LUND UNIVERSITY

LUND UNIVERSITY &

UNIVERSITY OF CALIFORNIA,
BERKELEY

MASTER'S THESIS

2016: E48

Construction of the Berkeley Innovation Index:
A Higher-Order Item Response Theory Model Approach

Authors:

Alexander Fred Ojala

Johan Eng Larsson

Supervisors:

Prof. Charlotta Johnsson
Automatic Control, Lund University

Prof. Ikhlaq Sidhu
Operations Research, UC Berkeley

Examiner:

Prof. Andreas Jakobsson
Mathematical Statistics, Lund University

A thesis submitted in fulfilment of the requirements
for the degree of Master of Science at

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

September 2016

LUND UNIVERSITY &
UNIVERSITY OF CALIFORNIA, BERKELEY

Abstract

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Master's Thesis

Construction of the Berkeley Innovation Index:
A Higher-Order Item Response Theory Model Approach

by Johan Eng Larsson and Alexander Fred Ojala

The Berkeley Innovation Index (BII) is a tool developed for assessment of individual innovation capability. The index is based on responses to a survey that constitutes of questions linked to domain abilities, i.e., sub-traits, that are hypothesized to govern an individual's overall innovation ability. The underlying algorithm for the BII produces a score representing the test-takers' proficiency on the domain ability continua as well as a score associated with their general innovation ability. In this thesis, the algorithm for the BII is constructed by applying a Higher-Order Item Response Theory model for hierarchical latent trait estimation. Simultaneous estimation of the vast amount of model parameters is done by employing a Markov Chain Monte Carlo (MCMC) method that utilizes a multi-level bayesian inference sampling technique. The validity, feasibility, and usefulness of the approach is analyzed throughout the thesis. The statistical relevance of the obtained results is evaluated by examining the Deviance Information Criteria, the Item Response Theory Information Criteria, the posterior predictive values, different convergence criteria for the MCMC chains etc. In order to reduce the amount of questions, and make the index more user-friendly, feature selection techniques are applied to explore the possibility of discarding items that contribute with the least amount of information. An easily implementable and scalable algorithm is presented, and the advantages/disadvantages of the acquired model are discussed. Lastly, recommendations on how to further improve the Berkeley Innovation Index are proposed.

Keywords:
Index Construction, Innovation Capability, Item Response Theory, Higher-Order Latent Trait Estimation, Bayesian Inference, Markov Chain Monte Carlo, Feature Selection

To handle failures is an important aspect of an innovative mindset.

— Ancient Swedish entrepreneurial proverb

The Authors during the indian summer of 2010

Acknowledgements

First and foremost we would like to thank our families and friends for their ever-present love and support.

We would also like to express our lifelong gratitude to our supervisor Charlotta Johnsson and our examiner Andreas Jakobsson for their invaluable help and encouragement throughout the whole process of this thesis.

We would also like to sincerely thank Ikhlaq Sidhu for his superb feedback, all the inspiring meetings, and for inviting us to be part of the BII research team. Furthermore, we would like to thank Holger Assenmacher for long and uplifting discussions, Jean-Etienne Goubet for being fun and French, and all the other wonderful people we had the opportunity to collaborate with at The Sutardja Center for Entrepreneurship & Technology at the University of California, Berkeley.

We would also like to acknowledge the indispensable help that we received from Jonathan Templin (Educational Psychology, The University of Kansas) and Hung-Yu Huang (Psychology, University of Taipei) when they explained the contemporary aspects of Item Response Theory. This work would not have been possible without their support. Moreover, we would like to thank Magnus Wiktorsson for his review of the MCMC chapter and Stefan Ingi Andalbjörnsson for his help with the variable reduction.

Finally, Alexander wants to thank Johan for his magical work! And, Johan wants to thank Alexander for being a wizard as well!

Contents

Abbreviations

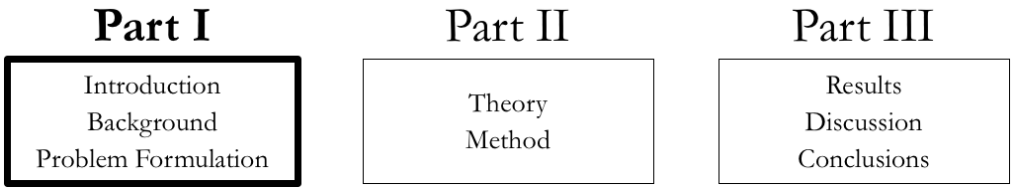
AIC	**A**kaike **I**nformation **C**riteria
BIC	**B**ayesian **I**nformation **C**riteria
BII	**B**erkeley **I**nnovation **I**ndex
CFA	**C**onfirmatory **F**actor **A**nalysis
CTT	**C**lassical **T**est **T**heory
CV	**C**oefficient of **V**ariation
DIC	**D**eviance **I**nformation **C**riteria
EFA	**E**xploratory **F**actor **A**analysis
EM	**E**xpectation-**M**aximization
GBM	**G**radient **B**oosting **M**achine
GPCM	**G**eneralized **P**artial **C**redit **M**odel
G-R	**G**elman-**R**ubin
GRM	**G**raded **R**esponse **M**odel
HO	**H**igher **O**rder
ICC	**I**tem **C**haracteristic **C**urve
IIF	**I**tem **I**nformation **F**unction
ICIC	**I**tem **C**ategory **I**nformation **C**riteria
IRT	**I**tem **R**esponse **T**heory
JAGS	**J**ust **A**nother **G**ibbs **S**ampler
ML	**M**aximum **L**ikelihood
MML	**M**arginal **M**aximum **L**ikelihood
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MZSA	**M**aximum **Z** score among **S**hadow **A**ttributes
PCA	**P**rincipal **C**omponent **A**nalysis
PL	**P**arameter **L**ogistic

PPC	**P**osterior **P**redictive **C**heck
PSRF	**P**otential **S**cale **R**eduction **F**actor
RMSE	**R**oot **M**ean **S**quare **E**rror
SD	**S**tandard **D**eviation
SE	**S**tandard **E**rror

Dedicated to all our loved ones — past, present, and future.

Part I

Measuring the Unmeasurable



Key takeaways: In *Part I* the importance of measuring individual innovation capability is presented. Different innovation capability measures are introduced and the past research on the Berkeley Innovation Index is highlighted. Furthermore, the requirements for the new Berkeley Innovation Index algorithm are postulated.

Chapter 1

Introduction

In the beginning...
—God (Gen.1:1)

In today's society, innovation is regarded as one of the key drivers for economic growth, it is the foundation for technological progress and the ability to innovate is one of the main competitive advantages for both individuals and organizations.

In November 2015, Brian Quinn (Senior Contributor at Forbes Magazine) wrote an article with the title *Why Measuring Innovation Matters*. In the article, Quinn highlights the old management saying "What gets measured gets done", i.e., if you can set a target and measure progress, then any type of goal will be attainable.

Quinn writes: "Without measuring these things [innovation], we're effectively driving without headlights — faintly hoping once again that innovation will deliver something useful rather than demanding it, and holding ourselves accountable for achieving it."

However, innovation is quite an abstract and vague concept. It is difficult to precisely define innovation and even more difficult to measure innovation capability. Therefore, in order to assess innovation capability one needs to define what it is, in what context it is applied, and what it ultimately is set out to deliver.

This thesis builds upon innovation research that has been carried out at the Sutardja Center for Entrepreneurship and Technology (SCET) at the University of California, Berkeley. At the SCET an international research group have identified characteristics linked to an individual's level of *innovation capability*. The research group has also developed a survey, composed of a set of questions, that measures individual innovation capability. The final product of this research is an index called the *Berkeley Innovation Index* (BII) which enables individuals to assess their level of innovation capability.

However, in order for the index to be relevant, there has to be a theoretically valid and statistically relevant algorithm that calculates the index scores with the use of answers given to the survey questions. The BII algorithm is what we aim to construct in this thesis.

1.1 Scope of the Thesis

In this thesis, we set out to quantify and validate the innovation capability scores given by the BII. We aim to construct an algorithm that is valid according to modern standards in statistics, psychometrics and applied mathematics. The algorithm should be scalable, easy to implement, and customizable so that future iterations of the questionnaire and new entries in the data set also can make use of the findings in this thesis.

As a result, we want to present an algorithm that can be used to evaluate an individual's BII scores instantly. We will also analyze the relevance and precision of the measures obtained as well as give recommendations on how the accuracy of the results can be improved.

1.2 Thesis Outline

The thesis is structured into three parts, as described below:

Part I contains the *Introduction*, *Background*, and *Problem Formulation* chapters. In this part, the topic of innovation assessment is presented. We summarize the past research that has been done and present the foundation for the thesis. Then, we state the problems we will try to solve as well as the limitations of the work.

Part II contains the *Theory* and the *Method* chapters. These two chapters have been merged in order to enhance the reading experience. Here, we will present the theory behind the methods utilized to construct the index as well as how we have applied the methods to our specific case. In this part, we will also derive some pre-results needed in order to conduct the main analysis.

Part III contains the *Results*, *Discussion*, and *Conclusions* chapters. Here, we present our findings and give our subjective view on the results. We will also give recommendations on how the BII can be improved.

Chapter 2

Background

Just as energy is the basis of life itself, and ideas the source of innovation, so is innovation the vital spark of all human change, improvement and progress.

—Prof. Theodore Levitt

2.1 The Importance of Innovation

Innovation capability is widely regarded as one of the most important assets for a company, an employee, a university or any type of organization to have in order for them to compete on the global market. Furthermore, innovation can be seen as the ability to come up with original ideas, be creative or to act as a pioneer. Innovation is accomplished through more effective products, processes, services, technologies, and/or business models. Rapid changes in market needs and the constantly evolving technological landscape pressure entities to present and implement novel solutions to both new and old problems in order to stay ahead of their competitors. Therefore, without the ability to be innovative an organization or an individual experience stagnation, and in a world that encourages constant progress the lack of innovation capability becomes a major disadvantage (Sidhu et al., 2016a).

Hana (2013) identified the most important characteristics of innovations in an organization as:

- A strong relationship between market performance and new products. New products help maintain market shares and improve profitability.
- Growth also by means of non-price factors (design, quality, individualisation, etc.).
- Ability to substitute outdated products (shortening product life cycles).
- Innovation of processes that lead to production time shortening and speed up new product development in comparison to competitors.

Developing successful technological innovations is essential for creating and sustaining competitive advantage and the expenditures on research and development alongside the ability to introduce innovations are some of the determining characteristics for gaining a dominant part of the market share. Also, if an organization is not capable of introducing innovations on an ongoing basis, it risks that it will lag behind in the competition and the initiative will be taken over by other market actors (Hana, 2013).

The fact that organizations have identified the importance of innovation is also reflected in the increase of global spending on Research and Development (R&D). Global R&D investments have increased significantly between the years 2005 to 2014 (with the only exception being 2010), see Figure 2.1

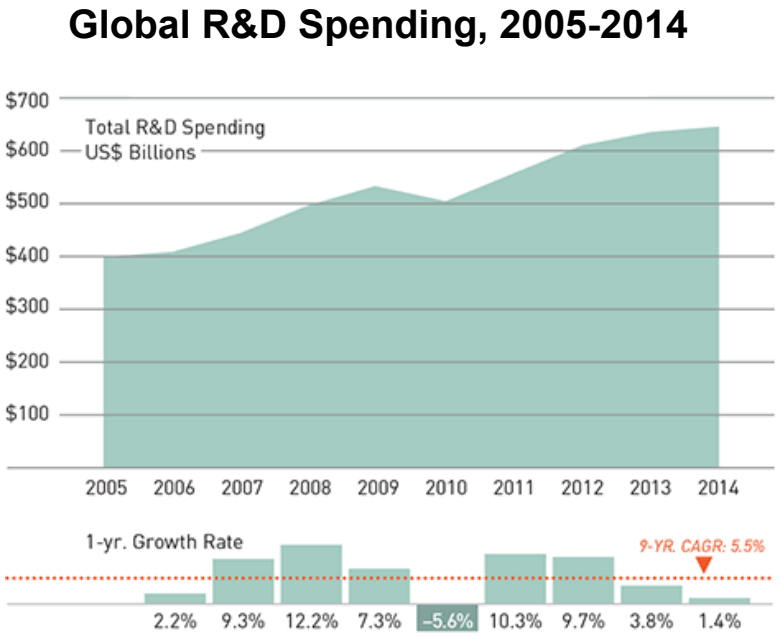


FIGURE 2.1: Global R&D Spending from 2005 to 2014.
Source: Bloomberg Data, Capital IQ data, Strategy and analysis

Innovation is recognized as a variable that organizations and individuals use as an asset, a competitive advantage and a factor that drives economic growth. Therefore, there is an evident need to be able to measure the degree of innovation in regards to a project, a work-group and an individual. This in order to identify behaviors and strategies that can be implemented in order to increase innovation capability, and develop innovative mindsets and cultures.

Martín-de Castro et al. (2013) found that the foundation for all innovation processes in an organization is brilliant, motivated, experienced, and creative employees. The innovation process is generally a collective achievement of the organization's members. An organization with a thriving innovation culture is critical in order for the firm to pursue technological advancements and constitutes the best incentive towards obtaining new knowledge and achieving innovations (Martín-de Castro et al., 2013).

Hence, the natural questions that arise are:
How can innovation, and especially innovation capabilitiy be measured?
What are the aspects of innovation that can be assessed, evaluated, and then improved?

2.2 Past Measures of Innovation

Innovation is a rather vague and abstract concept, therefore it is a difficult task to scientifically measure or quantify innovation capability. However, a broad range of metrics, scales, and indices are and have been used in order to quantify innovation capability and innovation performance of an *organization*, i.e., on a firm level. Indicators of a high degree of innovation could be the number of patents filed in a year, relative increase in labor productivity, R&D investments, revenues due to products launched in the last three years etc. However, most past innovation measures have not been insightful or holistic enough to help companies make the right decisions in order to become more innovative (Sidhu et al., 2016b). For example, the numbers of patents filed or the amount of money spent on R&D have not shown any significant relationship with an organizations' ability to be innovative nor to make profits (Jaffe, 1986). Overall, the existing variables for innovation assessment mostly use a quantitative approach inherited from financial analysis. Almost none of the current innovation metrics, frequently used by companies today, take the individuals' innovation capabilities into account (Sidhu et al., 2016a).

Even though not frequently used, there currently exists models and measures that attempt to quantify and assess individual innovation capability. A critical review of these measures were done by Menold et al. (2014) and they found that a comprehensive, rigorously validated psychometric instrument does not yet exist to assess the aptitudes, skills,

knowledge, personal traits, and behaviors that are indicative of individual innovation capability. In their study they state that innovativeness studies in personality research tend to evaluate individual innovativeness in terms of innate traits and characteristics. In the study they examined three measures of general, individual innovation capability:

- **Kirton's Adaption-Innovation Inventory (KAI)**, a psychometric instrument that exclusively measures an individual's cognitive style (Kirton, 2004).
- **The Jackson Personality Index (JPI)** assesses Innovation and Risk-Taking on two of its subscales. The JPI is a measure of attributes, or innate personality characteristics, mainly focusing on cognitive style (Jackson and Paunonen, 1996).
- **Hunter et al.'s Model of Innovativeness** analyzes individual personality traits from a human resources perspective based on the organizational standpoint of hiring innovative individuals to increase a company's overall innovativeness. Hunter et al. defined innovativeness in terms of innovative output (Hunter et al., 2012).

In summary, the general innovativeness instruments listed above either measured only internal attributes over external actions or, in the case of Hunter's model, did not provide a measurement instrument, but only descriptions. Menold et al. (2014) concluded that general measures of innovativeness do exist, but they fail to assess domain-specific traits, skills, knowledge, and behaviors. Hence, while valuable, these general measures of innovativeness are insufficient for the assessment of individual innovation capability.

Sidhu et al. (2016a) also notes that metrics derived from existing innovation assessment tools are often past-oriented and do not give a correct overview of the ability to be innovative in the future.

Evidently there is a need for a novel instrument that measures individual innovation capability.

2.3 The Berkeley Innovation Index

An international research team at the Sutardja Center for Entrepreneurship and Technology, UC Berkeley, recognized the need for a new tool that could measure individuals' overall innovation capabilities. The research team started to develop a new metric that aimed to assess overall innovation capability as well as six sub-traits linked to innovation (e.g., an individual's level of trust, resilience, perfection etc.). The measures are collectively called the Berkeley Innovation Index (BII).

Mindset and Description	***Psychological Construct***	***Trait***
Friend or foe Learn to trust others without expecting anything in return.	Social cohesion, honest behaviour	**Trust**
Plan to fail It is necessary to be wrong sometimes. Plan to experiment. Plan to fail (Fail fast). Analyze, adapt and repeat. The smarter you think you are, the harder this is going to be.	Grit, resilience, entrepreneurial failure	**Resilience**
Diversify Diversify your networks. Connect to people you would not normally, then go and listen, open up, and connect them to others.	Social capital	**Diversity**
Believe Believe that what you do can change the world	Self-efficacy	**Belief**
Good Enough Perfection is not good, but good enough is perfect.	Perfectionism	**Perfection**
Collaboration Individual vs. team and competitors vs. partners.	Cooperation	**Collaboration**

TABLE 2.1: List of the six traits identified to characterize the mindset of an innovative individual

The BII is based on previous findings in the fields of psychology, entrepreneurship, and innovation science. Sidhu et al. (2015) defines entrepreneurship as the act of combining resources in a novel way in order to create new industries and generate wealth. Sidhu et al. (2015) also presents the hypothesis that the mindset of entrepreneurship can be described by ten behavioral patterns (traits) and states that an entrepreneurial mindset and way of action is well correlated with being innovative. Menold et al. (2014) states that an entrepreneurial mindset constitutes of three sub-dimensions, namely: innovativeness, risk-taking, and pro-activeness. The BII aims at deconstructing the sub-dimensions of innovation capability. Six of the ten traits identified with an entrepreneurial mindset in Sidhu et al. (2015) were also found to be connected to a measurable psychological construct linked to an innovative mindset.

It is these six personality traits, and their associated psychological construct, that are used as the basis for the construction of the BII. The traits, together with their respective

psychological construct and mindset dimension (presented in Sidhu et al. (2015)), are shown in Table 2.1.

Based on these six psychological constructs the BII research group developed a survey, in the form of a questionnaire, that aims to measure and quantify each of these traits. The questionnaire was validated from a literature review that presented findings from the fields of social and organizational psychology (Sidhu et al., 2016b).

The BII is based on the assumption that individual innovation capability is something that can be improved by practice, i.e., it is not a static or constant trait, but a skill that can be refined and perfected. Therefore, the BII can be used to measure an individual's degree of innovation capability over time and track if any improvements have been made.

The first iteration of the BII algorithm can be found in Sidhu et al. (2016b). The algorithm calculates quantitative scores for each of the six domain abilities linked to overall innovation capability as well as an overall score indicating an individual's general innovation capability. However, this first iteration of the BII algorithm was not sufficient to make a scientifically valid analysis of an individual's innovation capability. The reason is that the first BII algorithm was built upon the test-takers' self assessed level of innovation capability, and it failed to take into account that different sub-traits might influence the overall innovation score differently etc.

As the current research on the BII is in its cradle, the underlying algorithm, the models, and the data structure need to be improved. The aim of this thesis is to construct a theoretically valid algorithm for the BII, such that it becomes a useful psychometric evaluation tool that assesses individuals' innovation capability.

2.4 Personality Assessment

Innovation, like other abstract concepts or mental properties e.g., intelligence, love, or kindness, is difficult to quantify. This is partly due to the fact that there is no exact definition of the concepts, but even if it was, there is no way to directly measure it as you can with for example mass or temperature. These traits are in that sense hidden, or latent. The desire to define and measure these types of concepts has been displayed numerous times with one example being the 80's band *Foreigner* and their despair in not knowing what love is. They asked an arbitrary person (called *you*) to show it, however what they do not mention is that actual progress to define love has been accomplished by researchers in the fields of test theory and psychometric evaluation.

One of the first widely used statistical approaches to measure latent traits is called *Classical Test Theory* (CTT) (Spearman, 1904). CTT evaluates a test taker's performance on a test and is based on the proposition that a test taker has an observed score and a true score and, however due to a random error an observable test score is not the true value of a subject's performance on the test. The main purpose of CTT is to determine in what degree the test scores are influenced by this random error, and in turn be able to more accurately measure how much of a specific personality trait that an individual possesses.

CTT has several limitations, and one of the most important ones being that test characteristics and examinee characteristics cannot be separated. This means that the test score of an individual is dependent on the other subjects taking the same test.

In order to address the limitations of CTT, a new method for the design, analysis, and scoring of tests and questionnaires called Item Response Theory (IRT) was developed (Lord (1953), Rasch (1960) et al.). CTT focuses on scoring the entire test by treating every question the same. IRT on the other hand focuses on the individual questions of the test, and IRT does not assume that every question is equally difficult.

The One Parameter Logistic (1PL) model (Rasch, 1960) was the first IRT model developed and it was used to compute latent traits from tests with binary item responses, i.e., questions with only two possible answers (oftentimes one answer is correct and one answer is wrong). Since then a framework of models that attempt to explain the connection between different types of observed item responses and an underlying construct, i.e., the latent trait, have been developed. Item responses can be discrete or continuous; be dichotomously (binary) or polytomously (more than two possible responses) scored; there can be one or many abilities that measure the test performance of the subject and there are many ways, i.e., models, in which the relationship between item responses and the latent trait(s) can be specified (Hambleton and Jones, 1993). Common for all IRT models is that they define mathematical relationships between a person's true ability and the person's probability of giving a certain response to an item.

The two IRT models mainly used in this thesis are the *Graded Response Model* (GRM) presented by Samejima (1969) and the *Generalized Partial Credit Model* (GPCM) introduced by Muraki (1992). Both these models allow for use of polytomously scored item responses and the models map categorical responses to a continuous latent trait scale.

The Berkeley Innovation Index is constructed under the assumption that the level of general innovation capability can be viewed as a combination of an individuals' ability in six different sub-domains (i.e., the traits presented in Table 2.1). For the BII it is

desirable to not only measure the overall trait, innovation capability, but also present the test-takers with scores indicating their ability in each of these sub-domains.

This type of hierarchical structure of the abilities is modeled with the use of *Higher-Order Item Response Theory* (HO-IRT), in which the scores for all the traits are estimated simultaneously. The work in this thesis is primarily inspired by the HO-IRT models developed in de la Torre and Song (2009) and Huang et al. (2013).

2.5 The BII Data Set

The data used as the foundation for the analysis in this thesis is the questionnaire responses to the BII survey. The responses were collected between November 2015 to March 2016. The survey was conducted online through the website `https://berkeleyinnovationindex.org/` and the full questionnaire can be found in Appendix A. The full data set constitutes of responses from 1029 test takers, or *subjects*. The questionnaire is constructed such that each sub-trait have four questions, or items, directly related to that specific domain ability.

The test takers also answered questions regarding their level of Comfort Zone, their Say-Do-Ratio, their level of past success in business, and their perceived level of innovation capability. Since these measures are not included in any of the psychological constructs presented earlier they will be disregarded when constructing the index. However, it provides material for further studies and extensions of the BII.

The questionnaire also collects demographic statistics related to the test taker's age, gender, country of origin, field of study/work, and if the subject works or are still in school. The demography of the BII data set is presented in table 2.2.

Age	
< 29	69%
$29 - 40$	16%
> 40	15%

Field of Study/Work	
Technical	48%
Management	37%
Arts/Humanitarian	15%

Geography	
North America	62%
Europe	21%
Rest of the World	17%

Gender	
Male	67%
Female	33%

Career stage	
School	65%
Work	35%

TABLE 2.2: Demography of the BII data set

Chapter 3

Problem Formulation

> *The mere formulation of a problem is far more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science.*
>
> —Albert Einstein

The aim of this thesis is to construct an algorithm, computing the scores of the BII, that has full theoretical support. We want to construct an index that is a valid measure of an individual's overall innovation capability, henceforth called *innovation*. We also want to construct six sub-scales indicating the individual's proficency in each of the six sub-trait domains, namely: *trust*, *resilience*, *diversity*, *belief*, *perfection* and *collaboration*.

3.1 Requirements of the Index

Since an index or a scale can be presented in many different (often arbitrary) ways the following requirements on the index and the sub-scales are postulated:

- The index and the sub-scales should be **ordinal**, i.e., a higher score reflects a higher level of proficiency on the latent trait continuum.
- The results of the index and the sub-scales should be cast on a **scale ranging from 1-10**.
- The index and the sub-scales should be **continuous**, or at least appear continuous if the scale only allows scores from a (very large) finite set of discrete scores.

- Extreme results, i.e., **the worst/lowest score (1) and the best/highest score (10) should be possible to obtain** if the subjects either have answered all the questions right or wrong.
- The scale should be **robust against outliers and bad samples**, e.g., subjects that have not given trustworthy or typical answers should not (greatly) influence the resulting output of the algorithm.
- Different items should be able to have a varying degree of impact on the sub-scores and the overall score. The score in the different sub-trait categories should also be able to impact the overall ability differently, i.e., the **weight of each item and sub-trait category must not be equal to the others**.

