the largest item-specific discrimination variances and zero for all of the threshold parameters. The average posterior probability of the null hypothesis was close to zero for all items except for those with the lowest discrimination parameter variance, indicating that invariance was not likely. In the item group where $\sigma^2_{a_k} = .04$, 4%–7% of the items were incorrectly indicated as having invariant discrimination parameters, and the average posterior probability of the null hypothesis was .08-.18. As this variance was very close to the specified δ values, this was to be expected.

All null hypotheses of invariant discrimination and threshold parameters were tested simultaneously and provided information about the possibility of partial measurement invariance. The assumption of full measurement invariance can be tested by comparing the DIC of the full measurement invariant model and the

measurement non-invariant model. In Table 3.1, the estimated DIC values are given. The DIC of the full measurement invariant model was higher than that of the measurement non-invariant model, favoring the last model and correctly rejecting measurement invariance. Although the number of effective parameters (pD) increased when using random item effects parameters, the deviance based on the fitted parameters decreased more to compensate for this.

3.6 European Social Survey: Attitude Towards Immigration

Response data from the European Social Survey (ESS round 1, 2002) were considered, in which 22 countries participated: Austria, Belgium, Switzerland, Czech Republic, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Hungary, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Poland, Portugal, Sweden and Slovenia. Respondents from an ethnic minority group or with a foreign nationality were excluded from the sample (7%). The resulting sample sizes per country varied from 850 to 2,646, where missing values were treated as missing at random.

In the 2002–2003 round, a large module about immigration was included, which consisted of several subscales. In the present analysis, eight dichotomized items concerning the perceived consequences and allowance of immigration were used to illustrate the proposed model. The items are described in Appendix A. An exploratory factor analysis on the tetrachoric correlations between the items showed a clear one factor solution with medium to high factor loadings for all eight items. A posterior predictive check for evaluating violations of unidimensionality assumption (e.g. Levy, 2011) supported the local independence assumption for most of the items. The object of this study was to detect random item variation and to test simultaneously the null hypotheses of measurement invariance without assuming the presence of anchor items.

3.6.1 Invariance Testing of the ESS Immigrant Items

Invariant and constrained non-invariant models were estimated on the eight immigration items. The MCMC estimation procedure was run for 10,000 iterations using a single long chain, with a burn-in of 1,000 iterations. No autocorrelations higher than .15 were found and the Geweke Z (Geweke, 1992) convergence diagnostic did not show values above three, indicating that the chains converged well and reached stationarity.

In Table 3.2, parameter estimates and test results are shown for the general full non-invariant model, which has a multilevel structure on the latent variable and random item parameters. The Bayes factors and posterior probabilities, for a δ value of .0025, showed strong support for invariant discrimination parameters of item 5 and 8. For the other items, the test results did not show enough evidence to reject the null hypothesis of invariant discrimination parameters. The posterior probabilities of invariant and non-invariant discrimination parameters were almost equal for item 6. For none of the threshold parameters measurement invariance was

Table 3.2: Example ESS. Posterior means and standard errors for the general item parameters ξ_k and the variance of the country-specific item parameters $\sigma^2_{\xi_k}$. Bayes factors (BF) with posterior probability of invariance $P(H_0 \mid \mathbf{y})$ and DIC tests for invariance.

	Mean	SD	$\sigma^2_{\xi_k}$	SD	BF	$P(H_0 \mid \mathbf{y})$
Discrimination Parameter						
1 Allow from poor countries	1.01	0.05	0.04	0.02	0.00	0.000
2 Allow from same ethnicity	0.92	0.05	0.03	0.02	0.00	0.000
3 Make worse country	1.30	0.07	0.08	0.04	0.00	0.001
4 Bad for economy	1.38	0.06	0.04	0.02	0.02	0.016
5 Undermine culture	1.13	0.04	0.02	0.01	4.22	0.808
6 Take away jobs	0.86	0.04	0.02	0.01	1.25	0.556
7 Take out more than put in	0.94	0.05	0.04	0.02	0.00	0.000
8 Worse crime rate	0.74	0.03	0.01	0.01	299.69	0.997
Difficulty Parameter						
1 Allow from poor countries	-0.30	0.09	0.17	0.06	0.00	0.000
2 Allow from same ethnicity	0.23	0.08	0.13	0.04	0.00	0.000
3 Make worse country	0.19	0.06	0.07	0.03	0.00	0.000
4 Bad for economy	0.26	0.08	0.15	0.05	0.00	0.000
5 Undermine culture	0.68	0.08	0.13	0.05	0.00	0.000
6 Take away jobs	0.12	0.08	0.14	0.05	0.00	0.000
7 Take out more than put in	-0.22	0.07	0.12	0.04	0.00	0.000
8 Worse crime rate	-0.96	0.08	0.12	0.04	0.00	0.000
		DIC		Dhat		p_D
Full Measurement Invariance		130322		106676		11655
Measurement Non-Invariance		130292		104716		12788
Partial Measurement Invariance		128632		103892		12370

supported by the Bayes factor. Therefore, it can be concluded that for each item in this subscale there is cross-national item threshold variation. Note that detecting non-invariance of all threshold parameters is not possible when anchor items are needed to identify the scale. Here, all marginal hypotheses of invariance were evaluated simultaneously, which led to the conclusion that each item threshold showed cross-national variation.

Examining the deviance information criterion estimates, the model with random item parameters had a slightly better fit than the model with invariant item parameters, as the increase in effective parameters was compensated for by a decrease in deviance for the estimated parameters. The fit of the partial measurement invariant model with the invariant discrimination parameters for item 5 and 8 fixed to be equal over countries was much better than the fit of both the full measurement invariant and the full measurement non-invariant model.

Table 3.3: Example ESS. Posterior means and standard errors of country-level covariate effects (% Immigrants, % Unemployed, GDP) on the country-specific item parameters.

	% Immigrants		% Unemployed		GDP	
	δ_{ξ_k}	SD	δ_{ξ_k}	SD	δ_{ξ_k}	SD
Discrimination Parameter						
1. Allow from poor countries	-.02	.06	.03	.08	.15	.09
2. Allow from same etnicity	.07	.06	.01	.06	.07	.07
3. Make worse country	-.02	.07	-.01	.08	.03	.09
4. Bad for economy	-.11	.07	.01	.08	-.04	.09
5. Undermine culture	.02	.06	-.08	.06	-.01	.07
6. Take away jobs	.01	.04	-.02	.05	**-.15**	.06
7. Take out more than put in	.06	.05	.01	.06	-.08	.06
8. Worse crime rate	-.01	.04	.04	.05	.05	.06
Difficulty Parameter						
1. Allow from poor countries	-.03	.12	.04	.14	.03	.16
2. Allow from same etnicity	.05	.10	.03	.12	-.02	.13
3. Make worse country	.00	.07	.05	.08	.01	.09
4. Bad for economy	.10	.10	-.01	.12	-.02	.13
5. Undermine culture	.00	.10	.03	.12	.00	.13
6. Take away jobs	**-.18**	.08	-.06	.09	**.32**	.10
7. Take out more than put in	.01	.09	-.06	.10	-.17	.11
8. Worse crime rate	.05	.09	-.03	.11	-.15	.12

3.6.2 Explaining Cross-National ESS Immigrant Item Variation

Background information was used to explore possible causes of differential item functioning. Therefore, the parameters described in Equations (3.9),(3.10) and (3.11) were added to the previous model. Following the suggestion of Welkenhuysen-Gybbels, Billiet and Cambre (2003) about explanations of differences in item parameters, the following variables were included in the analysis: The percentage of immigrants (% Immigrants) and the percentage of unemployment (% Unemployed) in the country at the time of the survey, and the gross domestic product (GDP) per capita (a measure of a country's overall economic output). These explanatory variables are also frequently used as possible predictors for country-level differences in attitude towards immigrants (Card, Dustmann & Preston, 2005; Malchow-Moller, Munch, Schroll & Saksen , 2009; Meuleman, Davidov & Billiet, 2009; Sides & Citrin ,2007).

In Table 3.3, the results are given of the fixed effects of the three covariates on the country-specific item parameters (see Equation (3.9) and (3.10)). Significant results were found for item six (immigrants take away jobs) indicating an effect of .32 and -.15 of GDP on the country-specific threshold and discrimination

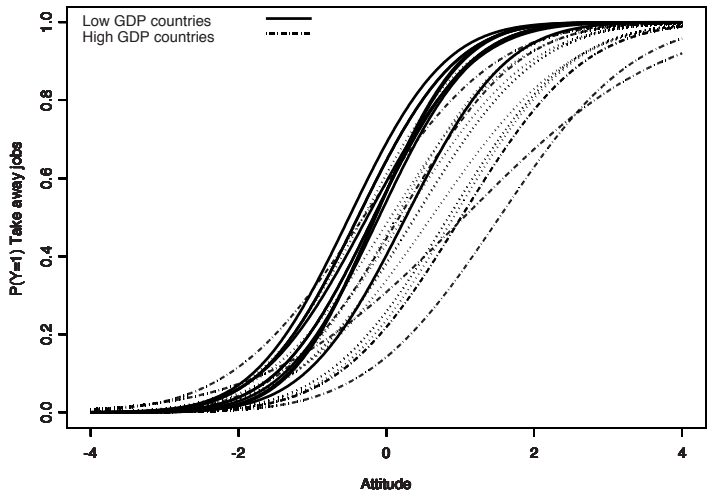

Figure 3.1: Country-specific item characteristic curves of item about immigrants taking away jobs.

parameters, respectively. This means that respondents from countries with high overall economic output (GDP) were less likely to agree with item six, while the respondents in those countries were also more alike in their responses. Furthermore, a negative effect of -.18 of the country-percentage of immigrants on the country-specific threshold parameters was found. This means that respondents from countries with a high percentage of immigrants were more likely to agree with item six in comparison to countries with lower percentages of immigrants.

In Figure 3.1, various country-specific characteristic curves of item six are plotted to illustrate the cross-national variation in item functioning (threshold and discrimination). Countries with low overall economic output (GDP), as indicated by the solid lines, had a lower threshold and a steeper slope than countries with a higher GDP. This illustrates that in countries with a low GDP, item six was both more relevant to the attitude and endorsed more frequently given attitude level.

3.7 Discussion

This study presented the utility of the random item effects MLIRT model for testing and modeling variance in cross-national response data. Bayesian tests were presented, which allow simultaneous testing of multiple invariance hypotheses without the need for anchor items. In addition, the Bayes factor only requires the estimation of the most general model.

The Bayes factor for nested models with an about equality constraint δ is easy to compute and represents the relative support for invariance over non-invariance

by the data. The simulation study showed that invariance of the discrimination and difficulty parameters was very well detected. It was shown that the posterior probability of each marginal null hypothesis can be computed, which gave a good indication of the strength of the evidence for invariance for both item parameters.

The DIC provides an overall measure of fit for each estimated model. The simulation studies showed that the DIC correctly favored the measurement non-invariant model. The DIC can be used to asses the overall difference in fit between a full measurement invariant model and other models. This does require the estimation of each model separately, however.

Several studies have reported the detection of cross-national item variation of social survey immigrant items. Welkenhuysen-Gybbels, Billiet and Cambre (2003) detected non-invariant items in the International Social Survey Programme (ISSP, 1995), and Billiet and Welkenhuysen-Gybbels (2004), Davidov, Meuleman, Billiet, and Schmidt (2008), and Meuleman, Davidov, and Billiet (2009) in the 2002–2003 ESS round. In this study, it was shown that the conditional model can be used to identify the effect of country differences on the item responses and to investigate why inhabitants from certain countries answer items differently. Information about the influence of country differences on item responses can be interesting in itself, but it can also be valuable in the process of test or survey creation.

After invariance has been detected, the invariant item parameters should be fixed to acquire a final model. The ESS example showed that this increased the model fit considerably.

The framework was laid out for Bayesian tests of measurement invariance. Future work can apply the basis provided here to extended models or different measurement situations. Cross-national surveys often make use of ordinal items instead of or in addition to binary items. The model and tests of measurement invariance presented here could be extended to mixed response item types. Furthermore, there has been increasing interest in differential item functioning over time in longitudinal data studies (e.g. Milsapp, 2010). The model and invariance tests presented in this paper could be modified to test for measurement invariance over time.

Chapter 4

Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses

Abstract

Longitudinal surveys measuring physical or mental health status are a common method to evaluate treatments. Multiple items are administered repeatedly to assess changes in the underlying health status of the patient. Traditional models to analyze the resulting data assume that the characteristics of at least some items are identical over measurement occasions. When this assumption is not met, this can result in ambiguous latent health status estimates. Changes in item characteristics over occasions are allowed in the proposed measurement model, which includes truncated and correlated random effects and a growth model for item parameters. In a joint estimation procedure adopting MCMC methods, both item and latent health status parameters are modeled as longitudinal random effects. Simulation study results show accurate parameter recovery. Data from a randomized clinical trial concerning the treatment of depression by increasing psychological acceptance showed significant item parameter shifts. For some items the probability of responding in the middle category versus the highest or lowest category increased significantly over time. The resulting latent depression scores decreased more over time for the experimental group than for the control group and the amount of decrease was related to the increase in acceptance level.

Adapted from: Verhagen, A. J., & Fox, J.-P. Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. Accepted for publication in *Statistics in Medicine*.

4.1 Introduction

Repeatedly administered questionnaires are increasingly used in clinical studies to assess the effects of interventions on or to track changes in reported physical or mental health status. Examples are quality of life questionnaires for patients with chronic diseases, psychiatric questionnaires to follow patients with psychiatric disorders, and cognitive ability questionnaires to track the onset of Alzheimer's disease. In this type of studies, a questionnaire is administered at each occasion to measure the underlying health status. The resulting data can be characterized as longitudinal multivariate response data designed to track intra-individual changes in latent physical or mental health status.

When the outcome consists of multiple item responses, it is common to use a latent variable to model the dependency between the item responses. The multivariate item responses are assumed to be conditionally independently distributed given a common latent variable. The latent variable and item characteristics are used to specify the probability of each item response, which defines the measurement model.

Examples are common factor models for continuous responses (e.g. Spearman, 1904), MIMIC models (e.g. Halberstadt, Schmitz and Sammel, 2012), item response models for discrete responses (e.g. Lord, 1980), and latent class models for discrete latent variables (e.g. Lazarsfeld, 1950; McHugh, 1956). In this paper we will focus mainly on IRT models, which are becoming more and more popular for health questionnaire data (e.g. Finkelman et al.,2001; Li & Baser, 2012; Klein Entink, Fox & van den Hout, 2011).

Mixed random effects models have been proposed for modeling individual change over time (Bryk and Raudenbusch, 1987; Goldstein, 1989; Laird and Ware, 1982). The specification of random effects accounts for the dependency between the within-subject measurements. Furthermore, random effects provide a very flexible way of handling missing data within subjects, as the number of measurement occasions per subject is allowed to vary. In addition, individual differences in the onset of the measurements and in the time span between measurement occasions can be taken into account (Hertzog and Nesselroade, 2003).

In longitudinal surveys consisting of multiple administered items per measurement occasion, latent growth models are often used to model longitudinal growth in the latent variable using mixed random effects (McArdle, 1986, Meredith and Tisak, 1990; Muthen and Muthen, 2001; Muthen et al., 2002). In latent growth curve models, for example, the random effects structure allows individual growth trajectories to vary randomly around a general growth trend. Longitudinal item response models (e.g. Blanchin et al., 2011; Liu, 2008) have appeared particularly useful for analyzing longitudinal health related questionnaires. Using a latent variable has the additional advantage that not all respondents necessarily have to answer all items at each occasion.

A common assumption in latent growth models is that the parameters of the measurement part, which connects the latent variable to the observed responses, are invariant over occasions. However, due to either the process of test administration, developmental changes between measurement occasions, or other occasion-dependent influences, the characteristics of the test can change over time. As a

result, the relation between the observed responses and the latent variable will differ per occasion. Consequently, a constant level of underlying health status leads to different response probabilities at sequential occasions.

Invariance over measurement occasions of item characteristics relating the latent variable to the response probabilities will be referred to as longitudinal measurement invariance, a form of measurement invariance (Meredith, 1993). In a factor analysis measurement model, for example, this implies that intercepts and factor loadings are invariant over occasions (Meredith and Horn, 2001). In the field of item response theory, non-invariance of discrimination and difficulty parameters over measurement occasions is known as item parameter drift (Meade, Lautenschlager and Hecht, 2005; Millsap, 2010). It is well-known that longitudinal measurement invariance is not self-evident and should be tested for (e.g. Hertzog and Nesselroade, 2003; Horn and McArdle, 1992; Meredith and Horn, 2001). Nevertheless, in most applications longitudinal measurement invariance of at least some items is assumed.

Implications of an erroneous assumption of longitudinal measurement invariance are that latent variable means and variances can be affected, rendering estimates of changes in latent health status ambiguous to interpret. To avoid bias in latent growth estimates due to violations of measurement invariance assumptions, latent growth models for longitudinal (health) questionnaire data should allow for occasion-specific measurement characteristics.

Procedures have been developed to assess measurement invariance for cross-sectional response data (e.g. Crane et al., 2006; Crane, van Belle & Larson, 2004), which are usually based on fixed item parameters. In the present approach, longitudinal measurement invariance will be assessed using a random item effects structure that allows the modeling of longitudinal growth in item characteristics. A nonlinear model for longitudinal multivariate responses will be introduced, which models growth in both the parameters of the measurement part and in latent health status simultaneously. The random item effects multilevel IRT model developed by Fox (2010) (see also De Jong, Steenkamp and Fox, 2007; Fox and Verhagen, 2010, Verhagen and Fox, 2012) will be extended with multivariate and covarying random item parameter effects as well as with growth structures to incorporate change over time in both latent health status and item parameters. In this model, the measurement and latent health parameters will be treated as crossed random effects (e.g., De Boeck, 2008; Cho and Rabe-Hesketh, 2011). Moreover, it is possible to include time-varying, person or item level covariates to explain variation in health status and item parameters. Markov Chain Monte Carlo methods will be used for inference.

Advantages of this approach are that it will enable the growth modeling of latent health given a very flexible occasion-specific measurement model, not assuming item parameters to be invariant over measurement occasions. The more realistic occasion-specific measurement models will increase the accuracy of the latent health status estimates. In addition, information acquired about item parameter shifts will provide more insight in measuring latent health status longitudinally. As a case study, data from a randomized clinical trial concerning the treatment of depression will be analyzed.

In section 2, the model will be described in more detail. In section 3, the

estimation procedure and ways to explore longitudinal invariance are described. In section 4, the results of a simulation study to assess parameter recovery are shown and the results of the randomized clinical trial concerning the treatment of depression are presented. A discussion is given in section 5.

4.2 A joint random effects growth model for longitudinal multivariate discrete responses

Health-related questionnaires often consist of a set of items with categorical or ordinal response categories. It will be assumed that the items under analysis measure one unidimensional latent construct, which represents some form of physical or mental health status.

The probability of each response Y_{ijk} can be perceived as a function of the latent health variable θ_{ij} of person $i = 1, ..., I$ at occasion $j = 1, ..., J$ and of the parameters of the measurement part of the model $\tilde{\boldsymbol{\xi}}_{kj}$ for item $k = 1, ..., K$ at occasion j.

To account for the cross-classified nested structure of the data, where occasions are nested in persons and items, random effects structures will be imposed to model the dependency structure. This results in two cross-cutting hierarchies: A three-level structure of observations Y_{ijk} within occasions j within items k referred to as the measurement part of the model and another three-level structure of observations Y_{ijk} within occasions j within persons i, referred to as the latent variable part of the model. Within these hierarchical structures, growth models can be implemented, as well as fixed or random covariates to explain variance within and between persons, items, or occasions (see Figure **??**).

4.2.1 Occasion-specific measurement models for categorical responses

The basis of the measurement model presented here was developed by Masters (1982) and Muraki (1992) for educational data and can be seen as an extension to the polytomous logistic model for multinomial responses or a discrete choice model (Agresti, 2002). The probability that a subject with latent health status θ responds with category c depends on threshold parameters of the item categories and on a discrimination parameter, which can be unique for each item.

In this generalized partial credit model (GPCM), the conditional probability that person i chooses the cth category over the $c-1$st category is modeled with a dichotomous logistic model (see also Li & Baser, 2012). In a longitudinal framework, the GPCM can be extended by including occasion j:

$$\frac{P(Y_{ijk} = c \mid Z_{ijck})}{P(Y_{ijk} = c \mid Z_{ijck}) + P\left(Y_{ijk} = (c-1) \mid Z_{ijck}\right)} = \frac{\exp^{Z_{ijck}}}{1 + \exp^{Z_{ijck}}},$$

which can also be written as:

$$\frac{P(Y_{ijk} = c \mid Z_{ijck})}{P\left(Y_{ijk} = (c-1) \mid Z_{ijck}\right)} = \exp^{Z_{ijck}}.$$

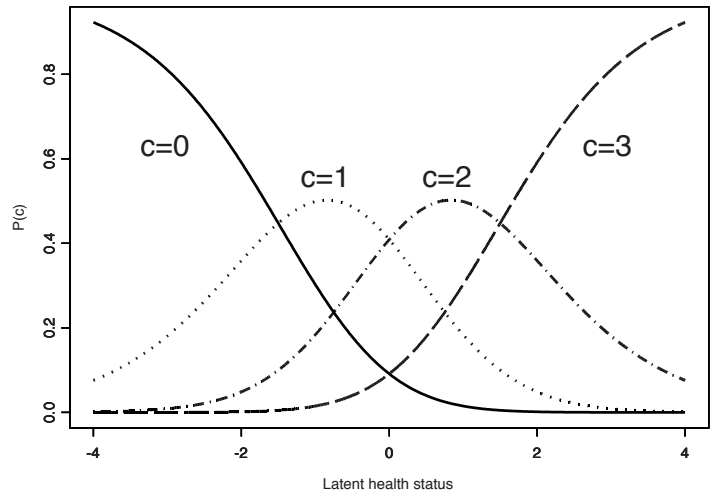

Figure 4.1: Item category response curves: Probability of a response c ($c = 0, 1, 2, 3$) given latent health status θ ($\theta \in [-4, 4]$).

As $\sum_0^C P(Y_{ijk} = c|Z_{ijck})$ equals 1, the probability of response c given Z_{ijck} is given by:

$$P(Y_{ijk} = c \mid Z_{ijck}) \quad = \quad \frac{\exp(\sum_0^c (Z_{ijck}))}{\sum_0^C \exp\left(\sum_0^c (Z_{ijck})\right)}. \tag{4.1}$$

The Z_{ijck} will be modeled in such a way that respondents with a higher latent score θ_{ij} are more likely to score in a higher response category. Furthermore:

$$Z_{ijck} = \tilde{a}_{kj}(\theta_{ij} - \tilde{b}_{ckj}). \tag{4.2}$$

The threshold parameters $\tilde{b}_{ckj}$ are the points on the latent scale at which the category response functions of $P(Y_{ijk} = c|\theta)$ and $P(Y_{ijk} = (c-1)|\theta)$ intersect (Figure 4.1). When the latent health level increases beyond this point, the probability of responding with c becomes higher than the probability of responding with $(c-1)$. The discrimination parameter $\tilde{a}_{kj}$ is related to the discrimination between categorical responses as the latent health level changes. The higher the discrimination parameter, the steeper the category response functions, and the crisper the response categories discriminate between higher and lower latent health levels.

To account for changes in the measurement characteristics over time, Equation (4.2) contains occasion-specific item parameters. For each item, these parameters will be modeled as random effects varying around general longitudinally invariant item parameters. In this way, the parameters of all items are allowed to vary over occasions. The occasion-specific item parameters are exchangeable given the general item parameters and form a second level between the observed responses (level 1) and the invariant general item parameters (level 3).

Since they relate to the same item, the occasion-specific parameters per item are assumed to be correlated. The random item parameters will be assumed to be multivariate normally distributed with an unstructured covariance matrix. This novel way of modeling the GPCM parameters captures within-item correlations. It follows that the random occasion-specific item parameters $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{b}_{0kj}, ..., \tilde{b}_{Ckj})$ are assumed to be multivariate normally distributed given the general item parameters $\boldsymbol{\xi}_k = (a_k, b_{0k}, ..., b_{Ck})$:

$$\tilde{\boldsymbol{\xi}}_{kj} \mid \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\tilde{\xi}_k} \quad \sim \quad \mathcal{N}(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\tilde{\xi}_k}), \tag{4.3}$$

where

$$\boldsymbol{\Sigma}_{\tilde{\xi}_k} \quad = \quad \begin{bmatrix} \sigma^2_{a_k} & \sigma_{a_k b_{0k}} & ... & \sigma_{a_k b_{Ck}} \\ \sigma_{b_{0k} a_k} & \sigma^2_{b_{0k}} & ... & \sigma_{b_{0k} b_{Ck}} \\ ... & ... & ... & ... \\ \sigma_{b_{Ck} a_k} & \sigma_{b_{Ck} b_{0k}} & ... & \sigma^2_{b_{Ck}} \end{bmatrix}. \tag{4.4}$$

Following a hierarchical prior structure, at the third level the general item parameters are assumed to be multivariate normally distributed given mean parameters $\boldsymbol{\mu}_\xi = (a_0, b_{00}, ..., b_{C0})$:

$$\boldsymbol{\xi}_k \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi). \tag{4.5}$$

The extension of the GPCM with occasion-specific random item parameters following a multivariate hierarchical structure is a natural extension in this Bayesian modeling framework. For model identification additional restrictions are necessary, which will be described in section 4.2.4. When the item parameters are longitudinally invariant, the occasion-specific parameters are all restricted to be equal to the general item parameters. In that case the item characteristics are a priori distributed according to Equation 4.5.

4.2.2 A growth model for item characteristic change

It is natural to assume that change in item parameters over occasions is not completely random, but follows a growth pattern. This change can be modeled with a random linear time effect of the average time passed at occasion j over subjects. Depending on the total number of occasions available, characteristics can be added to explain linear and higher order change in item parameters over occasions, for example through polynomial time effects. Let $\mathbf{v}_j$ denote a vector of explanatory information at occasion j, and $\boldsymbol{\delta}_k$ a vector of item-specific coefficients predicting the occasion-specific item parameters $\tilde{\boldsymbol{\xi}}_{kj}$. Then, a conditional growth model can be specified:

$$\tilde{\boldsymbol{\xi}}_{kj} \quad = \quad \boldsymbol{\xi}_k + \mathbf{v}_j \boldsymbol{\delta}_k + \boldsymbol{\epsilon}_{\xi_{kj}}, \ \boldsymbol{\epsilon}_{\xi_{kj}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\tilde{\xi}_k}), \tag{4.6}$$

where $\boldsymbol{\Sigma}_{\tilde{\xi}_k}$ is given by Equation 4.4. Its diagonal elements denote the conditional variance in occasion-specific item parameters given the explanatory information. The off-diagonal terms denote the within-item covariance between item characteristics.

4.2.3 A growth model for latent health status

The change over occasions in latent health status of person i can be modeled by a random intercept (β_{0i}) and random linear time effect (β_{1i}) of the time passed since the individual's starting point t_{i0},

$$\theta_{ij} = \beta_{0i} + (t_{ij} - t_{i0})\beta_{1i} + e_{ij}.$$

The growth model can be easily extended by time-varying covariates with random effects (including polynomial effects) denoted as $\mathbf{x}_{ij}$, time-varying covariates with fixed effects (including main time effects) denoted as $\mathbf{s}_{ij}$, and person-level covariates that do not vary over occasions, denoted as $\mathbf{w}_i$. Then, the more general representation of the latent variable growth part of the model is given by:

$$\begin{aligned} \theta_{ij} &= \mathbf{x}_{ij}^t\boldsymbol{\beta}_i + \mathbf{s}_{ij}^t\boldsymbol{\zeta} + e_{ij}, e_{ij} \sim \mathcal{N}(0, \sigma_j) \\ \boldsymbol{\beta}_i &= \mathbf{w}_i^t\boldsymbol{\gamma} + \boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_i \sim \mathcal{N}(0, \mathbf{T}). \end{aligned} \tag{4.7}$$

An occasion-specific residual variance parameter was specified to model the unexplained variability per measurement occasion. This residual variance parameter will be included in identifying restrictions described in Section 4.2.4.