3.2 Analysis of Model Relevance and Accuracy

The relevance of the models, and consecutive results, will be evaluated through different measures of model fit, and the models will be validated in regards to model assumptions postulated in their respective theoretical frameworks. Moreover, the accuracy of the algorithm will be evaluated through different test-statistics suitable for the tools used in the analysis.

3.3 Scalable Algorithm

The final algorithm, that is obtained after the full analysis has been conducted, should be (easily) implementable and the steps carried out in order to obtain index and sub-scale scores should be reproducible. Also, the final step in the algorithm, that produces the scores for the overall trait and the sub-traits, should be of low computational complexity such that a test taker's final score can be computed instantly. This allows for the algorithm to be easily modified when the data set gets larger or if the data structure becomes more complex in a future iteration of the BII.

3.4 Variable Reduction

Test takers are more willing to participate in and complete a shorter survey compared to a more comprehensive one. Therefore, one aim of the thesis is to identify the questions/items that contain low amounts of information, i.e., items that have a low correlation

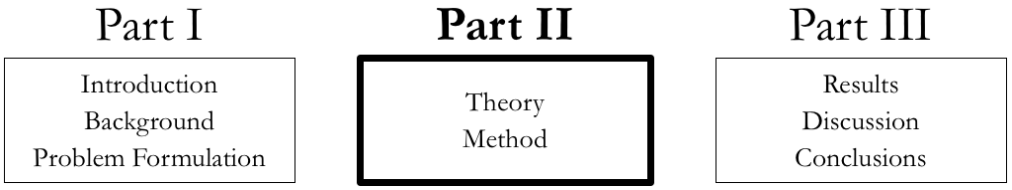
with the final scores for the overall trait and the sub-traits. Then, we want to present a recommendation if these items are to be eliminated from the questionnaire or not.

3.5 Formulate Recommendations to the BII Research Team

After the whole analysis has been conducted and the algorithm for the BII has been formulated, we aim to present our recommendations on how to improve the BII, i.e., how to make the BII as scientifically valid as possible, what the BII research team can do to improve the accuracy of their results, and what general improvements that can be made in light of the findings in the thesis.

Part II

Construction of the Berkeley Innovation Index Algorithm



Key takeaways: In *Part II* theoretical frameworks and methods are presented. Several latent trait models are introduced in order to determine what model that is applicable to the BII case. Simultaneous latent trait estimation is made possible with the use of *Higher-Order Item Response Theory* (HO-IRT) and the model parameters can be estimated through Markov Chain Monte Carlo. To assess the fit, accuracy, and relevancy of the models several error and convergence statistics are defined. Lastly, in order to (possibly) omit items/questions with low amounts of information, two feature selection algorithms are presented.

Chapter 4

An MCMC HO-IRT Approach to Measure Innovation

In the first place, the best way to convey to the experimenter what the data tell him about the model parameter(s), θ, is to show him a picture of the posterior distribution

—Prof. George E.P. Box and Prof. George C. Tiao

4.1 Data Preparation

In order for the algorithm to be as valid as possible it is important that the data set consists of "true" samples from individuals that have answered the questionnaire. Therefore, the data set was examined and cleaned from bad samples before the full analysis was conducted.

The total data set consists of 1029 samples where each subject, $s_{tot} = 1, ..., 1029$, have responded to 24 test items, $i = 1, ...24$. For each item, the response was given on a Likert scale, ranging from one to five, where each possible answer is mapped to an integer, as described below.

Integer values mapped to the responses in the BII questionnaire

1 = *Completely Disagree*

2 = *Disagree*

3 = *Don't Know*

4 = *Agree*

5 = *Completely Agree*

For 20 out of the 24 questions the answer **5** = *Completely Agree* represents the "correct" response, the four remaining questions are reversed and a "correct" response on those questions is given by answering **1** = *Completely Disagree*. Each subject's response to the 24 items constitutes the *full item response matrix*, $X^{tot}_{s_{tot},i} \in \mathbb{N}^{1029\times 24}$.

Removing bad samples

First, the data set was cleaned from bad samples. There are two categories of bad samples in the BII data set: *duplicate answers* and *bad response patterns*.

The first type of bad samples are duplicate or updated answers. These are responses submitted by the same individual either by mishap or to tweak their answers to achieve a higher index score. These samples can be identified in the data set by looking for duplicate entries in the email address column. Even though roughly 14% of the full data set are duplicate entries and more samples will create a more robust model, these entries were deemed to affect the result in a negative. This is due to the fact that an individual only can have a *true* ability level at one time instant. Therefore, the 147 responses provided by individuals already in the data set were removed before the analysis was conducted. However, it should be noted that, over time a person can improve his or her BII scores.

The second type of bad samples are response patterns where the subject has given the same response to every question in the questionnaire. These response patterns are deemed very unlikely, and they are probably provided by individuals who only wants to complete the test quickly in order to get a result. In the BII data set two subjects only responded **3** = *Don't Know* to each question and two subjects chose the answer **5** = *Completely Agree* to all 24 questions. These four samples were removed from the data set.

In total, after removing duplicate entries and improbable response patterns, the *item response matrix*, $\boldsymbol{X} \in \mathbb{N}^{878\times 24}$, used in the analysis consists of item responses from 878 subjects, $s = 1, ...878$. Each element, $x_{s,i}$, in $\boldsymbol{X}$ represents a response from subject s to item i.

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,24} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,24} \\ \vdots & \vdots & \ddots & \vdots \\ x_{878,1} & x_{878,2} & \cdots & x_{878,24} \end{pmatrix} \tag{4.1}$$

Handling NA entries in the data set

Some of the first subjects that provided answers to the test have left some questions unanswered (this was before a restriction was implemented that disabled the possibility of leaving a question unanswered). In total there are 45 subjects who have left at least one question unanswered. The subject that has left the most questions unanswered did not provide a response to 8 items. In total there were 78 item responses missing in the item response matrix $\boldsymbol{X}$. Although the method used to construct the index can handle missing responses the 78 unanswered questions, labeled NA in the data set, were mapped to the response **3** = *Don't know* to facilitate the analysis procedure.

Reversed items

Four of the items in the questionnaire, *QT2*, *QT4*, *QP1* and *QP3*, are *reversed* in relation to the other items. This means that it is assumed that the response indicating a high level of related ability for these questions is **1** = *Completely Disagree* (for all other items high proficiency is indicated by the response **5**=*Completely Agree*). To facilitate the construction of the index the item responses for these four questions were reversed, i.e. $[1, 2, 3, 4, 5] \rightarrow [5, 4, 3, 2, 1]$.

4.2 Latent Trait Models

Latent variables are hidden variables that are not directly observable, but instead inferred through other measurable variables. This means that the effect of the latent variable(s) can only be measured through observable *manifest variables*. In the case of the BII, the manifest variables are the responses given to the survey items, i.e., $x_{s,i}$. The basic idea with a latent trait model is to establish a relationship between the manifest and the latent variables to enable estimation of the latent variables.

There are four different types of latent variable models depending on if the manifest and the latent variables are continuous or categorical. The different types of latent variable models are summarized in Table 4.1.

Latent Variables	**Manifest Variables** *Continuous*	*Categorical*
Continuous	Factor Analysis	Item Response Theory
Categorical	Latent Profile Analysis	Latent Class Analysis

TABLE 4.1: Different types of Latent Trait Models

As described in section 3.1 of the *Problem Formulation*, the resulting indices should be continuous, hence the latent variables are continuous. The item responses from the questionnaire are given on a Likert scale, i.e., integers $x_{s,i}$ from the set $x_{s,i} \in \{1, 2, 3, 4, 5\}$, and these can be seen as both continuous and categorical. Therefore, the possible Latent Trait Models to use for the analysis of the BII data are Factor Analysis and *Item Response Theory*.

Factor Analysis is a linear model and thus a one-unit change in the latent variable relates to the same increase in the expected response. Implicitly it is therefore assumed that the response is a continuous variable. Since a Likert scale can, even though only five different responses are possible, be viewed as a continuous scale on the interval $I \in [1, 5]$, a Factor Analysis model could be used for the analysis. However, Factor Analysis can only describe a linear relationship between the item responses and the latent trait and, as McDonald (2013) noted, when applied to categorical responses it is merely a linear approximation of the non-linear Item Response Theory model. Thus, an Item Response Theory model is be better suited for the BII analysis.

4.2.1 Item Response Theory

Item Response Theory (IRT) models are a class of statistical models used to describe the relationship between a latent trait and the probability of certain responses to categorically scored items.

It is reasonable to believe that each test subject possess a certain level of the latent trait and that a higher level of this trait will generate a higher test score. Another way to look at this is to see that a person with a higher level of the latent trait also must have a higher probability of giving the "correct" item response to an item.

IRT models the response of a subject to an item with the item characteristic curve (ICC). For the IRT models with binary item responses the ICC is a monotonically increasing probability function that gives the probability of a subject answering an item correctly, given a certain level of the latent trait. A higher value of the subject's latent trait score, the greater the probability is that the subject answers the item correctly. The IRT model was originally developed using normal ogives as the ICC (Ferguson, 1942), but this

function was too computationally demanding in the 1960's so the standard was changed to use the similar logistic model, i.e., $f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$. Due to its simplicity the logistic function has become the preferred function to work with in IRT.

The most simple logistic IRT model is the 1PL model that is used for dichotomously scored data, i.e., binary responses where $Y = 1$ is the correct answer and $Y = 0$ is the wrong answer. This is also called the Rasch model named after the Danish mathematician George Rasch. The 1PL IRT model is given by

$$P_i(Y = 1|\theta_s) = \frac{1}{1 + e^{-\alpha(\theta-\beta_i)}} \tag{4.2}$$

Here, $P_i(Y = 1|\theta_s)$ is the probability that a subject s with ability level θ_s answers item i correctly. Here, α is called the *discrimination parameter* and in the 1PL model it is constant for all items. Thus, α is as a constant scaling factor usually set to 1 (or 1.7 if one wants to produce similar results to the ones obtained through the normal ogive model). Furthermore, b_i is the *item difficulty parameter* for item i.

An extension of the 1PL model is the 2PL model defined as

$$P_i(Y = 1|\theta_s) = \frac{1}{1 + e^{-\alpha_i(\theta-\beta_i)}} \tag{4.3}$$

In the 2PL model the 1PL model is extended by introducing a unique discrimination parameter for each item, α_i, which affects the slope of the ICC. For an IRT model with a binary item responses, different values of the parameters α_i and β_i affect the shape of the ICC as shown in Figure 4.1.

β_i indicates the value on the ability axis where the subject has a 50% chance to answer an item correctly, i.e., $P(Y = 1|\theta_s) = 0.5$. A higher value on β_i will result in a shift of the curve along the x-axis and thus the probability of giving the correct item response will be lowered over the whole latent trait interval. The α_i parameter determines the speed of the transition between low and high probability of "success". A higher value on α_i value will result in a steeper slope of the probability curve and thus a more sudden shift between low/high probability of answering the item correctly.

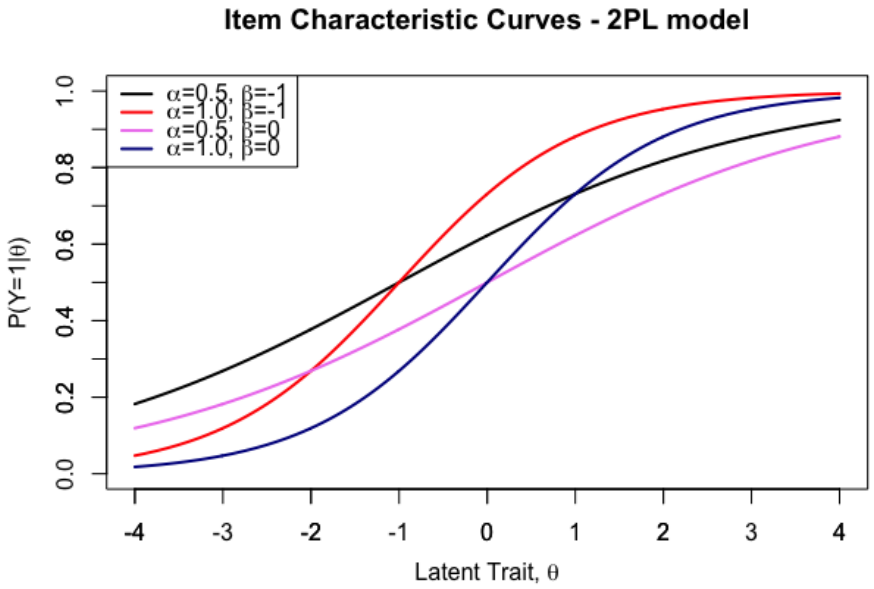


FIGURE 4.1: The Item Characteristic Curve for the 2PL model, with varying discrimination and difficulty parameters (α_i and β_i).

When working with IRT models one has to make sure that certain model assumptions are met in order obtain a valid measure of the latent trait(s). If the model assumptions are violated then the IRT estimates will not be trustworthy. The fundamental IRT model assumptions are listed below.

IRT Model Assumptions

Experimental Independence: The answers from each subject, s, are independent in regards to the answers given by other subjects (Lord et al., 1968)

Local Independence: Given θ for a subject, that subject's item responses are independent from one another. This means that the observed item scores are conditionally independent of each other given an individual's ability level.

Parameter Invariance: Item parameters are invariant over samples of subjects. The latent trait(s) are also invariant over test items.

The 1PL and 2PL models above only explain the relationship between latent trait and item responses for dichotomously scored responses (binary data). In the BII the responses are polytomously scored, given on an ordinal Likert scale which present the test taker

with five possible responses, $k = 1, 2, 3, 4, 5$. To handle item responses with polytomous data a number of models have been developed. One main difference in these models, compared to the binary ones, is that each item response has a unique ICC that represents its probability of over the latent trait continuum.

The two polytomous IRT models that lie within the scope of this thesis are the *Graded Response Model* (GRM) developed by Samejima (1969) and the *Generalized Partial Credit Model* (GPCM) developed by Muraki (1992). The main difference between the two models is that the GRM requires the possible item responses to be ordinal, i.e., that the responses follow a pattern where a higher level of latent trait implies an item response of higher order. The GPCM is more general in the sense that it does not impose the restriction that the item responses need to be ordered.

One of the goals with the thesis is to compare these two models to see which model is most applicable to the BII. The results produced by the GPCM and the GRM will generally agree very closely, unless one or more of the possible item responses are underused (Templin, 2014). Another difference is that the GRM will force the categories' boundary parameters to be ordered, the GPCM does not. The comparison between the two models will also be a way of confirming the goodness of fit in regards to the results obtained. For the reasons noted above, comparing the results of the GRM and the GPCM will further validate our results and the final algorithm.

4.2.2 Graded Response Model (GRM)

The Graded Response Model was derived by Samejima (1969). The basic idea of the model is to make use of the ordinal structure of the item responses by applying the 2PL model at each category boundary, i.e., each possible item response.

For each subject there is a response pattern $X_s = (x_1, x_2, ...x_n)$. x_i is a response to item i, where $i = 1, ..., n$. In the case of the BII, $n = 24$ and the responses are given on a Likert scale and therefore the response pattern can be expressed as a sequence of integers, i.e. $x_i \in \{1, 2, 3, 4, 5\}$.

Subjects with different ability levels have varying degrees of probability to give a certain response to item i. The operating characteristic is the probability of choosing a response, x_i, to an item i given a certain level of latent trait, θ. We define the operating characteristic as

$$P_{x_i}(\theta) = P(x_i|\theta) \tag{4.4}$$

To derive the operating characteristic we will first make use of the ordinal structure of the item responses by applying the 2PL model to each possible item response. $k_i = k = 1, 2, ..., 5$ denotes possible item responses to item i (the subscript i on k can be dropped, since each item in the BII data set has five possible item responses). Note the difference between k_i, which is the possible item responses, and x_i, which are the responses given by a subject. By describing each possible response as the binary probability of either $< k$ or $\geq k$ we can transform the problem to a set of linear combinations of the 2PL model. The cumulative probability of giving an item response greater or equal to k is given by

$$P^*_{i_k} = P(x_i \geq k|\theta) = \frac{e^{\alpha_i(\theta-\beta_{i_k})}}{1 + e^{\alpha_i(\theta-\beta_{i_k})}} \tag{4.5}$$

where x_i is the response given to item i, k is a possible item response, α_i is the discrimination parameter for item i, β_{i_k} is the difficulty parameter for response k to item i and θ the latent trait.

The probability for a subject, given θ, of responding $x_i = k$ on a given item, i, is obtained by subtracting the cumulative probability for that item response, k, with the cumulative probability of responses greater than k ($k' > k$), i.e.,

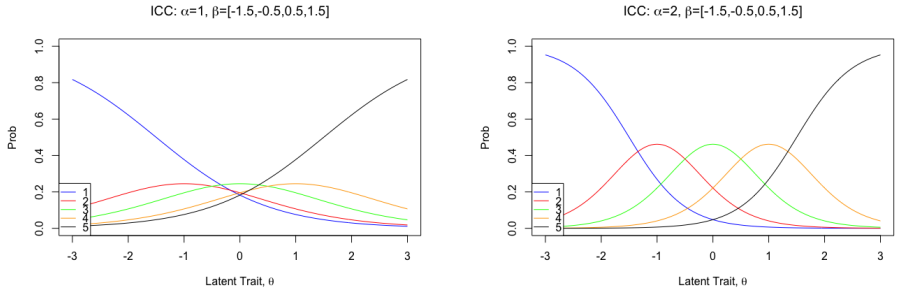
$$P_{i_k}(\theta) = P(x_i = k|\theta) = P(x_i \geq k|\theta) - P(x_i \geq k+1|\theta) = P^*_{i_k} - P^*_{i_{k+1}} \tag{4.6}$$

In the case of the BII, $k \in I[1, 5]$ and therefore $P^*_{i_1} = 1$ and $P^*_{i_6} = 0$. Thus, we can obtain all the probabilities for a subject's item response given their level of θ. These probabilities are the *Item Characteristic Curves* (ICCs) of the GRM, given as

$$P_{i_k}(\theta_s) = \begin{cases} 1 - \frac{e^{\alpha_i(\theta_s-\beta_{i_k})}}{1-e^{\alpha_i(\theta_s-\beta_{i_k})}} & \text{if } k = 1 \\ \frac{e^{\alpha_i(\theta_s-\beta_{i_k})}}{1-e^{\alpha_i(\theta_s-\beta_{i_k})}} - \frac{e^{\alpha_i(\theta_s-\beta_{i_{k+1}})}}{1-e^{\alpha_i(\theta_s-\beta_{i_{k+1}})}} & \text{if } 1 < k < 5 \\ \frac{e^{\alpha_i(\theta_s-\beta_{i_k})}}{1-e^{\alpha_i(\theta_s-\beta_{i_k})}} & \text{if } k = 5 \end{cases} \tag{4.7}$$

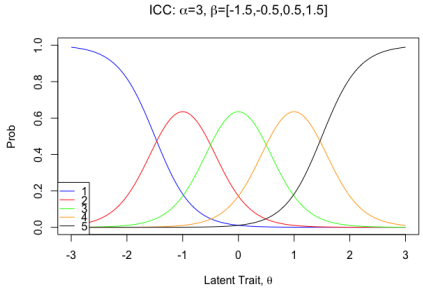
For each item, we will thus have five ICCs that define the probability for each item response given a subject's latent trait value. Three examples of the ICCs of an item given different values of the discrimination parameter, α, are shown in figure 4.2. As can be seen a higher α results in a steeper and more distinct probability curve for each separate response. Thus, it is intuitive to reason that items with a higher α value are able to better describe the distinction between subjects and therefore should contain more information compared to items with lower discrimination. The difficulty parameter, β

shifts the position of the curve as can be seen in figure 4.3. A more disperse distribution of the βs allows for more distinct ICCs and then it becomes easier to identify a subject's level of the latent trait given the item response.



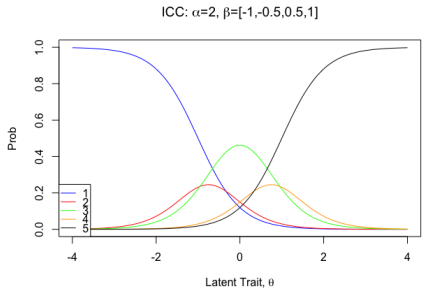
(a) Low α

(b) Medium α

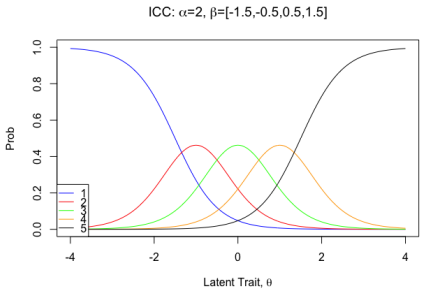


(c) High α

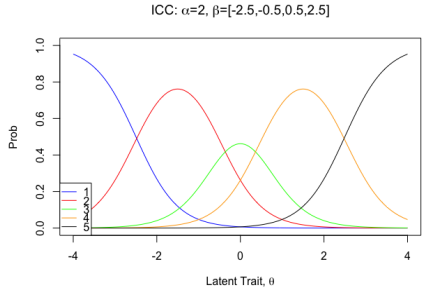
FIGURE 4.2: ICC for the GRM with different values of discrimination parameter, α. (a) Shows low discrimination, therefore it is difficult to differentiate between subject responses in the center of the latent continuum. (b) Shows medium discrimination. (c) Shows high discrimination, therefore it is easy to differentiate between subject responses around the center of the latent trait continuum



(a) Low dispersion of β



(b) Medium dispersion of β



(c) High dispersion of β

FIGURE 4.3: ICC for GRM with different sets of difficulty parameters, β. (a) Shows low dispersion of the difficulty of the item responses making it difficult to differentiate the trait levels of subjects with item response $x_i = 2, 3, 4$. (b) Shows a medium dispersion of difficulty parameters making it easier to differentiate the item's different characteristic curves. (c) Shows high dispersion of the difficulty parameters resulting in clearly defined item characteristic curves.

For each subject's specific response pattern X_s, we can assign m ICCs, one ICC for each item i. The *operating characteristic* of a response pattern is the joint probability of this specific response pattern given θ and is defined as

$$P_{X_s}(\theta) = P(X_s|\theta) \tag{4.8}$$

As noted before, local independence is assumed between all items, and thus the operating characteristic can be expressed as the product of all the item characteristics

$$P_{X_s}(\theta) = P([x_1, x_2, ..., x_n]|\theta) = P(x_1|\theta) \cdot P(x_2|\theta) \cdots P(x_n|\theta) = \prod_{x_i \in X_s} P_{i_k}(\theta) \tag{4.9}$$

This means that the response pattern given by a subject can be regarded as a sample of n independent observations from n (possibly) different distributions, with θ as the single unknown parameter. Since a subject's likelihood function is determined by the

joint probability of the response vector, the operating characteristic and the likelihood function, which will be used later for model parameter estimation, are equal in the GRM.

$$L(\theta|X_s) = \prod_{x_i \in X_s} P_{i_k}(\theta) = P(X_s|\theta) \tag{4.10}$$

4.2.3 Generalized Partial Credit Model (GPCM)

The Generalized Partial Credit Model (GPCM) was derived by Muraki (1992) and it is an extension of the Partial Credit Model (PCM) developed by Masters (1982) (which in turn is an extension of the 1PL model that works with polytomous data). The GPCM is derived from the 2PL model, which can be rewritten as

$$P_i(x_s = 1|\theta) = \frac{e^{\alpha_i(\theta-\beta_i)}}{1 + e^{\alpha_i(\theta-\beta_i)}}$$

where $P_i(x_s = 1|\theta)$ is the probability that a subject answers an item correctly. This can be extended for the BII data set, where each item has 5 possible item responses, or response categories, denoted by $k_i = k = 1, ..., 5$. For each adjacent response category (where we dichotomize each adjacent response category, i.e. 1-2, 2-3, 3-4, 4-5) the probability of a subject at ability level θ to score k over $k-1$, is given by the conditional probability

$$C_{i_k} = P_{i_k|k-1,k}(\theta) = \frac{P_{i_k}(\theta)}{P_{i_{k-1}}(\theta)+P_{i_k}(\theta)} = \frac{e^{\alpha_i(\theta-\beta_{i_k})}}{1+e^{\alpha_i(\theta-\beta_{i_k})}}, \quad k = 2, 3, 4, 5 \tag{4.11}$$

This can in turn be rewritten as

$$P_{i_k}(\theta) = \frac{C_{i_k}}{1 - C_{i_k} P_{i_{k-1}}(\theta)}$$

where $\frac{C_{i_k}}{1-C_{i_k}} = e^{\alpha_i(\theta-\beta_{i_k})}$ is the ratio of two conditional probabilities. For the BII, we can define the probabilities

$$
\begin{aligned}
P_{i_1}(\theta) &= \frac{1}{G} \\
P_{i_2}(\theta) &= \frac{\exp\left(\alpha_i(\theta - \beta_{i,2})\right)}{G} \\
P_{i_3}(\theta) &= \frac{\exp\left(\sum_{v=2}^{3} \alpha_i(\theta - \beta_{i_v})\right)}{G} \\
P_{i_4}(\theta) &= \frac{\exp\left(\sum_{v=2}^{4} \alpha_i(\theta - \beta_{i_v})\right)}{G} \\
P_{i_5}(\theta) &= \frac{\exp\left(\sum_{v=2}^{5} \alpha_i(\theta - \beta_{i_v})\right)}{G}
\end{aligned}
\tag{4.12}
$$

Also note that $\sum_{k=1}^{5} P_{i_k}(\theta) = 1$ and the normalizing factor G is equal to

$$
G = 1 + \sum_{c=2}^{5} \exp\left[\sum_{v=2}^{c} \alpha_i(\theta - \beta_{i_v})\right]
$$

This can be combined into the final probability expression for the GPCM

$$
P_{i_k}(\theta_s) = \frac{e^{\sum_{v=1}^{k} \alpha_i(\theta_s - \beta_{i_v})}}{\sum_{c=1}^{5} e^{\sum_{v=1}^{c} \alpha_i(\theta_s - \beta_{i_v})}} \tag{4.13}
$$

in which $P_{i_k}(\theta_s)$ is the probability that subject s scores k on item i. Note that $\beta_{i,1} = 0$. This value is arbitrarily chosen, as it is not a location (difficulty factor), and will be canceled both from the numerator and denominator, as

$$
P_{i_k}(\theta) = \frac{e^{Z_{i_1}(\theta)} \cdot e^{\sum_{v=2}^{k} Z_{i_v}(\theta)}}{e^{Z_{i_1}(\theta)} + \sum_{c=2}^{5} e^{[Z_{i_1}(\theta) + \sum_{v=2}^{c} Z_{i_v}(\theta)]}} = \frac{e^{\sum_{v=2}^{k} Z_{i_v}(\theta)}}{1 + \sum_{c=2}^{5} e^{\sum_{v=2}^{c} Z_{i_v}(\theta)}} \tag{4.14}
$$

where $Z_{i_k}(\theta) = \alpha_i(\theta - \beta_{i_k})$. The probability function given in equation (4.13) reduces to the 2PL item response model when $k \in \{1, 2\}$.