4.2.4 Model identification

Two issues need to be addressed. First, the latent scales for the different occasions need to be linked. Separate models for each occasion result in incomparable scales between occasions. Second, there is no unique solution to Equation 4.2, as multiple parameter combinations result in the same likelihood. For each occasion, a shift in the latent mean μ_j results in the same expected response probabilities as a shift of all threshold parameters $\tilde{b}_{ckj}$ in the opposite direction. A similar identification problem exists for the discrimination parameters $\tilde{a}_{kj}$ and the occasion-specific residual variance of the latent variable σ_j.

Assuming that an overall shift in the health responses over measurement occasions is more likely the result of a change in latent health status than of a mean change in threshold parameters over occasions, the following restriction is imposed: $\sum_k \sum_c \tilde{b}_{ckj} = 0$ for each j. For each occasion, the mean of the thresholds $\sum_k b_{kj}$ is fixed to the arbitrary value of zero, where $b_{kj} = \sum_c \tilde{b}_{ckj}$. Constraining the sum of the threshold parameters to be equal in both groups links the occasion-specific scales, while fixing this sum to zero identifies the model. For each occasion, the product of the discrimination parameters is fixed to one:$\prod_k \tilde{a}_{kj} = 1$ for each j. This expresses the assumption that it is more likely for the latent health variance to change over time than for all items to discriminate equally more or less. In addition to these elementary identification constraints, the random effects occasion-specific item parameters will shrink towards the general item parameters. The variance of item parameters over occasions will indicate the degree to which the items are non-invariant.

The identification constraints deviate from the usual constraints for the GPCM for incomplete designs. Traditionally, so-called "anchor" items that have identical parameters over occasions are used to link the scales. A reference occasion with

fixed mean and variance is used to identify the scales. Restricting the sum of the thresholds and the product of the discrimination parameters is a very natural way of identification in a random effects framework, however, as it allows the variance components to be estimated freely. This is not possible when restricting specific distributional parameters. Problems would arise with the specification of proper prior distributions and the results would be hard to rescale. In addition, unrestricted covariance components make it much easier to extend the model to include for example covariates to explain variance in both the latent variable and the item parameters.

4.3 Estimation and inference

Combining all parts of the model described previously, the implied conditional model can be defined by inserting Equation 4.6 and 4.7 into Equation 4.2 and 4.1. Hence, the likelihood model can be represented by:

$$P(Y_{ijk} = c|\boldsymbol{\xi}_{kj}, \theta_{ij}) \quad = \quad \frac{\exp(\sum_0^c (Z_{ijck}))}{\sum_0^C \exp\left(\sum_0^c (Z_{ijck})\right)},$$

where

$$Z_{ijck} \quad = \quad (a_k + \mathbf{v}_j \boldsymbol{\delta}_{ak} + \epsilon_{a_{kj}})((\mathbf{x}_{ij}\boldsymbol{\beta}_i + \mathbf{s}_{ij}\boldsymbol{\zeta} + e_{ij}) - (b_{ck} + \mathbf{v}_j \boldsymbol{\delta}_{ck} + \epsilon_{b_{ckj}})).$$

4.3.1 Estimation

In the hierarchical modeling approach, the parameters at each level are conditionally independent given the parameters on the higher level. The resulting full posterior is therefore a product of the likelihood and the hierarchical priors at each level:

$$p(\boldsymbol{\theta}, \tilde{\boldsymbol{\xi}}, . \mid \mathbf{Y}) \propto \left[\prod_i \left[\prod_j \left[\prod_k p(y_{ijk} \mid \tilde{\boldsymbol{\xi}}_{kj}, \theta_{ij}) p(\tilde{\boldsymbol{\xi}}_{kj} \mid .) \right] p(\theta_{ij} \mid .) \right] \right],$$

with conditional hierarchical priors for the latent variables $p(\theta_{ij} \mid .)$ and for the random item parameters $p(\tilde{\boldsymbol{\xi}}_{kj} \mid .)$. The hierarchical prior incorporating the mixed effects model on the latent variable is constructed as follows:

$$p(\theta_{ij} \mid .) = p(\theta_{ij} \mid \mathbf{x}_{ij}, \beta_i, \mathbf{s}_{ij}, \zeta, \sigma_j) p(\boldsymbol{\beta}_i \mid \mathbf{w}_i, \boldsymbol{\gamma}, \mathbf{T}) p(\boldsymbol{\zeta} \mid \boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta) p(\sigma_j),$$

and the hierarchical prior defining the growth model for the item parameter:

$$p(\tilde{\boldsymbol{\xi}}_{kj} \mid .) = p(\tilde{\boldsymbol{\xi}}_{kj} | \boldsymbol{\xi}_k, \boldsymbol{\delta}_k, \mathbf{v}_j, \boldsymbol{\Sigma}_{\tilde{\xi}_k}) p(\boldsymbol{\xi}_k \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) p(\boldsymbol{\delta}_k \mid \boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta) p(\boldsymbol{\Sigma}_{\tilde{\xi}_k}).$$

To estimate the model parameters, a Markov chain Monte Carlo sampling method will be used. The first sampling step for both person and item parameters will be a Metropolis Hastings step, which samples correlated and truncated group-specific item parameters. The hierarchical prior parameters will be sampled with a Gibbs sampler. Conjugate hyperprior distributions will be used, which define a hierarchy

of normal distributions. The means and variances of the mean parameters $\boldsymbol{\mu}_\zeta, \boldsymbol{\gamma}, \boldsymbol{\mu}_\xi$ and $\boldsymbol{\mu}_\delta$ will be drawn from normal-inverse Wishart distributions. The hyperpriors for the variance terms σ_j will be drawn from inverse gamma distributions and the hyperpriors for the covariance matrices $\boldsymbol{\Sigma}_{\tilde{\xi}_k}$ from inverse Wishart distributions. The full sampling scheme can be found in Appendix D. Software in the form of Splus or R code combined with Fortran .dll file is available on request from the corresponding author (a.j.verhagen@gw.utwente.nl).

4.3.2 Exploring longitudinal invariance

The proposed model offers many opportunities to explore whether the item parameters are longitudinally invariant. We will focus on two tests of longitudinal measurement invariance, a Bayes factor test focused on invariance of the separate items, and the DIC for the comparison of models with and without invariance restrictions.

Bayes Factor

The variance and covariance components of $\boldsymbol{\Sigma}_{\tilde{\xi}_k}$ can be used to evaluate whether the occasion-specific parameters are invariant over measurement occasions, and whether the discrimination and threshold parameters covary over time. To test longitudinal invariance of each item parameter, a Bayes factor can be computed to compare the marginal likelihood of the nested models with and without invariance of the parameter (see also Verhagen and Fox, 2012).

In case of nested models, the Bayes factor reduces to the ratio of the density (region) in accordance with the null hypothesis under the prior and posterior distribution for the most general model (Dickey, 1971; Verdinelli and Wasserman,1995). Under the conditions that both models share the same conditional distribution of observed data and the parameter space associated with the prior under the null hypothesis, denoted as Θ_0, is a subset of the parameter space associated with the prior under the alternative hypothesis, Θ_1, the Bayes factor can be evaluated as the posterior expectation of the ratio of prior densities.

Let $p\left(\sigma^2_{b_k} \mid H_0\right)$, $\sigma^2_{b_k} \in \Theta_0$ and $p\left(\sigma^2_{b_k} \mid H_1\right)$, $\sigma^2_{b_k} \in \Theta_1$, where $\Theta_0 \subseteq \Theta_1$, denote the prior under the null hypothesis and the prior under the alternative hypothesis, respectively. The Bayes factor in favor of the null hypothesis can be expressed as (Geweke, 2005)

$$\begin{aligned}
\text{BF} &= \frac{\int_{\Theta_0} p\left(\sigma^2_{b_k} \mid H_0\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}}{\int_{\Theta_1} p\left(\sigma^2_{b_k} \mid H_1\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}} = \frac{\int_{\Theta_0} p\left(\sigma^2_{b_k} \mid H_0\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right) d\sigma^2_{b_k}}{p\left(\mathbf{y} \mid H_1\right)} \\
&= \int_{\Theta_1} \left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)}\right] \frac{p\left(\sigma^2_{b_k} \mid H_1\right) p\left(\mathbf{y} \mid \sigma^2_{b_k}\right)}{p\left(\mathbf{y} \mid H_1\right)} d\sigma^2_{b_k} \\
&= \int_{\Theta_1} \left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)}\right] p\left(\mathbf{y} \mid \sigma^2_{b_k}, H_1\right) d\sigma^2_{b_k} \\
&= E\left[\frac{p\left(\sigma^2_{b_k} \mid H_0\right)}{p\left(\sigma^2_{b_k} \mid H_1\right)} \mid \mathbf{y}, \right]. \qquad (4.8)
\end{aligned}$$

The ratio of prior densities is evaluated using MCMC samples from the marginal posterior density of $\sigma^2_{b_k}$ under the alternative hypothesis, given that the ratio is bounded on Θ_1. Since the specific null hypothesis of a variance component of zero is on the boundary of the parameter space, null hypothesis will be specified as $H_0 : \sigma^2 < \delta$, where δ is a very small number which is chosen to represent a small difference in the context of the problem under analysis (Hoijtink, 2011; Klugkist and Hoijtink, 2007; Klugkist, 2008).

The Bayes factor is the ratio of the support for the two models given the data and the prior information. A Bayes factor higher than one indicates more evidence for the null hypothesis, and a Bayes factor lower than one indicates more support for the alternative hypothesis.

DIC

The deviance information criterion (DIC) of Spiegelhalter et al. (2002) will be used to compare the fit of models with and without invariant item parameters. The deviance function will be defined as:

$$D\left(\mathbf{\Lambda}\right) = -2\log\left[\prod_i\left[\prod_j\left[\prod_k p(Y_{ijk} \mid \tilde{\boldsymbol{\xi}}_{kj}, \theta_{ij})\right]\right]\right]. \tag{4.9}$$

The DIC consists of the posterior mean of the deviance corrected for the number of parameters in the model. In hierarchical models the number of parameters is hard to determine, which is solved in the DIC by computing the effective number of parameters p_D. The p_D is computed by subtracting the deviance at the posterior means from the posterior mean of the deviance.

The DIC is given by

$$\begin{aligned} \text{DIC} &= \overline{D\left(\mathbf{\Lambda}\right)} + \overline{D\left(\mathbf{\Lambda}\right)} - D\big(\hat{\mathbf{\Lambda}}\big) \\ &= \overline{D\left(\mathbf{\Lambda}\right)} + p_D, \end{aligned}$$

where $\overline{D\left(\mathbf{\Lambda}\right)}$ is the posterior mean deviance and $D\big(\hat{\mathbf{\Lambda}}\big)$ the estimated deviance at the posterior estimate of $\hat{\Lambda}$.

Longitudinal measurement invariance can be tested by comparing the DIC of the model with occasion-specific item parameters with the DIC of the measurement invariant model. Both models have to be estimated.

4.4 Results

4.4.1 Simulation study: Parameter recovery

To check whether parameter recovery is accurate under the proposed estimation procedure, a simulation study was performed. A data set containing 800 cases with 10 measurement occasions each was generated from a model with both latent (Equation 4.7) and item parameter (Equation 4.6) growth. Measurement occasion times were generated for each case, with varying starting points and time intervals between the occasions.

Table 4.1: Simulation study: True values and posterior means and standard errors of general item parameters ξ_k and correlations between the true values and posterior means of occasion-specific $(a_{kj}, b_{ckj}, \theta_{ij})$ and growth (δ_k, β_i) parameters

item	a_k	$\hat{a}_k$	SE a_k	b_{1k}	$\hat{b}_{1k}$	SE b_{1k}	b_{2k}	$\hat{b}_{2k}$	SE b_{2k}
1	1.20	1.16	0.10	-0.82	-0.79	0.08	0.52	0.51	0.08
2	1.20	1.24	0.06	-0.13	-0.11	0.06	1.00	1.02	0.08
3	0.98	1.04	0.07	-1.05	-1.12	0.09	-0.73	-0.79	0.06
4	0.80	0.76	0.08	-0.11	-0.06	0.07	1.37	1.33	0.08
5	0.98	0.93	0.06	-1.50	-1.52	0.07	1.58	1.61	0.07
6	0.80	0.87	0.08	-1.26	-1.25	0.09	-0.76	-0.87	0.06
7	1.04	1.09	0.08	-0.86	-0.81	0.06	1.55	1.56	0.07
8	0.98	0.94	0.08	-0.08	0.02	0.08	0.11	0.02	0.07
9	0.85	0.79	0.05	-0.53	-0.51	0.06	0.04	0.07	0.05
10	0.99	0.82	0.10	-1.54	-1.44	0.07	0.56	0.54	0.06
11	1.16	1.17	0.08	0.05	0.07	0.07	0.67	0.67	0.08
12	0.70	0.69	0.05	-0.08	-0.10	0.07	0.59	0.55	0.07
13	1.19	1.15	0.06	-0.24	-0.13	0.07	1.79	1.77	0.08
14	1.02	0.93	0.06	-0.74	-0.70	0.06	0.69	0.70	0.06
15	1.02	1.02	0.09	-1.24	-1.19	0.07	0.49	0.53	0.06
16	0.92	0.90	0.06	-0.69	-0.58	0.07	0.27	0.28	0.07
17	1.20	1.18	0.07	0.76	0.82	0.07	1.34	1.26	0.08
18	0.99	0.95	0.07	-1.37	-1.20	0.07	-0.12	-0.13	0.05
19	0.98	0.98	0.05	-0.47	-0.37	0.06	0.73	0.76	0.06
20	1.20	1.16	0.07	-0.26	-0.18	0.07	0.49	0.46	0.07
Correlations between true values and posterior means									
$\rho_{a_{kj}}$			0.94						
$\rho_{b_{ckj}}$			0.99						
ρ_{δ_k}			0.94						
$\rho_{\theta_{ij}}$			0.94						
ρ_{β_i}			0.97						

Normal distributions were used to draw person-specific means ($N(0, .5)$) and person-specific latent time effects ($N(.2, .5)$). Within-person latent variable values were drawn from normal distributions with the person-specific means and person specific variances from an inverse gamma distribution $IG(15, 1/15)$. The general item parameters were also drawn from normal distributions, with means equal to one for the discrimination parameters and means equal to $\pm 2/3$ for the threshold parameters. The occasion-specific item parameters were simulated to be normally distributed around the general item parameters with variances $\sigma^2_{b_{ckj}} = .02$ and $\sigma^2_{a_{kj}} = .04$. Time effects were randomly assigned to each item parameter, varying between zero and .5. A single long MCMC chain was run with 1000 burn-in iterations and 10000 final iterations.

In Table 4.1, an illustration of the results can be found. For all general item

parameters except for the lower category of item 18, the true value fell within 2 standard errors of the posterior mean and the posterior means did not differ systematically from the true values. The correlations between the true values and the posterior means of the occasion-specific parameters, as well as the correlations between the true values and the posterior means of the growth parameters were all above .94.

4.4.2 Application: Intervention effects on Depression level

The proposed model was applied to a study on the effects of guided self-help based on Acceptance Commitment Therapy (ACT) (Hayes et al., 2006). Participants were recruited through advertisements in Dutch newspapers requesting people who want more out of their life but are hindered by depressive or anxiety symptoms. Respondents with very few symptoms were excluded, as well as respondents with a severe disorder, respondents already receiving treatment and respondents with a high risk of suicide. The remaining 376 participants were randomly assigned to one of three condition's: ACT intervention with minimal or extensive email support (250) and a waiting list (116) condition.

In the two experimental conditions, questionnaires were administered at five moments during the study: at the start of the study and after 3, 6, 9, and 20 weeks. The respondents in the control condition only answered the questionnaires at the start of the study and after 9 weeks. It was assumed that missing occasion measurements (2 for 12 experimental and 1 for 21 experimental and 3 control group members) were missing randomly.

The aim of the study was to investigate whether the ACT intervention reduced depression and anxiety. To measure depression, the CES-D depression questionnaire (Radloff, 1977) was used, which consists of 20 items measuring symptoms of depression experienced in the last week on a four point scale (seldom or never i.e. less than 1 day, sometimes i.e. 1-2 days, often i.e. 3-4 days, almost always i.e. 5-7 days). For many items, few or no responses were observed in the highest answer category. Therefore, for all items, the category was collapsed with the third category. This last category now indicates the experience of the symptom on more than three days in the past week. One of the items generated almost no responses in the lowest category, and this item was removed for the analysis without threatening the test validity. The content of the CES-D items is given in Appendix A.

In addition to the effect of the ACT intervention, the hypothesis that the process of decreasing depression was mediated by a higher acceptance level (Bohlmeijer et al., 2011; Forman et al., 2007) was investigated. Acceptance is characterized by patients becoming more able to embrace and accept negative personal experiences instead of avoiding them (Hayes et al., 2006). Therefore, the construct acceptance was measured using the Acceptance and Action Questionnaire (AAQ-II) (Bond et al., 2011).

For all models, a single long MCMC chain of 50000 iterations was run, with a burn-in of 5000 iterations. The trace plots showed good convergence characteristics, and the convergence statistics (Geweke Z, autocorrelations) were satisfactory. The responses at the five measurement occasions were modeled with the occasion-

specific GPCM (Equation 4.1).

Latent growth model

First, the model component for the latent variable depression was modeled conditional on invariant item parameters, reducing the item parameter structure to the general item parameters (Equation 4.5).

In the first model, denoted as M1, the common structure on the latent variable for the experimental and control group consisted of a random intercept and fixed occasion means,

$$\theta_{ij} = \beta_{0i} + \zeta_j Occasion_j + e_{ij},$$

where $e_{ij} \sim N(0, \sigma_j)$ and $\beta_{0i} \sim \mathcal{N}(\gamma_{00}, T_{00})$ The random intercept varied over individuals and specified the between-subject variability in average depression levels conditional on occasion-specific mean levels. An occasion-specific residual variance parameter was specified to model the unexplained variability per measurement occasion.

In the second model, denoted as M2, a latent growth model was implemented. For individuals measured on more than two occasions, the latent growth model included a random intercept β_{0i} and a random slope for a first (β_{1i}) and second (β_{2i}) order polynomial time effect,

$$\theta_{ij} = \beta_{0i} + \beta_{1i} Time_{ij} + \beta_{2i} Time_{ij}^2 + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma_j).$$

The random effects were assumed to be multivariate normally distributed, where the random intercepts and first order slopes were defined conditional on membership of the control (Experimental = 0) or experimental (Experimental=1) group,

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \gamma_{01} Experimental_i + \upsilon_{0i} \\ \beta_{1i} &= \gamma_{10} + \gamma_{11} Experimental_i + \upsilon_{1i} \\ \beta_{2i} &= \gamma_{20} + \upsilon_{2i}, \end{aligned} \tag{4.10}$$

with $\boldsymbol{\upsilon}_i \sim N(0, \mathbf{T})$. The time-point zero corresponds with the first measurement occasion such that the random intercept variance specifies the between-subject variation in depression levels at the first measurement occasion. Furthermore, between-subject variation was specified over the subject-specific linear trend variable and quadratic time variable. The random effects were allowed to correlate using a common covariance matrix. For individuals assessed at just two measurement occasions (i.e. the control group), the latent trajectory was specified without the subject-specific quadratic time effect.

In the third model, denoted as M3, the subject-specific difference in acceptance between the first and fourth measurement occasion was used as an explanatory third-level variable, denoted as Acceptance. Subsequently, the random subject effects at level 3 are given by,

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \gamma_{01} Experimental_i + \gamma_{02} Acceptance_i + \upsilon_{0i} \\ \beta_{1i} &= \gamma_{10} + \gamma_{11} Experimental_i + \gamma_{12} Acceptance_i + \upsilon_{1i} \end{aligned}$$

Table 4.2: Deviance Information Criteria for models M1 to M6

Model	Specification	$\hat{D}$	$\bar{D}$	pD	DIC
Invariant item parameters					
M1	Occasion means	40191	41885	1694	43580
M2	M1 + Latent growth, Condition	40116	41081	965	42047
M3	M2 + Acceptance	40120	41082	962	42044
Non-invariant item parameters					
M4	Non-invariant	40228	41234	1007	42241
M5	Partially invariant	39968	40992	1024	42015
M6	M5 + Latent item trajectories	39909	40976	1066	42041

$$\beta_{2i} = \gamma_{20} + \upsilon_{2i}, \tag{4.11}$$

with $\boldsymbol{\upsilon}_i \sim N(0, T)$.

In the upper part of Table 4.2, the DIC of the three models are compared. The object is to select the best model given invariant item parameters such that in a next stage the assumption of longitudinal measurement invariance can be evaluated. Conditional on invariant item parameters, the second and third model have a substantially lower DIC than the first model, which indicates that the subject-specific latent variable trajectory significantly improves the model fit. According to the DIC, the fit of model M3 does not increase relative to model M2. However, the DIC measures the fit of the random effects at level 1, which does not really change by adding acceptance as an explanatory variable. The common effect of acceptance on the linear trend parameter of the latent variable trajectory of depression is significant and large, around 1.06 (.28), which shows that participants increasing their level of acceptance show a significant decrease in their level of depression. As a result, model M3 will be used to evaluate the longitudinal measurement invariance assumptions.

Investigating longitudinal measurement invariance

In model M3, item parameters were restricted to be measurement invariant over occasions, reducing the item parameter structure to the general item parameters (Equation 4.5). Given the latent growth structure of depression in model M3, this model was generalized by assuming all item parameters to be measurement non-invariant using the random item effects specification over time (Equation 4.3), which will be referred to as model M4. A restricted version of model M4, denoted as model M5, consisted of some items restricted to be invariant and was used to evaluate partial longitudinal measurement invariance. In model M6 latent trajectories were added to the identified non-invariant item characteristics of model M5.

In the lower-part of Table 4.2, the DICs of the generalized models from the full invariant model M3 are given. The assumption of full longitudinal measurement

non-invariance is not supported by the data, since the DIC of model M4 is lower than that of model M3. Subsequently, the measurement invariance assumption of each item was evaluated. Therefore, the multiple marginal null hypotheses of longitudinal measurement invariance were investigated using the Bayes factor specified in Equation (4.8), and by examining the item parameter variances over time.

Table 4.3: For model M4, posterior variance estimates of random item characteristics over time, and Bayes factor estimates concerning the marginal measurement invariance null hypotheses. For model M6, latent trajectory parameter estimates of the non-invariant item parameters.

Item	Model M4						Model M6		
	σ_{a_k}	BF	$\sigma_{b_{1k}}$	BF	$\sigma_{b_{2k}}$	BF	δ_{a_k}	δ_{1k}	δ_{2k}
1	.15	5.18	.14	6.58	.21	1.45		-.01(.05)	.09(.06)
2	.14	6.23	.14	6.59	.20	2.89		.02(.05)	.06(.06)
3	.18	3.87	.17	4.69	.20	2.89		-.06(.05)	.03(.06)
4	.14	6.77	.16	4.73	.15	4.78			
5	.16	4.02	.16	4.41	.24	**0.32**		-.03(.06)	.09(.06)
6	.29	1.88	.41	1.03	.23	3.30	.13(.08)	**-.28**(.10)	.12(.08)
7	.20	2.43	.17	4.69	.18	2.64	.08(.06)	**-.16**(.08)	.05(.06)
8	.14	6.77	.15	6.17	.14	6.79			
9	.17	4.21	.14	6.38	.17	4.00			
10	.14	7.30	.13	8.02	.17	4.18			
11	.16	2.59	.18	3.53	.18	2.34	.02(.03)	-.10(.06)	.06(.05)
12	.15	5.27	.16	4.79	.19	3.10			
13	.16	4.60	.19	3.06	.16	4.50			
14	.14	7.18	.14	6.43	.15	6.06			
15	.17	3.62	.21	3.45	.14	7.39			
16	.16	4.66	.14	6.46	.19	3.55			
17	.19	2.91	.15	6.28	.16	5.12	-.02(.05)		
18	.18	1.83	.14	6.98	.15	5.54	**-.07**(.03)		
19	.16	4.48	.27	0.62	.27	**0.16**		-.08(.05)	**.11**(.06)

In Table 4.3, for each item the estimated posterior standard deviation over time of the three item parameters are given under the label σ_{a_k}, $\sigma_{b_{1k}}$, $\sigma_{b_{2k}}$ respectively. A high Bayes factor value supports the null hypothesis of longitudinal measurement invariance. For items 5 and 19, the Bayes factors showed substantially more evidence for longitudinal variance in the highest item threshold parameters than for longitudinal invariance, as indicated by a Bayes Factor lower than .33. The measurement invariance assumption was at least three times more likely than measurement non-invariance for the parameters of item 2,8 to 10,12 to 15, and 16, for the threshold parameters of items 17 and 18, and for the discrimination parameters of item 1,2,3,5, and 19.

Model M5 was defined as a partially restricted measurement invariant model, where the measurement invariant item parameters (i.e., tested to be invariant un-

der model M4) were restricted to be invariant over time. This model M5 showed a better fit to the data than both the full invariant (M3) and the full non-invariant (M4) model according to the DICs represented in Table 4.2. The test result supports the joint hypothesis of partial measurement invariance.

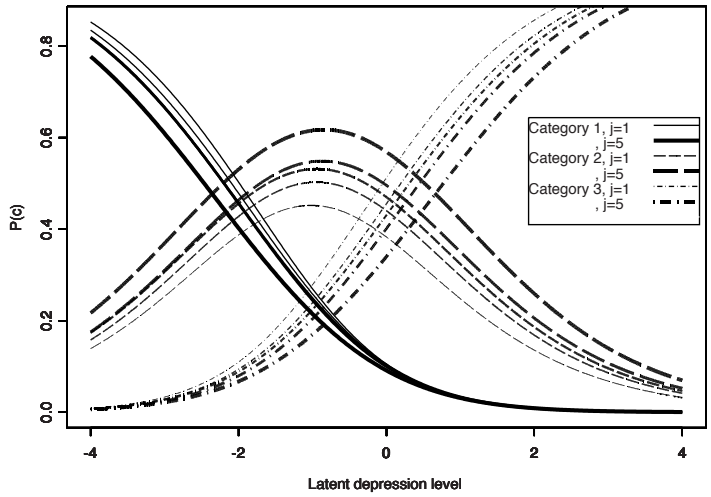

Figure 4.2: Item category response probabilities for item 19 on five measurement occasions.

In model M6, latent trajectories with a linear trend component were defined for the non-invariant item characteristics. According to Equation 4.6, a latent trajectory was defined with a random intercept, defining the mean item characteristic level over time, and a time variable that models the trend over time. The estimated posterior mean trend effects and the standard errors can be found in Table 4.3 under model M6. The discriminating effect of item 18 decreased significantly over time, while for item 6 and 7 the probability of responding with the second over the first category increased over time and for item 19 the probability of responding with the third over the second category decreased over time. In Figure 4.2, the significant shift in probabilities for the answer categories over time is illustrated for item 19, where the category bounds as a function of depression level become more bold over time. This illustrates the trend of the middle category to become more probable over time, as the lower threshold shifts downward and the upper threshold shifts upward for each level of depression.

Latent developmental trajectory of depression

In Model M6, the subject-specific latent trajectory of depression was modeled conditional on the partial measurement invariant item parameters. Model M7 corresponds with model M6 but the explanatory variable Acceptance was excluded from

the latent trajectory function. The mean latent trajectory parameter estimates of both models can be found in Table 4.4.

Table 4.4: Conditional on partial longitudinal measurement invariance, population parameter estimates and standard errors of the mean latent trajectory of depression with and without variable Acceptance.

		Model M7		Model M6	
	Parameter	EAP	SE	EAP	SE
Fixed Effects					
Intercept	γ_{00}	.13	.07	.13	.07
	γ_{10}	**-.28**	.09	-.15	.09
	γ_{20}	**.17**	.02	**.17**	.02
Experimental	γ_{01}	-.05	.08	-.06	.08
	γ_{11}	**-.84**	.11	**-.78**	.11
Acceptance	γ_{02}			.03	.23
	γ_{12}			**1.06**	.28
Random Effects					
Residual Variance Level 2					
	σ_1	.04	.01	.04	.01
	σ_2	.10	.03	.10	.03
	σ_3	.09	.03	.09	.03
	σ_4	.04	.01	.04	.01
	σ_5	.05	.02	.06	.02
Residual Variance Level 3					
	T_{00}	.52	.05	.52	.05
	T_{11}	.98	.13	.83	.13
	T_{22}	.12	.03	.11	.03

Model M7 shows a non-zero negative linear trend of depression for subjects in the control group ($\gamma_{10} = -.28, SE = .09$), with a steeper negative mean trend for subjects in the experimental group ($\gamma_{11} = -.84, SE = .11$). For the experimental group members, the negative linear mean trend is decelerated by a second order time effect ($\gamma_{20} = .17, SE = .02$). The estimated random effects variances show that there is heterogeneity in depression levels at the start of the study, and significant between-subject variation in trend effects and in the decelerating effects of the squared time variable.

The inclusion of the level-3 variable Acceptance to explain heterogeneity in the subject-specific trajectory parameters (intercept and trend), annihilates the first-order time effect for the control group ($\gamma_{10} = -.15, SE = .09$), which is no longer significantly different from zero. It also slightly attenuates the negative linear mean trend in depression for the experimental group ($\gamma_{11} = -.78, SE = .11$). The decrease of both effects indicates that part of the decrease of depression over time can be explained by an increase in level of acceptance. The same effect is also indicated by a decrease in the residual variance in first order slopes from .98 to .83, when conditioning on the change in acceptance. A strong positive effect was

found of change in acceptance on decrease in depression ($\gamma_{12} = 1.06, SE = .28$). The other variance components were unaffected by the inclusion of acceptance. The small residual variances at level 2 show that model M6 and M7 explain most of the heterogeneity per measurement occasion.

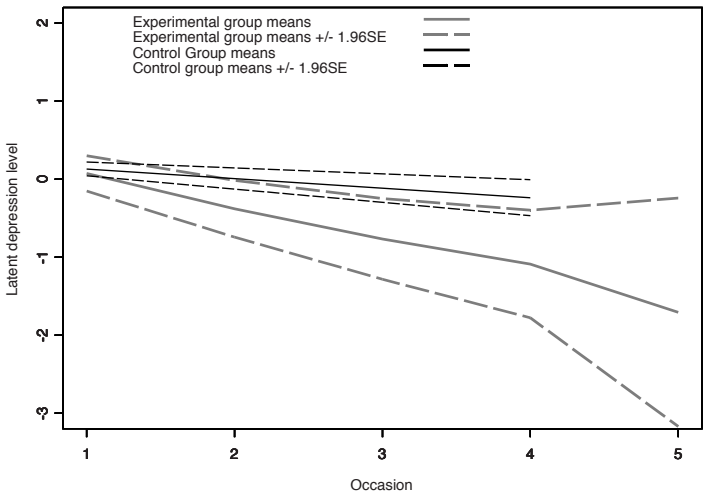

Figure 4.3: Mean latent growth patterns of depression for the experimental and control group.

Figure 4.3 shows the predicted occasion means of depression for the experimental and control group and 95% credible intervals. The mean depression level in the experimental group declined much steeper than that of the control group resulting in a lower average depression level at measurement occasion four.

4.5 Discussion

In this paper, a joint growth model with random occasion-specific parameters for both latent health and item parameters was proposed to measure changes in latent health status over time. Instead of assuming longitudinal invariance of the measurement model, we modeled variance in item parameters over measurement occasions with a growth model. The result is a multilevel growth model which reflects the effects of time characteristics on the occasion-specific item parameters and on the latent health construct simultaneously. Using the more flexible occasion-specific item parameters, change in the latent variable can be measured more accurately. In addition, the change in item parameters can provide insight into the process of longitudinal measurement.

A simulation study showed accurate parameter recovery for most of the general, the occasion-specific, and the growth parameters in both the latent and the item parameter growth models.

The analysis of a randomized trial on depression showed that the item parameters did change over measurement occasions. For some items, participants were more inclined to answer towards the middle category instead of the lowest or highest category over time. Conditional on this item parameter change, there was an overall decrease in depression level for the experimental group, taking subject specific change over time into account. The decrease in depression was stronger for the experimental group than for the control group. The change in depression level was steeper for the participants with a larger change in acceptance level.

A natural next step is to investigate what causes the shifts in item parameters and how to handle this in the best way. In addition, health questionnaires often consist of clusters of items, of which some may be more sensitive to change over time than others. To model this, it would be possible to include covariates with item characteristics for each item on the third level to predict differential time paths for items with, for example, specific content.

Chapter 5

Bayesian Item Response Theory Models for Measurement Variance

5.1 Introduction

When scores of several groups on tests or questionnaires (cognitive tests, psychological questionnaires, consumer surveys, attitude questionnaires) are compared, the scores for these groups should be on the same scale for the comparison to be valid. This can be achieved by ensuring that the measurement instrument (test, questionnaire) used to determine the scores is at least partially measurement invariant (see for an overview: Millsap, 2011). In another approach, the measurement variance is explicitly modeled such that comparable scores are obtained. This paper will cover Bayesian IRT models for a range of test situations, introducing a general Bayesian framework for modeling measurement variance and testing for measurement invariance.

Measurement invariance is defined as the situation in which persons from each group with the same underlying value of the measured construct have the same probability of endorsing an item. If there is measurement variance, there are differences between groups in the way the group members respond to items in the test or questionnaire. To investigate whether there is measurement variance and how it is expressed, it is necessary to model the responses on a test or questionnaire with measurement models which allow group differences in both the test scores and the characteristics of the test items. Those models will be referred to as multi-group measurement models.

Item response theory (IRT) models are a common choice for measurement models, especially in case of discrete responses. A variation on the basic Rasch model (Rasch, 1960), the one parameter normal ogive model (1PNO), will be used

This chapter has been written in collaboration with prof. dr. R.E. Millsap and dr. R. Levy from Arizona State University. An adapted version of this chapter will be submitted for publication.

to introduce the concepts of the Bayesian framework for measurement variance modeling. In the 1PNO model, the probability of a dichotomous response of person $i = 1, \ldots, N$ on item $k = 1, \ldots, K$ is modeled as a function of the threshold or difficulty of an item k, b_k (item parameter), and the score of a person on the underlying construct being measured θ_i (person parameter):

$$P(Y_{ik} = 1|\theta_i, b_k) = \Phi(\theta_i - b_k). \tag{5.1}$$

The model assumes unidimensionality and local independence of the item responses.

More recently, Bayesian versions of the well-known IRT models have been developed (Albert, 1992; Fox & Glas, 2001; Patz & Junker, 1999a, 1999b). In the Bayesian framework, item parameters are modeled to be random. The priors for the item parameters specify the variation among item characteristics.

Bayesian IRT models known as random item effects models (e.g. De Boeck, 2008; Janssen, Tuerlinckx, Meulders & De Boeck, 2000; Glas & van der Linden, 2003) assume the items in a test are a random sample from an item population, in the same way persons are usually assumed to be a random sample from a larger population. Applying this to the 1PNO model in Equation 5.1, the following distributions for the person and item parameters result:

$$\theta_i \mid \mu_\theta, \sigma^2_\theta \sim \mathcal{N}(\mu_\theta, \sigma^2_\theta)$$
$$b_k \mid \mu_b, \sigma^2_b \sim \mathcal{N}(\mu_b, \sigma^2_b),$$

such that the person parameters are assumed to be normally distributed with population mean μ_θ, and the item parameters to be normally distributed with population mean μ_b.

In this chapter, group-specific random item parameters will be used to model variation in item functioning on top of the variation across items. Bayesian IRT models are easily extended in this way to form a multi-group measurement model. Although random group-specific item parameters were originally used to model the nesting of items within testlets or item families (Bradlow, Wainer & Wang, 1999; Glas & Van der Linden, 2003; Glas, Van der Linden & Geerlings, 2010; Janssen et al., 2000; Sinharay, Johnson & Williamson, 2003), recently they have also been used to model the nesting of persons within groups with different item parameters, resulting in multi-group measurement models (De Boeck, 2008; Frederickx, Tuerlinckx, De Boeck & Magis, 2010; de Jong, Steenkamp & Fox, 2007; de Jong & Steenkamp, 2009; Fox, 2010; Fox & Verhagen, 2010, Verhagen & Fox, 2012).

The multi-group 1PNO model for persons $i = 1, \ldots, N$ nested in groups $j = 1, \ldots, J$ is given by:

$$P(Y_{ijk} = 1) = \Phi(\theta_{ij} - b_{kj}), \tag{5.2}$$

with group specific thresholds b_{kj}. In a similar way, more general IRT models can be extended to multi-group models by specifying group-specific item and person parameters (see Appendix E).

Response patterns do not uniquely define the group-specific parameters in a multi-group measurement model. Consequently, restrictions are needed to identify the model. In the Bayesian random effects framework, restricting the mean

of the threshold parameters to zero within all groups is a natural choice, which reflects the assumption of equal test difficulty across groups. Traditional methods to estimate multi-group IRT models and test for measurement invariance (e.g. likelihood ratio test, Thissen, Steinberg, & Wainer, 1993), based on maximum likelihood estimates of the item parameters, are usually identified based on a fixed person parameter mean μ_θ for a reference group and at least one anchor item with equal item parameters in all groups.

The goals of this chapter are to introduce a general Bayesian multi-group IRT model framework and to show its advantages for both measurement variance modeling and measurement invariance testing.

First, the random item parameter modeling framework is more flexible than traditional models for measurement variance. Not restricted by anchor items or reference groups, it is straightforward to estimate variance and covariance terms and to model differences between groups in either item or person parameters with covariates to explain variation in item parameters over items (e.g. to explain the difficulty by item content) and groups (e.g. to explain item parameter variance), and variation in person parameters (e.g. explain scores by background person- or group-level information). The flexibility of the framework is increased by the possibility to include latent classes (e.g. Vermunt, 2008) at several levels by specifying non-normal distributions of the random effects parameters.

A basic distinction is made between models for two situations. In the first situation, differences between group means and item parameters of a few specific groups are investigated, for example between boys and girls on an educational test. In the second situation, variation in group-specific parameters is investigated for groups which are assumed to be sampled from a larger population. These fixed or random groups can either be manifest and known beforehand, but they can also be latent to be exposed by a latent class analysis, for example when there is suspicion of groups with different response styles on an attitude questionnaire. As a result, the Bayesian multi-group IRT model framework is adaptable to a wide range of testing situations.

Second, within the Bayesian multi-group IRT model framework, Bayes factors for nested models are easily incorporated to identify which items are invariant. Multiple invariance hypotheses can be tested simultaneously, conditional on other model specifications and without the need to specify anchor items. In addition, the Bayes factor allows evaluation of the likelihood of both the null and the alternative hypothesis given the data, providing a balanced evaluation of the evidence for both the null and the alternative hypothesis. Bayes factor based measurement invariance tests will be described to test the item parameter variance over groups and to test specific differences between the item parameters of two groups. Another method to identify which items are invariant is to classify items as invariant or non-invariant using mixture models with latent classes for invariant and non-invariant item parameters. When items are found to be invariant, these items can be restricted to anchor items to decrease the number of parameters to be estimated and to increase model fit.

This chapter will start with an overview of Bayesian multi-group IRT models. Assumptions regarding the identification of multi-group IRT models will be described and explained in detail. Then, Bayesian estimation procedures and Bayes

factor tests for measurement invariance will be introduced. Simulation studies will evaluate the Bayes factor test for measurement invariance, and compare these results to a likelihood ratio test for measurement invariance (Thissen, Steinberg & Wainer, 1993). The presented Bayesian IRT models will be illustrated with examples concerning geometry items from the College Basic Academic Subjects Examination (CBASE) and the depression scale of the Survey of Health, Ageing and Retirement in Europe (SHARE) survey. The inferences made with Bayesian IRT models will be compared to the results from a traditional maximum likelihood estimation method. Appendix G and H provide WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000) and R code for the presented Bayesian IRT models and Bayes factor tests.

5.2 Bayesian multi-group IRT models

This section will give an overview of Bayesian multi-group IRT models. The overview will be structured by dividing the models into random versus fixed group models, latent versus manifest group models and single versus multiple cluster models.

Random group models assume that the groups under investigation are a random sample from a population of groups. This assumption is translated into a hierarchical or multilevel model on the group-specific parameters. Consider persons $i = 1, \ldots, N$ nested within random groups $j = 1, \ldots, J$. The hierarchical distribution of the group-specific person parameters of the IRT model in Equation 5.2 around the group means is given by:

$$\theta_{ij} \mid \mu_{\theta_j}, \sigma^2_{\theta_j} \sim \mathcal{N}(\mu_{\theta_j}, \sigma^2_{\theta_j}), \tag{5.3}$$

with the group means assumed to be a sample from a larger population with population mean μ_{θ_0}:

$$\mu_{\theta_j} \sim \mathcal{N}(\mu_{\theta_0}, \tau^2). \tag{5.4}$$

Fixed group models assume that the groups under investigation are a finite fixed number of groups of specific interest. Although in this situation the person parameters are still randomly distributed around their group means as in Equation 5.3, the group means are assumed to be independent and uninformative about each other. A possible prior for the group means in this situation is a normal prior with a large variance:

$$\mu_{\theta_j} \quad \sim \quad \mathcal{N}(0, M), \tag{5.5}$$

where M is a large number. Similar multilevel and fixed group structures can be specified for the group-specific item parameters b_{kj} in Equation 5.2.

The second distinction is made between models with manifest and with latent groups. In a generalized latent variable framework (e.g. Skrondal & Rabe-Hesketh, 2004), a categorical latent variable can be specified, indicating that the persons or items are part of a certain class $c = 1, \ldots, C$ with class probability $\boldsymbol{\pi}_c$ (e.g. Vermunt, 2008). The latent classes can be considered as "latent groups" with different or differently distributed group-specific parameters.

Table 5.1 gives an overview of known Bayesian multi-group IRT models with group structures on person parameters, item parameters or both. Below, basic fixed and random Bayesian multi-group IRT models will be described.

5.2.1 Multi-group IRT models for fixed groups

Three multi-group Bayesian IRT models for fixed groups will be described. The first model assumes a manifest group structure for both person and item parameters. The second model assumes a latent group structure for both item and person parameters. The third model extends the first model with an additional latent clustering for the item parameters, which indicates whether item parameters are invariant over groups or not. The code for estimation in WinBUGS (Lunn et al., 2000) is given in Appendix G.

Manifest groups for item and person parameters

In fixed manifest group models, differences in group-specific item parameters are modeled for a small number of groups (De Boeck, 2008). For the 1PNO IRT model with group-specific parameters as specified in equation 5.2, and person parameters as in equation 5.5, a model can be specified in which the item parameters $\tilde{b}_{kj}$ are random within each group over items:

$$\tilde{b}_{kj} \quad = \quad \mu_{b_j} + e_{kj},$$

where

$$\begin{aligned} \mathbf{e}_k \mid \mathbf{\Sigma}_b &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_b) \\ \mathbf{\Sigma}_b &= \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1 b_{\cdots}} & \sigma_{b_1 b_J} \\ \sigma_{b_{\cdots} b_1} & \sigma^2_{b_{\cdots}} & \sigma_{b_{\cdots} b_J} \\ \sigma_{b_J b_1} & \sigma_{b_J b_{\cdots}} & \sigma^2_{b_J} \end{bmatrix}, \end{aligned} \tag{5.6}$$

with hyperpriors $\mu_{b_j} = 0$ and $\mathbf{\Sigma}_b \sim IW(S, J)$ (See Appendix F for choice of S), respectively, and $\mathbf{e}_k = (e_{k1}, \ldots, e_{kJ})$.

The covariance structure of the group specific item parameters is modeled by the covariance matrix $\mathbf{\Sigma}_b$. The item parameter estimates will show shrinkage towards the item parameter mean μ_{b_j} within each group, but the group-specific estimates for the same item k will also shrink towards each other due to the covariances $\sigma_{b_j b_{j'}}$ $(j \neq j')$, depending on the information in the data. The model becomes less parsimonious as the number of groups increases, because the number of parameters to be estimated increases rapidly. In Appendix E, a fixed manifest groups model for the two parameter normal ogive (2PNO) and for a generalized partial credit (GPCM) IRT model are given.

Latent groups for person and item parameters

IRT models in which subjects are clustered in latent classes based on differential response patterns are known as mixture IRT models (Bolt, Cohen & Wollack, 2001; Cho & Cohen, 2010; Cohen & Bolt, 2005; Li, Cohen, Kim & Cho, 2009; Rost, 1990; von Davier & Yamamoto, 2004). Mixture IRT models are most useful

Table 5.1: Classification of Bayesian unidimensional multi-group IRT models

		Group structure on:		
		Person	Item	Person and Item
Fixed	Manifest	Bayesian Multiple group IRT model (Beguin & Glas, 2001; Azevedo et al., 2012)		Random Item Profiles (De Boeck, 2008)
	Latent	Latent class IRT (e.g. Hoijtink & Molenaar, 1997)		Mixture IRT models (e.g. Rost, 1990)
	Manifest person Latent item			RIM (Frederickx et al.,2010) (Soares et al. 2009)
Random	Manifest	Multilevel IRT model (Fox & Glas, 2001)	Testlet model (Bradlow, Wainer & Wang, 1999) Item family models (Glas & Van der Linden, 2003)	Random item effects multilevel IRT model (Fox, 2010) Longitudinal joint growth model (Verhagen & Fox, in press)
	Latent			
	Manifest person Latent item			Finite mixture multilevel IRT model (De Jong & Steenkamp, 2009)

with a considerable amount of (non-invariant) items and when there are substantial differences in item parameters between the latent groups (DeMars & Lau, 2011).

A mixture 1PNO model with two classes $G_i = 1, 2$ can be defined as:

$$\begin{aligned} P(Y_{ik} = 1 \mid \theta_i, \mathbf{b}_k, G_i = g) &= \Phi(\theta_{ig} - \tilde{b}_{kg}) \\ G_i &\sim Bern(\pi_g) \end{aligned} \tag{5.7}$$

where the probability of endorsing the item depends on the class-specific item parameters. As hyperprior for the class probabilities π_g, a beta distribution $\pi_g \sim B(1, 1)$ can be used. When more than two latent classes are assumed, the Bernouilli distribution for class membership g_i can be replaced by a multinomial distribution, with a Dirichlet prior for the class probabilities π_g.

For each class g, a different set of item parameters $\tilde{b}_{kg}$ is estimated, and the person parameters θ_{ig} are modeled to have a group specific mean and variance. These group specific means and variances can be modeled according to Equation 5.6, in the same way as the manifest group parameters, replacing j with g.

Latent groups for items, manifest groups for persons

Frederickx et al. (2010), and Soares and Gamerman (2009) have proposed models in which some items are classified to come from a class with invariant item parameters, and the other items to come from a class with group-specific item parameters. In this way, an indication of which items are invariant is acquired, without specifying anchor items in advance.

Adding such a latent class structure to the 1PNO IRT model with group specific parameters as specified in equation 5.2, the model on the item parameters becomes:

$$\tilde{b}_{kj} = \mu_{b_j} + e_{kj},$$

where

$$\begin{aligned} e_{k1} = \ldots = e_{kJ} \mid C_k = 0 &\sim \mathcal{N}(0, \sigma_b^2) \\ \mathbf{e}_k = (e_{k1}, \ldots, e_{kJ}) \mid C_k = 1 &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_b) \\ \mathbf{\Sigma}_b &= \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1 b_{\ldots}} & \sigma_{b_1 b_J} \\ \sigma_{b_{\ldots} b_1} & \sigma^2_{b_{\ldots}} & \sigma_{b_{\ldots} b_J} \\ \sigma_{b_J b_1} & \sigma_{b_J b_{\ldots}} & \sigma^2_{b_J} \end{bmatrix} \\ C_k &\sim Bern(\pi). \end{aligned}$$

The hyperprior for the class probability can be specified as $\pi \sim B(1, 1)$, and hyperpriors for the variance and covariance matrix could be $\sigma_b \sim IG(1, .1)$ and $\mathbf{\Sigma}_b \sim IW(S, J)$. The person parameters are still specified as in equation 5.5. By adding latent classes to the item parameters in this way, the number of parameters to be estimated is reduced, and a straightforward estimate is obtained for the probability that an item is invariant given the data.

5.2.2 Multi-group IRT models for random groups

Two random multi-group models will be described. The first model assumes a manifest group structure for both person and item parameters. The second model

extends the first model with two latent classes for the item parameters which indicate whether item parameters are invariant or not. A multi-group model with random latent groups is possible in theory, but has not been implemented so far. The code for estimation in WinBUGS (Lunn et al., 2000) can be found in Appendix G. An advantage of the multilevel structure in random group models is the ease with which explanatory information can be added to explain variance in person parameters, item parameters, or both. Models with explanatory information will be discussed in 5.2.2.

Manifest groups for item and person parameters

The random effects model for modeling measurement variance (De Jong, Steenkamp & Fox, 2007; Fox, 2010; Fox & Verhagen, 2010; Verhagen & Fox, 2012) is based on the assumption that the available groups are a sample from a larger population of groups. In this case, the group-specific item parameters can be modeled as normally distributed random effects around general parameters for each item.

For the 1PNO IRT model with group specific parameters as specified in Equation 5.2, the group means will be modeled as random effects around a common population mean, as specified in Equation 5.4. For the item parameters $\tilde{b}_{kj}$, a multilevel model can be specified in a similar way

$$\begin{aligned} \tilde{b}_{kj} &= \mu_{b_0} + u_k + e_{kj} \\ u_k &\sim N(0, \sigma_b^2) \\ e_{kj} &\sim N(0, \sigma_{b_k}^2), \end{aligned} \tag{5.8}$$

with hyperpriors $\mu_{b_0} = 0$ and $IG(1, .1)$ (see Appendix F) for both σ_b^2 and $\sigma_{b_k}^2$, respectively. This creates a compound symmetry covariance structure in which the group-specific parameters within an item are correlated, but the group-specific parameters for different items are unrelated. As the group-specific item parameters are random effects around the general item parameters, the group-specific parameters will shrink towards these general item parameters depending on the information in the data. The model is less useful for a small number of groups, since in that case the assumption of exchangeable, normally distributed group-specific (item) parameters is difficult to verify and estimates of variance components are based on just a few observations. In Appendix E, the corresponding model extensions for the two parameter normal ogive (2PNO) and for the generalized partial credit (GPCM) IRT models are given.

Manifest groups for persons, latent groups for items

The manifest random groups model in Equation 5.8 can also be extended with latent classes to account for variant and invariant items, without specifying anchor items in advance. This model has not been described before and will be introduced here.

The person parameters are modeled according to Equation 5.4. For the item parameters, an extension to the model in Equation 5.8 can be made:

$$\tilde{b}_{kj} = \mu_{b_0} + u_k + C_k e_{kj}$$

$$\begin{aligned} u_k &\sim N(0, \sigma_b^2) \\ e_{kj} &\sim N(0, \sigma_{b_k}^2) \\ C_k &\sim Bern(\pi) \end{aligned} \tag{5.9}$$

As hyperpriors $\pi \sim B(1,1)$ for the class probability, $IG(1,.1)$ (see Appendix F) for both σ_b^2 and $\sigma_{b_k}^2$ and $\mu_{b_0} = 0$ can be used.

In essence, the variance component is forced to equal zero for the items in the invariant class. This reduces the number of group-specific parameters to be estimated, and supports estimating the probability that an item is invariant given the data.

Explanatory information in random group models

The multilevel structure and the absence of fixed group-specific parameters makes it easy to extend the random group models for the person and item parameters to include explanatory information.

The multilevel IRT model (Fox & Glas, 2001) extends the random group model on the person parameters with explanatory information. Individual ($\boldsymbol{x}_{ij}$) and group ($\mathbf{w}_j$) level information accounts for background differences in the person parameters. It follows that,

$$\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma_{\theta_j}^2 \sim \mathcal{N}\left(\mathbf{x}_{ij}^t \boldsymbol{\beta}_j, \sigma_{\theta_j}^2\right) \tag{5.10}$$

$$\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T} \sim \mathcal{N}\left(\mathbf{w}_j \boldsymbol{\gamma}, \mathbf{T}\right), \tag{5.11}$$

where $\boldsymbol{\beta}_j$ are the (random) effects of individual covariates and $\boldsymbol{\gamma}$ are the effects of the group level covariates on $\boldsymbol{\beta}_j$.

A similar multilevel model with covariates can be introduced for item parameters. To explain variance in item parameters over groups, group-level explanatory covariates $\mathbf{v}_{kj}$ can be included (Verhagen & Fox, 2012):

$$\tilde{b}_{kj} \mid \mathbf{v}_{kj}, \boldsymbol{\delta}_{b_k}, \sigma_{b_k}^2 \sim \mathcal{N}\left(b_k + \mathbf{v}_{kj}^t \boldsymbol{\delta}_{b_k}, \sigma_{b_k}^2\right), \tag{5.12}$$

where the regression coefficients, $\boldsymbol{\delta}_{b_k}$, can be assumed fixed or random across items.

Longitudinal models can also be defined as multi-group models. Growth models can be used to model changes over time in the latent person scores, but also in the item parameters: the "item parameter drift" (Meade, Lautenschlager & Hecht, 2005; Millsap, 2010). Bayesian joint growth models for multivariate ordinal responses were proposed by Verhagen and Fox (in press). For the person parameters, the longitudinal growth model can include time-varying covariates with random effects (including polynomial effects) denoted as $\mathbf{x}_{ij}$, time-varying covariates with fixed effects (including main time effects) denoted as $\mathbf{s}_{ij}$, and person-level covariates, denoted as $\mathbf{w}_i$:

$$\begin{aligned} \theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_i, \mathbf{s}_{ij}^2, \boldsymbol{\zeta}, \sigma_{\theta_j} &\sim \mathcal{N}(\mathbf{x}_{ij}^t \boldsymbol{\beta}_i + \mathbf{s}_{ij}^t \boldsymbol{\zeta}, \sigma_{\theta_j}) \\ \boldsymbol{\beta}_i \mid \mathbf{w}_i, \boldsymbol{\gamma}, \mathbf{T} &\sim \mathcal{N}(\mathbf{w}_i^t \boldsymbol{\gamma}, \mathbf{T}). \end{aligned} \tag{5.13}$$

For the item parameters, a growth model can be specified as well, similar to the model in Equation 3.10. If $\mathbf{v}_j$ denotes a vector of explanatory information at occasion j, for example time passed since the first occasion, and $\boldsymbol{\delta}_k$ a vector of item-specific coefficients predicting the occasion-specific item parameters $\tilde{b}_{kj}$:

$$\tilde{b}_{kj} \mid b_k, \mathbf{v}_j, \boldsymbol{\delta}_k, \sigma^2_{b_k} \quad \sim \quad \mathcal{N}(b_k + \mathbf{v}_j\boldsymbol{\delta}_k, \sigma^2_{b_k}). \tag{5.14}$$

The possibility to add explanatory information provides opportunities to find out not only whether but also why parameters differ between groups, or occasions.

5.3 Identification of multi-group IRT models

Each measurement model containing both latent scores and thresholds has an identification problem. Several combinations of item parameters and latent variable values result in identical likelihood values, complicating parameter estimation. In single group settings, this identification problem is generally solved by fixing the latent variable to have a mean of zero. As the mean of the scale is arbitrary, this restriction has no implications for the interpretation of the model.

In a multiple group setting, however, this identification problem exists in each group. For the 1PNO IRT model:

$$P(Y_{ijk} = 1 \mid \theta_{ij}, \tilde{b}_{kj}) = \Phi(\theta_{ij} - b_{kj}) = \Phi((\mu_{\theta_j} + e_{ij}) - (\mu_b + e_{kj})).$$

Within the term $((\mu_{\theta_j} + e_{ij}) - (\mu_b + e_{kj}))$, there is a trade-off between the latent mean μ_{θ_j} for a group and the mean threshold value $\sum_k \tilde{b}_{kj}/N$ within a group. A common shift of μ_{θ_j} and $\sum_k \tilde{b}_{kj}/N$ in opposite directions would lead to the same values of $\theta_{ij} - b_{kj}$. As an example: When students from one group are more likely to give the correct answer to all test items than students from another group, this can be the result of a higher mean ability level in this group, but it can also be the result of all the test items being easier for this group.

In addition, a separate model with group-specific item and person parameters is estimated within each group. To estimate parameters on the same scale, the parameters of the groups have to be linked in such a way that the relation between the two scales is specified.

There are several ways in which the multi-group 1PNO IRT model can be identified. First, the scale has to be identified for one group by fixing at least one group-specific parameter. This can either be the group mean ($\mu_{\theta_1} = 0$) of the person parameters, the sum of the item thresholds for that group ($\sum_k \tilde{b}_{k1} = 0$), or one (or more) of the group-specific threshold parameters (e.g. $\tilde{b}_{k1} = 0$). Once the scale for one group has been identified, the scales of the other groups can be linked as well as identified by defining either at least one group-specific item parameter (e.g. $\tilde{b}_{k1} = \tilde{b}_{kj}$), the sum of the item parameters within the group ($\sum_k \tilde{b}_{k1} = \sum_k \tilde{b}_{kj}$), or the person parameter mean ($\mu_{\theta_1} = \mu_{\theta_j}$) to be equal to that of the first group.