The β_{i_k} parameters in equation (4.13) are called step parameters, and these are the points on the latent trait continuum where the ICCs for $P_{i_{k-1}}(\theta)$ and $P_{i_k}(\theta)$ intersect, i.e., where a subject's response to item i has equal probability to be either $k-1$ or k. This can only happen once on the θ axis. This intersection is attainable anywhere along the θ scale. Thus, we can form the relationships

$$
\begin{aligned}
P_{i_k}(\theta) &= P_{i_{k-1}}(\theta) \text{ if } \theta = \beta_{i_k} \\
P_{i_k}(\theta) &> P_{i_{k-1}}(\theta) \text{ if } \theta > \beta_{i_k} \\
P_{i_k}(\theta) &< P_{i_{k-1}}(\theta) \text{ if } \theta < \beta_{i_k}
\end{aligned}
\tag{4.15}
$$

This is under the assumption that $\alpha_i > 0$, which always should be the case. N.B. This indicates that the difficulty parameters (β_{i_k}) do not need to be ordered within item i, as in the GRM, because the parameter represents the relative magnitude of adjacent probabilities $P_{i_{k-1}}(\theta)$ and $P_{i_k}(\theta)$. Therefore the GPCM is more general then the GRM and can be applied to non-ordinal, as well as ordinal, polytomous data.

4.2.4 Information Criteria for the GRM and the GPCM

IRT provides additional reliability measures of test scores and in IRT it is assumed that the precision of a test is not uniform across the entire range of test scores, i.e., across the latent trait continuum. One of the most important reliability measures in the IRT framework is the *Item Information Criteria* (IIC). The IIC, $I_i(\theta)$, represents the information contributed by a specific item, i, over the latent continuum θ. Samejima (1974) defined the IIC for polytomous item response models as

$$\begin{aligned} I_i(\theta) &= \sum_{k=1}^{m_i} P_{i_k}(\theta)\Big[-\frac{\partial^2}{\partial\theta^2}\ln P_{i_k}(\theta)\Big] = \sum_{k=1}^{m_i} P_{i_k}(\theta)\left\{\left[\frac{\frac{\partial}{\partial\theta}P_{i_k}(\theta)}{P_{i_k}(\theta)} - \frac{\frac{\partial}{\partial\theta^2}P_{i_k}(\theta)}{P_{i_k}(\theta)}\right\}\right. \\ &= D^2\alpha_i^2\sum_{k=1}^{m_i} P_{i_k}(\theta)\Big\{\sum_{c=1}^{m_i} T_c^2 P_{i,c}(\theta) - \Big[\sum_{c=1}^{m_i} T_c P_{i,c}(\theta)\Big]^2\Big\} \\ &= D^2\alpha_i^2\Big\{\sum_{c=1}^{m_i} T_c^2 P_{i,c}(\theta) - \Big[\sum_{c=1}^{m_i} T_c P_{i,c}(\theta)\Big]^2\Big\} \\ &\overset{D=1}{=} \alpha_i^2\Big\{\sum_{c=1}^{m_i} T_c^2 P_{i,c}(\theta) - \Big[\sum_{c=1}^{m_i} T_c P_{i,c}(\theta)\Big]^2\Big\} \end{aligned} \tag{4.16}$$

where m_i is the highest possible response category to item i. In our case $m_i = 5$, $\forall i = 1, ..., 24$. T_k is the scoring function equal to the item score (i.e., in the case of the BII $T_k = k, \quad \forall k \in 1, 2, 3, 4, 5$). D is a scaling factor and in the case of the BII $D = 1$. $D \neq 1$ is normally used to be able to map the logit link used in the GPCM/ GRM to the normal ogive.

For the logistic form of the GRM, Baker (1992) algebraically defined equation (4.16) as

$$I_i(\theta) = \sum_{k=1}^{m_i} \frac{\Big[P_{i_k}^*(1 - P_{i_k}^*) - P_{i_{k+1}}^*(1 - P_{i_{k+1}}^*)\Big]^2}{P_{i_k}} \tag{4.17}$$

Due to the local independence assumption on IRT the *Item Information Functions* (IIF) are additive. Thus, the combined information of all items, $i = 1, ..., n$, in a domain is

called *test information*, $I(\theta)$. It is obtained by summarizing the IIC of each item in the domain, i.e.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{4.18}$$

An important feature of the definition of test information given in equation (4.18) is that the more items there are in the domain, the greater the amount of information.

For the GPCM, the information that is provided in item i for response category k can be partitioned into the expression in (4.19), i.e., the *Item-Category Information Criteria* (ICIC). N.B. The ICICs for $I_{i_{k-1}}(\theta)$ and $I_{i_k}(\theta)$ will intersect at the point for the step parameter β_{i_k} (where the two response categories have equal probability) on the θ axis. The ICIC in the GPCM is given by

$$I_{i_k}(\theta) = P_{i_k}(\theta) I_i(\theta) \tag{4.19}$$

4.2.5 Higher-Order Item Response Theory (HO-IRT)

The latent traits in the BII data set are expected to have a hierarchical structure, i.e., they are multidimensional and one overall ability (the latent trait innovation capability) is governed by six domain abilities (the sub-traits trust, resilience, diversity, belief, perfection and collaboration). The overall ability is also called the *second-order trait* and each of the domain abilities are referred to as *first-order traits*. It is assumed that a subject's second-order latent trait ability level is directly affected by the subject's first-order latent trait abilities.

When IRT models are fit to data that supposedly have a hierarchical underlying latent trait structure, such as the BII data set, Huang et al. (2013) have summarized five model approaches that can be used in order to conduct the analysis and output ability scores for the subjects in the data set. These models are presented in Figures 4.4(a) - 4.4(e).

The first model shown in figure 4.4(a) is a *consecutive unidimensional* approach in which the first-order latent traits are estimated through a unidimensional IRT model fitted to each sub-test (one per domain ability). Using this method it is not possible to directly obtain the second-order trait and a lot of test information is ignored due to the fact that the correlation between the first-order latent traits are not considered. The second model shown in figure 4.4(b) is called the *multidimensional* approach and it addresses the last problem by invoking a correlation between the first-order traits to improve their

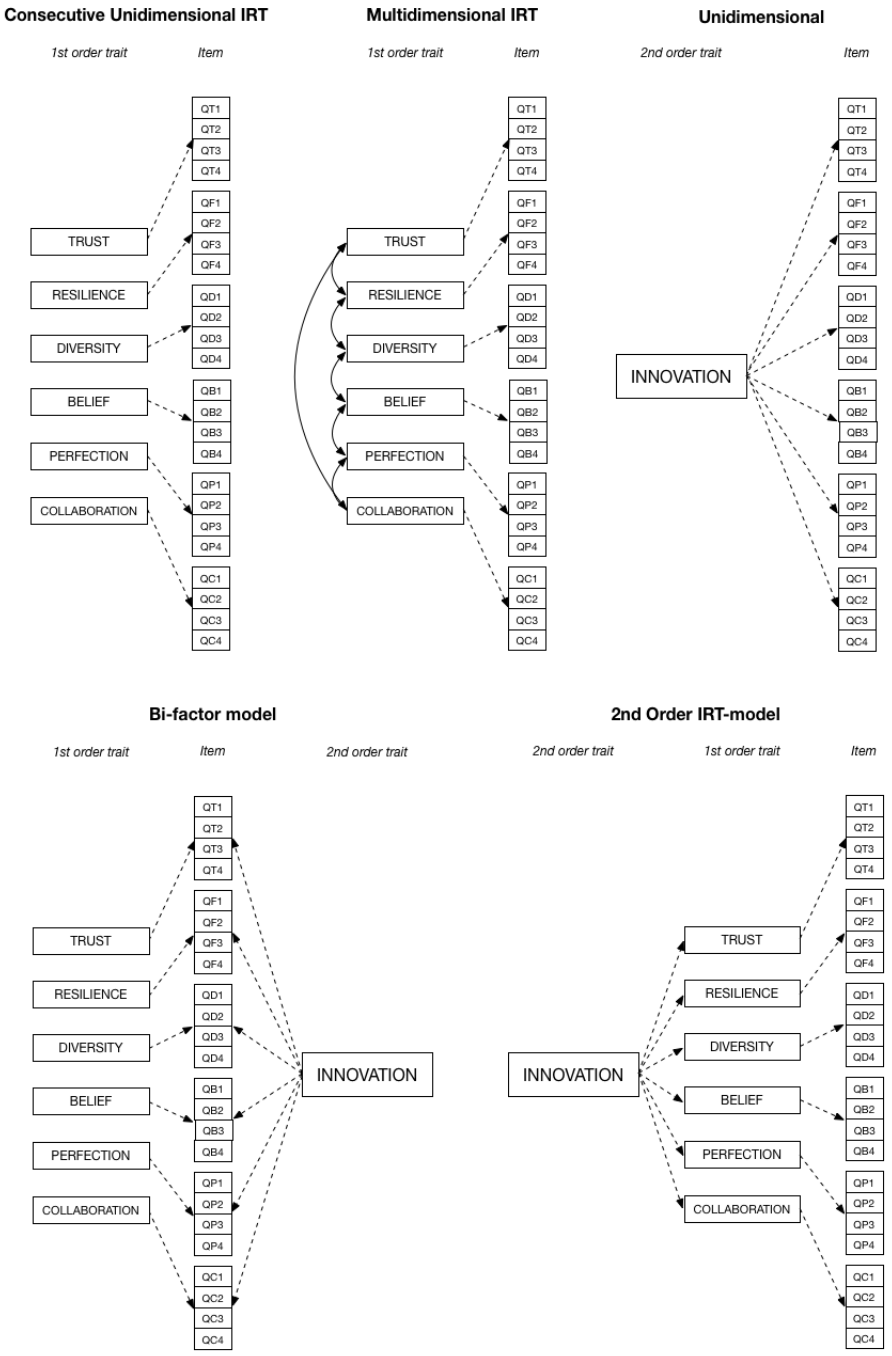


FIGURE 4.4: The possible models given the data structure of the BII. (a) Consecutive Unidimensional IRT (b) Multidimensional IRT (c) Unidimensional IRT (d) Bi-factor IRT (e) 2nd Order IRT (Higher-Order IRT)

estimates. However, as in the first model, the second-order trait can not be obtained directly.

In these two models the second-order latent trait can be indirectly estimated using *Confirmatory Factor Analysis* (CFA), a latent trait model that maps a set of continuous data (here the estimated values for the domain abilities) to a continuous latent trait (here overall innovation capability). However, the scores observed for the domain abilities contain measurement errors which the CFA method does not take into account when estimating the overall score. This, together with the fact that the CFA will treat the values of the first-order traits as observed values rather than estimates, may result in an over-estimation of the second-order trait.

Figure 4.4(c) shows the *composite unidimensional* approach where all items are expected to measure the same latent trait. It allows for a direct estimation of the second-order trait, but on the other hand this estimate may not be valid because of the extent to which the unidimensional assumption is violated. The items supposedly measure several different first-order traits at the same time (de la Torre and Song, 2009). In addition to this the composite unidimensional model does not enable estimation of the first-order traits.

Two models that deal with the disadvantages presented by the three former ones above is the *bifactor* model and the *Higher-Order IRT* (HO-IRT) model (Figure 4.4(d) and Figure 4.4(e), respectively). Both models enable the estimation of both an item specific latent trait (first-order), as well as an overall latent trait common to all items (second-order). It has been shown in Yung et al. (1999) that the bifactor model is mathematically equivalent to the HO-IRT model when there are two levels of latent trait, with one overall latent trait and several sub-traits. This is the case for the BII, hence it would be possible to work both with a bifactor approach and a HO-IRT approach. Due, mainly, to the previous work done on HO-IRT as well as its more intuitive interpretation this is the model chosen for the analysis conducted in the thesis.

The over-estimation of the second-order trait, which was a problem for the multidimensional and consecutive unidimensional model, is in the HO-IRT addressed by introducing an error term in the relationship between first and second-order latent traits (Huang et al., 2013). de la Torre and Hong (2010) also showed that the HO-IRT is superior to non-hierarchical approaches when the number of tests, i.e., number of first-order latent traits, is big and the size of each test is relatively small. Given the structure of the BII with six different domain abilities and only four items per domain ability this validates the choice of the HO-IRT model over the multidimensional and the consecutive unidimensional IRT models.

Inspired by the ideas of de la Torre and Song (2009), Huang et al. (2013) presented a HO-IRT model that enables the use of polytomously scored data. A model based on their work is the one used in this thesis.

de la Torre and Song (2009) states that the two orders of latent traits are associated through the linear relationship between the domain abilities and the overall ability as

$$\theta_{sj}^{(1)} = \lambda_j \theta_s^{(2)} + \epsilon_{sj} \tag{4.20}$$

where $\theta_{sj}^{(1)}$ is the first-order latent trait of domain j for subject s (where $j = 1, ...6$, and $1 = trust$, $2 = resilience$ etc.). $\theta_s^{(2)}$ is the second-order latent trait governing overall ability for subject s. λ_j is the latent coefficient (also called factor loading) regressing ability j on the overall ability, where $0 \leq \lambda_j \leq 1$. Mathematically λ_j can be negative, but here the domain abilities are hypothesized to be related to the overall ability, hence λ_j is expected to be nonnegative. Moreover, the constraint on λ_j guarantees that the overall and domain abilities are on the same scale. ϵ_{sj} is an error term that is independent of all other terms.

To obtain the ICCs of the Higher-Order GRM (HO-GRM) the relationship in equation (4.20) is inserted in the equation for the cumulative probability given in (4.5). Thus, the cumulative probability for the HO-GRM is given by

$$P_{i_k}^* = P(x_i \geq k|\theta^{(2)}) = \frac{e^{\alpha_i(\lambda_j\theta^{(2)} + \epsilon_j - \beta_{i_k})}}{1 + e^{\alpha_i(\lambda_j\theta^{(2)} + \epsilon_j - \beta_{i_k})}} \tag{4.21}$$

With the use of equation (4.21), instead of (4.5), the derivation of the item characteristic curve for the HO-GRM model is analogous to the method presented in section 4.2.2.

The ICCs for the Higher-Order GPCM (HO-GPCM) is directly obtained by inserting the relationship (4.20) in equation (4.13):

$$P_{i_k}(\theta^{(2)}) = P(x_i = k|\theta^{(2)}) = \frac{e^{\sum_{v=1}^{k} \alpha_i(\lambda_j\theta^{(2)} + \epsilon_j - \beta_{i_v})}}{\sum_{c=1}^{5} e^{\sum_{v=1}^{c} \alpha_i(\lambda_j\theta^{(2)} + \epsilon_j - \beta_{i_v})}} \tag{4.22}$$

The IIF for the HO-IRT models can be obtained following the derivation steps presented in section 4.2.4 (and making use of (4.21) and (4.22)).

In order to estimate the model parameters and the ability estimates some assumptions need to be imposed on the model. Therefore, the distribution of the model parameters are specified in accordance with the parameter distributions presented in de la Torre and Song (2009) and Huang et al. (2013).

$$\theta_s^{(2)} \sim \mathcal{N}(0,1)$$

$$\theta_{sj}^{(1)}|\theta_s^2, \lambda_j \sim \mathcal{N}(\lambda_j\theta_s^{(2)}, \sqrt{1-\lambda_j^2})$$

$$\epsilon_{sj} \sim \mathcal{N}(0, 1-\lambda_j^2)$$

Furthermore, it is assumed that all the domain-level abilities are independent conditional on $\theta^{(2)}$. The correlation between the first-order trait $\theta_j^{(1)}$ and the second-order trait is given by $\rho(\theta^{(2)}, \theta^{(1)}) = \lambda_j$ and the correlation between the first-order traits is given by $\rho(\theta_j^{(1)}, \theta_{j'}^{(1)}) = \lambda_j\lambda_{j'}$. de la Torre and Song (2009) also points out that if the number of domains $J \geq 4$, then this implies that there exists more correlations between the abilities than there are regression parameters (factor loadings). As a result the true correlation structure might be more complex than what the linear model can fit. Furthermore, de la Torre and Hong (2010) found that first-order latent traits are better estimated when they are highly correlated with the second-order trait, i.e., have a higher λ-value. However, a high correlation between latent traits does not indicate that they are the same latent trait (e.g., longevity and wealth are highly correlated, but they are totally different attributes).

de la Torre and Song (2009) also highlights that the first-order latent trait estimates in the HO-IRT model should mainly be used for within-person comparisons (i.e., the domain ability estimates are not comparable between subjects), whereas the second-order latent trait estimates can be used for between-person comparisons. This is due to the fact that the domain scores for the first-order latent traits make use of information from other domains, due to the correlational structure of the traits. The interpretation of the domain ability estimates can therefore be somewhat ambiguous, because a subject's ability level in one domain is influenced by his or her proficiency level in the other domains.

In order to estimate all the HO-IRT model parameters a *Markov Chain Monte Carlo* (MCMC) method is utilized. It is cast in a hierarchical Bayesian framework, which has been used successfully by Huang et al. (2013) and de la Torre and Song (2009) when estimating the parameters of HO-IRT models.

4.3 Markov Chain Monte Carlo (MCMC)

Before the MCMC method was introduced for IRT models the most common practice in order to recover IRT model parameters was to use the *Expectation-Maximization* (EM) algorithm (Bock and Aitkin, 1981). When solving for the model parameters the EM algorithm first marginalizes the likelihood with respect to the subject parameters (e.g.,

θ_s) and solve the Marginal Maximum Likelihood (MML) problem in order to estimate the item parameters (e.g., α_i). These parameters are then fixed in order to solve for the subject parameters. A problem with this approach is that the two-step nature of the procedure cannot truly incorporate uncertainty into the item parameter estimates in the calculations of the subject parameters, and there is no way of knowing to which extent the standard errors for the subjects are overly optimistic (Tsutakawa and Soltys, 1988).

The development of the MCMC method for IRT models was justified by the fact that it is beneficial to incorporate more uncertainty in the calculations, due to the fact that all parameters are estimated simultaneously, and it is still relatively straightforward to implement the model when the complexity increases compared to e.g., the EM algorithm (Patz and Junker, 1999).

4.3.1 Derivation of MCMC

The derivation of the MCMC method for IRT models starts by specifying the joint likelihood function for all test subjects. Due to the assumptions of independence of the item responses (given the latent trait for subject s, θ_s) the total joint likelihood for a general IRT model, given a set of observed item responses (X), is the product of all probabilities for each of the item responses and all the subjects as shown in equation (4.23)

$$P(X|\theta_s^{(2)}, \epsilon, \lambda, \beta, \alpha) = \prod_s \prod_i P(x_{si}|\theta_s^{(2)}, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i), \quad j = \lfloor \frac{i+3}{4} \rfloor \tag{4.23}$$

ϵ_{sj} is the error term that together with the factor loading λ_j defines the relationship given by equation (4.20). β_i is the difficulty parameter for item i given the observed response x_{si}. α_i is the discrimination parameter of item i and $\theta_s^{(2)}$ is the second-order latent trait for subject s. $\theta_j^{(1)}$ can easily be computed using equation (4.20) and thus only the second-order latent trait needs to be estimated with the MCMC. Therefore, the notation $\theta_s^{(2)} = \theta_s$ will be used in the derivation of the MCMC model.

The product over all subjects, s, is possible due to the assumption of *experimental independence* among subjects in IRT and the product over all items, i, is possible because of the assumption of *local independence*.

Using MCMC requires the inclusion of hyperparameters when specifying the prior distributions. Due to the potential complexity of the priors we will follow the example given by Patz and Junker (1999) and Huang et al. (2013), i.e., keep all the hyperparameters of the priors fixed which also allows us to exclude these from the notations.

The MCMC algorithm is used in order to estimate the model's joint posterior distribution:

$$P(\theta, \epsilon, \lambda, \beta, \alpha|X) \propto P(X|\theta, \epsilon, \lambda, \beta, \alpha)P(\theta, \epsilon, \lambda, \beta, \alpha) \tag{4.24}$$

Once the posterior distribution is estimated one can make inferences about each one of the HO-IRT model parameters.

To ease the notation in the following short explanation of the basics of MCMC we will use a joint posterior function of the two arbitrary parameters θ and β given an observed data set X i.e.

$$P(\theta, \beta|X) \tag{4.25}$$

The essential idea behind MCMC is to define a (stationary) Markov Chain, $M_0, M_1, M_2, ...$, with states $M_k = (\theta^k, \beta^k)$ and then simulate new observations from the Markov chain. The distribution will, under suitable conditions, converge to the chain's stationary distribution $\pi(\theta, \beta)$, and the simulated observations can then be used to make inferences about the parameters. To achieve this we want to define the Markov chain in such a way that the stationary distribution is the posterior distribution defining our parameters, i.e., $\pi(\theta, \beta) = P(\theta, \beta|X)$.

The behaviour of the Markov chain is determined by its transition kernel

$$t[(\theta^k, \beta^k), (\theta^{k+1}, \beta^{k+1})] = P[M_{k+1} = (\theta^{k+1}, \beta^{k+1})|M_k = (\theta^k, \beta^k)] \tag{4.26}$$

which is the proability of moving to the new state $M_{k+1} = (\theta^{k+1}, \beta^{k+1})$ given the old state $M_k = (\theta^k, \beta^k)$.

If it is feasible to define the kernel such that $\pi(\theta, \beta) = P(\theta, \beta|X)$, then after the first K steps of the Markov chain the observed states $M_{K+1} = (\theta^1, \beta^1), M_{K+2} = (\theta^2, \beta^2), \ldots, M_{K+L} = (\theta^L, \beta^L)$ will each be distributed as draws from the posterior distribution and thus give us information about the properties of the parameters. One can also say that the Markov Chain has converged to the stationary distribution.

The first K steps discarded from the chain are called *burn-in* iterations. MCMC algorithms often randomly choose a starting point and in high dimensions this point generally start at an area of low density for the posterior distribution, then after some iterations the

MCMC algorithm reaches an area of high density and this is generally a better starting point in order to make inferences about the posterior distribution.

4.3.2 Metropolis-Hastings within Gibbs Sampling

One of the most common MCMC algorithms is the *Gibbs Sampler* which makes use of the transition kernel in (4.27).

$$t[(\theta^k, \beta^k), (\theta^{k+1}, \beta^{k+1})] = P(\theta^{k+1}|\beta^k, X)P(\beta^{k+1}|\theta^{k+1}, X) \tag{4.27}$$

The kernel was first introduced by Geman and Geman (1984) and it produces a stationary distribution that is equal to the posterior distribution. The transition kernel is the product of all the parameters *full conditional probabilities*, i.e., a parameter distribution which is conditional on all other parameters. In (4.27) the full condtional probabilities are the densities $P(\theta^{k+1}|\beta^k, X)$ and $P(\beta^{k+1}|\theta^{k+1}, X)$. Thus an iteration of the *Gibbs Sampler* consists of drawing a new parameter from each of the parameters full conditional distributions. For the HO-IRT the updating scheme for iteration m is as follows

Pseudocode 1 BII HO-IRT updating scheme, Gibbs sampler (arbitrary iteration m)

1: Draw $\theta_s^m \sim P(\theta_s|\theta_{<s}^m, \theta_{>s}^{m-1}, \epsilon^{m-1}, \lambda^{m-1}, \alpha^{m-1}, \beta^{m-1}, X)$ ▷ $\forall s = 1, ..., 878$
2: **for** $j \leftarrow 1$ to 6 **do**
3: Draw $\epsilon_{sj}^m \sim P(\epsilon_{sj}|\theta^m, \epsilon_{s,<j}^m, \epsilon_{<s,j}^m, \epsilon_{>s,j}^{m-1}, \epsilon_{s,>j}^{m-1}, \lambda^{m-1}, \alpha^{m-1}, \beta^{m-1}, X)$ ▷ $\forall s = 1, ..., 878$
4: **end for**
5: Draw $\lambda_j^m \sim P(\lambda_j|\theta^m, \epsilon^m, \lambda_{<j}^m, \lambda_{>j}^{m-1}, \alpha^{m-1}, \beta^{m-1}, X)$ ▷ $\forall j = 1, ..., 6$
6: Draw $\alpha_i^m \sim P(\alpha_i|\theta^m, \epsilon^m, \lambda^m, \alpha_{<i}^m, \alpha_{>i}^{m-1}\beta^{m-1}, X)$ ▷ $\forall i = 1, ..., 24$
7: Draw $\beta_i^m \sim P(\beta_i|\theta^m, \epsilon^m, \lambda^m, \alpha^m\beta_{<i}^m, \beta_{>i}^{m-1}, X)$ ▷ $\forall i = 1, ..., 24$

where $< s = [1, ..., s-1]$ and $> s = [s+1, ..., 878]$ and the same is true for j and i.