Each combination of identification and linkage rules lead to models which are equally likely based on the data. Different linkage rules can lead to different parameter estimates though, and as a result the variance or invariance of the item

parameters over groups depends on the chosen way of linking the scales (unless the parameters are spaced in such a way that they are in agreement with all rules). The influence of the choice of linkage becomes smaller with more items, more groups, and more invariant items, however.

The way in which the scales are linked reflects assumptions about the data. Most multi-group measurement models identify the scale for a reference group by restricting the person parameter mean, and then link the scales of the other groups to the scale of the reference group through one or more anchor items. The anchor items will determine at which point the group scales are linked. This reflects the assumption that there are items with exactly the same item parameters, which are known beforehand. In the Bayesian IRT models, the mean of the threshold parameters is usually restricted to zero in one group, and the threshold parameter means in all groups are restricted to be equal, linking the scales at equal item parameter means. This way of linking the scales reflects the assumption that the measurement instrument has the same overall level for all groups. When (some) items in a test are more difficult for a certain group, the mean ability estimate for this group will be lower.

Restricting the person parameter mean and variance for the reference group and linking with anchor items is easy to implement, especially in a traditional maximum likelihood estimation procedure. There are some limitations to this way of model identification, though. Restricting the mean and variance for one group leaves no room for models on the person parameters for the measured construct, like explanatory covariates or multilevel (longitudinal) structures. Furthermore, restricting group means or variances arbitrarily can result in numerical problems when the parameters for this group are extremes on the scale. In addition, if the wrong items are chosen as anchor items, or if none of the items is invariant, this causes bias in the estimated latent scores and in the estimated group differences. In a Bayesian framework, restricting individual parameters to fixed values can lead to complex conditional prior specifications. Therefore, the restriction on the sum of the thresholds is more convenient. This leaves the variance components free to be estimated in a very flexible modeling framework. Explanatory models on both the item parameters and the latent group means are easily implemented.

When linking the scales of the groups using the restriction of equal mean threshold parameters, anchor items do not need to be known beforehand. However, anchor items can be used to link the scales in Bayesian IRT models as well, for example when anchor items are known, or when latent classes of anchor items are estimated. In this case, anchor items can either be used as an additional linkage rule, or to replace the linkage by equal item parameter means.

5.4 Bayesian estimation

Multi-group IRT models can be estimated in several ways. Traditionally, maximum likelihood based methods have been used to estimate multi-group IRT models. Recent methods usually consider the item parameters as fixed effects and the latent trait parameters θ as random effects. In this case, the item parameters can be estimated by maximizing the marginal maximum likelihood (with θ integrated out) using for example an Expectation Maximization (EM) algorithm (Bock &

Aitkin, 1981; Bock & Lieberman, 1970).

In the Bayesian approach to statistics, all parameters λ are assumed to be random variables with probability distributions. The posterior distributions of the parameters given the data, $p(\lambda \mid y)$ can be found using Bayes' rule to combine the likelihood $p(y \mid \lambda)$ and a prior distribution for the parameters $p(\lambda)$. According to this rule,

$$p(\lambda \mid y) = \frac{p(y \mid \lambda)p(\lambda)}{p(y)} \tag{5.15}$$

$$\propto p(y \mid \lambda)p(\lambda) \tag{5.16}$$

Since $p(y)$, the marginal probability of y, does not depend on θ, in the estimation of the posterior distribution of the parameters θ it is convenient to use the unnormalized posterior density function in Equation 5.16.

The prior distribution $p(\lambda)$ can contain prior information about the parameters, but can also be chosen rather non-informative. When the data contain a lot of information about the parameters in the data, the prior will be of minor influence, and the posterior distribution will be tight and close to the maximum likelihood estimate.

The posterior distribution of the parameters is often hard to compute analytically. Therefore, Markov Chain Monte Carlo (MCMC) methods are often used to approximate the posterior distribution. In the MCMC chains, parameters are sequentially drawn from their conditional posterior distributions given the other parameters. When conjugate priors are used, most of the conditional posterior distributions given the other parameters are known, in which case a Gibbs sampler (Geman & Geman, 1984) is used to draw samples from the conditional posterior distributions directly. When the priors are not conjugate, a Metropolis-Hastings step (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953) can be used, in which values from a proposal density around the previously sampled value are drawn. The new values are accepted when the likelihood of the new value is higher than that of the previous value, or with a probability based on the ratio of the likelihoods of the new and previous value. In the long run, the draws will converge to samples from the posterior distribution, and can be used to estimate for example the posterior mean and variance of a parameter, which can be used as point estimates and standard errors for the parameters in the model. More about Bayesian data analysis can be found in for example Gelman, Carlin, Stern and Rubin (2004).

5.5 Bayes Factors for nested models

In the practice of testing, it is often relevant to know which items are and which items are not invariant. To accomplish this, the hypothesis that the item parameters are equal in all groups should be tested. In this section, Bayes factor tests are introduced to test the equality of item parameters in fixed groups, and to test whether there is variance of item parameters over random groups.

In a Bayesian framework, hypotheses are ideally compared using Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor is the ratio of the support

for two models given the data and the prior information, expressed as the ratio of the marginal likelihoods. Bayes factors are not always easy to compute, but in case of nested models, it can be shown that the Bayes Factor in favor of the null hypothesis reduces to the ratio of the density or density region of the null hypothesis under the posterior and prior distribution of the most complex model. When the null hypothesis is a point hypothesis, this can be accomplished by computing the Savage Dickey density ratio of the density at the null hypothesis under the prior and posterior distribution (see Dickey, 1971; Verdinelli & Wasserman, 1995). To evaluate the relative support for the null hypothesis that the difference between the item parameters of two groups $d = b_{k1} - b_{k2}$ is equal to zero over the alternative hypothesis that there is a difference, the following Bayes factor can be defined:

$$BF_{01} = \frac{p(Y \mid H_0)}{p(Y \mid H_1)} = \frac{p(d = 0 \mid H_1, Y)}{p(d = 0 \mid H_1)} \tag{5.17}$$

Another possibility is that the null hypothesis concerns a region. For the null hypothesis that the variance in item parameters over groups is equal to zero, the point zero is not included in the parameter space for the variance parameters. In this case, it is possible to define the null hypothesis based on an "about equality" constraint δ, which corresponds closely with the original hypothesis when a very small value for δ is chosen (Klugkist and Hoijtink, 2007; Klugkist, 2008). Assuming both models share the same conditional distribution of observed data and the parameter space associated with the prior under the null hypothesis is a subset of the parameter space associated with the prior under the alternative hypothesis, the Bayes factor in favor of the null hypothesis $H_0 : \sigma^2_{b_k} < \delta$ can be expressed as the ratio of the density region containing the null hypothesis under the prior and posterior distribution (Geweke, 2005 ; Hoijtink, 2011; Klugkist and Hoijtink, 2007; Klugkist, 2008). The Savage Dickey density ratio is a special case (Wetzels, Grasman & Wagenmakers, 2010). To test the null hypothesis that the variance of item parameters over random groups $\sigma^2_{b_k}$ is equal to zero, the following Bayes factor can be computed:

$$BF_{01} = \frac{p(Y \mid H_0)}{p(Y \mid H_1)} = \frac{p(\sigma^2_{b_k} < \delta \mid H_1, Y)}{p(\sigma^2_{b_k} < \delta \mid H_1)} \tag{5.18}$$

This test of variance components can also be performed on the residual variance when covariates are added to the model (see section 5.2.2), to assess whether there is residual unexplained variance in item parameters.

Within an MCMC sampling scheme, there are several ways to compute these Bayes factors for nested models. One relatively easy computation method is to sample from both the prior and posterior distribution under the most complex model, and then compute the density or density region at the null hypothesis under both models, for example using WinBUGS (Lunn et al., 2000) output and R (Wagenmakers, Lodewyckx, Kuriyal & Grasman, 2010; Wetzels, Raaijmakers, Jakab & Wagenmakers, 2009) (Appendix (H).

The interpretation of the Bayes factor tests for measurement invariance is straightforward. The Bayes factor indicates how much more likely the data are to have occurred given the null hypothesis than given the alternative hypothesis. The conclusion will contain whether there is convincing support in favor of the null or the alternative hypothesis. In the results section of this chapter, the categorization proposed by Jeffreys (1961) will be used to decide whether there is substantial evidence for either hypothesis. A Bayes factor (BF_{01}) larger than three, implying that the data are three times more likely to have occurred under H_0 than under H_1, will be considered as substantial support for the null hypothesis H_0. A Bayes factor smaller than .33 will be considered as substantial support for the alternative hypothesis H_1 ($BF_{10} = 1/BF_{01}$ is larger than 3).

The density or density region at the null hypothesis, and therefore the result of the Bayes factor test, depend on the specific prior chosen. However, for the models under review here, priors can be chosen which reflect reasonable assumptions about the parameter values evaluated by the Bayes factor (Appendix F). The result of the Bayes factor test for variance components also depends on the specified "about equality" constraint δ. It is possible, however, to evaluate a range of "about equality" constraints, or to pick a constraint which represents a negligible difference on the scale at hand.

5.6 Results

Two simulation studies and two empirical data sets will illustrate and evaluate the models and tests described in this chapter. The simulation studies have the goal to evaluate Bayes factor tests for measurement invariance using a procedure in R based on WinBUGS output (Wagenmakers et al., 2010; Wetzels et al., 2009) (Appendix (H) (Appendix H). The first simulation study evaluates the Bayes factor for item parameter differences within the fixed manifest groups model. The second simulation study evaluates the Bayes factor test of the item parameter variance over groups within the random manifest groups model. A comparison will be made with a likelihood ratio test (Thissen, Steinberg & Wainer, 1993) procedure in IRTPRO (Cai, Thissen & du Toit, 2011), in which all other items are considered anchor items while testing each item for invariance.

To illustrate the use of the described models in real test situations, two data sets are analysed. The first data set consists of geometry items from the College Basic Academic Subjects Examination (CBASE) for males and females. Results from the Bayesian IRT model for fixed manifest groups and the model with additional latent classes for non-invariant and invariant items will be illustrated and compared to estimation results from a maximum likelihood based method restrictions (EM algorithm, Bock & Aitkin, 1981) with anchor item and reference group restrictions as implemented in IRTPRO (Cai, Thissen & du Toit, 2011). The second data set consists of items from the Survey of Health, Ageing and Retirement in Europe (SHARE) depression questionnaire for 12 European countries. For this example as well, results from the Bayesian IRT model for random manifest groups and the model with additional latent groups for non-invariant and invariant items will be illustrated and compared to maximum likelihood based results.

5.6.1 Simulation study 1: Evaluation of the Bayes factor test for item parameter differences in the fixed manifest group model

The first simulation study evaluated a Bayes factor for the difference between item parameters, implemented for the fixed manifest group model. The Bayes factor (BF_{01}) compares the null hypothesis of invariance $H_0 : d = 0$ to the alternative hypothesis that there is a difference between the item parameters of two groups $H_1 : d \neq 0$.

Data were generated consistent with the assumptions for both the Bayesian IRT model and the maximum likelihood estimation procedure: for each group j, the sum of the item thresholds $\sum_k b_{kj}$ equaled zero, the mean for the reference group μ_{θ_j} equaled zero and two (anchor) items were invariant. Two groups consisting of 1500 subjects answering 10 items were generated, where five pairs of items showed an increasing amount of difference d between the item parameters $(0, .1, .3, .5$ and $.7)$. The combined results of the analysis of 50 simulated data sets are presented in Table 5.2.

Table 5.2: Results for both the Bayes factor test for item parameter differences and the likelihood ratio chi squared test, and average estimation results (average parameter estimates, BIAS, MSE) for ten items with five varying amounts of DIF over 50 replicated data sets with two groups

BF_{01}						
d	$\overline{BF_{01}}$	$\%BF_{01} > 3$	$\%BF_{01} < .33$	EAP d	BIAS b_{kj}	MSE b_{kj}
0.00	5.63	0.84	0.00	0.01	0.06	0.01
0.10	3.74	0.59	0.06	0.10	0.06	0.01
0.30	0.14	0.01	0.90	0.31	0.06	0.01
0.50	0.00	0.00	1.00	0.50	0.06	0.01
0.70	0.00	0.00	1.00	0.69	0.07	0.01
χ^2						
d	$\overline{p}$		$\%p < .05$	$\hat{d}$	BIAS b_{kj}	MSE b_{kj}
0.00	0.54		0.01	0.01	0.07	0.02
0.10	0.35		0.17	0.09	0.08	0.02
0.30	0.01		0.95	0.30	0.07	0.02
0.50	0.00		1.00	0.47	0.08	0.02
0.70	0.00		1.00	0.65	0.09	0.03

In the first part of Table 5.2, results are shown from the fixed manifest group Bayesian IRT model, as described in section 5.2.1, with Bayes factors testing the difference between item parameters of the two groups. In the second part of the table, results acquired with the maximum likelihood based EM algorithm in IRTPRO (Cai, Thissen & du Toit, 2011) are shown, using a likelihood ratio (LR) chi-squared test to evaluate each item for invariance, with all other items as anchors.

The first three columns show the test results. For the *invariant items* (d =0)

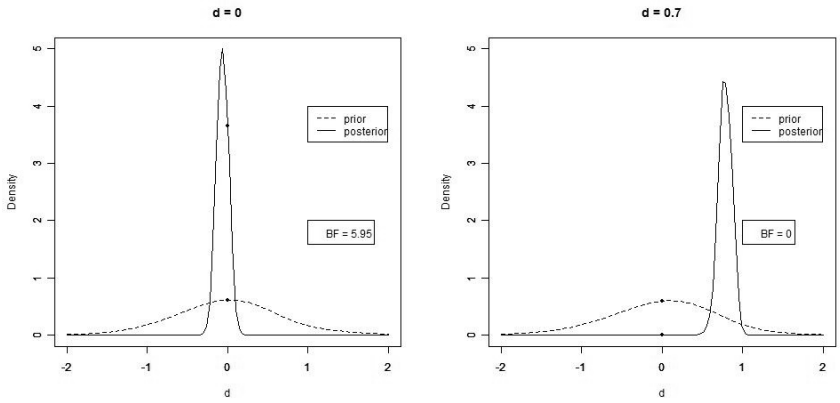

Figure 5.1: Illustration of the Bayes factor test for the difference in item parameters d. The dots indicate the density of the prior (dotted line) and posterior (solid line) at the null hypothesis $d = 0$, for an item with no difference $d = 0$ and for an item with a large difference $d = .7$ between the item parameters of two groups.

the average Bayes factor over all data sets and items was 5.63. For 84 percent of the invariant items, there was substantial evidence for the null hypothesis of invariance, as indicated by a Bayes factor larger than three. For none of the invariant items there was substantial evidence for the alternative hypothesis indicated by a Bayes factor lower than .33. This illustrates how the Bayes factor can be used to compare support for both hypotheses, instead of focusing only on rejection of the null hypothesis. The likelihood ratio (LR) chi-squared test resulted in a mean p-value for the invariant items of .54. Assuming that the null hypothesis of invariance will be rejected at a p-value lower than .05, only one percent of the invariant items was incorrectly designated as non-invariant.

For the items with a *small* difference of ($d = .1$) between the item parameters, the average Bayes factor was 3.74. For 59 percent of the items there was substantial evidence for the null hypothesis of invariance ($BF_{01} > 3$), and for 6 percent of the items there was substantial evidence for the alternative hypothesis ($BF_{01} < .33$). This left 35 percent of these items with no substantial evidence for either hypothesis. The likelihood ratio (LR) test rejected the null hypothesis of invariance for 17 percent of the items, with an average p-value of .35.

For the items with *larger* differences ($d \geq .3$) between the item parameters, the average Bayes factors were all below .33. Variance in item parameters ($BF_{01} < .33$) was correctly supported for 90 percent of the items with $d = .3$ and for all items with a larger difference. For one percent of the items with $d = .3$, the alternative hypothesis was incorrectly supported ($BF > 3$). The likelihood ratio (LR) chi-squared test showed a similar pattern, rejecting the null hypothesis of invariance for 95 percent of the items with a .3 difference, with an average p-value of .01, and for all the items with a larger difference, with an average p-value smaller than .01.

The bias and mean squared error, based on a comparison of the simulated and estimated item parameters, were slightly larger for the maximum likelihood based

method. As the likelihood ratio test can only reject invariance but not support variance, and the Bayes factor is the ratio of the amount of support for either hypothesis, the results should be compared with caution. Both tests perform perfectly in identifying large differences ($d \geq .5$) in item parameters. The Bayes factor is more conservative when decisions would be made based on substantial evidence for H_1, but less conservative when decisions would be made based on whether or not there is substantial evidence for H_0. For the smaller differences ($d = .1$, $d = .3$), the Bayes factor is undecided for a percentage of items, which is desirable for small differences. Overall the tests produced practically similar results, given that the assumptions were met for all models. An advantage of the Bayes factor is that it gives more detailed information about the support for both hypotheses. In addition, the models within which the Bayes factor can be applied have more possibilities for extensions, for example by including models on the person parameters.

Figure 5.1 gives an illustration of the way the Bayes factor test for a difference in item parameters works. In the figure on the left (for the situation in which $d = 0$), the density of the posterior distribution (solid line) at the null hypothesis $d = 0$ is higher than the density of the prior distribution (dotted line) at this point. The Bayes factor here is 5.95, indicating substantially more evidence for H_0 than H_1. In the figure on the right, where $d = .7$, the prior density at the null hypothesis is higher than the posterior density. The Bayes factor less than .001 indicates substantially more support for the alternative hypothesis H_1 than for H_0.

5.6.2 Simulation study 2: Evaluation of the Bayes factor test for variance components in the random manifest groups model

The second simulation study evaluated the performance of the Bayes factor test for the item parameter variance components, implemented for the random group Bayesian IRT model as described in Section 5.8. Data were generated consistent with the assumptions that the item difficulties summed to zero in all groups, that the mean for the reference group was zero, and that there were two anchor items. Five groups with 500 persons were simulated, answering two items with no DIF, and pairs of items with increasing variance over the five groups ($\sigma^2_{kj} = 0, .2, .11, .28, .56$), resulting in ten items in total. The combined results of the analysis of 50 data sets are presented in Table 5.3.

The likelihood ratio (LR) chi-squared test in IRTPRO (Cai, Thissen & du Toit, 2011) tests the equality of item parameters over groups with Helmert contrasts, comparing each group-specific parameter to the mean of subsequent group specific parameters. For this study, it will be assumed that the null hypothesis of invariance is rejected for an item when the minimum p-value of the four contrast tests is below .001, using a Bonferroni correction on the p-value for the number of contrasts for the full data set. Therefore, in Table 5.3 the mean minimum p-value over the 50 data sets and the percentage of minimum p-values under .001 are given.

The Bayes factor test found substantial evidence for invariance ($BF_{01} > 3$) for 94 percent of the *invariant* items ($\sigma^2 = 0$), with an average Bayes factor of 5.26.

Table 5.3: Simulation results for the Bayes factor test for variance components and the likelihood ratio chi-squared test, and average estimation results (BIAS, MSE) for ten items with five increasing item parameter variances over 50 replicated data sets.

BF					
$\sigma^2_{b_k}$	$\overline{BF}$	$\%BF > 3$	$\%BF < .3$	BIAS b_{kj}	MSE b_{kj}
0.00	5.26	0.94	0.00	0.07	0.01
0.02	3.29	0.57	0.00	0.07	0.01
0.11	0.53	0.01	0.48	0.08	0.01
0.28	0.07	0.00	0.98	0.08	0.01
0.53	0.02	0.00	1.00	0.09	0.01
χ^2					
$\sigma^2_{b_k}$	$\overline{min(p)}$		$\%min(p) < .001$	BIAS b_{kj}	MSE b_{kj}
0.00	0.20		0.00	0.08	0.03
0.02	0.06		0.08	0.08	0.02
0.11	0.01		0.88	0.10	0.03
0.28	0.00		1.00	0.08	0.03
0.53	0.00		1.00	0.10	0.03

For none of the items substantial evidence for variance ($BF_{01} < .3$) was found. The likelihood ratio test resulted in an average minimum p-value of .20 over the four contrasts. For three percent of the invariant items the null hypothesis of invariance was rejected. For 57 percent of the items with a *small* variance over the item parameters ($\sigma^2 = 0.02$), there was substantial evidence for the invariance hypothesis ($BF_{01} > 3$), while for the remaining 43 percent there was no clear support for either hypothesis. The likelihood ratio test rejected the invariance hypothesis for eight percent of the items. For the items with slightly *larger* variance ($\sigma^2 = 0.11$) over the item parameters, the Bayes factor test found substantial evidence for the alternative hypothesis for 48 percent of the items ($BF_{01} > 3$), while for one percent the invariance hypothesis was supported ($BF_{01} < .3$), and for the remaining half of the items substantial evidence for none of the hypotheses was found. The likelihood ratio test rejected invariance for 88 percent of those items. For all the items with *large* item parameter variances higher than .28, the Bayes factor was smaller than .3, correctly supporting the alternative hypothesis of variant item parameters. The likelihood ratio test also rejected the null hypothesis of invariance for all these items.

The bias and mean squared error, based on a comparison of the simulated and estimated item parameters, were again slightly larger for the maximum likelihood based method. As the likelihood ratio test can only reject invariance but not support variance, comparisons between the two testing procedures should be made with caution. The items with large variance were correctly indicated as non-invariant by both methods for almost all of the items. Again, the Bayes factor is more conservative when decisions would be made based on substantial evidence for H_1, but less conservative when decisions would be made based on whether

or not there is substantial evidence for H_0. For the smaller variances, the Bayes factor leaves the situation more often undecided. Both methods performed well in identifying invariant items, although the Bayes factor did not actively support invariance for six percent of the items. In sum, the tests produced practically similar results, given that the assumptions were met for both procedures. The combined interpretation of all contrasts in the likelihood ratio test is inconvenient, though, and the conditions for rejecting invariance depend on the chosen α levels. An additional advantage of the Bayes factor for testing variance components is that it can also be used to test residual variance components when explanatory information is added on the item parameters (Section 5.2.2), and that it can be used within models which include information on person parameters.

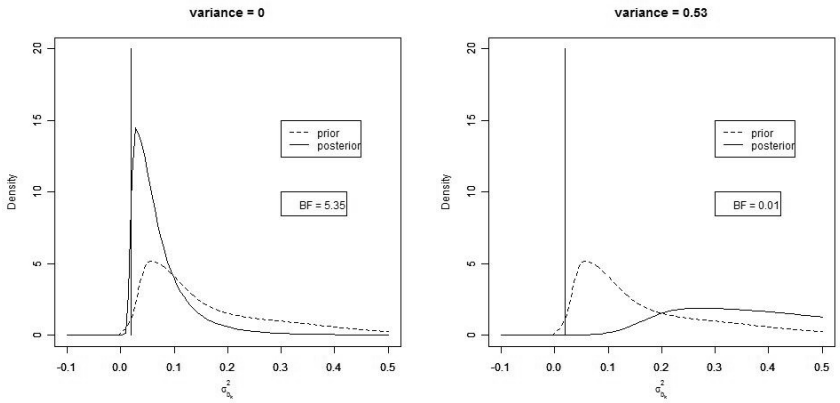

Figure 5.2: Illustration of the Bayes factor test for the item parameter variance over groups $\sigma^2_{b_k}$. The areas to the left of the vertical line under the posterior (solid line) and the prior (dotted line) distributions indicate the density regions containing the null hypothesis $\sigma^2 < .02$ for an item with no variance $\sigma^2_{b_k} = 0$ and for an item with a large variance $\sigma^2_{b_k} = .53$ of the item parameters over groups.

Figure 5.2 gives an illustration of the Bayes factor test for variance components. On the left, for the situation in which $\sigma^2 = 0$, the density region of the posterior distribution (solid line) containing the null hypothesis ($\sigma^2 < \delta$) is larger than the density region of the posterior distribution (dotted line) containing H_0. The Bayes factor here is 5.35, indicating substantially more evidence for H_0 than H_1. On the right, where $\sigma^2 = .56$, the density region under the prior density containing the null hypothesis is larger than under the posterior density. The Bayes factor of 0.01 indicates substantially more support for the alternative hypothesis H_1 than for H_0.

5.6.3 Empirical example 1: Geometry items for males and females (CBASE)

CBASE is an exam intended for students enrolled in college, assessing knowledge and skills in mathematics, English, science and social studies. The analysis will

focus on measurement invariance for 11 items from the geometry subtest of the mathematics test, comparing females ($N = 4452$) and males ($N = 1034$). First, Bayesian IRT models for fixed groups are estimated, first without and then with latent classes for the item parameters. The results will be compared to results from a traditional analysis based on maximum likelihood estimates, in which the scales are linked with anchor items instead of by equal average thresholds in both groups.

Bayesian IRT models

To illustrate models for fixed groups as described in section 5.2.1 both the basic Bayesian IRT model for fixed groups and the model with additional latent classes for invariant and non-invariant item parameters were estimated.

For the basic fixed groups Bayesian IRT model, after 5000 iterations, with a burn in of 500 iterations, convergence of the MCMC chains was reached, as indicated all lag 50 autocorrelations were below .1 and all Geweke Z statistics (Cowles & Carlin, 1996) below 2.

Table 5.4 shows the results for this model. The Bayes factor identifies item 1, 6, 7 and 11 as non-invariant items, whereas for item 4 and 9 there is substantial evidence for invariance. Item 1, 7 and 11 are indicated to be easier for male students, while item 6 is indicated to be easier for female students. Figure 5.3 illustrates the Bayes factor test for item 4 and 7.

Table 5.4: Item parameter estimates, results for the Bayes factor test for item parameter differences and posterior class probabilities for the non-invariant latent item class (CBASE example).

	Basic model, Bayes factor				Latent item class model			
	b_k M	b_k F	DIF	BF_{01}	b_k M	b_k F	DIF	$p(c_k = 1 \mid Y)$
item 1	-0.27	-0.03	-0.24	0.16	-0.25	0.00	-0.26	0.90
item 2	0.78	0.67	0.11	1.99	0.75	0.72	0.02	0.28
item 3	-0.66	-0.77	0.11	2.78	-0.69	-0.72	0.03	0.33
item 4	0.68	0.63	0.05	5.32	0.68	0.67	0.01	0.20
item 5	-1.09	-1.25	0.16	1.30	-1.11	-1.20	0.10	0.68
item 6	0.46	0.16	0.30	0.02	0.45	0.21	0.25	0.88
item 7	0.85	1.21	-0.36	0.01	0.86	1.25	-0.39	1.00
item 8	1.11	1.28	-0.17	0.59	1.17	1.31	-0.15	0.69
item 9	-0.61	-0.68	0.06	4.20	-0.62	-0.63	0.01	0.22
item 10	-0.87	-1.06	0.19	0.58	-0.90	-1.01	0.11	0.60
item 11	-0.38	-0.17	-0.21	0.25	-0.33	-0.13	-0.19	0.77
mean μ_{θ_j}	1.36	0.62			1.36	0.66		
sd σ_{θ_j}	1.47	1.28			2.19	1.64		

A second Bayesian IRT model which can be estimated on these data is the fixed manifest groups model in which the item parameters are classified as invariant or

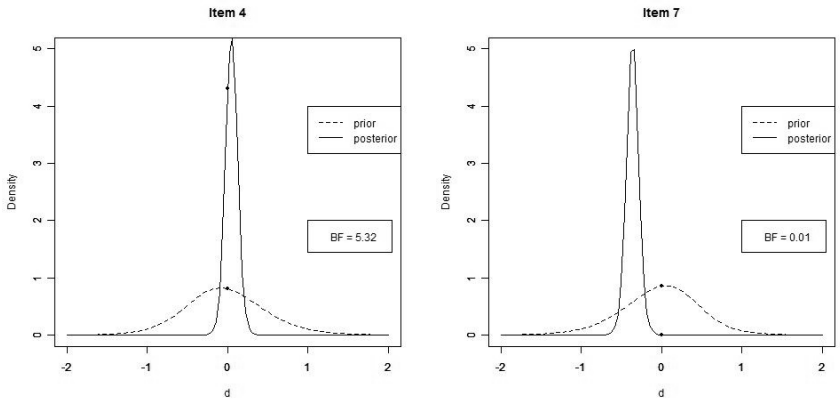

Figure 5.3: Illustration of the Bayes factor test for the difference in item parameters for items 4 and 7. The dots indicate the density of the prior (dotted line) and posterior (solid line) at the null hypothesis $d = 0$.

non-invariant by latent classes (Section 5.2.1). Convergence for this latent class model is slower than for the previous model, but after 5000 iterations convergence was reached for the male group, as indicated by the autocorrelations and Geweke Z statistic (Cowles & Carlin, 1996). The convergence patterns for the female group parameters are dependent on whether the items are sampled as invariant or non-invariant in each iteration, and hence can not be evaluated by standard convergence criteria.