The steps to derive the full conditional probabilities are analogous for all the parameters and therefore only the derivation of θ will be explained in detail. It is based on consecutive use of *Bayes Theorem* (4.28) which explains the relationship between the joint probability and the conditional probability between two, possibly multidimensional, variables:

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y}) \tag{4.28}$$

In the updating scheme in Pseudocode 1 it is implicated which MCMC iteration, m, a parameter belongs to and thus this notation will be excluded in the derivation. Note

that due to the *experimental independence* assumption among subjects and the *local independence* assumption among items we can assume that a parameter's full conditional distribution is not conditioned on parameters of the same type. It also implicates that all types of parameters are independent in regards to each other with the exception of ϵ which depends on λ. Thus, we can derive the *full conditional probability* for θ as

$$\begin{aligned} P(\theta_s|\theta_{<s}, \theta_{>s}, \epsilon, \lambda, \alpha, \beta, X) &= P(\theta_s|\epsilon, \lambda, \alpha, \beta, X) \\ &\propto P(\theta_s, \epsilon, \lambda, \alpha, \beta, X) \\ &\propto P(X|\theta_s, \epsilon, \lambda, \alpha, \beta)P(\theta_s, \epsilon, \lambda, \alpha, \beta) \\ &\propto P(X|\theta_s, \epsilon, \lambda, \alpha, \beta)P(\theta|\epsilon, \lambda, \alpha, \beta) \\ &= P(X|\theta_s, \epsilon, \lambda, \alpha, \beta)P(\theta) \end{aligned} \tag{4.29}$$

This means that the full conditional distribution for a parameter is its likelihood function, $P(X|\theta_s, \epsilon, \lambda, \alpha, \beta)$, multiplied with its prior distribution, in this case $P(\theta)$. The full conditional probabilities for the other parameters can be derived analogously . They are presented in equations (4.30) - (4.34).

$$P(\theta_s|rest) \propto \prod_{i=1}^{24} P(x_{si}|\theta_s, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i)P(\theta_s),\ \forall s = 1, ..., 878 \tag{4.30}$$

$$P(\epsilon_{sj}|rest) \propto \prod_{i \in j} P(x_{si}|\theta_s, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i)P(\epsilon_{sj}|\lambda_j),\ \forall s = 1, ..., 878;\ \forall j = 1, ..., 6 \tag{4.31}$$

$$P(\lambda_j|rest) \propto \prod_{s=1}^{878} \prod_{i \in j} P(x_{si}|\theta_s, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i)P(\lambda_j),\ \forall j = 1, ..., 6 \tag{4.32}$$

$$P(\alpha_i|rest) \propto \prod_{s=1}^{878} P(x_{si}|\theta_s, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i)P(\alpha_i),\ \forall i = 1, ..., 24 \tag{4.33}$$

$$P(\beta_i|rest) \propto \prod_{s=1}^{878} P(x_{si}|\theta_s, \epsilon_{sj}, \lambda_j, \beta_i, \alpha_i)P(\beta_i),\ \forall i = 1, ..., 24 \tag{4.34}$$

Here the notation *rest* indicates all other model parameters and $i \in j$ indicates all items i that belong to the domain j.

Gibbs sampling requires, in each iteration, a draw directly from the full conditional distribution and to be able to do this it is necessary that the distribution for each parameter can be derived in closed form (i.e., that it can be evaluated in a finite number of steps). If this is not possible the model can be extended using a *Metropolis-Hastings within Gibbs* approach which allows the model to perform Gibbs sampling when possible,

and when not it takes a so called *Metropolis-Hastings step*. A Metropolis-Hastings step (Chib and Greenberg (1995), Hastings (1970), Metropolis et al. (1953)) for an arbitrary parameter, τ, in the MCMC algorithm is carried out as in Psuedocode 2

Pseudocode 2 Metropolis-Hastings step for an arbitrary parameter τ

1: Draw $\tau_k^* \sim g(\tau_k|\tau_k^{m-1})$
2: Accept τ_k^* with an acceptance rate of $\alpha^* = \min\left\{\frac{f(\tau_k^*|rest)g(\tau_k^{m-1}|\tau_k^*)}{f(\tau_k^{m-1}|rest)g(\tau_k^*|\tau_k^{m-1})}, 1\right\}$
3: If accepted $\tau_k^m = \tau_k^*$, else $\tau_k^m = \tau_k^{m-1}$

In the first step of Pseudocode 2 τ_k^* is a proposal value which is drawn from an arbitrary convenient *proposal density* g. One common choice of proposal density is an *independent density*, i.e. $g(\tau_k^m|\tau_k^{m-1}) = g(\tau_k^m)$. In the second step it is determined if the proposal value is accepted, with an acceptance rate of α^* which is determined by the values of the proposal density, g, and the full conditional density, f, given the old value, τ_k^{m-1} and the proposal value, τ_k^*.

The implementation of the MCMC model is done with the use of JAGS (Just Another Gibbs Sampler) in R through the package `rJAGS`. JAGS allows the user to define the model and the model priors. In JAGS where conjugate distributions (when the posterior distributions are in the same family as the prior distribution) are used regular Gibbs sampling is done. When that is not the case, i.e., there is no closed form distribution for the parameters, then a Metropolis-Hastings step is used (adaptive rejection or slice sampling might also be used, but those methods are not included in the scope of the thesis). Earlier work on MCMC for IRT (particularly the GRM) done by Albert and Chib (1993), Cowles and Carlin (1996) and Kuo and Sheng (2015) has mainly focused on the development of a Metropolis-Hasting-within-Gibbs method.

When specifying the MCMC model in JAGS one of the major concerns is the *prior selection* since this is one of the most evident ways to influence the analysis and make an impact on the result. Therefore, we will conduct a comprehensive prior analysis in the section 4.3.3 to see how the choice of prior distributions influences the estimates of the model parameters.

4.3.3 Prior Selection

The priors are the the only way to adjust a specified MCMC model in order to affect the results. Hence, it is desirable to analyze the importance of the prior selection and determine which priors are the best fit for our model.

The distribution of ϵ and θ are known since they are defined by the model and the prior distribution of λ is chosen to cover the interval $I \in [0, 1]$ so that all correlations between the first- and second-order latent traits are attainable (Huang et al., 2013). Patz and Junker (1999) suggests the use of the lognormal distribution for α and normal distributions for the other model parameters. The constant prior hyperparameters for α and β are chosen according to the values presented in Sung and Kang (2006). The following priors were used as a starting point for the prior analysis:

$$\theta_s \sim \mathcal{N}(0, 1)$$
$$\lambda_j \sim \mathcal{N}(0.5, 0.2)I(0, 1)$$
$$\epsilon_{s,j} \sim \mathcal{N}(0, 1 - \lambda_j^2)$$
$$\alpha_i \sim ln\mathcal{N}(0, 1)$$
$$\beta_{i_k} \sim \mathcal{N}(0, 1)$$

N.B. The prior $\lambda_j \sim \mathcal{N}(0.5, 0.2)I(0, 1)$ was selected because the distribution covers the whole interval for the factor loading term. $I(0, 1)$ truncates the distribution $\mathcal{N}(0.5, 0.2)$ so that λ_j only takes values between 0 and 1.

The GRM model poses a requirement that the responses are ordered according to their difficulty level and thus the difficulty parameter β_i is required to have an ordered structure. Therefore, the priors for $\beta_{i,j}, j \neq 1$ are given a lower limit determined by $\beta_{i,j-1}$ (and no upper limit), i.e.

$$\beta_{i,1} \sim \mathcal{N}(0, 1)$$
$$\beta_{i,2} \sim \mathcal{N}(0, 1)I(\beta_{i,1},)$$
$$\beta_{i,3} \sim \mathcal{N}(0, 1)I(\beta_{i,2},)$$
$$\beta_{i,4} \sim \mathcal{N}(0, 1)I(\beta_{i,3},)$$

A prior analysis was conducted on a set of simulated data (see Section 4.6.2) in order to determine the effect the choice of prior has on the results. The reason why the analysis is conducted on simulated data, rather than on a real data set, is that it is possible to do a more general analysis (instead of a specific one) on the data structure of the BII data set. It also allows us to directly estimate the accuracy of the parameter estimates since we can compare them with the *true* values, i.e., the model parameters used when simulating the data set. For each simulation in the prior analysis only one prior distribution was changed with respect to the starting priors to determine how that specific prior affects the outcome. The simulated data set consists of $S = 1000$ subjects, $n = 24$ items, $J = 6$ domains and five possible item responses, $k = 1, ..., 5$. In order to make the analysis

	2^{nd} **order**	1^{st} **order**					
Prior Adjustment	$\rho^{\theta^{(2)}}$	$\rho_1^{\theta^{(1)}}$	$\rho_2^{\theta^{(1)}}$	$\rho_3^{\theta^{(1)}}$	$\rho_4^{\theta^{(1)}}$	$\rho_5^{\theta^{(1)}}$	$\rho_6^{\theta^{(1)}}$
None	0.930	0.869	0.874	0.892	0.903	0.899	0.915
$\alpha \sim ln\mathcal{N}(0, 0.5)$	0.930	0.869	0.874	0.892	0.903	0.899	0.915
$\alpha \sim ln\mathcal{N}(0, 2)$	0.929	0.868	0.874	0.892	0.903	0.899	0.915
$\alpha \sim ln\mathcal{N}(1, 1)$	0.929	0.869	0.875	0.892	0.903	0.899	0.915
$\beta \sim \mathcal{N}(0, 0.5)$	0.929	0.868	0.874	0.891	0.903	0.899	0.915
$\beta \sim \mathcal{N}(0, 2)$	0.930	0.869	0.875	0.892	0.903	0.899	0.916
$\beta \sim \mathcal{N}(-1, 1)$	0.929	0.869	0.874	0.892	0.904	0.899	0.916
$\beta \sim \mathcal{N}(1, 1)$	0.930	0.868	0.874	0.892	0.904	0.899	0.916
$\lambda \sim \mathcal{N}(0.5, 0.1)$	0.930	0.868	0.874	0.892	0.903	0.899	0.916
$\lambda \sim \mathcal{N}(0.7, 0.2)$	0.930	0.869	0.875	0.892	0.904	0.899	0.916
$\theta \sim \mathcal{N}(0, 0.5)$	0.930	0.869	0.875	0.892	0.903	0.899	0.916
$\theta \sim \mathcal{N}(0, 2)$	0.929	0.868	0.873	0.891	0.903	0.899	0.916

TABLE 4.2: Prior analysis for the HO-IRT GRM model conducted on simulated data

reproducible the *seed* in R (the starting point for the random number generator) was set to 71.

Due to the number of parameters no unique solution exists and a change of prior will lead to changes in the model parameters. Since the model parameters interact the estimated parameters might be cast on a different scale than the model parameters used in the data simulation. Therefore, one can not directly use the root mean square error (RMSE) of the estimated values as a mean of comparison between models.

Instead, in order to evaluate the goodness of fit of the models, we will make use of the correlation between the estimated values for the latent traits $\hat{\theta}$ and the "true" values of the latent traits θ obtained from the simulation, i.e. $\rho^{\theta^{(2)}} = \rho(\hat{\theta}^{(2)}, \theta^{(2)})$, $\rho_1^{\theta^{(1)}} = \rho(\hat{\theta}_1^{(1)}, \theta_1^{(1)})$ etc. This is also the preferred statistics used by de la Torre and Song (2009). The result of the prior analysis is presented in table 4.2

As can be seen by comparing the outcomes of the different choices of prior distribution it is apparent that the goodness of fit is more or less indifferent to (at least) small changes of the prior distributions. This means that the data, and not the chosen priors, has the biggest impact on the estimations. Worth noting is that only one run has been done for each prior adjustment and therefore the result is not statistically significant, but since all runs show a very low discrepancy in the correlation this gives a clear indication that the original priors are sufficient. Therefore, the choice of prior distributions in the thesis is the one used as a starting point for this analysis.

4.3.4 Parameter Estimation

After the burn-in period the iterations in the MCMC chain are assumed to constitute of draws from the stationary posterior distribution. For each parameter in the model, here labeled as τ, we are interested in using the MCMC output $\tau^{(m)}$, where $m = 1, ..., M$, to estimate its distribution and make inference about it.

Due to the law of large numbers the mean of the MCMC chain for each parameter will converge towards the true expectation of the parameter's posterior distribution, as shown in (4.35).

$$E[\tau|X] \approx \frac{1}{M} \sum_{m=1}^{M} \tau^{(m)} = \bar{\tau} \tag{4.35}$$

To determine the accuracy of the parameter estimates an uncertainty measure needs to be employed. There are mainly two uncertainty measures of interest when estimating the posterior qualities, the *Monte Carlo uncertainty* and the *posterior uncertainty*.

The *Monte Carlo uncertainty*, SE_{MCMC}, simply expresses the standard error between the *true* expected value of all parameters and the estimations of the expected value given by (4.35) and therefore this error is reduced by increasing the MCMC sample size.

The *posterior uncertainty*, SD_{post}, of a parameter is the standard deviation of its posterior distribution and is often used to make inferences and construct confidence intervals for a parameter. We can make the inference about the parameters more precise by collecting more data. Increasing the number of MCMC iterations after burn-in, M, can make our estimates of posterior mean and variance more precise (by reducing SE_{MCMC}), but it can not improve the precision of our inference about τ (Junker et al., 2016).

To compute the posterior standard error one first needs to estimate the variance of the sampled values of the Markov Chain, i.e. $SD_{post} = \sigma_\tau^2 = Var[\tau]$. If the Markov chain is not ergodic, i.e., every state of the chain cannot be reached from any other state in exactly N finite steps. Then one cannot make use of the naive estimator, because of the dependency of $\tau^{(m)}$. One method one might use instead is *overlapping batch means* (OLBM) (Flegal and Jones (2011)). First define the batch length, b_m, and then construct batches $B_1 = (\tau^{(1)}, \tau^{(2)}, ..., \tau^{(b_m)})$, $B_2 = (\tau^{(2)}, \tau^{(3)}, ..., \tau^{(b_m+1)})$ etc. Let

$$\bar{\tau}_j = \frac{1}{b_m} \sum_{\tau^{(m)} \in B_j} \tau^{(m)}, \text{ and} \tag{4.36}$$

$$\hat{\sigma}^2_{OLBM} = \frac{Mb_m}{(M-b_m)(M-b_m+1)} \sum_{j=1}^{M-b_m+1} (\bar{\tau}_j - \bar{\tau})^2 \tag{4.37}$$

Under the condition that $b_m \approx \sqrt{M}$ it is shown that $\hat{\sigma}^2_{OLBM}$ will be a consistent estimator of the variance of the posterior distribution. Then the standard error is obtained as

$$SE_{MCMC} \approx \frac{1}{\sqrt{M}} \sqrt{\hat{\sigma}^2_{OLBM}} \tag{4.38}$$

and the confidence interval is $\bar{\tau} \pm t^* \cdot SE_{MCMC}$ where t^* follows a t-distribution with $M - b_m$ degrees of freedom.

4.3.5 MCMC Convergence Diagnostics

Gelman et al. (2011) presents general recommendations for assessing if a Markov chain has converged. The criteria listed are:

- Simulate three or more chains in parallel with three different, crude estimates of the starting point.
- Check convergence by discarding the first part of the simulations (the burn-in) then monitor within-chain stationarity and between/within chains comparisons to monitor *mixing*. Good mixing indicates that the stationary distribution is reached (fairly quick) for all chains, starting from an arbitrary position.
- Once approximate convergence has been reached, mix all the simulations from the undiscarded parts of the chains together to summarize the target distribution.

The above recommendations are considered when running the MCMC simulations on the BII data set. For the MCMC chains 2000 iterations (at least) are chosen as the burn-in period to follow the recommendations given by Huang (2015). Another recommendation is to have a high dispersion of the initial values for each separate chain such that all the MCMC chains do not get stuck in a local maximum. The initialization is taken care of automatically by the R packages used in the analysis conducted for the BII.

Within-chain stationarity is examined through each chain's *trace plot*. A trace plot plots the parameter value for each iteration in the chain. When converged the chain only takes values drawn from the stationary distribution and thus each single trace plot should show that the drawn values vary around a specific mean. If the trace plots of

three parallel chains overlap one another to a great extent, after the burn-in period has been discarded, this indicates that the MCMC chains are mixing well (the dependence decays quickly in successive iterations). If the chains converge quickly and if it is not highly auto-correlated, then samples from the multiple chains can be pooled together to make valid inferences about the posterior distribution. Because of the vast numbers of model parameters being estimated, > 11000, it is impossible to present all trace plots and auto-correlation function plots. However, a vast subset of these have been analyzed and typical graphical results are shown in section 5.1.

One of the most commonly used convergence diagnostic tools is the Gelman-Rubin diagnostic which will be used in this analysis to further validate the convergence of the results obtained from the MCMC analysis.

4.3.5.1 Gelman-Rubin Diagnostics

The *Gelman-Rubin* (G-R) diagnostic (Gelman and Rubin, 1992) utilizes the characteristics of multiple chains with different initial values to check if they have reached convergence. When the chains have converged they should have similar appearance to one another. Failure to converge could indicate the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain.

The G-R diagnostic is a variance ratio test statistic. The convergence is assessed by comparing the estimated *between-chains variance*, B, and the *within-chain variances*, W, for each model parameter. A large difference between these variances indicates that the chains have not converged (Brooks and Gelman, 1998).

To briefly summarize the G-R diagnostic, suppose there are N chains, all of equal length M. The model parameter of interest is θ and $\{\theta_n^m\}_{m=1}^M$ is the n:th simulated chain, $n = 1, ..., N$. Let $\hat{\theta}_n$ and $\hat{\sigma}_n^2$ be the sample posterior mean as well as the variance of the n:th chain. The sample mean for the chain is $\hat{\theta}_n = \frac{1}{M}\sum_{m=1}^M \theta_n^m$ and the overall sample posterior mean is $\hat{\theta} = \frac{1}{N}\sum_{n=1}^N \hat{\theta}_n$. The definitions of the between-chains variance B and the within-chain variances W are presented in equations (4.39) and (4.40).

$$B = \frac{M}{N-1}\sum_{n=1}^{N}(\hat{\theta}_n - \hat{\theta})^2 \tag{4.39}$$

$$W = \frac{1}{N}\sum_{n=1}^{N}\hat{\sigma}_n^2 \tag{4.40}$$

where $\sigma_n^2 = \frac{1}{M-1}\sum_{m=1}^{M}(\theta_n^m - \hat{\theta}_n)^2$. An unbiased estimate of the marginal posterior variance of θ, $\hat{\text{Var}}(\theta)$ can then be calculated as

$$\hat{\text{Var}}(\theta) = \frac{M-1}{M}W + \frac{N+1}{NM}B \tag{4.41}$$

If all N chains have converged to their target distributions, then $\hat{\text{Var}}(\theta)$ should be close to W. The square root of this ratio is called the *Potential Scale Reduction Factor* (PSRF) where a large value indicates that B is greater than W and therefore more iterations in the MCMC chains are needed. If the chains have converged the PSRF (also called the G-R diagnostic) should be close to 1. The PSRF presented in Brooks and Gelman (1998) is defined as

$$\hat{R} = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{\text{Var}}(\theta)}{W}} \tag{4.42}$$

Here, $\hat{d}$ is the degrees of freedom estimate of a t distribution. If the stringent condition $\hat{R} < 1.1$ holds for all model parameters, one can be fairly confident that convergence has been reached Brooks and Gelman (1998).

4.4 Model Fit

To assess the accuracy and validity of the parameter estimations and to be able to compare different models with each other, there is a need for some statistics that measure a model's goodness of fit. To asses this, one can examine the variation of the parameter output as well as performing a *Posterior Predictive check* (PPC). To compare two different models with each other one can also use the PPC as well as the *Deviance Information Criteria* (DIC) and the *Coefficent of variation* (CV).

4.4.1 Coefficient of Variation (CV)

General model fit can be assessed by looking at the mean posterior variance for the parameters recovered. A lower variance indicates that the model has a better fit, but the magnitude of the estimated parameters will also influence the magnitude of the variance and thus it is not possible to compare the model fit between two different models by only examining the parameter variance. Therefore, a better measure of how much a parameter estimate fluctuates can be obtained by taking the ratio of the mean of the standard deviation and the mean of the absolute value of the model parameters. This

measure is called the *Coefficient of variation* (CV) and a low CV indicates a better estimate of the model parameters. The total CV for a latent trait $\hat{\theta}$, c_v^{tot}, is defined as

$$c_v^{tot} = \frac{\frac{1}{S}\sum_{s=1}^{S}\sigma(\hat{\theta}_s)}{\frac{1}{S}\sum_{s=1}^{S}|\hat{\theta}_s|} \tag{4.43}$$

Where $\frac{1}{S}\sum_{s=1}^{S}\sigma(\hat{\theta}_s)$ is the subject mean of the estimated trait's standard deviation and $\frac{1}{S}\sum_{s=1}^{S}|\hat{\theta}_s|$ is the subject mean of the absolute value of the estimated trait. Since the CV is a dimensionless measure, it allows for comparing different models with varying mean estimates.

N.B. the regular CV has been modified in order to be an adequate measure for the BII. The trait and parameter estimates for the BII are not cast on a positive scale, some latent trait values can be close to zero and then the CV will increase greatly. This is the reason why we take the absolute value of the subject mean estimate and sum over all the estimates in the denominator.

4.4.2 Deviance Information Criteria (DIC)

The *Deviance Information Criteria* (DIC) is a measure of model fit and useful in Bayesian model selection where the posterior distributions of the model parameters have been obtained through a Markov Chain Monte Carlo analysis. DICs are comparable only over models constructed from the same data set, but there is no need for the models to be nested (Spiegelhalter et al., 2002).

The deviance is defined as $D(\theta) = -2log(P(X|\theta))$ where X is the data, θ the unknown model parameter(s) and $P(X|\theta)$ the likelihood function. $D(\theta)$ is a function of θ and thus it can be seen as a posterior distribution with expectation $\bar{D} = \mathbf{E}^{\theta}[D(\theta)]$. $\bar{D}$ can be computed, given the law of large numbers, as the mean of the estimated deviance in each Monte Carlo step. High values of the deviance indicates low values of the log-likelihood and thus a poorer model fit.

More complex models almost always fit the observed data better and therefore produce a higher log-likelihood than simpler models. However this does not necessarily mean that more complex models are a better fit for unobserved data. Thus, the DIC takes the model's degrees of freedom, p_D, into account to obtain a better estimate of the true model fit. Hence, the DIC is defined as

$$DIC = \bar{D} + p_D \tag{4.44}$$

The p_D value used in JAGS is computed based on an approach suggested by Plummer (2008). The details of this approach is outside the scope of this thesis, but in short it is estimated by taking the sample mean of the Kullback-Leibler information divergence between the chains.

4.4.3 Posterior Predictive Check (PPC)

In Bayesian statistics in order to asses the model fit one can perform a *Posterior Predictive Check* (PPC). In PPC predicted data is simulated from the fitted model and then compared to the observed data in order to see how well the model's estimated parameters can replicate the original data set.

The predictions are made by draws from the *posterior predictive distribution*, which is the distribution of unobserved predictions conditional on the observed data and the estimated model parameters. In the case of the BII predictions are drawn from a categorical distribution following the probability distribution of response k to item i for subject s.

The predictions are then compared to the real data, where the mean of the predicted (replicated) subject responses $\mu(\boldsymbol{X}_s^{rep})$ is compared to the mean of the subject responses in the real data set $\mu(\boldsymbol{X}_s)$. An attempt to replicate the full data set is carried out once for every iteration in the MCMC chains, hence there will be in total 18000 replicated data sets, all with different parameter estimates, to make comparisons with.

Mean Posterior Predictive Checks The PPC presented in Gelman et al. (2000) are statistics realted to the mean:

- $p_1^{mean} : \mathbb{E}(\mathbb{1}_{\mu(\boldsymbol{X}_s^{rep})=\mu(\boldsymbol{X}_s)})$
- $p_2^{mean} : \mathbb{E}(\mathbb{1}_{\mu(\boldsymbol{X}_s^{rep})>\mu(\boldsymbol{X}_s)})$
- $p_3^{mean} : \mathbb{E}(\mathbb{1}_{\mu(\boldsymbol{X}_s^{rep})<\mu(\boldsymbol{X}_s)})$

and the standard deviation statistics of interest are:

- $p_1^{std} : \mathbb{E}(\mathbb{1}_{\sigma(\boldsymbol{X}_s^{rep})=\sigma(\boldsymbol{X}_s)})$
- $p_2^{std} : \mathbb{E}(\mathbb{1}_{\sigma(\boldsymbol{X}_s^{rep})>\sigma(\boldsymbol{X}_s)})$
- $p_3^{std} : \mathbb{E}(\mathbb{1}_{\sigma(\boldsymbol{X}_s^{rep})<\sigma(\boldsymbol{X}_s)})$

The values p_1^{mean} and p_1^{std} are where the row mean and/or standard deviation of the replicated data set is exactly equal to the real data set. This is quite unlikely as there are uncertainties in the estimates. However, if p_2^{mean} and p_3^{mean} as well as p_2^{std} and p_3^{std} are equally distributed — i.e., there is an equal probability that the row mean as well as the row standard deviation is slightly lower or higher than compared to the real data set — then one can conclude that the model is relatively good at predicting itself. If so, the model fit is adequate according to the PPC method.

Ideally the distributions of the replicated values for both the mean and the standard deviation for every subject converge towards a normal distribution. The means of these standard distributions should ideally be $\mathbb{E}(\mu(\boldsymbol{X}_s))$ and $\mathbb{E}(\sigma(\boldsymbol{X}_s))$, respectively. This can be graphically checked by plotting a histogram of the resulting statistics for the 18000 replicated data sets and drawing a vertical line for the value of $\mathbb{E}(\mu(\boldsymbol{X}_s))$ and $\mathbb{E}(\sigma(\boldsymbol{X}_s))$ in the plots.

A common argument against using the PPC method is that the data is used twice. The argument "using the data twice" means that you use your data for estimating the model, and then for checking if the model fits the data. Even though some argue that it would be better to validate the model with external data not used for the estimation, the PPC method is still an accepted method used to assess model fit. All in all, posterior predictive checks are helpful in assessing if the model yields "valid" predictions. However, it should be noted that it does not give a definite answer if the model is adequate or if it is better than another model Gelman et al. (2000).

4.5 Variable Reduction and Construction of the Index Algorithm

The HO-IRT model provides a theoretically valid estimation of the latent traits for each subject s, but it has practical limitations when applying it to new test-takers. It is possible to add the new test subjects' data to the existing set and redo the HO-IRT MCMC simulations, but this method is time consuming and it would not be possible to generate an index score instantly after the test-taker has completed the questionnaire (as the MCMC simulations take several hours to complete on a contemporary, high performance computer). To enable faster estimation of latent trait scores, even though the ease of computation comes at the expense of estimate accuracy, we therefore wish to fit the latent traits to a set of linear regression models, as presented in equation (4.45). Each model has one of the seven traits θ as its *dependent variable* and the BII data set with item responses, $\mathbf{X}$, as the *explanatory variables*.

$$
\begin{aligned}
\boldsymbol{\theta}^2 &= \boldsymbol{X}\boldsymbol{c}_0^T + \boldsymbol{\epsilon}_0 \\
\boldsymbol{\theta}_1^1 &= \boldsymbol{X}\boldsymbol{c}_1^T + \boldsymbol{\epsilon}_1 \\
\boldsymbol{\theta}_2^1 &= \boldsymbol{X}\boldsymbol{c}_2^T + \boldsymbol{\epsilon}_2 \\
\boldsymbol{\theta}_3^1 &= \boldsymbol{X}\boldsymbol{c}_3^T + \boldsymbol{\epsilon}_3 \\
\boldsymbol{\theta}_4^1 &= \boldsymbol{X}\boldsymbol{c}_4^T + \boldsymbol{\epsilon}_4 \\
\boldsymbol{\theta}_5^1 &= \boldsymbol{X}\boldsymbol{c}_5^T + \boldsymbol{\epsilon}_5 \\
\boldsymbol{\theta}_6^1 &= \boldsymbol{X}\boldsymbol{c}_6^T + \boldsymbol{\epsilon}_6
\end{aligned}
\tag{4.45}
$$

where the linear regression equations in (4.45) make use of the following parameter matrices:

$$
\boldsymbol{\theta}^2 = \begin{bmatrix} \theta_1^2 \\ \vdots \\ \theta_{878}^2 \end{bmatrix} \quad \boldsymbol{\theta}_*^1 = \begin{bmatrix} \theta_{*,1}^1 \\ \vdots \\ \theta_{*,878}^1 \end{bmatrix} \quad \boldsymbol{c}_*^T = \begin{bmatrix} c_{*,0} \\ c_{*,1} \\ \vdots \\ c_{*,24} \end{bmatrix} \quad \boldsymbol{\epsilon}_* = \begin{bmatrix} \epsilon_{*,1} \\ \vdots \\ \epsilon_{*,878} \end{bmatrix}
$$

$$
\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & .. & x_{1,24} \\ \vdots & \vdots & \ddots & \\ 1 & x_{878,1} & .. & x_{878,24} \end{bmatrix}
$$

where the subscript $*$ can take on any of the following values: $(0), 1, 2, 3, 4, 5, 6$ (which corresponds to the traits *(innovation), trust, resilience, diversity, belief, perfection, collaboration* respectively). $\boldsymbol{X}$ is the data set with all the answers to the questionnaire, where $x_{s,i}$ is subject s' answer to item i. The columns of the $\boldsymbol{X}$ matrix corresponds to the questions/items in the data set, namely $QT1, QT2, ...QC3, QC4$ (all the questions are specified in Appendix A).

One aim of the thesis is to analyze the possibility of a variable reduction, i.e. the possibility of reducing the number of items without too much loss of estimation accuracy. In other words, we want to reduce the number of questions such that only the most important ones are used in the analysis. A set of linear models with differing dependent variables, but with the same explanatory variable is called a *multivariate linear regression model* and there is no known method that can be used for variable reduction on this

model, where the same explanatory variables are deemed insignificant simultaneously in each of the separate linear regression models.

The method used here will therefore be based on an iterative use of *feature selection methods* over the seven regression models. For each model every variable is assigned an importance estimate. Thereafter a full feature ranking, for the whole set of models, will be obtained by ranking the variables given their average importance over all of the fitted models. The feature ranking methods that will be used are the *Boruta* and the *GBM* and thus two feature rankings will be produced.

For each feature ranking 24 sets of the seven linear regression models are evaluated with $l = 1, ..., 24$ regression variables. The model with only one variable, $l = 1$, will only regress on the variable that contains most information according to the feature ranking method employed. When $l = 2$ the second most important variable is added etc. The last model contains all possible explanatory variables, i.e. $l = 24$.

Each linear regression model is fitted using *cross-validation*. Cross-validation means that the data set is randomly split in two parts, here 90% of the data is used as *training data* to fit the model and 10% is used as *test data* to test the model fit. This is done 500 times and the model with the best fit, according to the *root mean square error* (RMSE) given the test data, is used as the final model.

The goodness of fit for each of the 24 sets of models are then then compared using the mean of the RMSE, the *Akaike Information Criteria* (AIC), and the *Bayesian Information Criteria* (BIC) over all the seven linear regression models in the set (4.45). The final variable reduction will be determined based on a trade-off between goodness of fit and the number of parameters in the model.

4.5.1 Additional Goodness of Fit Measures

The RMSE determines the size of the average error, i.e., the deviance between true and the estimated value of the trait, θ, and is given by

$$RMSE = \sqrt{\frac{\sum_{s=1}^{S}(\hat{\theta}_s - \theta_s)^2}{S}} \tag{4.46}$$

The DIC presented earlier is the hierarchical modeling generalization of the AIC and the BIC. The AIC and the BIC measures the relative goodness of fit of nested models.

When fitting a model it is possible to increase the likelihood by adding more parameters, but more complex models might perform worse when they are evaluated on new data

(the model is over fitted). Therefore, both the AIC and the BIC shown in (4.47) and (4.48) are computed by adding a penalty term, related to the number of coefficients, to the estimated deviance $D(\theta) = -2log(P(X|\theta))$. Therefore, nested models with differing numbers of model parameters may be more accurately compared.

$$AIC = D(\theta) + 2l \tag{4.47}$$

$$BIC = D(\theta) + l\log(S) \tag{4.48}$$

in which l is the number of parameters in the linear regression model and S is the number of observations in X, i.e. the number of subjects.

4.5.2 Feature Selection

As mentioned earlier the following feature selection methods are used to obtain the feature rankings:

Boruta: A wrapper around the random forest algorithm

GBM: The Gradient Boosting Method

Boruta

The Boruta algorithm in R is a wrapper built around the random forest classification algorithm. The algorithm determines relevance of features by comparing them to the relevance of random probes. The random forest classification gives a numerical estimate of the feature importance that can be used to compare the importance features in a data set (Kursa and Rudnicki, 2010).

The Boruta algorithm is performed by voting of multiple unbiased weak classifiers (decision trees). The importance measure, Z, is calculated as the accuracy loss of classification caused by random permutation of attribute values between objects. Each tree in the algorithm is given an attribute for classification and the average and standard deviation of the accuracy loss are computed. The Z score is defined as the importance measure based on the variations of the mean accuracy loss among trees.

Kursa and Rudnicki (2010) summarizes the Boruta algorithm with the following steps:

1. Produce copies of all attributes and put them in the information system.

2. Shuffle the copied attributes to remove their correlations with the response and label the shuffled copies as shadow attributes.
3. Run a random forest classifier on the extended information system and gather the Z scores.
4. Find the *Maximum Z Score among Shadow Attributes* (MZSA), and then assign an importance score to every attribute that scored better than MZSA.
5. For each attribute with undetermined importance perform a two-sided test of equality with the MZSA.
6. Deem the attributes which have importance significantly lower than MZSA as "unimportant" and permanently remove them from the information system.
7. Deem the attributes which have importance significantly higher than MZSA as "important".
8. Repeat the procedure until an importance score Z is assigned for all the attributes, or the algorithm has reached the user defined limit of random forest runs.

The Boruta algorithm is run on all seven linear regression models in (4.45). The importance of every question is then labeled Z_p^i, e.g. Z_0^{QT1} is the importance of the question $QT1$ in the linear regression equation that has overall innovation capability, θ^2, as the dependent variable. Following the subscript notations introduced in (4.45) the subscript notation $p = 0$ corresponds to the importance value of some question for θ^2, $p = 1$ corresponds to θ_1^1 etc.

In order to compare the different features for all the seven linear regression models we also form the mean importance $\bar{Z}$ for every feature. E.g., the total Z score for $QB3$ is $\bar{Z}^{QB3} = \frac{1}{7}\sum_{p=0}^{6} Z_p^{QB3}$.

Gradient Boosting Method (GBM)

The Gradient Boosting Method (GBM) utilizes a prediction model based on decision trees in order to classify the importance of features. It can be seen as a boosted extension of the random forest algorithm. It distinguishes weak learners (that are not highly correlated with the true classification) with strong learners (that are highly correlated with the true classification).

The GBM algorithm implemented in the R package `gbm` is explained thoroughly in Ridgeway (2007).

Briefly the algorithm finds a regression function $\hat{f}(\boldsymbol{x})$ that minimizes the expectation of a loss function $\Psi(\theta, f)$.

The regression function $f(\boldsymbol{x})$ is assumed to be a function with a finite number of parameters, β. Estimations are carried out by selecting the values that minimize the loss function over a training sample of N observations for $(\theta, \boldsymbol{x})$:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(\theta_i, f(\boldsymbol{x}_i; \beta))$$

The steps performed in the GBM package implemented in R are shown in Psuedocode 3.

GBM initialization:
Select a loss function (distribution) Ψ, the number of iterations T, the depth of each decision tree,K, the sub-sampling rate p, and specify the shrinkage (learning rate) parameter λ.

Pseudocode 3 Steps performed for feature ranking with the GBM

Initialize $\hat{f}(\boldsymbol{x})$ as constant, $\hat{f}(\boldsymbol{x}) = \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(\theta_i, \rho)$

1: **for** $t \leftarrow 1$ to T **do**
2: Compute $z_i = -\frac{\partial}{\partial f(\boldsymbol{x}_i)} \Psi(\theta_i, f(\boldsymbol{x}_i))\Big|_{f(\boldsymbol{x}_i)=\hat{f}(\boldsymbol{x}_i)}$
3: Randomly select $p \times N$ cases from the data set
4: Fit regression tree with K terminal nodes, select $g(\boldsymbol{x}) = \mathbb{E}(z|\boldsymbol{x})$ from random observations
5: Compute $\rho_k = \underset{\rho}{\operatorname{argmin}} \sum_{\boldsymbol{x}_i \in S_k} \Psi(\theta_i, \hat{f}(\boldsymbol{x}_i) + \rho)$
6: Update $\hat{f}(\boldsymbol{x})$ as $\hat{f}(\boldsymbol{x}) \leftarrow \hat{f}(\boldsymbol{x}) + \lambda_{\rho_k(x)}$
7: **end for**

In the above Pseudocode 3 ρ is the optimal terminal node predictions. S_k is the set of $\boldsymbol{x}$:s that define terminal node k. $k(\boldsymbol{x})$ indicates the index of the terminal node with features $\boldsymbol{x}$.

The GBM model above is fitted to all the seven linear regression equations specified in (4.45). In order to assess the importance of each feature in our data set, we employ repeated cross-validation on the fitted GBM models. For each iteration, the model is trained on 90% of the data, selected at random. The prediction accuracy is recorded as the RMSE at each iteration. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all iterations, and normalized by the standard deviation. Then the sum of all the importance values is

calculated over all the boosting iterations. N.B. if the standard error is equal to 0 for a variable, the division is not done.

4.5.3 Scaling of the Index

After the variable reduction is done, and the latent trait scores for each subject has been calculated via the linear regression models, the final algorithm scores can be obtained by casting the regression results on the interval $I = [1, 10]$. The reason to map the latent trait estimates to a new interval is partly to meet the requirements of the index, but also to facilitate comparisons between the latent trait scores.

The HO-IRT model is constructed such that each latent trait estimate is almost equally distributed with mean $\mu_\theta = 0$. This means that the worst and the best score in each domain will be roughly equal for all different traits, which is contradictory to the fact that the mean item response for some traits differ considerably. Thus, the different trait scores produced by the HO-IRT model are cast on different scales and to enable comparisons of the scores between domains they must be cast on a common scale.

It should be noted that, as mentioned in Section 4.2.5, a comparison between first-order traits should only be done for a within-person comparisons and that a between-person comparison is only valid for the second-order trait.

The lowest latent trait score should be obtained when a subject responds $x_i = 1$, $\forall i = 1, ..., 24$ and the lowest first-order latent trait score for domain j should be obtained when a subject responds $x_i = 1$, $\forall i \in j$. Analogously the highest score for the second-order trait should be obtained when $x_i = 5$, $\forall i = 1, ..., 24$ and max for each first-order trait when $x_i = 5$, $\forall i \in j$.

The best and worse scores for the variable θ are annotated $\min(\theta)$ and $\max(\theta)$ and they are used to map θ values to a desired interval, i.e. $[\min(\theta), \max(\theta)] \rightarrow [1, 10]$. The algorithm used to map an arbitrary latent score θ to the interval $I[1, 10]$ is presented in equation (4.49).

$$\theta_{new} = \frac{(\theta_{old} - \min(\theta))(10 - 1)}{(\max(\theta) - \min(\theta))} + 1 \tag{4.49}$$

4.6 Further Analysis

4.6.1 Exploratory Analysis

In the BII it is assumed that the model structure is as shown in Figure 4.4 in section 4.2.5, but it is important that we can support this claim regarding the structure of the data. This can partly be done with an exploratory analysis. The goal of the exploratory analysis is to confirm, or discard, the assumptions we have made about the data structure. The structure assumptions on our data is that there exist one second-order trait and six first-order traits, and that each test item belongs to a domain governed by a specific first-order trait. E.g., the first-order trait *trust* can best be explained by the responses to the items *QT1*, *QT2*, *QT3* and *QT4*. The full explanation of the methods used in the exploratory analysis are outside the scope of this thesis, but short summaries will be presented below.

To confirm the assumption about the number of latent traits measured by the BII we will use *Principal Component Analysis* (PCA) and *Exploratory Factor Analysis* (EFA) to confirm or reject the hierarchical model structure.

The goal of the PCA is to find a set of k *principal components* where k is much smaller than the dimension of the original data set, but accounts for nearly all of the variability (information) of the data set. We can see the item response matrix as a set of data vectors described by i dimensions, in this case items. PCA transforms these vectors into a set of i new orthogonal vectors called *principal components*. The first principal component contains the most information and the following $i - 1$ components contains as much information as possible while fulfilling the requirement of being orthogonal to the former components. The first k *principal components* will contain most of the information of the data and thus we can answer the question of how many components are needed to accurately represent the data.

The results obtained from the PCA when run on the BII data set is presented in Figure 4.5.

In the latent trait model framework the PCA is used to confirm, or reject, the assumption of the number of underlying traits. The hypothesis is that these latent variables are those modeling the subjects' response patterns. Due to the hierarchical structure assumption, the hypothesis is that we will find one component with a lot of information (the second-order trait, i.e. innovation) and six components with less information, but still significant (corresponding to each of the six first-order traits). In the PCA a high level of information is equivalent with a high eigenvalue.

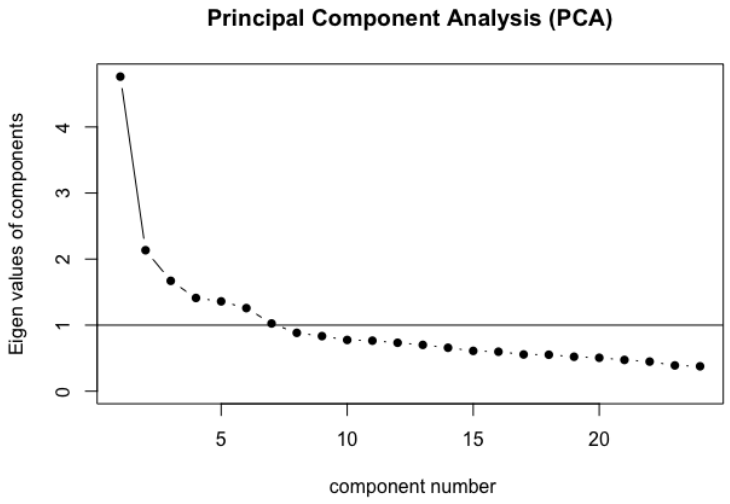


FIGURE 4.5: Eigenvalues of principal components

One critique against using the PCA method for exploratory analysis is that there does not exist any statistic to determine the number of high information variables. Therefore, the interpretation of the results is deemed to be very subjective. However, without making a big conclusion, what we can see is that the underlying data structure at least does not reject the idea of one overall latent trait, which accounts for a lot of information, and a set of six first-order latent traits. The interpretation of what the high information components represents is also highly subjective, therefore we will not draw the conclusion that the components actually represents the specified latent traits, but simply state that we can not reject this assumption.

The Exploratory Factor Analysis (EFA) models the latent data structure as a set of simultaneous linear models

$$\mathbf{X}_s = \boldsymbol{\mu}_X + \boldsymbol{\Lambda}\mathbf{F}_s^T + \boldsymbol{\epsilon}_p \tag{4.50}$$

where $\mathbf{F}_s$ is, for subject s, a set of n latent variables called factors, $\mathbf{X}_s$ is the response vector for subject s (the subject's manifest variables), $\boldsymbol{\mu}_X$ is the vector containing a mean parameter for each linear model, $\boldsymbol{\Lambda}$ is a matrix of *factor loadings* and $\boldsymbol{\epsilon}$ is the vector of residuals for each linear model (one model for each item). The model parameters are estimated using Maximum Likelihood.

The goal is to estimate how the model would look like if a set of n hypothetical latent factors had constructed the data set. The ordinary EFA assumes that the factors are uncorrelated and it assumes a non-hierarchical model structure, but it can be extended to the higher-order EFA. The higher-order EFA method used here is described in McDonald

(1985) and in short one makes an ordinary exploratory factor analysis on the data, rotate the factors to allow for correlation and then performs a Schmid-Leiman transformation (Schmid and Leiman (1957)) of the rotated factors (i.e., attribute the variation from the first-order factors to the second-order factors).

The fact that the EFA model is based on a linear assumption makes it less suited for our analysis, but it allows us to obtain a crude estimate of the factor loadings (given a set of n factors) and thus identify which domain an item supposedly belongs to and then confirm or reject the assumed hierarchical model structure. The model structures given $n = 3, 4, 5, 6$ are presented in figure 4.6(a)-(d). From the results we can draw the conclusion that six factors, or first-order latent traits, seems reasonable for the BII data set and the questions in the data set are also grouped together in the same domain as defined in the model assumptions (i.e., QT1,..QT4 belong to the trust domain).

4.6.2 Simulated Data

A study of the models on different sets of simulated data have been carried out in which the goal was twofold. The first reason was to ensure that our model yields a satisfying result under the assumption made on the data set, i.e., that the data follows the assumed hierarchical structure. The other aim was to, with simulated data sets of differing structure, enable an analysis of how different data structures affect the accuracy of the results. This analysis is conducted to provide a foundation for further development of the BII.

The simulation of the data sets are done in MatLab and the distribution of the model parameters are the following:

- $\theta^{(2)} \sim \mathcal{N}(0, 1)$
- $\epsilon \sim \mathcal{N}(0, 0.5)$
- $\lambda = [x : x = 0.6 + 0.3n/J, n \in (0, 1, ..., J)]$
- $\alpha \sim \mathcal{N}(1.8, 0.25)$
- $\beta = [(\beta_0 - 1.5, \beta_0 - 0.5, \beta_0 + 0.5, \beta_0 + 1.5), \beta_0 \sim \mathcal{U}(-1.5, 1.5)]$

The $\theta^{(1)}$ parameters are computed as in equation (4.20). The simulated response matrix is also analyzed manually and if it differs too much from the real data set the simulation is discarded (in order to obtain results that are relevant for the true data set).

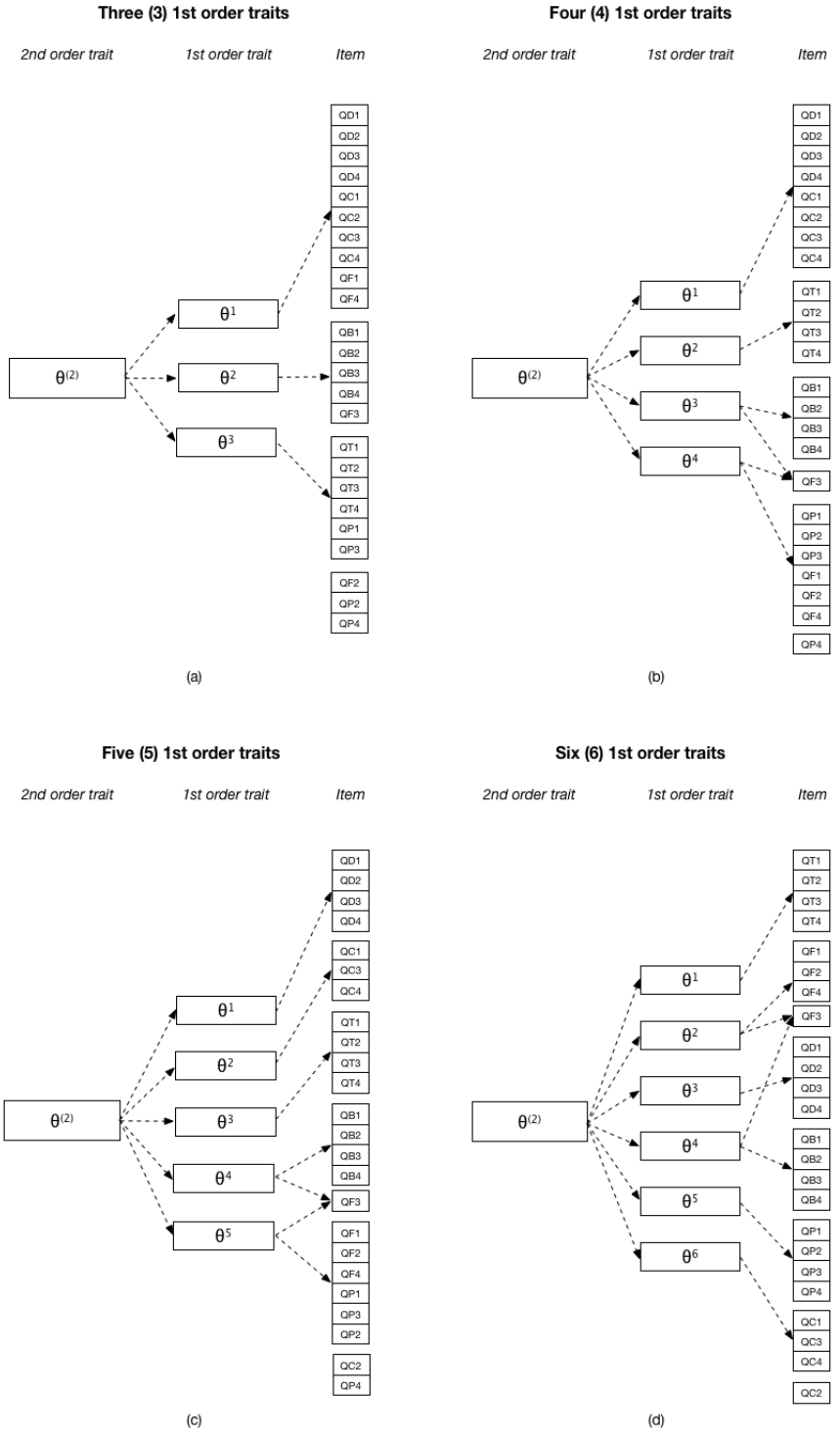


FIGURE 4.6: Higher-order factor analysis model structure for (a) three first-order latent traits, (b) four first-order latent traits, (c) five first-order latent traits and (d) six first-order latent traits

Since the β parameters are evenly spaced in the simulated data this will guarantee that the β parameter will not lower the amount of information given by an item. This will result in a data structure that is easier to model and the correlations will be higher.