The results are given in Table 5.4. The items for which invariance was supported by the Bayes factor (4 and 9) are also the items which have a high posterior probability to be assigned to the invariant class, while the items with the largest difference between item parameters (1, 6, 7, 11) are the items with the highest posterior probability to be assigned to the non-invariant item class.

The estimated item parameters for the two models are very similar. The parameter estimates of males and females are closer together in the latent class model, especially for items which have a high posterior probability of belonging to the invariant class. As these items are restricted to be anchor items in many of the iterations, there is an additional linkage between the two scales. In general the female item parameters shift a bit downward, which is compensated by an upward shift of the person parameter mean for the females.

Comparison to an ML based procedure with anchor item restriction

The same CBASE dataset was analyzed with standard maximum likelihood based procedures (EM algorithm, Bock & Aitkin, 1981) with anchor item and reference group restrictions as implemented in IRTPRO (Cai, Thissen & du Toit, 2011).

First, all items were tested for invariance, using all other items as anchors. The first two columns of Table 5.5 show the results of this likelihood ratio test. Even though the restrictions on which the parameters and parameter differences for these tests are based were different from the restrictions for the Bayes factor

Table 5.5: ML estimates, likelihood ratio test results and rescaled ML estimates for CBASE example

	LR test		ML estimates (anchors)			Rescaled parameters		
	χ^2(df)	p	b_{k1}	b_{k2}	DIF	b_{k1}	b_{k2}	DIF
item 1	5.20 (1)	0.02	-1.17	-0.95	-0.22	-0.28	-0.04	-0.24
item 2	2.30 (1)	0.13	-0.45	-0.45	0	0.72	0.67	0.05
item 3	1.20 (1)	0.28	-1.46	-1.46	0	-0.68	-0.76	0.09
item 4	0.70 (1)	0.40	-0.48	-0.48	0	0.68	0.63	0.05
item 5	1.90 (1)	0.17	-1.79	-1.79	0	-1.14	-1.23	0.10
item 6	12.20 (1)	0.00	-0.63	-0.81	0.18	0.47	0.16	0.31
item 7	13.50 (1)	0.00	-0.36	-0.07	-0.29	0.85	1.22	-0.37
item 8	2.40 (1)	0.12	-0.05	-0.05	0	1.28	1.24	0.04
item 9	0.50 (1)	0.47	-1.40	-1.40	0	-0.59	-0.68	0.08
item 10	3.20 (1)	0.07	-1.66	-1.66	0	-0.96	-1.05	0.09
item 11	3.90 (1)	0.05	-1.24	-1.04	-0.2	-0.37	-0.17	-0.21
mean μ_{θ_j}			0.00	-0.49				
sd σ_{θ_j}			1.06	0.90				

test in Table 5.4, the results are remarkably similar. For item 1, 6, 7 and 11, invariance is rejected, while for item 4 and 9 the p-values are the highest.

Next, the items for which invariance was not rejected were used as an anchor set to estimate item parameters for all items. Comparing these parameter estimates (in column 3-6 of Table 5.5) with the estimates from the Bayesian IRT model (Table 5.4), it is clear that the mean and variance of the scales are different, and as a result, the maximum likelihood item parameter estimates are higher and less spread out. This is a direct result of the identification restrictions: the sum of the item parameters was set to zero in the Bayesian IRT model, while the mean of the θ scale for males was set to zero in the maximum likelihood method .

The last columns of the table show the parameters rescaled to the scale of the Bayesian IRT estimates, by substracting within each group the mean difficulty from each group-specific item parameter, and multiplying by the ratio of the standard deviations of θ in that group. When compared to Table 5.4 the rescaled parameters are almost equal to the Bayesian IRT model estimates. This shows that the large amount of data dominates the posterior information and the priors are not influential.

The rescaled anchor items are not exactly equal for males and females anymore, however, and this is a direct consequence of the point at which the scales are linked, reflecting either a set of anchor items or an equal overall difficulty of the test. As there are many items and a relatively large amount of invariant items in this example, the invariance tests give exactly the same result.

Another difference is that the discrimination parameter is set to 1 in the Bayesian IRT models but is estimated in IRTPRO, resulting in the different variances

There are situations, however, in which different linkage rules can lead to different test results. In this case, in which there are often many non-invariant items, the equal overall difficulty restriction creates the possibility for an overall picture of differences in item parameters between groups independent of the chosen anchor items, and the possibility of including explanatory information directly into the model. When the aim is to identify and use anchor items, the average difficulty restricted Bayesian IRT models can be used as a base to start exploring which items are invariant. Parameters can then be restricted to invariance in a second step. In this example, items 4 and 9 are clearly indicated to be invariant, and could be used as anchor items in a second estimation round for Bayesian IRT models, using an anchor item restriction instead of or in addition to the equal average difficulty restriction (see Chapter 3 and 4).

5.6.4 Empirical example 2: SHARE depression questionnaire in 12 countries

To illustrate models for random groups as described in section 5.2.2, data from the Survey of Health, Ageing and Retirement in Europe (SHARE) held in 2004 were used. More than 45,000 individuals aged 50 or over, from 12 countries ranging from Scandinavia (Denmark and Sweden) through Central Europe (Austria, France, Germany, Switzerland, Belgium, and the Netherlands) to the Mediterranean (Spain, Italy and Greece) and Israel (collected in 2005) participated in this survey, financed by the European Union. Data on health, socio-economic status and social and family networks were collected, with the main goal of informing public policies.

As part of the survey, a depression questionnaire was administered (Appendix A). Of the 28813 complete cases, ranging from 962 to 3552 cases per country, one third per country was randomly selected for analysis to decrease estimation time.

First, the basic Bayesian IRT model for random groups was run. After 5000 iterations, with a burn in of 500 iterations, convergence of the MCMC chains was reached, as indicated all lag 50 autocorrelations were below .1 and all Geweke Z statistics (Cowles & Carlin, 1996) below 3. Then, a Bayesian IRT model for random groups with latent classes for invariant and non-invariant items was estimated for these data. Although convergence took longer than for the model without latent classes, the autocorrelations and Geweke Z statistics indicated that convergence was reached eventually. Finally, the data were analyzed with the maximum likelihood based method (EM algorithm, Bock & Aitkin, 1981) with anchor item and reference group restrictions as implemented in IRTPRO (Cai, Thissen & du Toit, 2011).

This paper uses data from SHARE 2004. The SHARE data collection was primarily funded by the European Commission through the Fifth Framework Program (project QLK6-CT-2001-00360 in the thematic program Quality of Life). Additional funding came from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01, and OGHA 04-064). SHARE data collection in Israel was funded by the U.S. National Institute on Aging (R21 AG025169), by the German-Israeli Foundation for Scientific Research and Development, and by the National Insurance Institute of Israel. The Israel Gerontological Data Center was established by a grant from the Ministry of Science, and is supported by the Ministry of Senior Citizens.

Table 5.6: SHARE depression questionnaire. General (b_k) and average ($\overline{b}_{kj}$) item parameter and item parameter variance estimates. Bayes factor (BF), number of contrasts smaller than α ($\#p < \alpha$) (LR test) and posterior class probabilities ($p(C_k = 1 \mid Y)$).

	Bayes factor test			LR test			Latent item classes		
	b_k	σ^2_{kj}	BF	$\overline{b_{kj}}$	$s^2_{b_{kj}}$	$\#p < \alpha$	b_k	$\sigma^2_{b_{kj}}$	$p(C_k = 1 \mid Y)$
1 Irritable	-0.59	0.11	0.20	1.36	0.06	1	-0.72	0.11	1.00
2 Fatigue	-1.06	0.08	0.55	0.98	0.03	1	-1.21	0.07	1.00
3 Tearfulness	-0.80	0.09	0.14	1.19	0.05	3	-0.93	0.10	1.00
4 Hopes	0.32	0.25	0.00	2.10	0.13	4	0.14	0.25	1.00
5 Suicidal	1.18	0.08	0.80	2.77	0.06	0	1.04	0.13	0.89
6 Sleep	-1.03	0.14	0.00	1.00	0.07	3	-1.19	0.14	1.00
7 Interest	0.89	0.08	1.02	2.57	0.05	0	0.75	0.09	0.93
8 Appetite	0.93	0.05	8.47	2.58	0.03	0	0.80	–	0.12
9 Concentration	0.36	0.10	0.33	2.13	0.05	3	0.25	0.11	1.00
10 Reading	0.12	0.10	0.23	1.94	0.06	2	-0.02	0.11	1.00
11 Enjoyment	0.27	0.14	0.01	2.04	0.07	3	0.15	0.16	1.00
12 Guilt	-0.51	0.33	0.00	1.42	0.19	6	-0.59	0.33	1.00

Table 5.6 shows the results of the Bayes factor test for variance in item parameters, the likelihood ratio tests considering all other items as anchors, and the Bayesian IRT model with latent classes for non-invariant and invariant items, as well as the general or average item parameter estimates and item parameter variance estimates for each item.

All methods indicate item 8 as invariant: with a Bayes factor larger than 3; by none of the contrasts being significant at a Bonferroni corrected α level of .0004; and by a probability of .12 of the item to belong to the non-invariant class. Item 5 and 7 are also indicated as invariant by the likelihood ratio test. The Bayes factor, although higher for item 5 and 7 than for the other items, does not provide substantial evidence for invariance for those items, and the posterior probability of belonging to the non-invariant class is higher than the posterior probability of belonging to the invariant class. The items indicated as varying over groups by the Bayes factor test ($BF < .33$) are 3, 4, 6, 9, 10, 11 and 12, the same items which also have 2 or more significant contrasts according to the likelihood ratio test. As in simulation study 2, the Bayes factor is more often undecided for items with small variance in item parameters. Item 5 and 7 would have been indicated as invariant by the likelihood ratio test, while item 1 and 2 would have been indicated as non-invariant. The posterior class probabilities are in agreement with the results from the Bayes factor test. Figure 5.4 illustrates the Bayes factor test for item 4 and 8.

Table 5.7 shows the item parameters for item 12 about feeling guilty, as estimated by the Bayesian IRT model for random groups and by the maximum

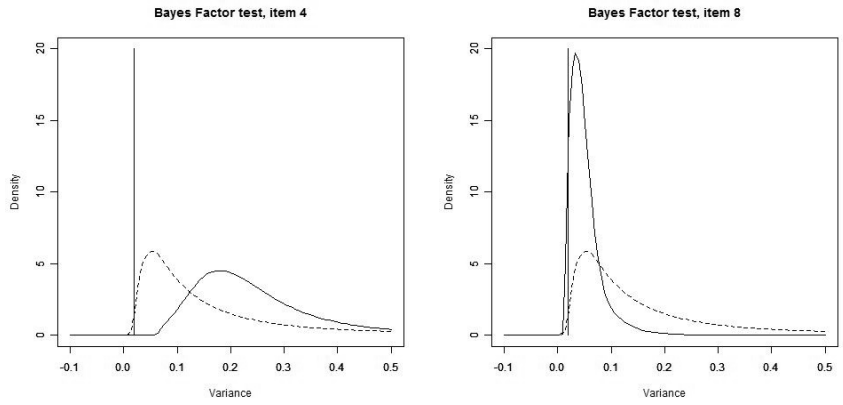

Figure 5.4: Illustration of the Bayes factor test for the item parameter variance over groups for items 4 and 8. The areas to the left of the vertical line under the posterior (solid line) and the prior (dotted line) distributions indicate the density regions containing the null hypothesis $\sigma^2 < .02$.

Table 5.7: Bayesian IRT model (BIRT), maximum likelihood (ML) and rescaled maximum likelihood item parameter estimates for item 12 per country (SHARE)

	BIRT b_{12j}	ML b_{12j}	Rescaled ML b_{12j}
Germany	-0.61	1.4	-0.62
Switzerland	-0.85	1.06	-0.9
Denmark	-1.27	0.91	-1.3
Austria	0.01	2.08	0.07
Netherlands	-0.58	1.27	-0.59
Sweden	-1.39	0.77	-1.4
Greece	-0.21	1.98	-0.2
Belgium	-0.47	1.62	-0.47
France	-0.72	1.36	-0.72
Spain	0.52	2.67	0.57
Israel	-0.72	1.37	-0.72
Italy	-0.14	1.69	-0.13

likelihood based procedure. The last column shows the maximum likelihood item parameter estimates rescaled to have a mean of zero and adjusted for the different variance of the scales. Again the rescaled parameter estimates are very similar to the estimates from the Bayesian IRT model, indicating that the large amount of data dominated the posterior information and the priors were not influential. The variance in item characteristic curves for item 12 is illustrated in Figure 5.5, where it is shown that in Sweden and Denmark on average a lower level of depression leads to feelings of guilt than in Norway and Spain.

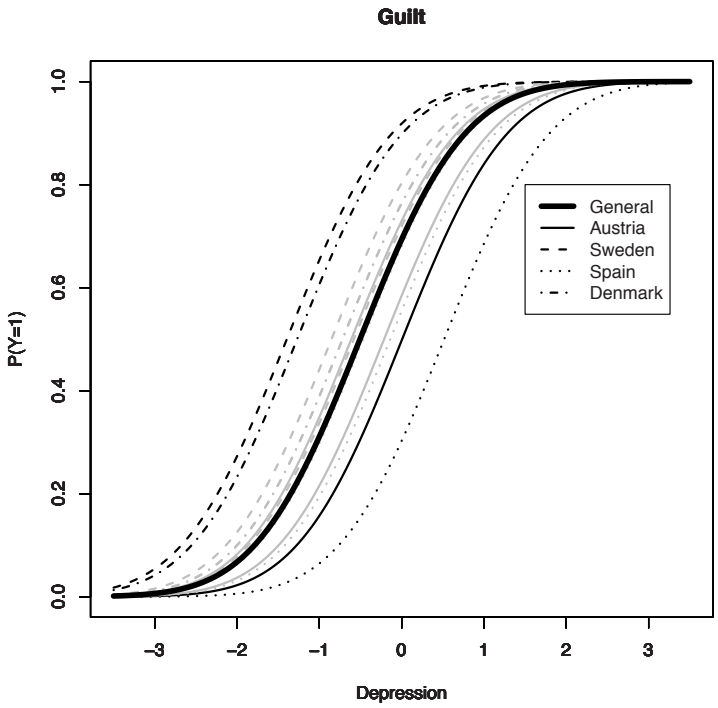

Figure 5.5: SHARE: Item characteristic curves for item 12.

As the item parameters for this item were shown to vary over countries, it would be interesting to add explanatory information about the countries to explain this variation (Section 5.2.2). The Bayes factor can then be used to evaluate whether there is residual variance over items.

5.7 Discussion

In this chapter, an overview was presented of Bayesian multi-group IRT models, and Bayes factors were introduced to test for measurement invariance.

It was shown that Bayesian IRT models create a flexible framework for modeling and testing for measurement variance, in which additional information is easily introduced. First, models were described for a small, fixed number of groups, with manifest or latent group structures. Then, models were described for a higher number of groups randomly sampled from a larger population. Possibilities for adding latent classes to classify items as invariant or non-invariant, and to add information to explain differences in group-specific parameters were introduced.

Two Bayes factor tests for invariance were introduced. The first Bayes factor test evaluated differences between the item parameters of a small number of fixed groups. The second Bayes factor evaluated the variance in item parameters over

groups for a large number of groups. Simulation studies evaluating both tests showed results similar to the results from a likelihood ratio (LR) chi-squared test. The results from the two testing procedures should be compared with caution, however, as the result of the likelihood ratio test comprises rejection of the null hypothesis or not, while the Bayes factor evaluates evidence for both the null and alternative hypothesis. The Bayes factor is more conservative than the likelihood ratio test when decisions would be made based on substantial evidence for H_1, but less conservative when decisions would be made based on whether or not there is substantial evidence for H_0. An advantage of the Bayes factor is that it gives a more differentiated view: there is substantial evidence for an item parameter to be invariant, non-invariant, or there is no substantial evidence for either hypothesis. Another advantage of the Bayes factor is that all parameters can be tested for invariance simultaneously without assuming anchor items. This is especially an advantage in situations with a large number of groups, as in this case many contrasts for the group-specific parameters need to be examined in a likelihood ratio testing procedure. In addition, the Bayes factor can be implemented in a wide range of Bayesian multi-group IRT models, evaluating invariance conditional on explanatory information about person and item parameter variance, where the likelihood ratio test can only evaluate invariance for basic multi-group IRT models.

The CBASE and SHARE examples illustrated the fixed and random group Bayesian IRT models, with and without latent item classes for invariant and non-invariant items. It was shown that the models with latent item classes performed well in identifying invariant items, as the results were similar to the results from the Bayes factor tests. The Bayesian IRT models produced the same (rescaled) estimates and approximately the same Bayes factor test results as the maximum likelihood based estimation and likelihood ratio tests for basic multi-group IRT models. This showed that, with data sets as large as the ones in both examples, the large amount of data dominates the posterior information and the priors are not influential.

Although the results of maximum likelihood and Bayesian estimation were very similar in the presented examples, it was also shown that different linkage rules can produce (small) differences in the estimated parameters.

Bayesian IRT models can be estimated with either linkage rule, which is not possible in the common maximum likelihood estimation procedures. The choice of linkage rule has to be governed by theoretical as well as practical arguments. If the aim is to acquire a better understanding of the diversity in both item responses, person parameters, and the factors associated with this variety, the equal threshold linkage offers the flexibility to incorporate all this information in a comprehensive framework. When the aim is to have as many invariant items in a measurement model as possible, or if the goal is to create a scale without DIF items, the anchor item linkage should be preferred. When no anchor items are known beforehand, a two-step procedure can be implemented. The equal threshold restriction in combination with the Bayes factor test can be used to indicate which items are most likely to be invariant without relying on anchor items. Then, in a second step, the anchor item restriction can be used to estimate the item parameters in the final model.

In sum, the Bayesian IRT modeling and testing framework provides many op-

portunities for investigating measurement variance in a wide variety of situations. The possibilities to extend the basic multi-group IRT model with explanatory covariates on person and item parameters and with latent classes makes it more flexible than traditional maximum likelihood estimation procedures. Bayes factor tests for invariance can be implemented within all these models, and give evidence for both the null and alternative hypothesis given the data.

Future research could extend the framework even further, with for example multi-dimensional IRT models or to measurement instruments with a mixed number of answer categories. Issues which have been encountered while investigating these models, like the prior sensitivity of the Bayes factors (Appendix F) and the effect of linkage rules under different conditions (the size of differences between parameters, the number of non-invariant items, the amount of groups) could be investigated in more detail.

Chapter 6

Discussion

Items in tests or questionnaires do not always measure the construct in the same way in all groups the measurement instrument is intended for. In this thesis, the use of Bayesian Item Response Theory models was explored for situations in which the measurement instrument does not function in the same way in all groups. Two aspects of models for measurement variance were investigated. On the one hand, tests have been developed to diagnose whether differences between groups in measurement instruments exist. On the other hand, models have been developed which take these differences into account to enable valid score comparisons and to gain insight into the nature of these differences.

This chapter will start with an overview of the models for measurement variance discussed in this thesis. Then, it briefly reviews the proposed tests for measurement invariance. The third section will provide some reflections regarding prior choice and linkage restrictions encountered during the thesis. The chapter closes with recommendations for future research.

6.1 The Bayesian IRT modeling framework

The first goal of this thesis was to develop models to take differences in item parameters into account, to enable valid score comparisons and to gain insight into the nature of these differences. The starting point for the thesis was the Bayesian multi-group IRT model with random item parameters, as proposed by Fox (2010), in which groups are assumed to be a random sample from a larger population of groups. This thesis showed that Bayesian IRT models form a very flexible framework for measurement variance modeling, as it is adaptable to a wide variety of testing situations.

Three extensions were made to incorporate explanatory information about group or occasion differences in person parameters, item parameters, or both into the Bayesian IRT models for random groups. Chapter 2 extended the model with the possibility to explain variation in person parameters with person and group-level covariates, while estimating group-specific item parameters. In this way, meaningful comparisons can be made across groups, including information on background variables, but without the need for invariant item parameters.

Chapter 3 showed that explanatory covariates can also be included to explain differences in the item parameters. It is possible to include covariates on the group level, covariates on the item level, and interactions of group level and item level covariates. This creates possibilities to investigate measurement variance in more detail, as the influence of group characteristics on item functioning can be studied. In chapter 4, Bayesian IRT models were developed for modeling joint growth over time in both person and item parameters. This provides the opportunity to account for and explain item parameter shifts over measurement occasions, while at the same time modeling and explaining growth in the measured construct.

Where the first two chapters focused on two-parameter models for dichotomous items, in chapter 4 an extension was made to a model for polytomous items. It was shown that shifts in the category thresholds of items with an ordinal rating scale can occur over time, in this example resulting in a higher probability of endorsing the middle category over time.

The last chapter 5 showed how the Bayesian IRT model framework can also include models for a smaller number of fixed instead of random groups. In these models, the item parameters within each group vary around a group-specific mean, with group-specific variance over items and covariance between the item parameters of the same item in different groups.

A last extension in chapter 5 introduced the possibility of adding latent classes to the models. Latent groups instead of manifest groups can be specified to identify groups with different response behavior. Another application of latent classes is to classify items as invariant or non-invariant. Examples showed that these models successfully classified items as invariant.

6.2 Bayesian tests for measurement invariance

The second goal of this thesis was to develop tests within the Bayesian IRT framework to diagnose whether differences between groups in measurement instruments exist. Three ways of testing for measurement invariance have been proposed.

Bayes factor tests for invariance, as introduced in chapter 3 and extended in chapter 5, can handle a set of multiple individual marginal invariance hypotheses concerning the invariance of each item parameter. The first Bayes factor test evaluated whether the variance of group-specific item parameters over groups was equal to zero for each item. After specifying the null hypothesis with an "about equality constraint", defining the invariance hypothesis as $\sigma^2_{b_k} < \delta$ (with δ a small value approaching zero) a Bayes factor for nested models was used to evaluate the support for this hypothesis given the data and the prior distribution. A second Bayes factor test evaluated whether the difference between the item parameters of two groups was zero for each item. A Bayes factor for nested models based on the Savage Dickey density ratio was implemented to test the null hypothesis of measurement invariance. Simulation studies in chapter 3 showed accurate detection of invariance and non-invariance in both discrimination and difficulty parameters, and a simulation study in chapter 5 showed accurate detection of invariance and non-invariance in threshold parameters, comparable to the performance of a likelihood ratio test (Thissen, Steinberg & Wainer, 1993) in IRTPRO (Cai, Thissen & du Toit, 2011). An advantage of Bayes factor tests over the likelihood ratio test

is that the amount of evidence for both hypotheses is indicated. In addition, all parameters can be tested for invariance simultaneously without assuming anchor items. This is especially an advantage in situations with a large number of groups, in which many contrasts need to be examined in the likelihood ratio testing procedure. Furthermore, the Bayes factor test can be implemented in a wider variety of models. This makes it possible, for example, to test for invariance conditional on explanatory covariates on the item parameters.

The use of the Deviance Information Criterion applied to test for measurement invariance in Bayesian IRT models was examined in chapter 3. The deviance information criterion provides an overall measure of fit for each estimated model, and was used to compare more general with restricted models to test measurement invariance hypotheses. The simulation study in chapter 3 showed that the DIC correctly favored the measurement non-invariant model. The examples in chapter 3 and 4 showed that the DIC favored the partial invariant models in which the items indicated as invariant by the Bayes factor were restricted to invariance. The DIC is therefore a useful tool for determining the final model with the best overall fit to the data.

In Appendix C, a highest posterior density (HPD) region test (Box and Tiao, 1973) for measurement invariance was investigated. This HPD region test evaluates whether the point where all country-specific item parameters are equal to the general item parameters is contained in the $(1 - \alpha)$ highest posterior density region of the joint posterior distribution of the country-specific item parameters (Appendix C). Unfortunately, this test for item parameter invariance had an inflated Type I error, as the point of equal means for all countries falls outside the HPD region rather quickly. This makes the test less useful, especially when the number of cases per country is large, as the within-group information becomes stronger and the HPD region tightens around the posterior mean.

A combination of the Bayes factor test for invariance of individual items with the DIC to evaluate overall model fit is therefore recommended to test the invariance of a measurement instrument.

6.3 Reflections on priors and linkage restrictions

6.3.1 Choice of priors

A common topic of debate surrounding Bayesian inference is the choice of prior distributions. With regard to the models discussed in this thesis, prior sensitivity tests (chapter 2) and a comparison with maximum likelihood estimation in chapter 5 showed that with a sufficiently large data set, estimates of especially the group-specific item parameters are not very sensitive to the choice of prior distribution.

For the computation of the Bayes factor tests for invariance, however, the ratio of density or density regions of the null hypothesis under the posterior and prior distributions is influenced by the specific prior distributions chosen. In addition, sufficiently large data sets are not always available.

Therefore, priors should be carefully chosen based on prior information. Prior information is often present, however, on the values the parameters are expected to take. For the models investigated in this thesis, the values the parameters can take

are restricted by the model, as the item and person parameters are estimated on a standard normal scale (see Appendix F). When the range of values the parameters are expected to take is known, it is often possible to choose a (conjugate) prior distribution which has a relatively uniform high density within this range, and low density outside this range. In this way, the influence of the prior distribution will be minimized.

6.3.2 Linkage restrictions

Another issue with multi-group IRT models concerns the linkage restrictions used to link the measurement scales of the groups. Linking the measurement scales of groups by one or more anchor items reflects the assumption that there are items with exactly the same item parameters in each group, and that these items are known beforehand. Linking the scales by restricting the average or the sum of the threshold parameters to be equal for each group reflects the assumption that items in the measurement instrument have the same average level or difficulty for all groups. In an educational setting, this would mean that when (some) items in a test are more difficult for a certain group, the estimated ability mean for this group will be lower. Both assumptions are plausible in some situations and have drawbacks in other situations.

As the scales are linked at a different point, choice for one or the other linkage rule can result in different parameter estimates and therefore can point to different items as non-invariant items. However, in situations with many (invariant) items and groups, the differences are often minor (chapter 5).

Both linkage rules can be used in the Bayesian IRT models described in this thesis. It is up to the researcher to determine which rule to choose, based on theoretical as well as practical arguments. If the aim is to acquire a better understanding of the diversity in both item responses and person parameters, and of the factors associated with this diversity, the equal threshold restriction offers the flexibility to incorporate all this information in a comprehensive model. When there is specific interest in cultural differences regarding the item parameters, for example, an equal threshold restriction can be used to enable explanatory information on the item parameters to be included in the model. Or, in a pilot study for a large cross-national survey, the equal threshold restriction can be used to enable exploration of the variance in item and person parameters, without assuming anchor items beforehand.

If the aim is to have as many invariant items as possible in a measurement instrument, or if the goal is to create a measurement instrument without non-invariant items, anchor items are the preferred way of linkage. Here, Bayesian IRT models provide an advantage over traditional models when there are no anchor items known beforehand. A two-step procedure can be implemented. First, the equal threshold restriction in combination with the Bayes factor test can be used to indicate which items are most likely to be invariant. Then, in a second step, the anchor item restriction can be used to estimate the item parameters in the final model.

6.4 Future directions

This thesis covered a wide array of Bayesian IRT models to investigate and model measurement variance and several tests for measurement invariance. Although the framework for Bayesian IRT models described in this thesis includes models for a large variety of testing situations, it is possible to extend the framework even further.

Models and tests have been developed for measurement instruments consisting of all dichotomous items, or consisting of polytomous items with the same number of answer categories. It would be useful to extend these models and tests to analyze measurement instruments with items with a mixed number of answer categories. The proposed models and tests could also be adapted to situations in which the measurement instrument measures several constructs, by including multi-dimensional multi-group IRT models to the framework, building on the work of De Jong and Steenkamp (2009).

Although the results for the Bayes factor tests for invariance with the chosen priors were good, the influence of different prior distributions for the variance and item parameters on the results of the Bayes factor tests for invariance could be investigated further. Some work in this direction has already been done by Mulder, Hoijtink and Klugkist (2010) for the about equality constraint and Wetzels, Grasman and Wagenmakers (in press) for the Savage Dickey density ratio.

In chapter 5, it was shown that the linkage rule chosen to link the scales of several groups can affect the resulting parameter estimates and the results of the measurement invariance tests. To get a better grasp of the influence of the specific linkage restrictions, the effect of linkage restrictions under different conditions such as the size of differences between parameters, the number of non-invariant items, and the amount of groups could be investigated, as well as the influence in IRT models with more than one item parameter.

A step has been made towards making the estimation of Bayesian IRT models and Bayes factor tests for invariance widely available by creating Winbugs and R code for the 1PNO model. Unfortunately, models with more than one item parameter are more difficult to specify in WinBUGS and take a long time to run. To make these Bayesian IRT models widely available as well, the Fortran code which was used for the 2PNO and GPCM models in chapter 3 and 4 of this thesis could be extended to incorporate the full range of Bayesian IRT models described in this thesis and be made more widely available, for example in an R package.

Appendix A

Questionnaires

A.1 Attitude towards immigration in the ESS (Ch. 3)

The attitude towards immigration scale consists of items about the perceived consequences and allowance of immigration items from the European Social Survey (ESS, 2002). The items were measured on 4 point (1,2) and 10 point (3-8) scales, but were dichotomized to enable the analysis in Chapter 3. The following items were used in the analysis:

1. To what extent do you think [country] should allow people of the same race or ethnic group as most people in [country] to come and live here? many–some(0) few–none(1)
2. To what extent do you think [country] should allow people from the poorer countries outside Europe to come and live here? many–some(0) few–none(1)
3. Is [country] made a worse or a better place to live by people coming to live here from other countries? better(0) worse(1)
4. Would you say it is generally bad or good for [country]s economy that people come to live here from other countries? good(0) bad(1)
5. Would you say that [country]s cultural life is generally undermined or enriched by people coming to live here from other countries? enriched(0) undermined(1)
6. Would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs? create(0) take(1)
7. Most people who come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out? put in(0) take out(1)
8. Are [country]s crime problems made worse or better by people coming to live here from other countries? better(0) worse(1)

A.2 CES-D depression questionnaire (Ch. 4)

The CES-D depression questionnaire (Radloff, 1977):

Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past week:

1. Rarely or none of the time (less than one day)
2. Some or a little of the time (1-2 days)
3. Occasionally or a moderate amount of time (3-4 days)
4. Most or all of the time (5-7 days)

During the past week:

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt that i was just as good as other people.
5. I had trouble keeping my mind on what i was doing.
6. I felt depressed.
7. I felt that everything i did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I talked less than usual.
13. I felt lonely.
14. People were unfriendly.
15. I enjoyed life.
16. I had crying spells.
17. I felt sad.
18. I felt that people dislike me.
19. I could not get "going".

Removed: I was happy.

A.3 SHARE depression questionnaire (Ch. 5)

Below, the items from the depression questionnaire from the Survey of Health, Ageing and Retirement in Europe (SHARE) are given.

This paper uses data from SHARE 2004. The SHARE data collection was primarily funded by the European Commission through the Fifth Framework Program (project QLK6-CT-2001-00360 in the thematic program Quality of Life). Additional funding came from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01, and OGHA 04-064). SHARE data collection in Israel was funded by the U.S. National

1. Have you been irritable recently?
2. In the last month, have you had too little energy to do the things you wanted to do?
3. In the last month, have you cried at all?
4. What are your hopes for the future? (any/no mentioned)
5. In the last month, have you felt that you would rather be dead?
6. Have you had trouble sleeping recently?
7. In the last month, what is your interest in things?
8. What has your appetite been like?
9. How is your concentration? (television/radio)
10. Can you concentrate on something you read?
11. What have you enjoyed doing recently? (any/no activity)
12. Do you tend to blame yourself or feel guilty about anything?

Institute on Aging (R21 AG025169), by the German-Israeli Foundation for Scientific Research and Development, and by the National Insurance Institute of Israel. The Israel Gerontological Data Center was established by a grant from the Ministry of Science, and is supported by the Ministry of Senior Citizens.

Appendix B

Computation of the Bayes factor (Chapter 3)

In a Bayesian framework, hypotheses are ideally compared using Bayes factors. The Bayes factor is the ratio of the support for two models given the data and the prior information, expressed as the ratio of the marginal likelihoods. In Chapter 3, a Bayes factor for nested models was described, where the null hypothesis $\sigma^2_{\xi_k} < \delta$ (with δ a small value approaching zero) is evaluated by the ratio of the density region of the null hypothesis under the posterior and prior distribution of the more complex alternative hypothesis H_1.

To compute this Bayes factor from MCMC samples, multiple strategies can be applied. One can use the MCMC chains to approximate the posterior distribution, and determine how many of the values sampled from the posterior distribution are in agreement with the about equality constraint. This will give the correct solution when the number of iterations is large, but the approach can give problems when for example none of the samples is in agreement with the constraint. A more efficient approach is to compute the conditional posterior density region containing the null hypothesis within each iteration, and then average over all iterations to obtain the expected density region containing the null hypothesis under the posterior density.

To do this, in each iteration M, the probability of H_0 under the conditional posterior density of $\sigma^2_{\xi_k}$ in that iteration of the MCMC chain is computed. Using the inverse gamma prior distribution $\sigma^2_{\xi_k} \sim IG(g1, g2)$, the conditional posterior distribution of $\sigma^2_{\xi_k}$ in iteration m given the sampled general and group-specific item parameters $\tilde{\xi}^{(m)}_{kj}$ and $\xi^{(m)}_k$ is:

$$\sigma^2_{\xi_k} \mid \tilde{\xi}^{(m)}_{kj}, \xi^{(m)}_k \sim IG(g1 + J/2, \sum_j (\tilde{\xi}_{kj} - \xi_k)^2/2 + g2)$$

The probability of the density region $H_0 : \sigma^2_{\xi_k} < \delta$ under the posterior density can now be computed from the cumulative distribution function of this Inverse Gamma distribution. As the prior density is the same in each iteration, the density region under the prior density remains constant over iterations. The resulting

Bayes factor $\mathrm{BF}_{01}^{(m)}$ for iteration m is:

$$BF_{01}^{(m)} = \frac{p(\sigma^2_{\xi_k} < \delta \mid H_1, Y)}{p(\sigma^{2(m)}_{\xi_k} < \delta \mid H_1)} = \frac{\int_0^\delta p(\sigma^2_{\xi_k} \mid \tilde{\xi}^{(m)}_{kj}, \xi^{(m)}_k) d(\sigma^2_{\xi_k})}{\int_0^\delta p(\sigma^2_{\xi_k}) d(\sigma^2_{\xi_k})}.$$

The final result of the Bayes factor test is now given by the average Bayes factor over all iterations

$$BF_{01} = \sum_1^M \mathrm{BF}_{01}^{(m)} / M.$$

Appendix C

HPD test for measurement invariance (Chapter 3)

C.1 Introduction

This chapter is written as an extension to Chapter 3, in which Bayesian tests for measurement invariance based on random item effects models have been proposed. Another possibility for measurement invariance testing in random item parameter multilevel IRT models is through a highest posterior density region (HPD) test. This Appendix will describe this test and give results from a simulation study and the European Social Survey (ESS) data described in Chapter 3.

C.2 HPD Region Testing

Highest posterior density (HPD) region testing (Box and Tiao, 1973) is a procedure which is based on the posterior density of the parameters of interest. The procedure evaluates the posterior density of the parameter region under the null hypothesis given the posterior density function. The null hypothesis is rejected when the region has low posterior density given a significance level α.

To test the assumption of measurement invariance for item k, let H_0 state that each of the j country-specific item parameter vectors $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{b}_{kj})^t$ equal their international item parameter vector $\boldsymbol{\xi}_k = (a_k, b_k)^t$ such that $\tilde{\boldsymbol{\xi}}_{kj} = \boldsymbol{\xi}_k$ for each country j. The HPD test for a given item k now consists of evaluating whether the point where all country-specific item parameters are equal to the international item parameters is contained in the $(1-\alpha)$ highest posterior density region of the joint posterior distribution of the country-specific item parameters.

This point $\left(\tilde{\boldsymbol{\xi}}_{k1} = \tilde{\boldsymbol{\xi}}_{k2} = \ldots = \tilde{\boldsymbol{\xi}}_{kJ} = \boldsymbol{\xi}_k\right)$ is included in the $(1-\alpha)$ HPD region of the country-specific item parameters if and only if

$$P\left(p\big(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}\big) > p\big(\tilde{\boldsymbol{\xi}}_{k1} = \tilde{\boldsymbol{\xi}}_{k2} = \ldots = \tilde{\boldsymbol{\xi}}_{kJ} = \boldsymbol{\xi}_k \mid \mathbf{y}\big) \mid \mathbf{y}\right) < 1 - \alpha, \tag{C.1}$$

where $p\big(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}\big)$ is the joint posterior distribution of all country-specific item

parameters. The posterior probability in Equation(C.1) represents the HPD region under this posterior distribution that just includes the point of equal country-specific item parameters.

Following the derivation in section C.6, the conditional posterior probability of the HPD region that just includes the point of equal country-specific item parameters can be expressed as

$$P\Big(\chi^2_{2J} \leq \sum_j \Big(\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}\Big)^t \boldsymbol{\Omega}^{-1}_{\tilde{\xi}_{kj}} \Big(\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}\Big) \mid \mathbf{z}_{kj}, \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_{\tilde{\xi}}, \boldsymbol{\xi}_k\Big),$$

with conditional posterior mean and variance

$$\begin{aligned} \boldsymbol{\Omega}^{-1}_{\tilde{\xi}_{kj}} &= \mathbf{H}^t\mathbf{H} + \boldsymbol{\Sigma}^{-1}_{\tilde{\xi}} \\ \boldsymbol{\mu}_{\tilde{\xi}_{kj}} &= \boldsymbol{\Omega}_{\tilde{\xi}_{kj}} \Big(\mathbf{H}^t\mathbf{z}_{kj} + \boldsymbol{\Sigma}^{-1}_{\tilde{\xi}}\boldsymbol{\xi}_k\Big). \end{aligned}$$

This conditional posterior probability is evaluated in each iteration of the MCMC algorithm. After M iterations, the mean value is the marginal posterior probability contained by the HPD region that just includes the point of invariant parameters. When this posterior probability is less than $1-\alpha$, it is concluded that item k is measurement invariant.

This approach provides an easy way to test the invariance assumption for the discrimination and difficulty parameters of item k simultaneously. However, it is also possible to test the discrimination parameters separately from the difficulty parameters or to test the measurement invariance assumption for a subset of countries or for multiple items simultaneously. This flexibility follows from the well-known fact that each subset of a sum of squared standard normal random variables is chi-square distributed.

In a similar way, the hypothesis of invariant latent country means, $\beta_{0j} = \gamma_{00}$ for $j = 1, \ldots, J$, can be tested given the conditional joint posterior distribution of the latent group means. The HPD test equals

$$P\left(p\big(\boldsymbol{\beta}_0 \mid \mathbf{y}\big) > p\big(\beta_{01} = \ldots = \beta_{0J} = \gamma_{00} \mid \mathbf{y}\big) \mid \mathbf{y}\right) < 1-\alpha,$$

which can be transformed (in a similar way as for the item parameters) to the conditional HPD test

$$P\Big(\chi^2_J \leq \sum_j \left(\gamma_{00} - \mu_{\beta_{0j}}\right)^t \Omega^{-1}_{\beta_{0j}} \left(\gamma_{00} - \mu_{\beta_{0j}}\right) \mid \boldsymbol{\mu}_{\beta_0}, \boldsymbol{\Omega}_{\beta_0}\Big) < 1-\alpha.$$

with $\mu_{\beta_{0j}}$ the posterior mean and $\Omega_{\beta_{0j}}$ the posterior variance of the latent country mean β_{0j}. The marginal posterior probability covered by the HPD interval that just contains the point of equal latent country means is the average of the conditional probabilities, which are computed in each MCMC iteration.

The assumption of invariant within-country variances is tested by defining $(J-1)$ linear independent contrasts $\Delta_j = \log\sigma^2_{\theta_j} - \log\sigma^2_{\theta_J}$. The point $\boldsymbol{\Delta}_0 = \mathbf{0}$ corresponds to the event that $\sigma^2_{\theta_1} = \ldots = \sigma^2_{\theta_J}$. The point $\boldsymbol{\Delta}_0 = \mathbf{0}$ is included in a $(1-\alpha)$ HPD region if and only if

$$P\left(p\left(\boldsymbol{\Delta} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) > p\left(\boldsymbol{\Delta}_0 \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) < 1-\alpha. \tag{C.2}$$

Now, consider the conditional distribution of the within-country variances,

$$\begin{aligned} p\left(\sigma^2_{\theta_1}, \ldots, \sigma^2_{\theta_J} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) &\propto \prod_j \left(\sigma^2_{\theta_j}\right)^{-((g_1+n_j)/2+1)} \exp\left(-\frac{s_j^2 + g_2}{2\sigma^2_{\theta_j}}\right) \\ &\propto \prod_j \left(\sigma^2_{\theta_j}\right)^{-\left(n'_j/2+1\right)} \exp\left(-\frac{n'_j s'^2_j}{2\sigma^2_{\theta_j}}\right), \end{aligned}$$

where

$$s'^2_j = \frac{\sum_{i|j} \left(\theta_{ij} - \beta_{0j}\right)^2 + g_2}{n_j + g_1},$$

using an inverse gamma prior with shape and scale parameters $g_1/2$ and $g_2/2$, respectively. Box and Tiao (1973, pp. 133–136) showed that, for $n_j \to \infty$, the conditional probability statement in Equation (C.2) is approximately equal to the probability statement

$$P\left(\chi^2_{J-1} \leq -\sum_{j=1}^{J} n_j \left(\log s'^2_j - \log \overline{s}'^2\right) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) < 1 - \alpha, \tag{C.3}$$

where $\overline{s}'^2$ is the weighted average variance across nations. The conditional posterior probability in Equation (C.3) is evaluated in each MCMC iteration and the averaged posterior probability is considered to be an estimate of the marginal posterior probability covered by the HPD interval that just contains the point $\boldsymbol{\Delta}_0$.

The HPD region tests are described for the unconditional random item effects MLIRT model but they are easily adjusted for the conditional model. When predictors are used to explain variance, the test results are to be interpreted conditionally on this background information, which means for example that the invariance assumption of within-country variances holds conditional on explanatory information.

C.3 Simulation study

Data were simulated using parameters drawn from the prior distributions. The latent group means μ_j were sampled from a normal distribution with mean zero and between-group variance .50. Individual latent variable values were drawn from a normal distribution with mean μ_j and a variance generated from an inverse gamma distribution such that the within-group variances ranged from .40 to 2.00. International difficulty and discrimination parameters were generated from normal distributions with mean zero and one and variance .30 and .10, respectively. Group-specific item parameters were generated from normal distributions with the international parameters as the mean values. The cross-national item parameter variances varied from zero to .13 for the discrimination parameters and from zero to .30 for the difficulty parameters across five groups of four items. As a result, the group-specific discrimination parameters ranged between .50 and 1.50, and the difficulty parameters ranged between -2.00 and 2.00.

The measurement invariance assumption was tested for each item nested in five item groups. The items in group one were simulated to be measurement invariant and the items in the other groups to have increasingly varying item parameters. The HPD test is focused on significant cross-national differences in item parameter estimates which implies that the HPD test evaluates item parameter differences between the groups in the sample. In addition, the invariance of the latent means and variances was tested.

Table C.1: Marginal invariance testing: Percentage of invariant parameters detected and average HPD interval just containing the null hypothesis over 50 data sets

H_0	HPD	
	$\alpha = .05$	p_0
$\sigma^2_{a_k}$=0		
.00	0.43	0.91
.04	0.06	0.98
.07	0.02	1.00
.10	0.02	0.99
.13	0.00	1.00
$\sigma^2_{b_k} = 0$		
.00	0.82	0.77
.05	0.01	1.00
.10	0.00	1.00
.20	0.00	1.00
.30	0.00	1.00

The highest posterior density region test rejects the null hypothesis when the HPD region that just includes the point of invariant parameters covers more than .95 of the posterior probability. In Table C.1, the results of the highest posterior density region test for groups of 100 persons are given. The item-level results are averaged for each item group. In the first column, for each item group the percentage of items across the 50 data sets for which the null hypothesis is not rejected is shown. The second column presents for each item group the average size of the HPD region that just includes the point of equal item parameters (labeled as p_0).

For the groups of non-invariant items, the HPD region test correctly rejected the null hypotheses on both item parameters for almost all items. The average posterior probability p_0 of the HPD region that just included the point of equal item parameters across nations was nearly one for these groups, indicating that equality of parameters over groups was very unlikely to be true. The HPD test of invariant discrimination parameters did not reject the null hypothesis for only 43% of the invariant items over the 50 data sets. The average size of the HPD region that just included the point of equal item discriminations was .91, indicating that the point of equal discrimination parameters was on average in the upper area of the 95% HPD region. With respect to the invariant difficulty parameters the

results were slightly better, as 82% were correctly indicated as invariant by the HPD test. For this group, the average size of the HPD region that just included the null hypothesis was .77.

These results show that the HPD test for item parameter invariance has an inflated Type I error, especially for the discrimination parameters. For this relatively large number of countries, the point of equal means for all countries falls outside the HPD region rather quickly. This becomes even more problematic when the number of cases per country increases, as the within-group information becomes stronger and the HPD region tightens around the posterior mean.

The HPD test for testing the latent factor means and the latent factor variances showed significant results: both points of invariance were not included in the 99% HPD regions. In addition, the Bayes factors for testing these assumptions were close to zero. The assumption of factor variance invariance and factor mean invariance were correctly rejected by those tests.

C.4 Example: ESS

As in Chapter 3, response data of 22 countries from the European Social Survey (ESS round 1, 2002) were considered, from which eight dichotomized items concerning the perceived consequences and allowance of immigration were used. The parameters were estimated with the random item parameter multilevel IRT model.

Invariant and constrained non-invariant models were estimated on the eight immigration items. Based on 10,000 MCMC samples with a burn-in of 1,000, no autocorrelations higher than .15 were found and the Geweke Z convergence diagnostic did not show values above three, indicating that the chains converged well and reached stationarity.

For all items, the points of equal group-specific item parameters were in the outer tail of the joint posterior distributions, even outside the 99% HPD region. The relatively large number of respondents per country resulted in sharply peaked posterior densities and very accurate posterior mean estimates, which led to significant between-country differences when evaluating the HPD regions.

C.5 Conclusion

The HPD test summarizes information from the joint posterior distributions of the country-specific parameters and evaluates the probability of parameter equality in all countries. The HPD test showed significant type I errors for the large number of countries used. This test is therefore not recommended when investigating measurement invariance for a large number of countries. Further research is needed to investigate the performance of this test for a smaller number of countries.

C.6 Derivation HPD test

From the HPD test introduced by Box and Tiao (1973), the point $\left(\tilde{\boldsymbol{\xi}}_{k1} = \tilde{\boldsymbol{\xi}}_{k2} = \ldots = \tilde{\boldsymbol{\xi}}_{kJ} = \boldsymbol{\xi}_k\right)$ is included in the $(1-\alpha)$ HPD region of the country-specific item

parameters if and only if

$$P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}) > p(\tilde{\boldsymbol{\xi}}_{k1} = \tilde{\boldsymbol{\xi}}_{k2} = \ldots = \tilde{\boldsymbol{\xi}}_{kJ} = \boldsymbol{\xi}_k \mid \mathbf{y}) \mid \mathbf{y}\right) < 1 - \alpha, \tag{C.4}$$

where $p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y})$ is the joint posterior distribution of all country-specific item parameters. The posterior probability in equation (C.4) represents the HPD region under this posterior distribution that just includes the point of equal country-specific item parameters.

The evaluation of equation (C.4) is complicated, since the analytical expression of the marginal posterior density of the country-specific item parameters, $p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y})$, is not known. First, we will use the auxiliary variable Z, which is defined in equation 3.2 in Chapter **??**. Then, following a.o. Held (2004) we will condition the country-specific item parameters on other model parameters to obtain an analytical expression. Subsequently, it will be shown that the posterior probability in C.1 can be obtained by integrating out the other model parameters in the expression using MCMC.

By definition of the augmentation scheme,

$$\begin{aligned} & P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}) > p(\mathbf{1}\boldsymbol{\xi}_k \mid \mathbf{y}) \mid \mathbf{y}\right) \\ &= \int P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}, \mathbf{z}) > p(\mathbf{1}\boldsymbol{\xi}_k \mid \mathbf{y}, \mathbf{z}) \mid \mathbf{y}, \mathbf{z}\right) p(\mathbf{z} \mid \mathbf{y}) d\mathbf{z} \\ &= \int P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}) > p(\mathbf{1}\boldsymbol{\xi}_k \mid \mathbf{z}) \mid \mathbf{z}\right) p(\mathbf{z} \mid \mathbf{y}) d\mathbf{z}. \end{aligned} \tag{C.5}$$

The analytical form of the conditional density $p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z})$ given the augmented data is also unknown. However, the conditional density $p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}, \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_{\tilde{\xi}}, \boldsymbol{\xi}_k)$ given the augmented data and country-specific parameters is known to be normal. Let $\boldsymbol{\Lambda} = \boldsymbol{\theta}_j, \boldsymbol{\Sigma}_{\tilde{\xi}}, \boldsymbol{\xi}_k$. Then,

$$p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}) = \int p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}, \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda} \mid \mathbf{z}) d\boldsymbol{\Lambda} = \int \Phi(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}, \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda} \mid \mathbf{z}) d\boldsymbol{\Lambda}, \tag{C.6}$$

where $\Phi()$ denotes the normal density function.

With the result of C.6, the inequality statement in C.5 can be rewritten such that:

$$\begin{aligned} P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}) p(\mathbf{1}\boldsymbol{\xi}_k \mid \mathbf{y}) \mid \mathbf{y}\right) &= \int P\left(\Phi(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{z}, \boldsymbol{\Lambda}) > \Phi(\tilde{\boldsymbol{\xi}}_{kj} = \mathbf{1}\boldsymbol{\xi}_k \mid \mathbf{z}, \boldsymbol{\Lambda}) \mid \mathbf{z}, \boldsymbol{\Lambda}\right) \\ &\qquad p(\boldsymbol{\Lambda} \mid \mathbf{z}) d\boldsymbol{\Lambda} \\ = \int P\left((\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\mu}_{\tilde{\xi}_{kj}})^t \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1} (\tilde{\boldsymbol{\xi}}_{kj} - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}) \leq (\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}})^t \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1} (\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}) \mid \mathbf{z}, \boldsymbol{\Lambda}\right) \\ &\qquad p(\boldsymbol{\Lambda} \mid \mathbf{z}) d\boldsymbol{\Lambda} \\ = \int P\left(\chi^2_{2J} \leq \sum_j \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}\right)^t \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1} \left(\boldsymbol{\xi}_k - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}\right) \mid \mathbf{z}, \boldsymbol{\Lambda}\right) & p(\boldsymbol{\Lambda} \mid \mathbf{z}) d\boldsymbol{\Lambda}, \end{aligned}$$

with conditional posterior mean and variance

$$\boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1} = \mathbf{H}^t \mathbf{H} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1}$$

$$\boldsymbol{\mu}_{\tilde{\xi}_{kj}} = \boldsymbol{\Omega}_{\tilde{\xi}_{kj}} \left(\mathbf{H}^t \mathbf{z}_{kj} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1} \boldsymbol{\xi}_k \right),$$

and $\mathbf{H} = \left(\boldsymbol{\theta}_j, -\mathbf{1}_{n_j} \right)$.

Based on the properties of MCMC, it follows that:

$$P\left(p(\tilde{\boldsymbol{\xi}}_k \mid \mathbf{y}) > p(\tilde{\boldsymbol{\xi}}_{k1} = \tilde{\boldsymbol{\xi}}_{k2} = \ldots = \tilde{\boldsymbol{\xi}}_{kJ} = \boldsymbol{\xi}_k \mid \mathbf{y}) \mid \mathbf{y} \right) = \lim_{M \to \infty} M^{-1} \sum_{m=1}^{M} P\left(\chi^2_{2J} \leq T\left(\boldsymbol{\Lambda}^{(m)} \right) \right),$$

where

$$T\left(\boldsymbol{\Lambda}^{(m)} \right) = \sum_j \left(\boldsymbol{\xi}_k^{(m)} - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}^{(m)} \right)^t \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1(m)} \left(\boldsymbol{\xi}_k^{(m)} - \boldsymbol{\mu}_{\tilde{\xi}_{kj}}^{(m)} \right),$$

with

$$\begin{aligned} \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{-1(m)} &= \mathbf{H}^{(m)t} \mathbf{H}^{(m)} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1(m)} \\ \boldsymbol{\mu}_{\tilde{\xi}_{kj}}^{(m)} &= \boldsymbol{\Omega}_{\tilde{\xi}_{kj}}^{(m)} \left(\mathbf{H}^{t(m)} \mathbf{z}_{kj}^{(m)} + \boldsymbol{\Sigma}_{\tilde{\xi}}^{-1(m)} \boldsymbol{\xi}_k^{(m)} \right), \end{aligned}$$

and $\mathbf{H}^{(m)} = \left(\boldsymbol{\theta}_j^{(m)}, -\mathbf{1}_{n_j}^{(m)} \right)$.

Appendix D

MCMC Algorithm for the Longitudinal Generalized Partial Credit Model (Chapter 4)

The combination of truncated and correlated random effects with the identification restrictions described in §2.4 makes the model unfortunately not suitable for estimation in programs like Winbugs or Mplus. The sampler was written in Fortran and can be called from Splus. It is available from the authors.

To sample the parameters of the joint growth model, an MCMC sampling scheme has been developed with a Metropolis-Hastings step for the sampling of the occasion-specific item and person parameters and a Gibbs sampler for the the higher level parameters.

At the $m+1$th iteration:

1. For each k, j and c, sample a proposal $\tilde{\xi}^{*}_{ckj}$ $(c = 1, ..., C)$ from $N(\tilde{\xi}^{(m)}_{ckj}, \sigma^2_{mh\xi})$, where $\sigma^2_{mh\xi}$ is tuned during the process to acquire an acceptance rate between .3 and .5. Let $\tilde{\boldsymbol{\xi}}^{-c}_{kj} = \left(\tilde{\xi}_{1kj}, \ldots, \tilde{\xi}_{(c-1)kj}, \tilde{\xi}_{(c+1)kj}, \ldots, \tilde{\xi}_{Ckj}\right)$. The acceptance ratio R is the posterior probability ratio of the proposed value $\tilde{\xi}^{*}_{ckj}$ and the previously sampled value $\tilde{\xi}^{(m)}_{ckj}$:

$$R = \frac{p\left(\mathbf{y}_{kj} \mid \tilde{\xi}^{*}_{ckj}, \boldsymbol{\theta}^{(m)}, \tilde{\boldsymbol{\xi}}^{(m)}_{kj}\right) p\left(\tilde{\xi}^{*}_{ckj} \mid \xi^{(m)}_{ck}, \delta^{(m)}_{ck}, \mathbf{v}_j, \tilde{\boldsymbol{\xi}}^{-c}_{kj}, \boldsymbol{\Sigma}^{(m)}_{\tilde{\xi}_k}\right)}{p\left(\mathbf{y}_{kj} \mid \tilde{\xi}^{(m)}_{ckj}, \boldsymbol{\theta}^{(m)}, \tilde{\boldsymbol{\xi}}^{(m)}_{kj}\right) p\left(\tilde{\xi}^{(m)}_{ckj} \mid \xi^{(m)}_{ck}, \delta^{(m)}_{ck}, \mathbf{v}_j, \tilde{\boldsymbol{\xi}}^{-c}_{kj}, \boldsymbol{\Sigma}^{(m)}_{\tilde{\xi}_k}\right)}.$$

A random uniform number u_{ckj} is drawn, and the proposal is accepted when $u_{ckj} \leq R$.