The MCMC simulations were carried out in JAGS through the R-package `RJags`. Each simulation was run with three parallel chains over 6000 iterations, including 3000 burn in iterations.

4.6.3 Outlier Analysis

To determine the robustness of the model and how outliers might affect the results two leave-one-out analyses were conducted where "extreme" values from the BII data set were omitted. The data entries discarded in the two analyses were the lowest and the highest (row) means of the subjects' responses, i.e., $\mathbb{E}(\boldsymbol{X}_s) = \frac{1}{24}\sum_{i=1}^{24} x_{s,i}$. The highest scoring subject had a row mean value equal to 5, and the lowest scoring subject had a row mean value equal to 2.375.

Subject responses omitted from the data set in the outliers analysis

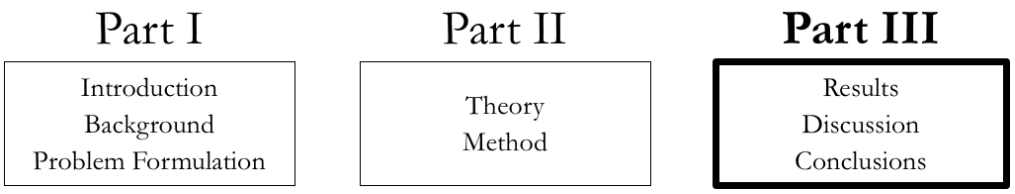
- $\mathbb{E}(\boldsymbol{X}_{375,i}) = 5 \quad \forall i$
- $\mathbb{E}(\boldsymbol{X}_{205,i}) = 2.375 \quad \forall i$

The two data sets cleaned from outliers, $\boldsymbol{X}'_{hi}$ and $\boldsymbol{X}'_{low}$, now has the maximum and minimum row mean value of $\max(\mathbb{E}(\boldsymbol{X}'_{hi})) = 4.875$ and $\min(\mathbb{E}(\boldsymbol{X}'_{low})) = 2.667$ respectively.

The full HO-IRT MCMC analysis was run on both the reduced data sets, $\boldsymbol{X}'_{hi} \in \mathbb{N}^{877\times 24}$ and $\boldsymbol{X}'_{low} \in \mathbb{N}^{877\times 24}$, in order to assess how these "extreme" values in the data set affect the results.

Part III

Consequences of Measuring the Unmeasurable



Key takeaways: In *Part III* the model selection analysis concluded that the *Higher-Order Graded Response Model* (HO-GRM) is the model of choice for the BII algorithm. The resulting index also seems to promote items with less variability. The variable reduction analysis showed that all items are relevant in the model, however an alternative reduced question set with the 17 most relevant items is also presented. In the Discussion chapter the results are analyzed and eventual doubts and inaccuracies are highlighted. Recommendations as well as improvements that can be made are presented and possible future research is put forward.

Chapter 5

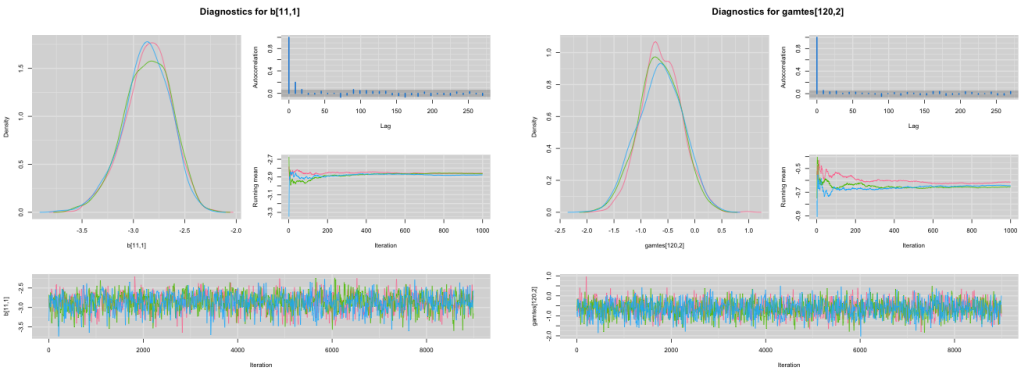
Results and Analysis

Happy people plan actions, they don't plan results.
—Denis Waitley, Motivational Speaker

5.1 Model Selection (HO-GRM vs HO-GPCM)

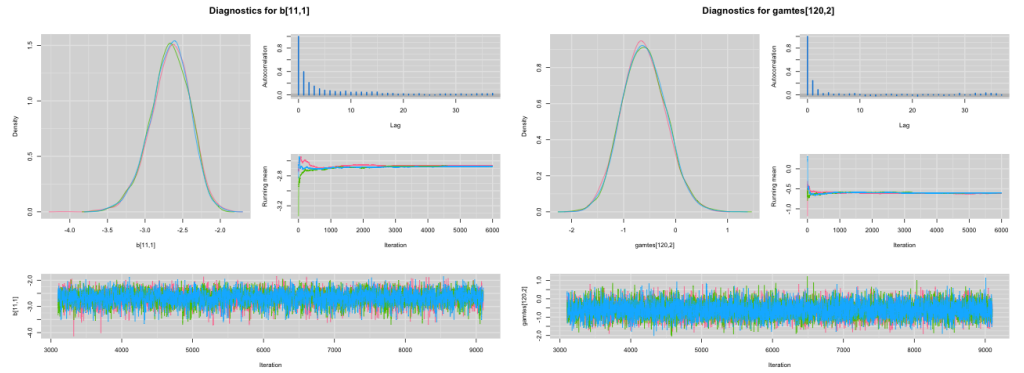
To check for convergence of the MCMC chains two of the most common methods to use are graphical analysis of chain convergence and the Gelman Rubin (G-R) diagnostic, i.e. the $\hat{R}$-value. The graphical analysis is done by inspecting the trace plots, the estimated posterior distribution of the parameters, the ACF (auto-correlation function) and the running mean.

Due to the vast number of estimated model parameters it is not feasible to present the full graphical convergence analysis for the HO-GRM and the HO-GPCM. Instead general results are presented in figure 5.1 and 5.2 which shows all the convergence graphs for the parameters $\beta_{11,1}$ and $\theta^{(1)}_{2,120}$ for the HO-GRM (figure 5.1a and 5.1b) and for the HO-GPCM (figure 5.2a and 5.2b). N.B. The top-left graph in each subplot is the estimated parameter distribution, the top-right plot is the autocorrelation function, the middle-right plot is the parameter running mean, and the bottom plot is the trace plot (after the burn-in has been discarded).



(a) (b)

FIGURE 5.1: Graphical analysis of the convergence for the parameters (a) $\beta_{11,1}$ and (b) $\theta^{(1)}_{2,120}$ in the HO-GRM model.



(a) (b)

FIGURE 5.2: Graphical analysis of the convergence for the parameters (a) $\beta_{11,1}$ and (b) $\theta^{(1)}_{2,120}$ in the HO-GPCM model

As noted before this is just a (very small) sample to give the reader a notion of the appearance of the different graphs. The graph in the bottom is the trace plot that shows each chain over the last 9000 iterations. Since the values of the chain vary around a mean with more or less the same deviance for the three parallel chains this indicates that the draws are made from a stable distribution. The running mean for both parameters slowly converges toward a value that is assumed to be the mean of the parameter posterior distribution. The ACF indicates that there is little to no correlation between sequential draws from the distribution and thus the chains mix well, i.e. successive draws from the distribution are (almost) independent of one another. Overall, the full graphical analysis for the whole simulation showed no signs of lack of convergence for any parameter.

To further confirm that the chains have converged the G-R diagnostic, $\hat{R}$, is computed and the maximum value of this statistic for each type of parameter is presented in 5.1. If $\hat{R} < 1.1$ for all parameter types that indicates that the chains have converged. Thus,

we can conclude that all the MCMC chains for both the HO-GRM and the HO-GPCM have fully converged.

	$\max(\hat{R})$					
Model	α	β	λ	$\theta^{(2)}$	$\theta^{(1)}$	ϵ
HO-GRM	1.008	1.008	1.013	1.002	1.004	1.005
HO-GPCM	1.007	1.005	1.011	1.005	1.002	1.005

TABLE 5.1: Maximum value of the G-R statistic, $\max(\hat{R})$, for each parameter type for the HO-GRM and the HO-GPCM

To assess which model, HO-GRM or HO-GPCM, is the best one for the BII data set the goodness of fit measures obtained for both the models are compared. In the comparison the DIC and the mean variance of the latent trait estimates are employed. Since the same data set have been used for both models the DIC is a valid measure of comparison. The DIC and the mean variance for the second-order latent trait, $\theta^{(2)}$, and the first-order latent traits, $\theta_j^{(1)}$, for the HO-GRM and the HO-GPCM are presented in table 5.4.

		Variance, σ^2						
Model	**DIC**	$\theta^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	$\theta_6^{(1)}$
HO-GRM	48497.7	0.214	0.218	0.208	0.203	0.163	0.374	0.214
HO-GPCM	49704.0	0.232	0.224	0.231	0.220	0.183	0.398	0.236

TABLE 5.2: DIC and mean variance of the latent trait estimates for the HO-GRM and the HO-GPCM

To compare the accuracy of the latent trait estimates we will also compare the total Coefficient of variation (CV), c_v^{tot}, as defined in (4.43). These values can be found in Table 5.3

	Coefficient of Variation, c_v^{tot}						
Model	$\theta^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	$\theta_6^{(1)}$
HO-GRM	0.749	0.706	0.707	0.691	0.593	1.032	0.744
HO-GPCM	0.754	0.718	0.724	0.701	0.621	1.064	0.766

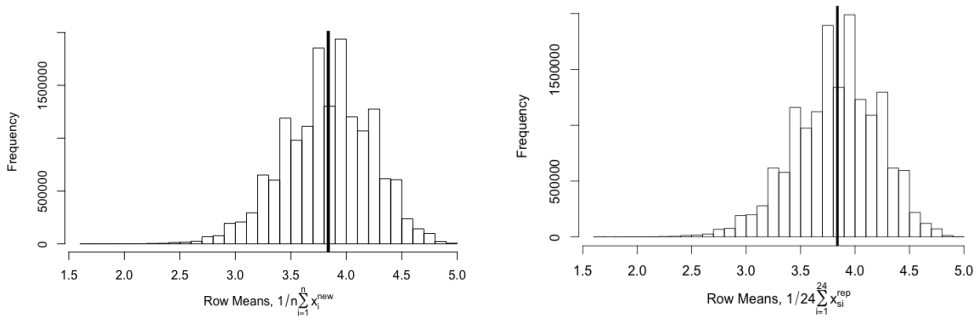
TABLE 5.3: c_v^{tot} for the latent trait estimates obtained from the HO-GRM and the HO-GPCM models

Furthermore, to assess the model fit we examine the Posterior Predicitive Check (PPC). PPC estimates how well the model can reconstruct the original data set, $\boldsymbol{X}$, into a new set of data $\boldsymbol{X}^{rep}$, given the model parameters estimated in each step of the MCMC

simulation. Even though the PPC should not be used to give a definite answer if one model is better than another, it will be used here as an indicator if either of the models are bad or if both yield a satisfactory result. In figure 5.3 the histogram of the frequencies of the mean item response for each subject at every iteration in the MCMC chains, $\mathbb{E}(\boldsymbol{X}_s^{rep}) = \frac{1}{n}\sum_{i=1}^{n} x_{si}^{rep}$, is shown. In figure 5.4 the histogram of the frequencies of the standard deviation of the item responses for each subject at every iteration of the MCMC chains, $\sigma(\boldsymbol{X}_s^{rep})$, is presented.

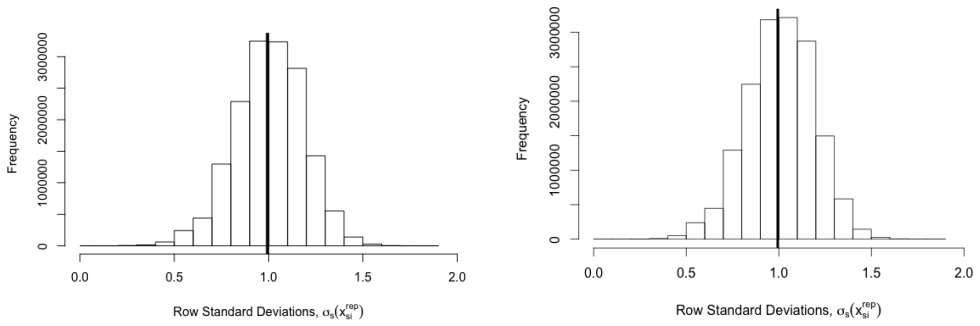
	$\mathbb{E}(\boldsymbol{X}_s^{rep})$			$\sigma(\boldsymbol{X}_s^{rep})$		
Model	p_1	p_2	p_3	p_1	p_2	p_3
HO-GRM	0.078	0.466	0.455	0.008	0.465	0.527
HO-GPCM	0.079	0.465	0.456	0.007	0.461	0.532

TABLE 5.4: PPC p-values of the mean item responses and subject standard deviation for the HO-GRM and the HO-GPCM



(a) HO-GRM (b) HO-GPCM

FIGURE 5.3: Histogram showing the frequencies of each subject's mean item response given the data set $\boldsymbol{X}_{rep}$ which is reconstructed from the model parameters at each step of the MCMC simulation. The vertical line represents the mean item response of all subjects in the original data set $\boldsymbol{X}$.



(a) HO-GRM (b) HO-GPCM

FIGURE 5.4: Histogram showing the frequencies of each subject's standard deviation given the data set $\boldsymbol{X}_{rep}$ which is reconstructed from the model parameters given in each step of the MCMC simulation. The vertical line represents the mean standard deviation of all subjects in the original data set $\boldsymbol{X}$.

All of these statistics indicate that the HO-GRM model performs better on the BII data set.

Since the data set will change when more data is added we also want to further assess the general performance of the models given new data with a structure that resembles the data structure in the BII. To do this we will first simulate a data set given the model structure of the HO-GRM and then fit both the HO-GRM model and the HO-GPCM to the same simulated data set. The same, but reversed, is done with data simulated from the HO-GPCM model. The performance is then evaluated based on the DIC and the correlation between the simulated "true" values for the latent traits and the estimated values. The results are presented in table 5.5

		2^{nd} **order**	1^{st} **order**					
Model (Data)	**DIC**	$\rho^{\theta^{(2)}}$	$\rho_1^{\theta^{(1)}}$	$\rho_2^{\theta^{(1)}}$	$\rho_3^{\theta^{(1)}}$	$\rho_4^{\theta^{(1)}}$	$\rho_5^{\theta^{(1)}}$	$\rho_6^{\theta^{(1)}}$
GRM (GRM)	63306.3	0.930	0.869	0.874	0.892	0.903	0.899	0.915
GPCM (GRM)	63699.6	0.928	0.867	0.868	0.891	0.902	0.895	0.911
GRM (GPCM)	59487.6	0.937	0.924	0.920	0.931	0.939	0.935	0.948
GPCM (GPCM)	59484.4	0.940	0.917	0.935	0.927	0.943	0.934	0.945

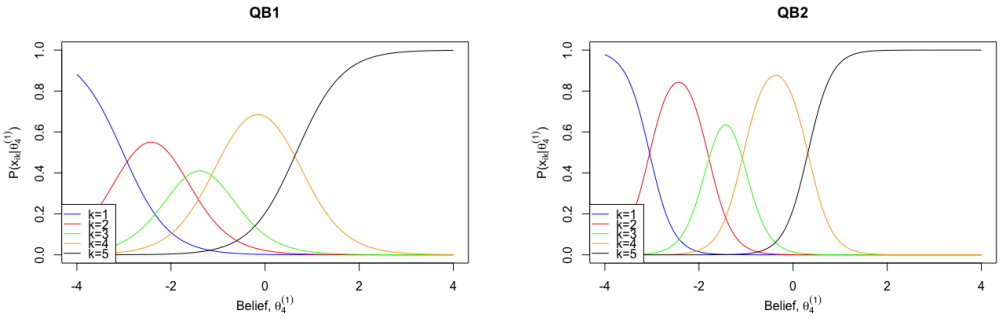
TABLE 5.5: Comparison of model fit when the HO-GPCM and the HO-GRM models estimated the model parameters from data generated by the HO-GRM. As well as the case when the HO-GRM and HO-GPCM estimated model parameters from data generated by the HO-GRM.

The result shows that, under the assumption that the hierarchical model assumption of the BII is correct, both models yield a satisfying result. It also further strengthens the indications that the HO-GRM model is the best fit for this type of data structure.

Therefore, HO-GRM is the model chosen for the construction of the BII algorithm and it is the results for this model that is presented throughout the rest of the results chapter.

5.2 HO-GRM Model Results

The full model specifications, i.e. all the model item parameters, of the HO-GRM is presented in table B.1 in Appendix B together with all the Item Characteristic Curves (ICC), Figures B.1-B.3, and the Item Information Functions, Figure B.5. In this section we will only present the curves for one domain of the model, namely *Belief*, to give the reader an overview of the resulting model from the MCMC. The domain Belief has been chosen since it contains items with both high and low item information and thus presents ICCs that best represents typical results. The ICCs are presented in Figure 5.5 and the IIC for the items in domain *Belief* are presented in Figure 5.6.



(a) ICC for item *QB1*

(b) ICC for item *QB2*

(c) ICC for item *QB3*

(d) ICC for item *QB4*

FIGURE 5.5: The Item Characteristic Curves for the items in the domain *Belief* given by the HO-GRM model

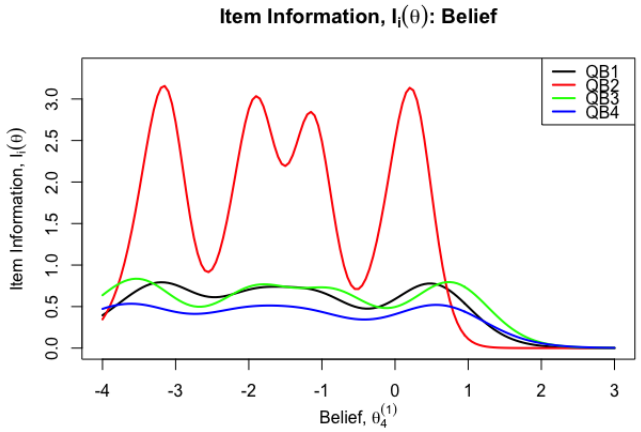


FIGURE 5.6: The Item Information Curves for the items in the domain *Belief* given by the HO-GRM model

5.3 HO-GRM Ability Estimates

In figure 5.7 the distribution of the second-order trait *innovation capability* is shown. The distribution of the MCMC estimates of all the traits can be found in Figure B.4 in Appendix B.

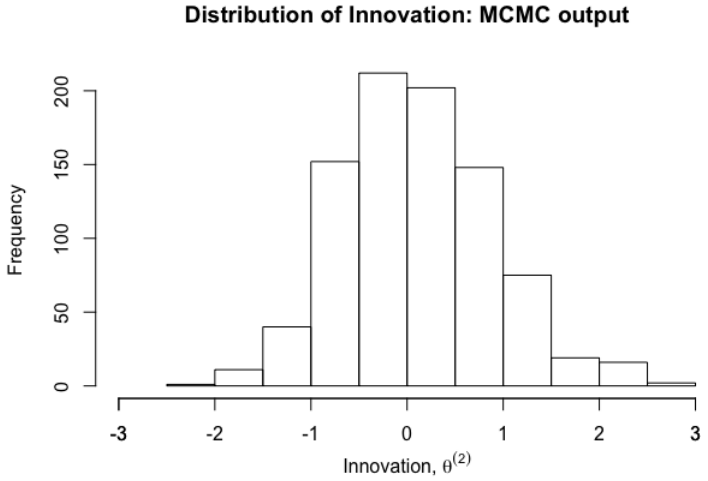


FIGURE 5.7: Histogram showing the distribution of the parameter estimates of the second-order trait *innovation capability*, $\theta^{(2)}$

Analytically, the correlation between the first-order traits j and j' is given by $\rho(\theta_j^{(1)}, \theta_{j'}^{(1)}) = \lambda_j \lambda_{j'}$ and the correlation between the first-order trait j and second-order trait is as the regression parameter λ_j. The estimated values for all the λs are given in table 5.6 and the estimated correlations between the traits are presented in table 5.7.

Parameter Estimate, (Standard Deviation, σ)					
$\hat{\lambda_1}(\sigma_{\lambda_1})$	$\hat{\lambda_2}(\sigma_{\lambda_2})$	$\hat{\lambda_3}(\sigma_{\lambda_3})$	$\hat{\lambda_4}(\sigma_{\lambda_4})$	$\hat{\lambda_5}(\sigma_{\lambda_5})$	$\hat{\lambda_6}(\sigma_{\lambda_6})$
0.51 (0.04)	0.857 (0.03)	0.80 (0.03)	0.67 (0.03)	0.42 (0.05)	0.75 (0.03)

TABLE 5.6: Mean of the parameter estimates and the SD_{post} for the HO-GRM model parameters λ_j where $j = 1, ...6$

	$\theta^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	$\theta_6^{(1)}$
$\theta^{(2)}$	1	-	-	-	-	-	-
$\theta_1^{(1)}$	0.58	1	-	-	-	-	-
$\theta_2^{(1)}$	0.94	0.47	1	-	-	-	-
$\theta_3^{(1)}$	0.88	0.45	0.77	1	-	-	-
$\theta_4^{(1)}$	0.74	0.33	0.67	0.57	1	-	-
$\theta_5^{(1)}$	0.51	0.36	0.43	0.39	0.31	1	-
$\theta_6^{(1)}$	0.84	0.46	0.72	0.69	0.51	0.42	1

TABLE 5.7: The estimated correlation matrix of the latent traits estimated by the HO-GRM model

We can conclude that the estimated correlation does not exactly follow the analytical assumption, but at least follows the same pattern. As mentioned in section 4.2.5 the domains with the highest λ, i.e., the domain abilities that have the highest correlation with the overall ability, also have the highest estimation accuracy. This is further validated and confirmed in the Structure Analysis presented in section 5.6. Thus, we can conclude that the domains *Resilience* and *Diversity* are best at estimating the latent trait score.

Table 5.8 compares the estimated latent trait estimates with the item responses of five selected subjects. The idea is to get an intuitive sense of the correctness of the estimated traits. If the estimated latent traits of a subject are very high it is hopefully correlated with generally high item responses. Two of the subjects, $s = 95$ and $s = 375$, are presented because their estimated second-order trait score was lowest and highest respectively. The three other subjects were chosen randomly.

One interesting result is that in the domain *Diversity* subjects $s = 95$ and $s = 375$ have identical item responses, but the level of their first-order trait $\theta_3^{(1)}$ is not equal. In fact it is much higher for subject $s = 375$. This is due to the fact that the first-order latent traits are correlated with one another as well as with the second-order latent trait. In the model the correlation between the domains j and j' are given by $\lambda_j \lambda'_j$ and this affects the estimates of the first-order traits. Therefore, all the item responses given by a subject in every sub-trait domain affects the estimates of all the first-order traits. Since subject $s = 95$ has scored lower in all the other domains this affects the result of $\theta_3^{(1)}$ negatively. However, as noted earlier, the first-order trait should only be used for within-person comparisons, and not between-person comparisons.