2. For each k and j and for each discrimination parameter $c = 0$, sample $\tilde{\xi}^*_{0kj}$ from $N(\tilde{\xi}^{(m)}_{0kj}, \sigma^2_{mh\xi})I(\tilde{\xi}^*_{0kj} > 0)$. Since the proposal density is truncated, the acceptance ratio R is:

$$R = \frac{p\left(\mathbf{y}_{kj} \mid \tilde{\xi}^*_{0kj}, \boldsymbol{\theta}^{(m)}, \tilde{\boldsymbol{\xi}}^{(m+1)}_{kj}\right) p\left(\tilde{\xi}^*_{0kj} \mid \xi^{(m)}_{ck}, \delta^{(m)}_{ck}, \mathbf{v}_j, \tilde{\boldsymbol{\xi}}^{(m+1)}_{kj}, \boldsymbol{\Sigma}^{(m)}_{\tilde{\xi}_k}\right)}{p\left(\mathbf{y}_{kj} \mid \tilde{\xi}^{(m)}_{0kj}, \boldsymbol{\theta}^{(m)}, \tilde{\boldsymbol{\xi}}^{(m+1)}_{kj}\right) p\left(\tilde{\xi}^{(m)}_{0kj} \mid \xi^{(m)}_{ck}, \delta^{(m)}_{ck}, \mathbf{v}_j, \tilde{\boldsymbol{\xi}}^{(m+1)}_{kj}, \boldsymbol{\Sigma}^{(m)}_{\tilde{\xi}_k}\right)} \frac{\boldsymbol{\Phi}\left(\frac{\tilde{\xi}^{(m)}_{0kj}}{\sigma_{mh}}\right)}{\boldsymbol{\Phi}\left(\frac{\tilde{\xi}^*_{0kj}}{\sigma_{mh}}\right)}.$$

A random uniform number u_{0kj} is drawn, and the proposal is accepted when $u_{0kj} \leq R$.

3. For each i, sample θ^*_{ij} from $N(\theta^{(m)}_{ij}, \sigma^2_{mh_\theta})$. The acceptance ratio R is:

$$R = \frac{p(\mathbf{y}_{ij} \mid \theta^*_{ij}, \tilde{\boldsymbol{\xi}}^{(m+1)}) p(\theta^*_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}^{(m)}_i, \mathbf{s}_{ij}, \boldsymbol{\zeta}^{(m)}, \sigma^{(m)}_j)}{p\left(\mathbf{y}_{ij} \mid \theta^{(m)}_{ij}, \tilde{\boldsymbol{\xi}}^{(m+1)}\right) p\left(\theta^{(m)}_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}^{(m)}_i, \mathbf{s}_{ij}, \boldsymbol{\zeta}^{(m)}, \sigma^{(m)}_j\right)}.$$

A random uniform number u_{ij} is drawn, and the proposal is accepted when $u_{ij} \leq R$.

4. For each k, sample the general item parameters $\boldsymbol{\xi}^{(m+1)}_k$ and time coefficients $\boldsymbol{\delta}^{(m+1)}_k$ from the full conditional

$$\text{vec}\left[\begin{pmatrix} \xi_{0k} \\ \boldsymbol{\delta}_{0k} \end{pmatrix}^{(m+1)} \quad \dots \quad \begin{pmatrix} \xi_{Ck} \\ \boldsymbol{\delta}_{Ck} \end{pmatrix}^{(m+1)} \right] \mid \cdot \sim \mathcal{N}\left(\boldsymbol{\mu}^*, \Omega^*\right),$$

where

$$\begin{aligned}
\boldsymbol{\Omega}^{*-1} &= \boldsymbol{\Sigma}^{-1}_{\tilde{\xi}_k} \otimes \left(\begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix}^t \begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix} \right)^{-1} + \boldsymbol{\Sigma}^{-1}_{\xi\delta}, \\
\boldsymbol{\mu}^* &= \boldsymbol{\Omega}^* \left(\boldsymbol{\Sigma}^{-1}_{\tilde{\xi}_k} \otimes \left(\begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix}^t \begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{\xi_{0k}} \\ \hat{\boldsymbol{\delta}_{0k}} \\ \dots \\ \hat{\xi_{Ck}} \\ \hat{\boldsymbol{\delta}_{Ck}} \end{bmatrix} + \boldsymbol{\Sigma}^{-1}_{\xi\delta} \begin{bmatrix} \mu_{\xi_0} \\ \boldsymbol{\mu}_{\delta_0} \\ \dots \\ \mu_{\xi_C} \\ \boldsymbol{\mu}_{\delta_C} \end{bmatrix} \right), \\
\text{with } \begin{bmatrix} \hat{\boldsymbol{\xi}_k} \\ \hat{\boldsymbol{\delta}_k} \end{bmatrix} &= \left(\begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix}^t \begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{1} & \mathbf{v} \end{bmatrix}^t \tilde{\boldsymbol{\xi}}_k, \\
\text{and } \boldsymbol{\Sigma}_{\xi\delta} &= (\sigma_{\xi_0} \oplus \boldsymbol{\Sigma}_{\delta_0}) \oplus \dots \oplus (\sigma_{\xi_C} \oplus \boldsymbol{\Sigma}_{\delta_C}).
\end{aligned}$$

5. For each k, sample $\boldsymbol{\Sigma}^{(m+1)}_{\tilde{\xi}_k}$ from the full conditional

$$\begin{aligned}
\boldsymbol{\Sigma}^{(m+1)}_{\tilde{\xi}_k} &\mid \tilde{\boldsymbol{\xi}}^{(m+1)}, \boldsymbol{\xi}^{(m+1)}_k, \boldsymbol{\delta}^{(m+1)}_k, S_{0\xi_k}, n_0 \sim \mathcal{IW}\left((n_0 + J/2), (S_{0\xi_k} + S)\right), \\
\text{where } S &= \left(\tilde{\xi}_k - (\xi_k + \mathbf{v}\boldsymbol{\delta}_k)\right)' \left(\tilde{\xi}_k - (\xi_k + \mathbf{v}\boldsymbol{\delta}_k)\right).
\end{aligned}$$

6. Sample $\boldsymbol{\mu}_\xi^{(m+1)}$ and $\boldsymbol{\Sigma}_\xi^{(m+1)}$ from the full conditionals

$$\begin{aligned}
\boldsymbol{\mu}_\xi^{(m+1)} &\mid \boldsymbol{\xi}_k^{(m+1)}, \boldsymbol{\Sigma}_\xi^{(m)}, K_0 \sim \mathcal{N}\left(\frac{K}{K_0+K}\overline{\boldsymbol{\xi}}, \frac{\boldsymbol{\Sigma}_\xi}{K_0+K}\right), \\
\boldsymbol{\Sigma}_\xi^{(m+1)} &\mid \boldsymbol{\xi}_k^{(m+1)}, \nu, K_0, \mathbf{S}_{0_\xi}, \sim \mathcal{IW}\left(K+\nu, \boldsymbol{\Sigma}^*\right), \\
&\text{with} \\
\overline{\boldsymbol{\xi}} &= \sum_k \boldsymbol{\xi}_k/K, \\
\boldsymbol{\Sigma}^* &= \mathbf{S}_{0_\xi} + K\left((\boldsymbol{\xi}_k - \overline{\boldsymbol{\xi}})(\boldsymbol{\xi}_k - \overline{\boldsymbol{\xi}})^t\right) + \frac{KK_0}{K+K_0}\overline{\boldsymbol{\xi}}\overline{\boldsymbol{\xi}}^t.
\end{aligned}$$

7. Sample $\boldsymbol{\mu}_\delta^{(m+1)}$ and $\boldsymbol{\Sigma}_\delta^{(m+1)}$ from the full conditionals

$$\begin{aligned}
\boldsymbol{\mu}_\delta^{(m+1)} &\mid \boldsymbol{\delta}_k^{(m+1)}, \boldsymbol{\Sigma}_\delta^{(m)}, K_0 \sim \mathcal{N}\left(\frac{K}{K_0+K}\overline{\boldsymbol{\delta}}, \frac{\boldsymbol{\Sigma}_\delta}{K_0+K}\right), \\
\boldsymbol{\Sigma}_\delta^{(m+1)} &\mid \boldsymbol{\delta}_k^{(m)}, \nu, K_0, \mathbf{S}_{0\delta} \sim \mathcal{IW}\left(K+\nu, \boldsymbol{\Sigma}^*\right), \\
&\text{with} \\
\overline{\boldsymbol{\delta}} &= \sum_k \boldsymbol{\delta}_k/K, \\
\boldsymbol{\Sigma}^* &= \mathbf{S}_{0\delta} + K\left((\boldsymbol{\delta}_k - \overline{\boldsymbol{\delta}})(\boldsymbol{\delta}_k - \overline{\boldsymbol{\delta}})^t\right) + \frac{KK_0}{K+K_0}\overline{\boldsymbol{\delta}}\overline{\boldsymbol{\delta}}^t.
\end{aligned}$$

8. For each i, sample $\boldsymbol{\beta}_i^{(m+1)}$ from the full conditional

$$\begin{aligned}
\boldsymbol{\beta}_i^{(m+1)} &\mid \boldsymbol{\theta}_i^{(m+1)}, \sigma_j^{(m)}, \mathbf{T}^{(m)}, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\zeta}^{(m)} \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right), \\
&\text{where} \\
\boldsymbol{\mu}_\beta &= \boldsymbol{\Sigma}_\beta\left(\sigma_j^{-2}(\boldsymbol{x}_i^t\boldsymbol{x}_i)\hat{\beta}_i + \mathbf{T}^{-1}\mathbf{w}_i^t\boldsymbol{\gamma}\right), \\
\boldsymbol{\Sigma}_\beta &= \left(\sigma_j^{-2}(\boldsymbol{x}_i^t\boldsymbol{x}_i) + \mathbf{T}^{-1}\right)^{-1}, \\
\text{with } \hat{\beta}_i &= (\boldsymbol{x}_i^t\boldsymbol{x}_i)^{-1}(\boldsymbol{x}_i^t(\boldsymbol{\theta}_i - \mathbf{s}_i\boldsymbol{\zeta})).
\end{aligned}$$

9. Sample $\boldsymbol{\gamma}^{(m+1)}$ from the full conditional

$$\begin{aligned}
\boldsymbol{\gamma}^{(m+1)} &\mid \boldsymbol{\beta}_i^{(m+1)}, \mathbf{T}^{(m)}, S_{0\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma), \\
&\text{where} \\
\boldsymbol{\mu}_\gamma &= \boldsymbol{\Sigma}_\gamma \sum_i \mathbf{w}_i^t\mathbf{T}^{-1}\boldsymbol{\beta}_i, \\
\boldsymbol{\Sigma}_\gamma &= \left(\sum_i \mathbf{w}_i^t\mathbf{T}^{-1}\mathbf{w}_i + S_{0\gamma}^{-1}\right)^{-1}.
\end{aligned}$$

10. Sample $\mathbf{T}^{(m+1)}$ from the full conditional

$$\mathbf{T}^{(m+1)} \mid \boldsymbol{\beta}_i^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, S_{0T} \sim \mathcal{IW}\left((n_{0T}+I), (S_{0T} + \sum_i (\boldsymbol{\beta}_i - \mathbf{w}_i\boldsymbol{\gamma})(\boldsymbol{\beta}_i - \mathbf{w}_i\boldsymbol{\gamma})^t)\right).$$

11. Sample $\boldsymbol{\zeta}^{(m+1)}$ and $\boldsymbol{\Sigma}_{\zeta}^{(m+1)}$ from the full conditionals

$$\begin{aligned}
\boldsymbol{\zeta}^{(m+1)} &\mid \boldsymbol{\theta}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \Sigma_{\zeta}^{(m)} \sim \mathcal{N}\left(\frac{K}{K_0+N}\hat{\boldsymbol{\zeta}}, \frac{\boldsymbol{\Sigma}_{\zeta}}{K_0+N}\right), \\
\boldsymbol{\Sigma}_{\zeta}^{(m+1)} &\mid \boldsymbol{\zeta}^{(m)}, \nu, K_0, \mathbf{S}_{0\zeta} \sim \mathcal{IW}\left(K+\nu, \boldsymbol{\Sigma}^{*}\right), \\
\text{with } \hat{\boldsymbol{\zeta}} &= (\mathbf{s}^t\mathbf{s})^{-1}\left(\mathbf{s}^t(\boldsymbol{\theta}-\mathbf{x}\boldsymbol{\beta})\right), \\
\text{and } \boldsymbol{\Sigma}^{*} &= \mathbf{S}_{0\zeta} + N\left((\boldsymbol{\theta}-\mathbf{x}\boldsymbol{\beta})-\mathbf{s}\boldsymbol{\zeta}\right)^t\left((\boldsymbol{\theta}-X\boldsymbol{\beta})-\mathbf{s}\boldsymbol{\zeta}\right) + \frac{NK_0}{N+K_0}(\hat{\boldsymbol{\zeta}})^t(\hat{\boldsymbol{\zeta}}).
\end{aligned}$$

12. Sample $\sigma_j^{(m+1)}$ from the full conditional

$$\begin{aligned}
\sigma_j^{(m+1)} &\mid \boldsymbol{\theta}_j^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\zeta}^{(m+1)}, n_0, s_{0\sigma} \sim \mathcal{IG}\left(n_j+n_0, s^{*}+s_{0\sigma}\right), \\
\text{where } s^{*} &= \sum_{1}^{n_j}\left((\theta_{ij}-\mathbf{x}_i\boldsymbol{\beta}_i)-\mathbf{s}_i\boldsymbol{\zeta}\right)^2.
\end{aligned}$$

Appendix E

Extensions of fixed and random group models to the 2 Parameter Normal Ogive and Generalized Partial Credit Model (Chapter 5)

E.1 Extension to the 2 parameter normal ogive model (2PNO)

For a person $i = 1, \ldots, N$ in group $j = 1, \ldots, J$, the probability of endorsing item $k = 1, \ldots, K$ in the multi-group 2PNO model can be defined as follows:

$$P(Y_{ijk} = 1 \mid \theta_{ij}, \tilde{a}_{kj}, \tilde{b}_{kj} \quad = \quad \Phi(\tilde{a}_{kj}\theta_{ij} - \tilde{b}_{kj})$$

In both the random and fixed group models, the model is identified by restriction of the sum of the threshold parameters to zero and the product of the discrimination parameters to one within each group.

E.1.1 2PNO for random groups

For a random groups model, the person parameters can be specified as in Equations 5.3 and 5.4. The item parameters can be specified having independent univariate distributions:

$$\begin{aligned}
\tilde{a}_{kj} &= \mu_{a_0} + u_{a_k} + e_{a_{kj}} \\
u_{a_k} \mid \sigma_a &\sim \mathcal{N}\left(0, \sigma_a^2\right) I(a_k > 0), \\
e_{a_{kj}} \mid \sigma_{a_k} &\sim \mathcal{N}\left(0, \sigma_{a_k}^2\right) I(a_{kj} > 0), \\
\tilde{b}_{kj} &= \mu_{b_0} + u_{b_k} + e_{b_{kj}}
\end{aligned}$$

$$\begin{aligned} u_{b_k} \mid \sigma_b &\sim \mathcal{N}\left(0, \sigma_b^2\right), \\ e_{b_{kj}} \mid \sigma_{b_k} &\sim \mathcal{N}\left(0, \sigma_{b_k}^2\right), \end{aligned}$$

with hyperpriors $\mu_{a_0} \sim N(1,1)I(\mu_{a_0} > 0)$ and $\mu_{b_0} \sim N(0,1)$ and Inverse Gamma priors for $\sigma_{b_k}^2$, σ_b^2 , $\sigma_{a_k}^2$, and σ_a^2 (see also Appendix F).

The parameters $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{b}_{kj})$ can also be specified to come from a multivariate normal distribution:

$$\begin{aligned} \tilde{\boldsymbol{\xi}}_{kj} &= \boldsymbol{\mu}_{\xi_0} + \mathbf{u}_k + \mathbf{e}_{kj} \\ \mathbf{u}_k \mid \boldsymbol{\Sigma}_\xi &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\xi) I(a_k > 0) \\ \mathbf{e}_{kj} \mid \boldsymbol{\Sigma}_{\xi_k} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\xi_k}) I(a_{kj} > 0), \end{aligned}$$

where

$$\boldsymbol{\Sigma}_{\xi_k} = \begin{bmatrix} \sigma_{a_k}^2 & \sigma_{a_k b_k} \\ \sigma_{b_k a_k} & \sigma_{b_k}^2 \end{bmatrix},$$

and as hyperpriors $\boldsymbol{\mu}_{\xi_0} = (\mu_{a_0}, \mu_{b_0}) \sim N((0,1), \Omega) I(a_0 > 0)$ and Inverse Wishart priors for $\boldsymbol{\Sigma}_\xi$ and $\boldsymbol{\Sigma}_{\xi_k}$ (see also Appendix F).

E.1.2 2PNO for fixed groups

For a fixed groups model, the person parameters can be specified as in Equation 5.5. When independent distributions for the discrimination and threshold parameters are assumed, the following specification for the item parameters results.

For the threshold parameters:

$$\tilde{\mathbf{b}}_k \mid \boldsymbol{\mu_b}, \boldsymbol{\Sigma}_b \sim MVN(\boldsymbol{\mu_b}, \boldsymbol{\Sigma}_b), \tag{E.1}$$

where $\tilde{\mathbf{b}}_k = (\tilde{b}_{k1}, ..., \tilde{b}_{kj})$, $\boldsymbol{\mu_b} = (b_1, ..., b_j)$ and $\boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma_{b_1}^2 & \sigma_{b_1 b_{...}} & \sigma_{b_1 b_J} \\ \sigma_{b_{...} b_1} & \sigma_{b_{...}}^2 & \sigma_{b_{...} b_J} \\ \sigma_{b_J b_1} & \sigma_{b_J b_{...}} & \sigma_{b_J}^2 \end{bmatrix}$.

For the discrimination parameters:

$$\tilde{\mathbf{a}}_k \mid \boldsymbol{\mu_a}, \boldsymbol{\Sigma}_a \sim MVN(\boldsymbol{\mu_a}, \boldsymbol{\Sigma}_a) I(a_{kj} > 0),$$

where $\mathbf{a}_k = (a_{k1}, ..., a_{kj})$, $\boldsymbol{\mu_a} = (a_1, ..., a_j)$ and $\boldsymbol{\Sigma}_a = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_{...}} & \sigma_{a_1 a_J} \\ \sigma_{a_{...} a_1} & \sigma_{a_{...}}^2 & \sigma_{a_{...} a_J} \\ \sigma_{a_J a_1} & \sigma_{a_J a_{...}} & \sigma_{a_J}^2 \end{bmatrix}$.

And on a second level, hierarchical normal priors are chosen with ($\boldsymbol{\mu_b} = 0; \boldsymbol{\mu_a} = 1$) and Inverse Wishart priors for the covariance matrices $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_a$ (See also Appendix F).

E.2 Extension to the Generalized Partial Credit Model (GPCM)

In the multi-group GPCM model, the probability of a response ($Y_{ijk} = c$) ($c = 0, \ldots, C$) of a person $i = 1, \ldots, N$ in group $j = 1, \ldots, J$ on item $k = 1, \ldots, K$ is defined as:

$$\begin{aligned} P(Y_{ijk} = c \mid Z_{ijck}) &= \frac{\exp(\sum_0^c (Z_{ijck}))}{\sum_0^C \exp\left(\sum_0^c (Z_{ijck})\right)}, \\ Z_{ijck} &= \tilde{a}_{kj}(\theta_{ij} - \tilde{b}_{ckj}). \end{aligned}$$

In both the random and fixed group models, the model is identified by restriction of the sum of the threshold parameters over categories and items to zero and the product of the discrimination parameters to one within each group.

E.2.1 GPCM for random groups

In the random groups Bayesian IRT model, the random item parameters $\tilde{\boldsymbol{\xi}}_{kj} = (\tilde{a}_{kj}, \tilde{b}_{0kj}, ..., \tilde{b}_{Ckj})$ are assumed to be multivariate normally distributed given the general item parameters $\boldsymbol{\xi}_k = (a_k, b_{0k}, ..., b_{Ck})$:

$$\tilde{\boldsymbol{\xi}}_{kj} \mid \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\xi_k} \sim \mathcal{N}(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_{\xi_k}) I(\tilde{a}_{kj} > 0),$$

where

$$\boldsymbol{\Sigma}_{\xi_k} = \begin{bmatrix} \sigma^2_{a_k} & \sigma_{a_k b_{0k}} & ... & \sigma_{a_k b_{Ck}} \\ \sigma_{b_{0k} a_k} & \sigma^2_{b_{0k}} & ... & \sigma_{b_{0k} b_{Ck}} \\ ... & ... & ... & ... \\ \sigma_{b_{Ck} a_k} & \sigma_{b_{Ck} b_{0k}} & ... & \sigma^2_{b_{Ck}} \end{bmatrix}.$$

The hyperprior for the covariance matrix $\boldsymbol{\Sigma}_{\xi_k}$ is defined as an Inverse Wishart prior.

Following a hierarchical prior structure, on a higher level the general item parameters are assumed to be multivariate normally distributed given the overall mean parameters $\boldsymbol{\mu}_{\xi_0} = (a_0, b_{00}, ..., b_{C0})$:

$$\begin{aligned} \boldsymbol{\xi}_k \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi &\sim \mathcal{N}(\boldsymbol{\mu}_{\xi_0}, \boldsymbol{\Sigma}_\xi). \\ \boldsymbol{\Sigma}_\xi &\sim IW(R, K). \end{aligned}$$

E.2.2 GPCM for fixed groups

In the fixed groups Bayesian IRT model, when $\tilde{\boldsymbol{\xi}}_{ckj} = (\tilde{a}_{kj}, \tilde{b}_{0kj}, ..., \tilde{b}_{Ckj})$ and $\boldsymbol{\xi}_c = (\xi_{c1}, ..., \xi_{cJ})$:

$$\begin{aligned} \tilde{\boldsymbol{\xi}}_{ckj} &\sim MVN(\boldsymbol{\xi}_c, \Sigma_{\xi_c}) \\ \boldsymbol{\Sigma}_{\xi_c} &= \begin{bmatrix} \sigma^2_{\xi_{c1}} & \sigma_{\xi_{c1}\xi_{c...}} & \sigma_{\xi_{c1}\xi_{cJ}} \\ \sigma_{\xi_{c...}\xi_{c1}} & \sigma^2_{\xi_{c...}} & \sigma_{\xi_{c...}\xi_{cJ}} \\ \sigma_{\xi_{cJ}\xi_{c1}} & \sigma_{\xi_{cJ}\xi_{c...}} & \sigma^2_{\xi_{cJ}} \end{bmatrix}, \end{aligned}$$

where there is a separate covariance matrix $\boldsymbol{\Sigma}_{\xi_c}$ for the discrimination and each category threshold parameter $c = 1, \ldots, C$ to describe their variance over items.

The hyperpriors for these covariance matrices are Inverse Wishart distributed $\mathbf{\Sigma}_{\xi_c} \sim IW(R, (J))$.

On a higher level, hierarchical normal priors are chosen

$$\begin{aligned} \xi_c &\sim N(\xi_{c0}, \Sigma_\xi) \\ \Sigma_\xi &\sim IW(R, (C+1)). \end{aligned}$$

Appendix F

Choosing priors for variance components (Chapter 5)

The ratio of density or density regions of the null hypothesis under the posterior and prior distributions, and therefore the result of the Bayes factor test, depend on the priors chosen for the parameters under evaluation. For the Bayes factor tests described in this chapter, evaluating either differences between group-specific item parameters or the variance of an item parameter over groups, priors can be chosen which reflect reasonable assumptions about the parameter values, however.

As the normal ogive model $P(Y_{ik} = 1|\theta_i, b_k) = \Phi(\theta_i - b_k)$ puts $(\theta_i - b_k)$ on a standard normal scale, the person and threshold parameters are defined on this scale. The location of the zero point will be defined by the mean of the threshold parameters, which is set to zero within each group (see also Section 5.3).

To compute the Bayes factor test for the difference d between item parameters of two groups for the fixed groups Bayesian IRT model, within each MCMC sample, a sample of the prior distribution for the difference between the item parameters is taken (see Appendix H). The prior for the group-specific item parameters is set to a normal distribution, with mean $\mu_{b_j} = 0$ (as is part of the restriction) and a covariance matrix $\mathbf{\Sigma}_b$, which is estimated. Therefore, the prior specified for $\mathbf{\Sigma}_b$ is important for the results of the Bayes factor.

Now how to choose a prior for this covariance matrix $\mathbf{\Sigma}_b$? First, the values which the components of the matrix are expected to take need to be determined. As the item parameters are on a standard normal scale, it is assumed that the variance of the item parameters will be somewhere between 0 and 2, with values between .5 and 1.5 more likely than more extreme values. Second, a conjugate prior is desirable to make the computation process less complicated. A conjugate prior for the covariance of a multivariate normal distribution is an Inverse Wishart prior $IW(S, J)$. When an S matrix with ones on the diagonal and zeros elsewhere is chosen, the density for the diagonal variance components is high in the region between 0 and 2, with rapidly declining density for higher values. This prior $\mathbf{\Sigma}_b \sim IW(S, J)$ (Figure F.1) therefore has high density exactly in the area where we expect the item parameter variances to be, while it has low density elsewhere.

To compute the Bayes factor test for the variance of item parameters over

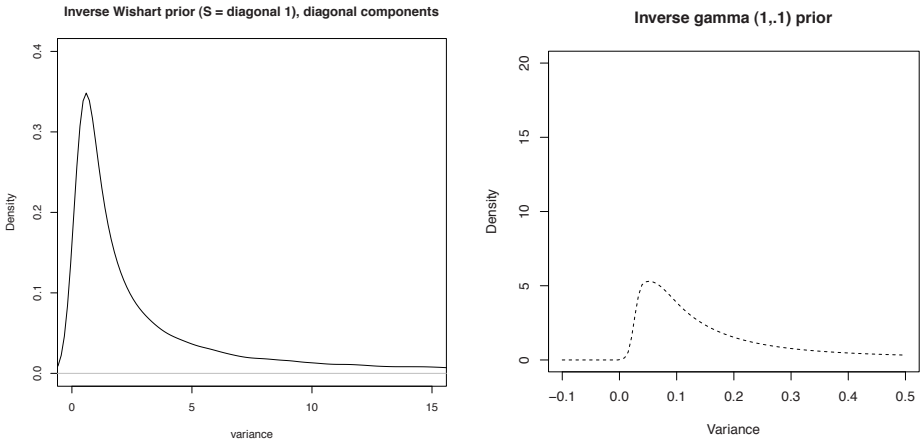

Figure F.1: Illustration of Inverse Wishart(S=diagonal 1,J) and Inverse Gamma (1,.1) priors

groups $\sigma^2_{b_k} < \delta$ in the random groups Bayesian IRT model, the prior distribution for $\sigma^2_{b_k}$ is of influence. Unlike the variance of item parameters over items, the variance of the group-specific threshold parameters over groups is not expected to span the whole range of the scale. Most of the differences between groups would be expected between 0 and .5, leading to a variance between 0 and .25, with higher variances possible but less likely to occur. A conjugate prior distribution which has most mass between 0 and .25, slowly decreasing in density for higher values of $\sigma^2_{b_k}$, is the Gamma(1,.1) distribution (Figure F.1).

Appendix G

Model specification in WinBUGS (Chapter 5)

G.1 Fixed multi-group IRT models

G.1.1 Manifest groups for item and person parameters

This section will present the WinBUGS (Lunn et al., 2000) code of the fixed manifest groups model for a data set Y with J groups j, K items k, and N persons i, stacked in such a way that $njl[1]$ is the first person for group j and $njh[J]$ is the last person for group j.