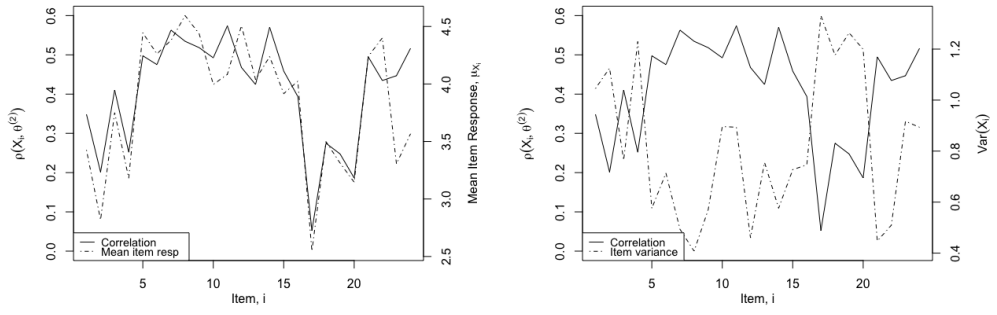
	Item Responses											
Subject, s	QT1	QT2	QT3	QT4	QF1	QF2	QF3	QF4	QD1	QD2	QD3	QD4
95	1	4	5	4	4	4	5	5	5	5	5	5
205	2	1	2	1	2	1	3	2	3	4	3	3
375	5	5	5	5	5	5	5	5	5	5	5	5
629	4	4	5	4	4	4	4	4	5	2	4	5
730	4	2	4	1	5	5	3	5	4	2	3	5

Subject, s	QB1	QB2	QB3	QB4	QP1	QFP2	QP3	QP4	QC1	QC2	QC3	QC4
95	4	5	4	4	2	3	2	3	4	4	4	4
205	1	1	1	3	3	5	1	4	3	4	1	3
375	5	5	5	5	5	5	5	5	5	5	5	5
629	4	4	3	4	2	2	2	2	4	5	2	2
730	5	5	3	4	1	2	4	2	5	5	3	4

	Parameter estimates of latent traits (standard deviation)						
Subject, s	$\hat{\theta}^{(2)}(\sigma_{\theta^{(2)}})$	$\hat{\theta}_1^{(1)}(\sigma_{\theta_1^{(1)}})$	$\hat{\theta}_2^{(1)}(\sigma_{\theta_2^{(1)}})$	$\hat{\theta}_3^{(1)}(\sigma_{\theta_3^{(1)}})$	$\hat{\theta}_4^{(1)}(\sigma_{\theta_4^{(1)}})$	$\hat{\theta}_5^{(1)}(\sigma_{\theta_5^{(1)}})$	$\hat{\theta}_6^{(1)}(\sigma_{\theta_6^{(1)}})$
95	0.48 (0.46)	0.41 (0.63)	0.28 (0.44)	1.00 (0.51)	0.42 (0.38)	-0.41 (0.54)	0.23 (0.43)
205	-2.38 (0.42)	-1.90 (0.43)	-2.47 (0.37)	-1.62 (0.38)	-3.089 (0.42)	0.04 (0.69)	-1.73 (0.43)
375	2.90 (0.66)	2.69 (0.67)	2.53 (0.74)	2.42 (0.74)	2.26 (0.69)	2.61 (0.72)	2.55 (0.67)
629	-0.56 (0.44)	0.95 (0.47)	-0.69 (0.39)	-0.23 (0.44)	-0.48 (0.37)	-1.14 (0.59)	-0.78 (0.46)
730	-0.01 (0.45)	0.00 (0.48)	0.07 (0.47)	-0.58 (0.42)	0.39 (0.41)	-0.91 (0.63)	0.48 (0.45)

TABLE 5.8: Parameter estimations and posterior standard deviation of the first- and second-order latent trait for five subjects given the their item responses

To give the reader a sense of what type of item characteristics that define an item's correlation with the second-order trait the plot in Figure 5.8a and 5.8b was produced. Figure 5.8a shows the correlation between the item response vector of an item i (X_i) and the second-order latent trait (θ^2), together with the mean item response, $1/N \sum_{i=1}^{N} X_i$. Figure 5.8b shows the same correlation together with the variance of the item response vector, $Var(X_i)$.



(a) (b)

FIGURE 5.8: Left axis (solid line): The correlation between second-order trait and item response vector X_i, $i = 1, ..., 24$. Right axis (dashed line): (a) Mean of item response, $\sum X_i/24$; (b) Variance of item response, $Var(X_i)$

The response pattern of the BII data is heavily weighted towards higher values, as can be seen on the mean of the item responses. It is natural to assume that "hard" questions

are characterised by a high variance in the responses, which can be found to be true if we compare figure 5.8b and 5.8a. Intuitively one might argue that these "hard" questions should have a high influence on the score of the index, but we clearly see that this is not the case. Questions with a high mean item response also have the highest correlation with the second-order trait. The main reason for this is the skewed data set. Since the majority of questions are "easy", i.e., a majority of the subjects give item response $x_i = 4$ or $x_i = 5$, the score for these subjects will also be quite high in the domains and overall. If some people give a high response to all items except for a few, and these few questions are the same for many people, these few low item responses will not lower the overall score remarkably. The result is that the correlation between the second-order trait score and the items with a higher difficulty/variance is significantly lower. This relationship will be relevant in the upcoming feature selection analysis since the feature selection techniques are based on the correlation between the dependent variable, i.e. the first- and second-order traits, and the explanatory variables, i.e. the item responses. Due to this, difficult questions will not be assigned a high value for their regression coefficient compared to easier questions and therefore the difficult questions will not be as important in the final result.

Another way to interpret this result is that it is crucial to give the correct answers to the "easy" questions in order to achieve a high innovation score. If the person does not answer these fundamental questions correctly it will affect the result severely since these questions are the foundation of the "innovative mindset". The more difficult items represents traits that are good to have, but they are not as crucial for an innovative mindset.

5.4 Variable Reduction

The results of the feature ranking analyses, done with the R packages `Boruta` and `GBM`, are presented below where $X_1 = QT1, X_2 = QT2, ..., X_{24} = QC4$, i.e. the item vectors are ordered as the questions in the survey (see Appendix A).

Boruta

$$X = [X_{14}, X_{11}, X_7, X_{24}, X_8, X_9, X_{21}, X_{10}, X_5, X_1, X_{23}, X_6, X_{12}, X_{15}, X_3, X_{13}, X_{22}, X_{18}, X_{16}, X_{19}, X_{20}, X_4, X_2, X_{17}]$$

GBM

$$X = [X_{14}, X_1, X_{11}, X_{21}, X_{18}, X_7, X_{24}, X_3, X_8, X_9, X_5, X_{23}, X_6, X_{12}, X_{15}, X_{10}, X_{20}, X_{19}, X_{13}, X_{22}, X_{16}, X_{17}, X_2, X_4]$$

The feature rankings form the basis of the variable reduction analysis. The first item in each vector is the one that contains most information according to the corresponding feature selection method. The number of regression variables, k, is the same as the length of the feature ranking vector where the last $24 - k$ elements have been removed. The goodness of fit of the resulting models given k variables is presented in table 5.9.

	Boruta			**GBM**		
k	$\frac{1}{7}\sum_{l=1}^{7} RMSE_l$	$\frac{1}{7}\sum_{l=1}^{7} AIC_l$	$\frac{1}{7}\sum_{l=1}^{7} BIC_l$	$\frac{1}{7}\sum_{l=1}^{7} RMSE_l$	$\frac{1}{7}\sum_{l=1}^{7} AIC_l$	$\frac{1}{7}\sum_{l=1}^{7} BIC_l$
1	0.540	1633.20	1647.21	0.535	1634.40	1648.42
2	0.473	1464.99	1483.68	0.480	1450.95	1469.64
3	0.453	1369.01	1392.37	0.426	1303.71	1327.07
4	0.407	1214.09	1242.12	0.394	1171.61	1199.64
5	0.390	1120.67	1153.37	0.359	1058.76	1091.46
6	0.373	1034.01	1071.38	0.338	948.64	986.02
7	0.342	935.15	977.20	0.317	838.20	880.25
8	0.337	862.78	909.50	0.294	725.58	772.30
9	0.321	792.76	844.15	0.275	628.86	680.25
10	0.280	617.29	673.36	0.257	534.33	590.39
11	0.273	506.44	567.18	0.247	466.17	526.91
12	0.253	441.48	506.88	0.230	357.87	423.28
13	0.250	368.71	438.79	0.219	287.56	357.64
14	0.234	282.81	357.57	0.216	231.60	306.36
15	0.218	169.27	248.70	0.200	152.50	231.93
16	0.214	96.76	180.85	0.186	35.89	119.98
17	0.203	9.74	98.50	0.173	-30.22	58.55
18	0.171	-127.83	-34.39	0.163	-137.18	-43.74
19	0.161	-184.23	-86.12	0.151	-208.07	-109.95
20	0.148	-241.52	-138.73	0.141	-297.68	-194.90
21	0.139	-361.08	-253.63	0.136	-359.33	-251.88
22	0.131	-422.60	-310.47	0.125	-523.09	-410.96
23	0.125	-522.29	-405.49	0.114	-615.42	-498.62
24	0.111	-683.05	-561.58	0.112	-683.36	-561.89

TABLE 5.9: The mean of the RMSE, AIC and BIC of the seven linear models given k explanatory variables ordered according to the ranking given by the adjusted Boruta and GBM feature selection methods

As we can see in table 5.9 both the RMSE, AIC and BIC improves faster with each added explanatory variable given the feature ranking of the GBM compared to Boruta. The reason for this is that, in the GBM, the difference of the information statistic is generally

higher between variables with high correlation to the dependent variable compared to variables with low correlation. When summing the information over each of the seven linear regression the GBM thus promote variables that have a high correlation with either of the dependent variables. In practice this means that for example item *QP2*, X_{18}, receives a higher ranking in the GBM compared to the Boruta. Since each of the linear models affect the mean of the RMSE, AIB and BIC equally, lower values of these statistics are obtained by having at least one highly correlated explanatory variable in each of the linear models. This is a purpose of the design of the feature selection algorithm to achieve a difference of accuracy between the latent trait estimates.

Our recommendation to the developers of the BII is to keep all the variables since they all contribute with information to the full model and yield a more accurate result. The results of the RMSE, AIC and BIC all strengthen this notion. The HO-GRM estimates of the traits have a mean standard deviance of $1/7 \sum_{k=1}^{7} \sigma_k = 0.48$. To add even more inaccuracy than needed to this estimate is not desirable. On the other hand, if we remove seven of the lowest information items of the feature ranking vector obtained through GBM we will only add $\Delta(RMSE) = 0.173 - 0.112 = 0.051$. Still the AIC and the BIC gets higher with each removed parameter which indicates that the variability lost with that parameter is not compensated by having a smaller set of explanatory variables. This can also be confirmed when examining the graphical result of the Boruta analysis, given the full set of items with the second-order trait as dependent variable, in Figure 5.9. This analysis shows that all the variables are deemed to be important for the regression by the Boruta (important variables are shown as green box-plots in the figure).

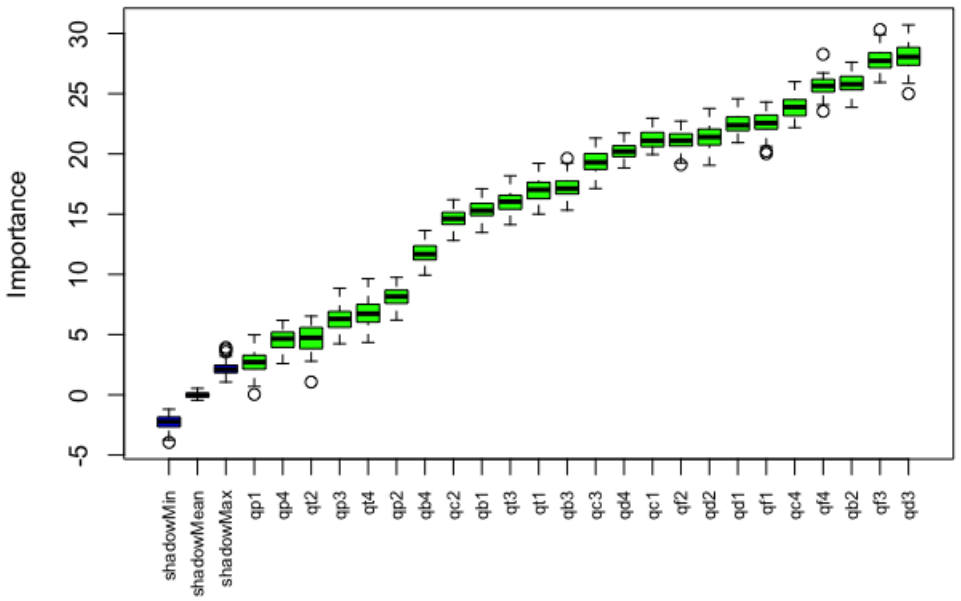


FIGURE 5.9: The result of the Boruta analysis with the second-order trait as the dependent variable, given the full set of items. It shows that all items are deemed to be relevant for the linear estimation of *innovation capability*

Despite the results, one of the goals of the thesis is to conduct a variable reduction and thus we will remove the seven items with the least information according to the GBM feature ranking vector (since this yielded a good trade-off between reduced number of items and loss of information). This will also enable us to compare the results from the reduced model with the results from the full model. The following items will be kept after the variable reduction has been performed:

Trust	**Resilience**	**Diversity**	**Belief**	**Perfection**	**Collaboration**
QT1	QF1	QD1	QB2	QP2	QC1
QT3	QF2	QD2	QB3	QP4	QC3
	QF3	QD3			QC4
	QF4	QD4			

TABLE 5.10: Items left in the model after the variable reduction

5.4.1 Final Algorithm

Once the variable selection is done we can obtain the linear regression coefficients for all the linear models specified in (4.45). In each linear regression, $\theta = \boldsymbol{\beta X}$ the parameters

are given by:

$$\boldsymbol{X} = [X_0, X_1, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{18}, X_{20}, X_{21}, X_{23}, X_{24}]$$
$$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{14}, \beta_{15}, \beta_{18}, \beta_{20}, \beta_{21}, \beta_{23}, \beta_{24}]$$

where $\boldsymbol{X}$ is a vector representing the response vectors of the items presented in table 5.10 and the linear regression intercept vector is $X_0 = [1, ..., 1]$. β_i is the regression parameter corresponding to item response vector X_i.

The estimates of the regression parameters for each of the models in (4.45) is presented in table 5.11 and the RMSE for each of the seven models is presented in table 5.12 together with the RMSE for the models fitted to the non-reduced data set.

Trait, θ	β_0	β_1	β_3	β_5	β_6	β_7	β_8	β_9	β_{10}
$\theta^{(2)}$	-7.573	0.073	0.066	0.112	0.121	0.174	0.169	0.110	0.113
$\theta_1^{(1)}$	-4.521	0.445	0.430	0.028	0.017	0.014	0.035	0.028	0.023
$\theta_2^{(1)}$	-7.525	0.032	0.036	0.227	0.220	0.315	0.315	0.056	0.077
$\theta_3^{(1)}$	-6.724	0.021	0.034	0.029	0.058	0.070	0.064	0.277	0.234
$\theta_4^{(1)}$	-5.767	0.014	0.006	0.008	0.036	0.08	0.047	0.003	0.029
$\theta_5^{(1)}$	-3.993	0.052	0.029	0.091	0.043	0.015	0.040	0.055	0.012
$\theta_6^{(1)}$	-6.253	0.026	0.047	0.024	0.038	0.087	0.070	0.042	0.059

Trait, θ	β_{11}	β_{12}	β_{14}	β_{15}	β_{18}	β_{20}	β_{21}	β_{23}	β_{24}
$\theta^{(2)}$	0.104	0.115	0.173	0.094	0.060	0.023	0.149	0.099	0.114
$\theta_1^{(1)}$	0.020	0.012	0.072	0.008	0.001	-0.001	0.030	0.019	0.043
$\theta_2^{(1)}$	0.048	0.063	0.096	0.057	0.042	0.002	0.078	0.059	0.073
$\theta_3^{(1)}$	0.273	0.269	0.064	0.045	0.028	0.007	0.049	0.049	0.06
$\theta_4^{(1)}$	0.046	0.036	0.674	0.338	0.016	0.006	0.031	0.023	0.023
$\theta_5^{(1)}$	0.022	-0.005	-0.001	-0.007	0.414	0.263	0.024	0.014	0.050
$\theta_6^{(1)}$	0.034	0.060	0.065	0.031	0.029	0.007	0.444	0.264	0.283

TABLE 5.11: Regression parameters for final algorithm on the reduced data set

When comparing with $RMSE_{full}$ we notice that the RMSE for both the second order trait and domains highly correlated with the second-order trait does not differ much compared to the reduced set. One reason might be that in the reduced set a lot of the variables removed where not highly correlated with either of these traits. Since the new parameters does not contribute with a lot more new information to explain the

Trait	θ	$RMSE_{full}$	$RMSE_{red}$
Innovation	$\theta^{(2)}$	0.122	0.130
Trust	$\theta_1^{(1)}$	0.081	0.186
Resilience	$\theta_2^{(1)}$	0.140	0.144
Diversity	$\theta_3^{(1)}$	0.139	0.144
Belief	$\theta_4^{(1)}$	0.109	0.227
Perfection	$\theta_5^{(1)}$	0.064	0.232
Collaboration	$\theta_6^{(1)}$	0.097	0.159

TABLE 5.12: RMSE of the final models for the reduced and the full data set

variability of these traits the RMSE will not improve significantly. We notice that all the items in the domains *resilience* and *diversity* are kept in the variable reduction and thus it might not be possible to improve the RMSE significantly by adding items related to other domains.

The conclusion is that, even though the variable reduction does not affect the RMSE of the second-order trait, innovation, it creates a big dispersion of the RMSE for the first-order traits. This further supports the recommendation to not reduce the number of items in the BII data set.

Cast results on 1-10 scale

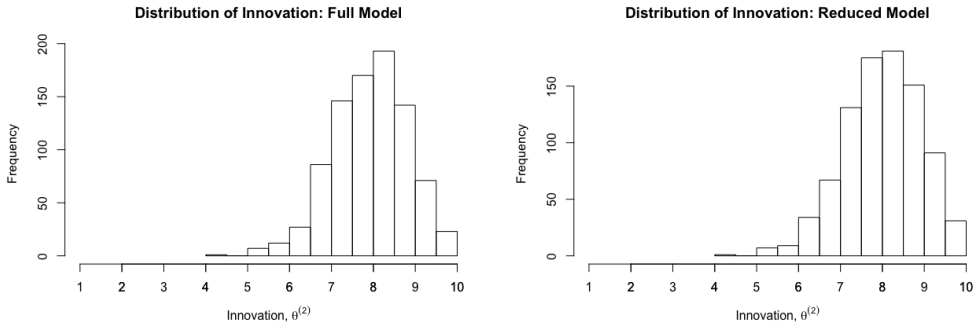
To fulfill the requirement given in Chapter 3 the interval of the index scores produced by the linear models are mapped to the new interval $I = [1, 10]$ as shown in equation (4.49) in 4.

The scaling parameters for each of the seven linear models are presented in table 5.13.

	X_{full}		X_{red}	
Trait	$\min(\theta)$	$\max(\theta)$	$\min(\theta)$	$\max(\theta)$
$\theta^{(2)}$	-5.898	1.916	-5.704	1.772
$\theta_1^{(1)}$	-3.464	1.986	-3.297	1.599
$\theta_2^{(1)}$	-5.890	1.534	-5.729	1.455
$\theta_3^{(1)}$	-5.130	1.474	-5.093	1.431
$\theta_4^{(1)}$	-4.608	1.496	-4.351	1.313
$\theta_5^{(1)}$	-3.235	2.078	-2.882	1.562
$\theta_6^{(1)}$	-4.920	1.841	-4.636	1.797

TABLE 5.13: Scaling parameters for full and reduced model for mapping $I_{old} = [\min(\theta), \max(\theta)] \rightarrow I_{new} = [1, 10]$. N.B. $\min(\theta)$ and $\max(\theta)$ is the values calculated by the linear regression equations if a subject answered either 1 on all items (that results in $\min(\theta)$ i.e. $x_i = 1, \forall i$) or if a subject answered 5 on all items (that results in $\max(\theta)$ i.e. $x_i = 5, \forall i$).

The distribution of the final score for the second-order trait innovation capability calculated by the linear models for the full and the reduced model is presented in Figure 5.10a and 5.10b respectively. All the distributions of the linear estimates of the first-order latent traits, for the full and the reduced model, can be found in Figures B.6 and B.7 in Appendix B.



(a) (b)

FIGURE 5.10: The distribution of the linear approximation of the second-order trait, innovation capability, given by (a) the full data set and (b) the reduced data set

5.5 Outlier Analysis

The result of the outlier analyses is presented in table 5.14. N.B. In HO-GRM$_{low}$ the lowest row mean value in the data set has been discarded, in HO-GRM$_{hi}$ the highest row mean value has been omitted, in accordance with the procedure presented in section 4.6.3.

		Variance, σ^2						
Model	**DIC**	$\theta^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	$\theta_6^{(1)}$
HO-GRM$_{full}$	48497.7	0.214	0.218	0.208	0.203	0.163	0.374	0.214
HO-GRM$_{hi}$	49009.0	0.223	0.222	0.217	0.210	0.167	0.380	0.219
HO-GRM$_{lo}$	48883.1	0.221	0.221	0.217	0.209	0.169	0.373	0.217

TABLE 5.14: DIC and mean variance of the latent trait estimates for the outlier analyses compared with the full model

The same data sets are not used and thus the DIC cannot be used to compare the model fits. However, the DIC at least indicates that the model fit is worse in both cases when the extreme values of the data set has been eliminated. The variance of the latent trait estimates shows that the uncertainty is slightly bigger for the models applied to the reduced data sets and the conclusion is thus that the full model is not

affected by any outliers. One reason for this is that the occurrence of true outliers is very improbable since the values in the data set are drawn from a small discrete sample space, i.e. $x_{si} \in [1, ..., 5]$.

The coefficient of variation is also higher for all ability estimates in the HO-GRM$_{hi}$ and HO-GRM$_{low}$ models as compared to the HO-GRM$_{full}$ model. When performing a posterior predictive check on the outlier models' the bayesian p-values were about the same as for the full model.

5.6 Structure Analysis

In this section the result of the analysis of different data structures will be presented. Each data set was simulated according to the procedure presented in section 4.6.2. As in the case with the prior analysis the method used for estimating model fit is the correlation between the estimated traits given by the MCMC simulation and the 'true' values used to simulate the data sets. Since different data structures, and therefore different sets of data, are compared the DIC cannot be used for comparisons. The result of this analysis is presented in table 5.15.

Given this result we can see some of the effect that the model structure has on the accuracy of the estimates. Note that only one simulation has been done for each of the different models and that the simulated data is not the same in any of the models. Thus, we can not argue that the result is statistical significant, but at least the indications are consistent over all of the different model structures. The most dominant patterns that can be obtained from the result are the following:

- More *subjects* yield a more accurate estimate for all parameters.
- More *items per domain* yield a more accurate estimate for first-order traits.
- More *domains* yield a more accurate result for the second-order trait.

The result that more items per domain yields more accuracy is in line with the definition of the test information presented in Chapter 4.

The 'true' λ_j parameters used for the simulation are constructed such that the they are evenly spaced out over the interval $I \in [0.6, 0.9]$ with $\lambda_1 = 0.6$ and $\lambda_J = 0.9$. Thus, our results confirms the findings of de la Torre and Hong (2010) that domains that are highly correlated with the second-order trait, i.e., have a high $\rho(\theta^{(2)}, \theta_j^{(1)}) \approx \lambda_j$, will be more accurately estimated.

	Model Specification			**Correlation**	
	Subjects, N	Items, n	Domains, J	$\rho^{\theta^{(2)}}$	$\frac{1}{J}\sum_j^J \rho^{\theta_j^{(1)}}$
1	1000	18	3	0.894	0.918
2	1000	18	6	0.918	0.882
3	1000	24	3	0.891	0.925
4	1000	24	6	0.923	0.889
5	1000	24	8	0.931	0.879
6	1000	36	3	0.904	0.949
7	1000	36	6	0.941	0.920
8	1000	36	9	0.958	0.915
9	3000	24	6	0.943	0.927

	Correlation, first-order trait j, $\rho^{\theta_j^{(1)}} = \rho_j$								
	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9
1	0.904	0.916	0.934	-	-	-	-	-	-
2	0.858	0.866	0.877	0.885	0.905	0.898	-	-	-
3	0.911	0.928	0.934	-	-	-	-	-	-
4	0.867	0.867	0.895	0.894	0.897	0.911	-	-	-
5	0.850	0.863	0.867	0.882	0.882	0.895	0.892	0.901	-
6	0.938	0.951	0.959	-	-	-	-	-	-
7	0.907	0.908	0.916	0.926	0.930	0.934	-	-	-
8	0.897	0.908	0.897	0.909	0.918	0.917	0.921	0.928	0.936
9	0.911	0.915	0.921	0.933	0.938	0.942	-	-	-

TABLE 5.15: Results of the structure analysis on varying simulated data sets. The models are compared by analyzing the correlation between estimated and 'true' first- and second-order traits given different model specifications.

Chapter 6

Discussion

Remember: Rewards come in action, not discussion.

—Tony Robbins

The constructed algorithm and the results obtained from it are deemed satisfactory in light of the given problem formulation. The aim of the thesis was to construct a valid measure of individual innovation capability based on the responses in the BII data set. In this section we will present our subjective view on the results as well as further analyze the constructed algorithm and its measures, e.g., the confidence interval of the latent trait scores, the scalability of the model, future improvements that can be made etc.