1. Basic model

```
for (j in 1:J){
    for (i in njl[j]:njh[j]){
        for (k in 1:K){
            logit(p[i,k]) <- theta[i]-Rbeta[k,j]
            Y[i,k] ~ dbern(p[i,k])
            }
        theta[i] ~ dnorm(mut[j],prec[j])
}}
```

2. Priors for group means

```
for (j in 1:J){
    mut[j] ~ dnorm(0,1)
    prec[j] ~ dgamma(1,.1)
    sigmat[j] <- 1/prec[j]
}
```

3. Item parameters

```
for (k in 1:K){
    beta[k,1:J] ~ dmnorm(mu[],Prec[,])
    priorbeta[k,1:J] ~ dmnorm(mu[],Prec[,])
}
```

4. Rescale item parameters

```
for (j in 1:J){
    meanb[j] <- mean(beta[1:K,j])
for (k in 1:K){
    Rbeta[k,j] <- beta[k,j] - meanb[j]
}}
```

5. Model DIF for Bayes Factor, 2 groups

```
for (k in 1:K){
    dif12[k] <- Rbeta[k,1]- Rbeta[k,2]
    difprior12[k] <- priorbeta[k,1]- priorbeta[k,2]
}
```

6. Priors for the item parameters (prior chosen for our analyses: R= $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, see Appendix F)

```
for (j in 1:J){mu[j] <- 0}
Prec[1:J,1:J] ~ dwish(R[1:J,1:J],J)
Sigma[1:J,1:J] <- inverse(Prec[1:J,1:J])
```

G.1.2 Manifest groups for persons, latent groups for items

- Specify 1 and 2 as in section G.1.1
- Item parameters

  Item parameters for unrestricted items:

```
for (k in 1:K){
     beta[k,1:J] ~ dmnorm(mu[],Prec[,])
}
```

  Item parameters for restricted items:

```
for (k in 1:K){
     betaR[k,1] ~ dnorm(mub,precb)
     for (j in 2:J){
     betaR[k,j] <- betaR[k,1]
}}
```

Assign item parameters according to class:

```
for (k in 1:K){
     for (j in 1:J){
     betaM[k,j]<-(cl[k]*beta[k,j])+(1-cl[k])*betaR[k,j]
}}
```

Restrict the sum of beta's to zero:

```
for (j in 1:J){
     mb[j] <- mean(betaM[1:K,j])
}

for (k in 1:K){
     Rbeta[k,1] <- (betaM[k,1] - mb[1])
     for (j in 2:J){
     c[j] <- (mb[1]-mb[j])*sum(1-cl[1:K])/sum(cl[1:K])
     means[k,j] <- (1-cl[k])*mb[1]+cl[k]*(mb[j]+c[j])
     Rbeta[k,j] <- betaM[k,j] - means[k,j]
}}
```

- Sample class and class probability

```
for (k in 1:K){cl[k] ~ dbern(.5)}
```

- Specify 5 and 6 as in section G.1.1
- Additional priors for restricted class

```
mub ~ dnorm(0,precb)
precb ~ dgamma(1,.1)
sigb <- 1/precb
```

G.1.3 Latent groups for person and item parameters

- Model specification with classes

```
for  (i in 1:N){
      for  (k in 1:K){
           logit(p[i,k]) <- (theta[i] - Rbeta[k,cl1[i]])
            Y[i,k] ~ dbern(p[i,k])
      }
      theta[i] ~ dnorm(mut[cl1[i]],prec[j])
}
```

- Priors for group means (The mean of the second group is restricted to be higher than the mean of the first group to avoid label switching):

```
for (j in 1:J){
        prec[j] ~ dgamma(1,.1)
        sigmat[j] <- 1/prec[j]
}
mut[1] ~ dnorm(0,1)
mut[2] ~ dnorm(0,1)   I(mut[1],)
```

- Class probabilities:

```
for (i in 1:N){
    class[i] ~ dbern(q)
    cl1[i] <- class[i] + 1
}
q ~ dbeta(1,1)
```

- Specify 3-6 as in section G.1.1.

G.2 Random multi-group IRT models

G.2.1 Manifest groups for item and person parameters

This section will present the WinBUGS (Lunn et al., 2000) code of the random manifest groups model for a data set Y with J groups j, K items k, and N persons i, stacked in such a way that $njl[1]$ is the first person for group j and $njh[J]$ is the last person for group j.

1. Model specification:

```
for (j in 1:J){
    for (i in njl[j]:njh[j]) {
        for (k in 1:K){
            logit(p[i,k]) <- theta[i] - (bkj[k,j] )
            Y[i,k] ~ dbern(p[i,k])
}}}
```

2. Multilevel specification for the person parameters:

```
for (j in 1:J){
    for (i in njl[j]:njh[j]) {
        theta[i] <- mu0 + ut[j] + er[i]
        er[i] ~ dnorm(0,prec[j])
    }
    ut[j] ~ dnorm(0,invtau)
    prec[j] ~ dgamma(1,.1)
    sigmat[j] <- 1/prec[j]
}
```

3. Priors for person parameter means and variances:

```
mu0 ~ dnorm(0,1)
invtau ~ dgamma(1,1)
tau <- 1/invtau
```

4. Multilevel specification of item parameters:

```
for (k in 1:K){
    for (j in 1:J){
        vbkj[k,j] <-  m0 + ek[k] + uj[j,k]
}}

for (k in 1:K){
    ek[k] ~ dnorm(0,precbk)
    for (j in 1:J){
        uj[j,k] ~ dnorm(0,precbjk[k])
}}
```

5. Rescaling of difficulty parameters to sum to zero

```
for (j in 1:J) {
    meanb[j] <- sum(vbkj[1:K,j])/K
}
for (k in 1:K){
    for (j in 1:J){
        bkj[k,j] <- vbkj[k,j]- meanb[j]
}}
```

6. Priors for item parameter means and variances

```
m0 ~ dnorm(0,1)
precbk ~ dgamma(1,.1)
sigbk <- 1/precbk

for (k in 1:K){
    precbjk[k] ~ dgamma(1,.1)
    sigbkj[k] <- 1/precbjk[k]
}
```

G.2.2 Manifest groups for persons, latent groups for items

- Follow the steps in G.2.1
- In step 4, replace the multilevel specification with:

```
for (k in 1:K){
    for (j in 1:J){
        vbkj[k,j] <-  m0 + ek[k] + class[k]*uj[j,k]
}}
```

- Add to step 4:

```
for (k in 1:K){
    class[k] ~ dbern(.5)
}
```

Appendix H

Bayes factor computation based on WinBUGS output in R (Chapter 5)

H.1 Bayes factor test for item parameter differences

The Bayes factors comparing nested models with regard to the difference in item parameters for all items simultaneously can be specified in R based on the index (ind) and coda (coda) files of WinBUGS (Lunn et al., 2000) output with XG iterations and BI as the number of burn in iterations. The sampled prior and posterior values of the differences in item parameters are specified in the variable "difchains", after which the density at $d = 0$ under both distributions under the logspline approximation of the density is computed similar to Wagenmakers et al. (2010).

```
library(polspline)

lo  <- which(rownames(ind) == "dif12[1]")
hi  <- lo+K-1
lop <- which(rownames(ind) == "difprior12[1]")
hip <- lop+K-1
it  <- XG-BURN

difchains       <- matrix(0,it,K)
difchains[,1:K]<-matrix(coda[ind[lo,2]:ind[hi,3],2],it,)
difchains[,(K+1):(K*2)]<-
                  matrix(coda[ind[lop,2]:ind[hip,3],2],it,)

posterior <- matrix(0,10,1)
```

```
prior      <- matrix(0,10,1)
BF01       <- matrix(0,10,1)

for (k in 1:K){
    fit.prior       <- logspline(difchains[,(k+10)])
    fit.posterior   <- logspline(difchains[,k])
    posterior[k]    <- dlogspline(0, fit.posterior)
    prior[k]        <- dlogspline(0, fit.prior)
    BF01[k]         <- posterior[k]/prior[k]
}
```

H.2 Bayes factor test for variance components

The Bayes factors comparing nested models with regard to the variance in item parameters for all items simultaneously can be specified in R based on the index (ind) and coda (coda) files of WinBUGS (Lunn et al., 2000) output with XG iterations and BI as the number of burn in iterations. The sampled posterior values of the item parameter variances are specified in the variable "sigchains", after which the area for $\sigma^2 < .02$ under the posterior distributions under the logspline approximation of the density is computed similar to Wagenmakers et al. (2010). The area for $\sigma^2 < .02$ under the prior distribution can be computed directly from the inverse gamma distribution.

```
library(polspline)
library(pscl)

lo  <- which(rownames(index) == "sigbkj[1]")
hi  <- lo+K-1
it  <- XG-BURN

sigchains        <- matrix(0,it,K)
sigchains[,1:K] <- matrix(coda[ind[lo,2]:ind[hi,3],2],it,)

posterior <- matrix(0,K,1)
prior     <- matrix(0,K,1)
BF01      <- matrix(0,K,1)

for (k in 1:K){
    fit.posterior = logspline(sigchains[,k],lbound=0)
    posterior[k]  = plogspline(0.02, fit.posterior)
    prior[k]      = pigamma(.02,1,.1)
    BF01[k]       = posterior[k]/prior[k]
}
```

Bibliography

Adams, R. J., Wilson, M. and Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variable Regression. *Journal of Educational and Behavioral Statistics, 22*, 47–76.

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, New Jersey: John Wiley and Sons.

Aitkin, M., and Longford, N. (1986). Statistical Modelling in School Effectiveness Studies. *Journal of the Royal Statistical Society A, 149*, 1-43.

Albers, W., Does, R. J. M. M., Imbos, Tj., and Janssen, M. P. E. (1989). A Stochastic Growth Model Applied to Repeated Tests of Academic Knowledge. *Psychometrika, 54*, 451–466.

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

Azevedo, C. L. N., Andrade, D. F., and Fox, J.-P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational Statistics and Data Analysis,12*, 4399–4412.

Bguin, A. A., and Glas, C. A. W. (2001) MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541–561.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*, 385–402.

Billiet, J. , and Welkenhuysen-Gybels, J. (2004). Assessing cross-national construct equivalence in the ESS: The case of six immigration items. *Paper presented at the Sixth International conference on Social Science Methodology, Amsterdam*.

Blanchin, M., Hardouin, J.-B., Le Neel, T., Kubis, G., Blanchard, C., Miraill, E. , and Sbille, V. (2011). Comparison of CTT and IRT based-approach for the analysis of longitudinal Patient Reported Outcome. *Statistics in Medicine,30*, 825–838.

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

Bock, R. D., and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179–197.

Bohlmeijer, E.T., Fledderus, M., Rokx, T. A. J. J. , and Pieterse, M. E. (2011). Efficacy of an early intervention based on Acceptance and Commitment Therapy for adults with depressive symptomatology: Evaluation in a randomized controlled trial. *Behavior Research and Therapy,49*, 62–67.

Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K.C., Guenole, N., Orcutt, H.K., Waltz, T. , and Zettle, R. D. (2011). Preliminary psychometric properties of the acceptance and Action questionnaire-II: A revised measure of psychological flexibility and acceptance. *Behavior Therapy,42*, 676–688.

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

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian inference in statistical inference*. New York: Wiley & Sons.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian Random Effects Model for Testlets *Psychometrika, 64*, 153–168.

Bryk, A. S., and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin,101*, 147–158.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications.

Cai, L., Thissen, D., and du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago, IL: Scientific Software International.

Card, D., Dustmann, C., and Preston, I. (2005). Understanding attitudes to immigration: The migration and minority module of the first European Social Survey. *Center for Research and Analysis of Migration Discussion Paper*.

Chiu, M. M., and Klassen, R. M. (2009). Relations of Mathematics Self-concept and its Calibration with Mathematics Achievement: Cultural Differences Among Fifteen-year-olds in 34 Countries. *Learning and Instruction, in press*.

Cho, S.-J., and Cohen, A. S. (2010). A multilevel mixture IRT model with applications to DIF. *Journal of Educational and Behavioral Statistics, 35*, 336–370.

Cho, S.-J., and Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis,55*, 12–25.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association, 91*, 883-904.

Crane, P. K., Gibbons, L. E., Jolley, L. M. S., and van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care,44*, 115–123.

Crane, P. K., van Belle, G., Larson, and E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.

Davidov, E., Meuleman, B., Billiet, J., and Schmidt, P. (2008). Values and support for immigration: A cross-country comparison. *European Sociological Review, 24*, 583–599.

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

DeMars, C. E., and Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*, 597–616.

De Jong, M. G. and Steenkamp, J. B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*, 3–32.

De Jong, M. G., Steenkamp, J. B. E. M., and Fox, J. -P. (2007). Relaxing Cross-national Measurement Invariance Using a Hierarchical IRT Model. *Journal of Consumer Research, 34*, 260–278.

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J-P., and Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research, 104*, 104–115.

De Leeuw, J., and Kreft, I. G. G. (1986). Random Coefficient Models for Multilevel Analysis. *Journal of Educational and Behavioral Statistics, 11*, 57–86.

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

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association, 96*, 1151–1160.

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

ESS round 1: European social survey round 1 data (2002). Data file edition 6.1. Norwegian Social Science Data Services, Norway - Data Archive and distributor of ESS data.

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

Finkelman, M., Green, J. G., Gruber, M. J., and Zaslavsky, A. M. (2001). A Zero- and K-Inflated Mixture Model for Health Questionnaire Data. *Statistics in Medicine, 30*, 1028–1043.

Forman, E. M., Herbert, J. D., Moitra, E., Yeomans, P. D., and Geller, P. A. (2007). A randomized controlled effectiveness trial of acceptance and commitment therapy and cognitive therapy for anxiety and depression. *Behavior Modification, 31*, 772–799.

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

Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software, 20*, 1–16.

Fox, J.-P. and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271–288.

Fox, J.-P., and Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In: Davidov, E., Schmidt, P. and Billiet, J. (editors), *Cross-cultural analysis: Methods and applications* London: Routeledge Academic, pp. 467–488.

Frederickx, S., Tuerlinckx, F., De Boeck, P., and Magis, D. (2010). RIM: a random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*, 432–457.

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

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid and A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169–193). Oxford: Clarendon Press.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*. New York: Wiley-Interscience.

Glas, C. A. W., and Van der Linden, W. J. (2003). Computerized Adaptive Testing With Item Cloning. *Applied Psychological Measurement, 27*, 247–261.

Glas, C. A. W., van der Linden, W. J., and Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of Adaptive Testing*(pp. 289–314). New York: Springer.

Goldstein, H. (1989) Models for multilevel response variables with an application to growth curves. In: Bock, R. D. (editor), *Multilevel Analysis of Educational Data* San Diego, CA: Academic Press, pp. 107–125.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd ed.). London: Edward Arnold.

Goldstein, H. (2004). International Comparisons of Student Attainment: Some Issues Arising from the PISA Study. *Assessment in Education, 11*, 319–330.

Halberstadt , S. M., Schmitz, K. H., and Sammel, M. D. (2012). A joint latent variable model approach to item reduction and validation. *Biostatistics, 13*, 48–60.

Hambleton, R. K., Merenda, P. F., and Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *(*Biometrika, 57), 97–109.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., and Lillis, J. (2006). Acceptance commitment therapy: Model, process and outcomes. *Behaviour Research and Therapy, 44*, 1–25.

Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hertzog, C., and Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging, 18*, 639–657.

Hoijtink, H. J. A. (2011). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists.*. Boca Raton: Chapman and Hall/CRC.

Hoijtink, H., and Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171–189.

Horn, J. L., and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

ISSP Research Group, International Social Survey Programme (ISSP): National-Identity Data Set, 1995. Distributor: GESIS Cologne Germany, ZA2880.

Janssen, R., Tuerlinckx, F., Meulders, M., and De Boeck, P. (2000). A Hierarchical IRT Model for Criterion-Referenced Measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306.

Jeffreys, H. (1961). *Theory of probability, 3rd.* Oxford: Oxford University Press.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G., and Sörbom, D. (1996). LISREL 8.14. *Chicago, Scientific Software*.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement, 38*, 79–93.

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

Klein Entink, R. H., Fox, J.-P., and van den Hout, A. (2011). A mixture model for the joint analysis of latent developmental trajectories and survival. *Statistics in Medicine, 30*, 2310–2325.

Klugkist, I. (2008). Competing theories based on (in) equality constraints. In H. Hoijtink, I. Klugkist, and P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 53–83). New York: Springer Verlag.

Klugkist, I., and Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis, 51*, 6367–6379.

Lazarsfeld, P. F. (1950). *Measurement and Prediction*. Princeton, NJ: Princeton University Press

Levy, R. (2011). Posterior Predictive Model Checking for Conjunctive Multidimensionality in Item Response Theory. *Journal of Educational and Behavioral Statistics, 36*, 672–694.

Li, Y., and Baser, R. (2012). Using R and Winbugs to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine*

Li, F., Cohen, A. S., Kim, S.-H., and Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*, 353–373.

Liu, L. C. (2008). A model for incomplete multivariate ordinal data. *Statistics in Medicine, 27*, 6299–6309.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307–330.

Malchow-Moller, N., Munch, J. R., Schroll, S., and Skaksen, J. R. (2009). Explaining cross-country differences in attitudes towards immigration in the EU-15. *Social Indicators Research, 91*, 371–390.

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

May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics, 31*, 63–79.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavioral Genetics, 16*, 163-200.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605.

Mchugh, T. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21**, 331–347.

Meade, A. W., Lautenschlager, G. J., and Hecht, J. E. (2005). Establishing measurement equivalence/invariance in longitudinal data with item response theory. *International Journal of Testing, 5*, 279–301.

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

Meuleman, B., Davidov, E., and Billiet, J. (2009). Changing attitudes toward immigration in Europe, 2002–2007: A dynamic group conflict theory approach. *Social Science Research, 38*, 352–365.

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

Meredith, W., and Horn, J. (2001). The role of factorial invariance in modeling growth and change. In: Collins, L. M. and Sayer, A. G. (editors), *New methods for the analysis of change: Decade of behavior* Washington, DC: American Psychological Association, pp. 203-240.

Meredith, W., and Millsap, R. E. (1992). On the Misuse of Manifest Variables in the Detection of Measurement Bias. *Psychometrika, 57*, 289–311.

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-22.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*, 1087–1092

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*, 5–9.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance.* New York, NY: Routledge.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17*, 297–334.

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician, 61*, 22–27.

Mulder, J., Hoijtink, H., and Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 140*, 887–906.

Muraki, E. (1992). A generalized partial credit model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muthén, B., Brown, C. H., Booiljo, K. M., Khoo, S-T, Yang, C.-C., Wang, C-P, Kellam, S. G., Carlin, J. B., and Liao, J. (2002). General Growth Mixture Modeling for randomized preventative interventions. *Biostatistics, 3*, 459-475.

Muthén, L. K. and Muthén, B. O. (2001). Mplus. *Los Angeles, CA: Muthén and Muthén* .

Organization for Economic Cooperation and Development. (2004). *Learning for Tomorrow's World: First Results from PISA 2003.* Paris: OECD Publications.

Organization for Economic Cooperation and Development. (2005). *Pisa Technical Report.* Paris: OECD Publications.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association, 94*.

Patz, R. J., and Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.

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

Rabe-Hesketh, S. P., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69*, 167–190.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Raudenbush, S. W., and Sampson, R. J. (1999). Ecometrics: Toward a Science of Assessing Ecological Settings, With Application to the Systematic Social Observation of Neighborhoods. *Sociological Methodology, 29*, 1–41.

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

Reise, S. P., Widaman, K. F., and Pugh, R. H. (1993). Confirmatory factor-analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Rensvold, R. B., and Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. *Research in Management, 1*, 21–50.

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

Scott, J. G., and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference, 136*, 2144–2162.

Sides, J., and Citrin, J. (2007). European opinion about immigration: The role of identities, interests and information. *British Journal of Political Science, 37*, 477–504.

Sinharay, S., Johnson, M. S., and Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*, 295–313.

Sireci, S. G., Patsula, L., and Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, NJ: Lawrence Erlbaum.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton: Chapman and HallCRC.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage: London.

Soares, T. M., Goncalves, F. B., and Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics, 34*, 348–377.

Spiegelhalter, D. J., Best, N. G., and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Unpublished manuscript. Available from http://yaroslavvb.com/papers/spiegelh bayesian.pdf.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 583–616.

Steenkamp, J. B. E. M., and Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*, 78–90.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 479–498.

Teresi, J. A. (2006). Overview of quantitative measurement methods- Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44*, S39–S49.

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Van der Linden, W. J., and Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer-Nijhoff.

Van de Vijver, F. J. R., and Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Mawhaw, NJ: Lawrence Erlbaum.

Van de Vijver, F. J. R., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 54*, 119–135.

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.

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

Verhagen, A. J., and Fox, J.-P. (2012). Bayesian tests of measurement invariance. *The British Journal of Mathematical and Statistical Psychology.*

Verhagen, A. J., and Fox, J.-P. (in press). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. Accepted for publication in *Statistics in Medicine*.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel datasets. *Statistical Methods in Medical Research, 17*, 33–51.

Von Davier, M., and Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389–406.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive psychology, 60*, 158–189.

Welkenhuysen-Gybels, J., Billiet, J., and Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34*, 702–722.

Wetzels, R., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2010). An Encompassing Prior Generalization of the Savage-Dickey Density Ratio Test. *Computational Statistics & Data Analysis, 54*, 2094–2102.

Wetzels, R., Grasman, R. P. P. P., and Wagenmakers, E.-J. (in press). A default Bayesian hypothesis test for ANOVA designs. *The American statistician*.

Wetzels, R., Raaijmakers, J., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A fl exible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review, 16*, 752–760.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*, 1097–1126.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley-Interscience.

Zwinderman, A. H. (1991). A Generalized Rasch Model for Manifest Predictors. *Psychometrika, 56*, 589–600.

Samenvatting

Testen en vragenlijsten zijn overal om ons heen: cognitieve testen om intelligentie of studiesucces te meten, psychologische vragenlijsten om stoornissen of persoonlijkheidskenmerken vast te stellen en vragenlijsten om achter de wensen van de klant of de mening van het Nederlandse volk te komen. Steeds vaker is er ook interesse in het vergelijken van test of vragenlijst scores tussen verschillende groepen, bijvoorbeeld tussen verschillende landen, tussen groepen van verschillende leeftijden of tussen mannen en vrouwen. Een belangrijke voorwaarde hierbij is dat de vragen het construct in iedere groep op dezelfde manier meten. Formeler gesteld is de belangrijke voorwaarde dat het meetinstrument gelijk functioneert in alle groepen. Wanneer dit niet het geval is, zijn de resulterende scores niet direct vergelijkbaar.

Vooral wanneer men scores wil vergelijken tussen een groot aantal groepen is het lastig om er zeker van te zijn dat al de vragen het onderliggende construct op dezelfde manier meten. Sommen in een wiskunde test kunnen moeilijker zijn voor leerlingen uit een land waarin het curriculum dat specifieke onderwerp niet bevat. Voor mannen is het bevestigend antwoorden op de vraag of ze wel eens huilbuien hebben een relevantere indicatie voor depressie dan voor vrouwen. Wanneer er verschillen zijn tussen groepen in karakteristieken van vragen, zoals de moeilijkheid of de relevantie van een vraag, is er sprake van een gebrek aan meetinvariantie.

In dit proefschrift worden Bayesiaanse Item Respons Theorie (IRT) modellen gepresenteerd die rekening houden met meetvariantie: verschillen tussen groepen in de manier waarop vragen (items) een construct meten. Dit heeft een tweeledig doel. Aan de ene kant worden binnen deze modellen invariantie testen voorgesteld om vragen die niet gelijk functioneren over groepen te identificeren. Aan de andere kant worden deze verschillen opgenomen in het model, zodat ondanks meetvariantie vergelijkbare scores geschat kunnen worden en ook informatie kan worden opgenomen om verschillen te verklaren.

Uitgangspunt voor dit proefschrift is het Bayesiaanse IRT model, ontwikkeld door Fox (2010), waarin voor elke vraag groeps-specifieke item parameters worden gemodelleerd als willekeurige afwijkingen van de algemene item parameters. Door deze manier van modelleren is het niet nodig om zogenoemde "anker items" met gelijke item parameters over groepen te veronderstellen, wat het model flexibeler maakt dan de traditionele meetmodellen. Verder is dit model zeer geschikt voor het modelleren van meetvariantie in een groot aantal groepen.

Dit model is in dit proefschrift op verschillende manieren uitgebreid voor uiteenlopende test situaties. In hoofdstuk 2 wordt beschreven hoe verklarende in-

formatie over personen en groepen meegenomen kan worden bij het schatten van persoonsscores, terwijl geen meetinvariante item parameters verondersteld worden. Op deze manier kunnen betekenisvolle vergelijkingen gemaakt worden tussen groepen, waarbij gecorrigeerd wordt voor achtergrond variabelen. Als voorbeeld wordt gekeken naar de invloed van onder andere tijd besteed aan huiswerk en geslacht op wiskunde scores van middelbare scholieren in het internationale onderwijsonderzoek PISA. Deze scores worden geschat met item parameters die varieren over landen.

In hoofdstuk 3 wordt beschreven hoe op eenzelfde manier ook verklarende informatie op groeps- of itemniveau over verschillen in item parameters tussen groepen meegenomen kan worden. Op deze manier kan meer inzicht worden verkregen in de achtergrond van meetinvariantie. Als voorbeeld wordt gekeken naar de invloed van onder andere het bruto binnenlands product van een land op de relevantie van items in een vragenlijst die de houding ten opzichte van immigranten meet in Europese landen (onderdeel van het European Social Survey (ESS)). Het item over immigranten die banen wegnemen bleek relevanter voor het meten van de houding ten opzichte van immigranten in armere dan in rijkere landen.

In hoofdstuk 4 worden de voorgaande modellen uitgebreid voor longitudinale onderzoeksdata. Op deze manier wordt rekening gehouden met het feit dat personen op een andere manier op items kunnen reageren wanneer ze deze steeds opnieuw moeten beantwoorden. Latente groeitrajecten voor zowel de persoonsscores als voor de item parameters over de tijd worden gemodelleerd. Een andere uitbreiding in hoofdstuk 4 is die van modellen voor dichotome items naar modellen voor items met meerdere antwoordcategorieën. In een onderzoek naar depressie werd van het symptoom "ik kon niet op gang komen" na een aantal metingen (bij gelijk depressie niveau) vaker aangegeven dat dit "1 of 2 keer" ervaren was in de afgelopen week dan dat het "nooit" of "meer dan 2 keer" was ervaren.

Hoofdstuk 5 tenslotte laat zien hoe binnen dit raamwerk ook modellen voor een kleiner aantal groepen kunnen worden ontwikkeld, door de item parameters voor de verschillende groepen geen afwijkingen te laten zijn van algemene item parameters maar van algemene groepsparameters. Ook wordt uitgelegd hoe latente klassen kunnen worden toegevoegd aan de eerder beschreven modellen om items als invariant te kunnen classificeren.

Er is gekozen voor een Bayesiaanse schattingsmethode om deze complexe modellen te kunnen schatten. Door de item parameters met een hiërarchische structuur te modelleren wordt het model vrij eenvoudig te schatten met Markov Chain Monte Carlo methoden.

Een tweede doel van dit proefschrift was het ontwikkelen van Bayesiaanse testen voor meetinvariantie. In hoofdstuk 3 wordt uiteengezet hoe meetinvariantie per item getest kan worden door middel van een Bayes factor test voor geneste modellen, waarin de nulhypothese dat de variantie van de item parameters over groepen gelijk is aan nul wordt getest door te evalueren hoe waarschijnlijk het is dat de variantie kleiner is dan een klein getal δ onder de prior en de posterior verdeling. Een simulatie studie toonde aan dat deze test goed werkt. Daarnaast blijkt een Deviance Information Criterion een goede manier te zijn om het uiteindelijk beste model te bepalen, door de fit van modellen met meer en minder invariante items met elkaar te vergelijken. Een highest posterior density test zoals beschreven in

appendix C is minder succesvol in het detecteren van invariante items.

In hoofdstuk 4 worden de Bayes factor test en de DIC ingezet voor het detecteren van longitudinale invariantie in modellen met multivariaat nomaal verdeelde item parameters voor items met meerdere categorieën. In hoofdstuk 5 wordt een tweede Bayes factor test voor verschillen tussen specifieke item parameters beschreven, en wordt een implementatie van beide Bayes factor testen in WinBUGS en R beschreven. Een simulatie studie waarbij beide methodes worden vergeleken met traditionele maximum likelihood methodes om invariantie te detecteren vond vergelijkbare resultaten. Door de Bayes factor wordt bewijs voor zowel de nulhypothese als de alternatieve hypothese geëvalueerd, in plaats van bewijs tegen de nulhypothese te verzamelen, waardoor een meer gedifferentieerde conclusie kan worden getrokken. Daarnaast is de Bayes factor test voor de variantie component makkelijk te implementeren in uitgebreidere modellen waarbij bijvoorbeeld verklarende informatie over de item parameters is toegevoegd, blijft hij makkelijk te berekenen en te interpreteren bij een groot aantal groepen, en zijn er geen anker items nodig.

Het resultaat van dit proefschrift is een uitgebreid en flexibel raamwerk waarin het mogelijk is te testen of er sprake is van meetinvariantie, maar waarin het ook mogelijk is meetvariantie te modelleren zodat vergelijkbare scores geschat kunnen worden en meer inzicht verkregen kan worden in oorzaken van meetinvariantie.