6.1 Fulfillment of the Requirements

In the requirements, specified in section 3.1, it is stated that the index should be **ordinal**, i.e., a higher score on each latent trait continuum should reflect a higher level of proficiency on that specific latent trait scale. In the algorithm developed for the thesis, the second-order latent trait is an ordinal measure. Therefore, the subjects in the data set can be ranked in regards to their overall innovation capability. The first-order latent trait is not ordinal in regards to other subjects, however it is ordinal when compared to that individual's other domain abilities. I.e., the second-order latent trait is be used for between-person comparisons and the first-order latent trait is used for within-person comparisons. Thus, an individual can rank/order his or her proficiency on the different sub-trait domains.

Due to the choice of model, HO-IRT, which maps categorical manifest variables to continuous latent traits, the MCMC parameter estimates will be **continuous**. The final algorithm however produces a score through a linear relationship with a finite set of

possible explanatory variables vectors. Thus, the final scale/index will be discrete and not continuous, with 5^{24} possible scores for each latent trait. However, since the scale is cast on a **1-10 scale including score 1 and score 10**, the large set of possible results relative to the quite compact scale will yield results that in practice almost can be seen as continuous.

The HO-IRT model is deemed to be **robust against outliers** as shown in the Results. When two leave-one-out analyses were run (where the lowest and highest row mean score were omitted from the data set respectively) all the model fit indicators were inferior to the one for the full model in which no outliers had been discarded. N.B. In our opinion the data set does not contain any true outliers. This is due to the fact that the sample space of the possible responses is both small and discrete.

In order to make the model **robust against bad samples** a number of criteria were listed that specified what a bad sample is. It was postulated that bad entries in the data set were ones that only had the same response to every question, or entries from individuals that already had taken the test. In total 151 bad samples were discarded from the data set (about 15% of the total number of samples) before any model was constructed or analyzed.

The requirements stated that the different sub-trait categories of innovation should have a **varying degree of impact on the overall innovation capability score**. The HO-IRT models allows for this through the varying factor loading term λ_j.

The method provided in this thesis, to construct the BII algorithm, is **scalable** in the sense that it can be directly applied to any data set consisting of polytomous questionnaire responses on a Likert scale (as long as each item is only measuring one first-order latent trait). Therefore, the model and the construction steps are also valid if more data is added or if the questionnaire is changed.

6.2 Strengths and Uncertainties of the Algorithm

What are we measuring?

One of the major issues needed to be addressed is what we are really measuring and if it really is innovation. As in all social science it is important to note that what we measure can only be interpreted in the context of the questions that have produced the data. The data is only a matrix of numbers and does not represent anything by itself. For example, an identical set of data, contrived from a completely different set of questions not related to innovation, could produce the exact same results as the ones stored in the BII data

set. In order to make a distinction between these two data sets one has to put the data in the context of the questions/items that produced the data set.

Therefore, what the BII is measuring is the interpretation of what innovation capability is that is presented in Sidhu et al. (2016b) and Sidhu et al. (2016a). This is also under the assumption that the BII questions accurately reflect the psychological traits they are constructed to quantify/measure.

The model is built upon the assumption that the subjects answer the questions truthfully and unbiased. However, since the questionnaire is constructed as a self-assessment it is not guaranteed that this is always the case. A biased response from a subject will lead to a difference in the interpretation of his/her score while a lot of untruthful responses will impact the whole model.

For example if the "correct" answer to a question is too obvious one may argue that the index score might not reflect the true level of innovation capability, but rather how good a person is at identifying the best response to an item. The exploratory analysis in some sense confirm that our current data set is not too corrupted by bad samples or bad data points, but it would be preferable to limit the possibility to answer questions untruthfully.

There are two methods to deal with this problem:

1. Include an honesty measure in the evaluation procedure. E.g., a question could be asked two or three times, but with different phrasings in order to confirm the consistency of the responses given by a subject.
2. Increase the difficulty of "easy" questions (i.e., questions with a high mean item response, see Figure 5.8a). An easy question could be rephrased or negated so that the subject has to provide a reverse response.

Even though easy questions might give subjects the incentive to answer untruthfully, the easy questions — with a high mean item response — also have the highest correlation with the final innovation score. While harder questions — with a low mean item response — have a very low correlation. As explained in 5.3 this is due to the fact that the distribution of item responses is skewed towards high values and since the majority of the questions are "easy" the algorithm will favor these over the harder ones since they will be more correlated with the overall innovation capability trait.

Because of the great differences in the difficulty level of items it is complicated to draw conclusions about each domain's importance in the estimation process of the overall level of innovation capability. We cannot tell for sure if *trust* and *perfection* really are less

important dimensions of an innovative mindset or if this result is just derived from the current construction of the questionnaire and the answers given to it stored in the BII data set. Therefore, we will refrain from drawing any conclusions about the importance of certain domain abilities. Nonetheless, when a HO-IRT model is applied to the current data set then the resulting model indicates that some domains have a higher correlation with overall innovation capability.

Inaccuracies of the BII Scores

To see how credible our results are we can form a credible interval for the ability estimates (the estimated posterior means) for the latent trait parameters derived from the MCMC analysis , i.e. $\hat{\theta}$. To do this one can utilize the standard deviation of the posterior means. The credible interval gives an idea of how biased the final estimates of the latent trait scores are. All the latent trait estimates are spread out on the approximate interval $I \sim [-2.75, 3]$. Given that the mean standard deviation for all the latent trait estimates' posterior distributions is ≈ 0.5 and given that the posterior distribution of the latent traits are approximately normally distributed, then the 95% credible interval for all the ability estimates is approximately $\hat{\theta} \pm 1.96 * 0.5 = \hat{\theta} \pm 0.98$. This means that the 95% credible interval covers almost a third of the total ability continuum. One can argue if this is good enough.

The standard deviation used in this calculation of the credible interval is the mean of all the latent trait estimates' standard deviation derived from their specific posterior distribution. The standard deviations for the latent trait estimates near the boundaries of the parameter space are higher, reaching almost 0.8 for a certain subject's estimated domain ability. This implies that these estimates, close to the boundary, are even more uncertain. On the contrary, scores near the mean of the ability estimates have a greater certainty.

When fitting the seven linear regression models, specified in equation (4.45), a lot of information is lost. This is due to the fact that the estimates obtained from the MCMC analysis contain measurement errors and when fitting the linear regression models these estimated ability measures are seen as observed in order to conduct supervised model training. Moreover, since the sample space of the independent variables consists of only five integer values it is difficult to fit a linear model mapping the categorical responses to continuous ability estimates.

One notable difference is the size of the distribution space given the parameter estimates of the MCMC and the ones obtained from the linear regressions. While the approximate interval for the estimated latent traits from the MCMC are $I \sim [-2.75, 3]$ it is

approximately $I \sim [-2, 2]$ after the linear regression (prior to rescaling). This highlights that the accuracy of the linear approximation has most information loss near the latent continuum boundaries.

The Feature Selection Method

As mentioned in Chapter 4 the variable reduction methods used in this thesis do not have a solid theoretical foundation. When researching the topic of variable reduction and feature selection we came across several sources that stated that no method has been developed for simultaneous feature selection for multivariate linear regression models that were applicable to our case.

In order to perform the variable reduction we constructed an ad hoc method based on the sum of the importance values (for each independent variable in the seven linear regression models, see equation (4.45)) obtained from the feature selection algorithms. Both the feature selection methods are statistically valid, but the item importance measure is not necessarily cast on the same scale for each regression model. Therefore, the sum of the importance measures is an ad hoc solution.

We should also mention that apart from the feature selection methods presented in the thesis an additional method for variable selection/reduction was considered. The method made use of consecutive LASSO regressions over the models to assess which variables were most relevant. The regression coefficients were manually analyzed as they gradually converged to zero (when the penalty parameter was increased) to determine which variables were most relevant to all models.

The reason this method is mentioned here is the fact that it produced very similar results as the feature selection methods detailed in the thesis, i.e., the variables that were deemed to be important were more or less the same. Even though it might not necessarily prove anything, it is still an indication that our variable reduction results obtained from the ad hoc methods seem reasonable.

Prior Analysis

The same set of simulated data is used for each model in the prior analysis and thus the goodness of fit of the models can be compared by inspecting the models' DIC value. However, the DIC values only measure the model fit for the simulated data set so even though the DIC might indicate that the model fit is better for specific priors these specific prior selections might not be the best ones for the real data set.

The reason why simulated data is used in the prior analysis, instead of real data, is that it enables us to analyze the correlation between the estimated and the "true" model parameters as a goodness of fit measure. This would obviously not be possible with real data since there are no "true" values for the the model parameters. Furthermore, a prior analysis performed on the real data set would only be applicable on that data set, and thus the model would lose generality.

An even more extensive prior analysis could be conducted, e.g., by using different sets of simulated data, analyzing the impact of *extreme* prior distributions, adjusting multiple prior distributions simultaneously etc. The reason why this has not been done is that the time-consuming computations have been a great limiting factor.

Structure Analysis

The results presented in table 5.15 come from one single simulation for each type of data structure. This should be regarded as a shortcoming of the analysis. The reason why multiple simulations for each type of data structure have not been carried out is, once again, due to time constraints and long computational times (the structure analysis presented in the thesis took over 40 hours to complete). Consequently, these results should serve as a mere indication of a possible underlying pattern, rather than a significant proof of any hypothesis.

One should also note that the simulated data sets follow the theoretical structure defined by the HO-GRM model and that this data structure assumption might not hold true for the real BII data set (even though the exploratory analysis does not reject our model structure assumption). Furthermore, the simulated data does not contain any "bad" samples and the factor loadings, λ, all have relatively high values. As a result the correlations between the true and the observed values presented in table 5.15 are most likely higher than the accuracy of the parameter estimates obtained for the real data set. Therefore, this result should not be used directly to draw any major conclusions about the accuracy of the parameter estimates. However, the outcome of the simulation study still provides an indication on how the index can be improved in future iterations of the BII.

6.3 Recommendations to the BII Developers

The variable reduction analysis indicated that the model fit was worse whenever an item was removed from the data set since the RMSE, AIC and BIC of the reduced models

increased, sometimes quite significantly. The two feature selection techniques applied, i.e. Boruta and GBM, also showed that all items present in the current data set were relevant. Therefore, in order for the index to be as accurate as possible we recommend the developers of the BII to either keep all the existing questions, reformulate them or add new questions that could replace the least relevant ones. This notion is also confirmed in the structure analysis, where the accuracy of the sub-trait domain estimates increases when the number of questions related to that domain increases.

There is an evident trade-off between user-friendliness (fewer items make the test more accessible to individuals as the the time it takes to complete the questionnaire is reduced when questions are omitted) and the accuracy of the index scores. However, in our opinion, if the innovation capability assessment is to be of true value the question set has to be somewhat extensive.

As discussed earlier we can not draw any major conclusion about the true importance of each domain in regards to the overall innovation capability level. To enable future iterations of the BII to more accurately determine the effect of each domain on the overall score we recommend that the difficulty levels of the questions are evened out and preferably that the difficulty level of the "easy" questions is increased. As can be seen by examining the ICCs (Figure B.1-B.2) more difficult questions, e.g. *QT2* and *QP1*, might result in less discrimination, i.e. lower α values. On the other hand the probability of different item responses is more spread out over the whole latent trait continuum which is not the case for easier questions (e.g. *QF4*).

We also recommend the BII developers to include items of differing difficulty in each separate domain and it might be valuable to add a "neutral" question where it is beneficiary to answer "Don't Know", instead of "Totally Agree" or "Totally Disagree". This would probably increase the dispersion of the responses and in turn increase the discrepancy among subjects. A good mixture of difficulty levels among the questions would result in lower values of the latent trait scores and therefore it would, we believe, result in a more accurate estimation of the importance of each separate domain.

The four questions with the highest response pattern variation are also the four *reversed* questions, i.e., where the response **1**=*Completely disagree* is the "correct" answer. If the response variation is higher due to the nature of the question or plainly because the question is reversed is not possible to tell, but it indicates that reversing a question might increase the variability of the answers. Reversing a question is also an effortless procedure that is easy to implement for the BII developers in order to potentially increase the variability of the answers provided to certain questions.

The BII algorithm is easily implemented as an individual's answers to the questions are mapped to BII scores with the use of the linear regression models presented in the Results. In order to maintain this algorithm and keep it up-to-date when new entries are added to the data set we recommend that the developers of the BII set certain milestones associated with the number of samples collected. When a certain milestone is reached the full HO-GRM MCMC analysis should be re-run and seven new linear regression models should be fitted to this result. By redoing the full analysis on a larger data set new and more precise estimates are acquired (as indicated by the simulation study). Example of milestone thresholds could be: 1,000, 3,000, 5,000, and 10,000 samples.

6.4 Future Work

Even though our recommendation for the BII is to keep all the current questions in the questionnaire, it would still be of great use in the future development of the index if a theoretically valid variable reduction method for the multivariate regression model was developed. The reason is that it would facilitate to identify if a new item is good or redundant early, when a few answers to the new question have been provided. Tools to assess if an item is good or bad already exist in the IRT framework, e.g., the ICC and the IIC. These tools are good to analyze separate items, but not to assess the validity of a question in light of the complete model when all data is taken into account. Hence, a theoretically valid variable reduction method suitable for the BII case would be of great use in the future development of the index.

One way to deal with biased item responses would be, as mentioned earlier, to implement some kind of "honesty" measure in the BII. This is quite common in personality assessments and the basic idea is to measure the truthfulness and the consistency of the subjects' item responses. This could be done by reversing questions and asking them again. An honesty measure like this would be able to identify biased and bad samples, and modify the model so that this is taken into account. Another idea is to implement a Graded Response Model based on the 3PL-IRT model, which would include a *guessing parameter*, that adds a lower asymptote to the item characteristic function so that even the person with the lowest ability estimate has a chance to answer an item correctly (due to guessing).

The formulation of the HO-GRM model assumes multi-unidimensionality. This means that an item can only belong to one domain. The result of the exploratory analysis (Figure 4.6) indicates that for the six domains only one item, i.e. *QF3*, may belong to more than one domain. Nevertheless, a possible future extension of the index would be to build a model that can handle items that belong to more than one domain. The model

presented in the thesis takes into account the correlation between different domains, but not the direct impact of an item on several domains.

The model in the thesis can also be slightly edited in order to handle missing item responses. It would be interesting to analyze the affect of this alteration on the current data set, and it also allows future iterations of the BII to make use of both the data sets with new items as well as the current data set. There already exists at least one reduced BII data set where ten items have been removed and two new items have been added. A revised model that handles NA-values could be used to perform the BII analysis on these two merged data sets.

To further confirm the hierarchical model structure one could analyze if the six first-order latent traits are related to one or several second-order latent traits that are related to a third-order latent trait (i.e. overall innovation capability). This analysis could be performed on the BII data set and then be compared with the current model.

Latent class analysis (see table 4.1) allows one to map categorical manifest variables to categorical latent classes. An extension of the BII could therefore include an algorithm that categorizes the test subjects as different types of innovation personas based on the entries in the BII data set. This type of clustering model could work as a complementary analysis.

One interesting analysis to conduct would be to analyze potential item bias in the different strata of the question set. Item bias emerges if the answers to certain items have significantly different response patterns for a specific subgroup of subjects. In the IRT framework this implies that the item characteristic curves of the different subgroups do not coincide. Bias can for e.g. be related to sex, academic background, regional differences. These three demographic statistics are currently stored in the BII data set. By performing an item bias analysis the BII developers can for example identify if certain questions are easier to answer for a specific demographic subgroup. If that is the case, then the items that show bias should be reviewed and potentially rephrased or discarded from the questionnaire. N.B. For the item bias analysis to be truly relevant each subgroup needs to have a large sample of subject responses. Hence, before making any big statements about item bias we recommend the BII developers to collect more data so that eventual biased response patterns, for the different strata, are statistically relevant.

The BII research group also hypothesized that the ability to work and be effective outside ones Comfort Zone would influence an individual's level of innovation capability. The structure analysis also implied that the inclusion of Comfort Zone as a first-order trait could have a positive impact on the accuracy of the second-order latent trait estimate.

Therefore, to include Comfort Zone as a variable might be an interesting future research topic for the BII developers.

Overall, the complexity of the model and the accuracy of the index could be increased in a plethora of different ways. One could for example include non-linear relationships between different items, apply non-linear relationship between items and traits etc. All these areas are open for exploration and could potentially increase the precision of the estimated index.

Chapter 7

Conclusion

Standing on the paving by the office building.
They've got so much to do, never time for you.
—Henrik Berggren

The aim of the thesis was to construct a scientifically valid measure of individual innovation capability in regards to the subject responses given to the Berkeley Innovation Index questionnaire.

The solution proposed is based on a Higher-Order Item Response Theory approach that utilizes a Graded Response Model. In order to provide each subject in the data set with an estimate of their overall innovation capability as well as their proficiency in each sub-trait domain a Markov Chain Monte Carlo method was employed. The bayesian inference approach made it possible to simultaneously estimate the vast amount of model parameters.

In order to confirm the proposed HO-GRM model it was compared to a HO-GPCM model (that produced similar, but less accurate results). In order to further validate the proposed model for the BII algorithm several test-statistics (e.g., the Deviance Information Criterion, the Posterior Predicitve Check, the Gelman-Rubin convergence criteria etc.) were employed.

To reduce the computational complexity and streamline the scoring procedure a multivariate linear regression model was fitted to the parameters estimated by the HO-GRM model. The regression parameters of these regression models constitutes the final BII algorithm.

This thesis only takes the first step in the creation process of the Berkeley Innovation Index algorithm and we strongly believe that once the shortcomings identified in this

work have been addressed the resulting index will be greatly improved. The presented algorithm for the BII is very flexible which opens up the possibility for future iterations and adjustments of the questionnaire, the data structure and the analysis procedure.

The work on the Berkeley Innovation Index has come a long way and we are very confident that the index will become a globally distinguished and commonly used tool for estimating individual innovation capability.

Bibliography

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Assocation*, 88(442):669–679, 1993.

F.B. Baker. *Item response theory: Parameter estimation techniques.* Marcel Dekker, Inc., 1992.

R.D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrica*, 46(4):443–459, 1981.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

S. Chib and E Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

M.K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics :a comparative review. *Journal of the American Statistical Assocation*, 91(434):883–904, 1996.

J. de la Torre and Y. Hong. Parameter estimation with small sample size a higher-order IRT model approach. *Applied psychological Measurement*, 34(4):267–285, 2010.

J. de la Torre and H. Song. Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 23(8):620–639, 2009.

G.A. Ferguson. Item selection by the constant process. *Psychometrika*, 7(1):19–29, 1942.

J. Flegal and G. Jones. Implementing MCMC: Estimating with confidence. In A. Jones G. Meng X-L. Brooks, S. Gelman, editor, *Handbook of Markov Chain Monte Carlo*, chapter 7, pages 175–197. Chapman Hall/CRC, 2011.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

A. Gelman, Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics*, pages 247–268, 2000.

A. Gelman, S. Gelman, K. Shirley, et al. Inference from simulations and monitoring convergence. *Handbook of Markov chain Monte Carlo*, pages 163–174, 2011.

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

R.K. Hambleton and R.W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47, 1993.

U. Hana. Competitive advantage achievement through innovation and knowledge. *Journal of Competitiveness*, 5:82–96, March 2013.

W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

H-Y. Huang. A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement*, 39(5):362–372, 2015.

H-Y. Huang, W-C. Wang, P-H. Chen, and Su C-M. Higher-order item response models for hierarchical latent trait models. *Applied Psychological Measurement*, 37(8):619–637, 2013.

S.T. Hunter, L. Cushenbery, and T. Friedrich. Hiring an innovative workforce: A necessary yet uniquely challenging endeavor. *Human Resource Management Review*, 22(4): 303–322, 2012.

D.N. Jackson and S.V. Paunonen. The jackson personality inventory and the five-factor model of personality. *Journal of Research in Personality*, 30(1):42–59, 1996.

A. Jaffe. Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *The American Economic Review*, 76:984–1001, December 1986.

B.W. Junker, R.J. Patz, and N.M. VanHoudnos. Markov chain Monte Carlo for item response models. *Handbook of Item Response Theory, Volume Two: Statistical Tools*, 21:271–325, 2016.

Michael J. Kirton. *Adaption-innovation: In the context of diversity and change*. Routledge, 2004.

T-C. Kuo and Y. Sheng. Bayesian estimation of a multi-unidimensional graded response IRT model. *Behaviormetrika*, 42(2):79–94, 2015.

M.B. Kursa and W.R. Rudnicki. Feature selection with the boruta package, 2010.

F.M. Lord. The relation of test score to the trait underlying the test. *Educational and Psychological Measurements*, 13(0):517–548, 1953.

F.M. Lord, M.R. Novick, and A. Birnbaum. Statistical theories of mental test scores. 1968.

G. Martín-de Castro, M. Delgado-Verde, J.E. Navas-López, and J. Cruz-González. The moderating role of innovation culture in the relationship between knowledge assets and product innovation. *Technological Forecasting and Social Change*, 80:351–363, 2013.

G.N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47:149–174, 1982.

R.P. McDonald. *Factor analysis and related methods.* Lawrence Erlbaum, 1985.

R.P. McDonald. *Test theory: A unified treatment.* Psychology Press, 2013.

J. Menold, K. Jablokow, S. Purzer, D.M. Ferguson, and M.W. Ohland. A critical review of measures of innovativeness. In *121st ASEE Annual Conference & Exposition*, 2014.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1091, 1953.

E. Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176, 1992.

R.J. Patz and B.W. Junker. A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2):146–178, 1999.

M. Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9 (3):523–539, 2008.

G. Rasch. Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*, 1960.

G. Ridgeway. Generalized boosted models: A guide to the gbm package, 2007. URL http://www.saedsayad.com/docs/gbm2.pdf. [Online; accessed 4-September-2016].

F. Samejima. Estimation of latent ability using a rsponse pattern of graded scores. *Psycometric Monograph*, 17, 1969.

F. Samejima. Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39(1):111–121, 1974.

J. Schmid and J.M. Leiman. The development of hierarchical factor solutions. *Psychmoetrika*, 22(1):53–60, 1957.

I. Sidhu, K. Singer, C. Johnsson, and M. Suoranta. Introducing the Berkeley Method of Entrepreneurship - a game-based teaching approach. 2015.

I. Sidhu, J-E. Goubet, H. Weber, A. Fred Ojala, C. Johnsson, and J.C. Pries. Berkeley Innovation Index: An approach for measuring and diagnosing individuals' and organizations' innovation capabilities, 2016a.

I. Sidhu, J-E. Goubet, and Y. Xia. Measurement of innovation mindset. In *IEEE ICE TEMS Norway*, 2016b.

C. Spearman. "General intelligence", objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

H.J. Sung and T. Kang. Choosing a polytomous IRT model using bayesian model selection methods. In *Annual Meeting of the National Council on Measurement in Education, San Francisco, CA*, 2006.

J. Templin. Item response theory workshop — lecture 1, 2014. URL http://jonathantemplin.com/files/irt/irt14icpsr/irt14icpsr_lecture01.pdf. [Online; accessed 23-August-2016].

R.K. Tsutakawa and M.J. Soltys. Approximation for bayesian ability estimation. *Journal of Educational and Behavioral Statistics*, 13(2):117–130, 1988.

Y.F. Yung, D. Thissen, and L.D. McLeod. On the relationship between the higher-order model and the hierarchical factor model. *Psychometrika*, 64(2):113–128, 1999.

Appendix A

Appendix A. The Questionnaire

TRUST

QT1	Most people can be trusted
QT2	Most people tell a lie when they can benefit by doing so
QT3	I trust other people
QT4	Those devoted to unselfish causes are often exploited by others

RESILIENCE

QF1	I can accept failures as part of a learning process
QF2	Failures often lead to positive outcomes in the long run
QF3	I overcome setbacks to conquer important challenges
QF4	Failures allow opportunities for reflection and consideration

DIVERSITY

QD1	It is important to me to interact with people, that are different from me
QD2	I frequently come in contact with people that are different from me
QD3	I feel comfortable to talk to people that are different from me
QD4	Interacting with other persons makes me interested in things that happen outside of my field

BELIEF

QB1	I can succeed at any endeavor to which I set myself
QB2	I am able to successfully overcome many challenges
QB3	When facing difficult tasks, I am certain I will accomplish them
QB4	I have been able to achieve most of the goals I set for myself

PERFECTION

QP1	I consider myself a perfectionist
QP2	I would prefer to hand in a product on time rather than making it perfect
QP3	In general, quality and perfection are more important than effectiveness
QP4	I would rather create something that is cost effective than the highest possible quality

COLLABORATION

QC1	There are times when it makes sense to collaborate with my competitors
QC2	An active cooperation with my collaborators is important to me
QC3	A cooperation with one of my enemies would be very important to my firm
QC4	There are times when I would be open to share resources and information with my competitor

RESPONSE ALTERNATIVES TO EACH QUESTION

1	Completely Disagree
2	Disagree
3	Don't Know
4	Agree
5	Completely Agree

Appendix B

Appendix B. Additional Results

In this appendix complete results for the HO-GRM model applied to the BII data set, that was not included in chapter 5, are presented:

- HO-GRM item parameter estimations (α_i, β_{ik}): Table B.1
- Item Characteristic Curves for all items: Figure B.1 (Trust and Resilience), Figure B.2 (Diversity and Belief), Figure B.3 (Perfection and Collaboration)
- Item and Domain Information Functions: Figure B.5
- The distributions of the first order latent trait estimates obtained from the
  - MCMC simulation: Figure B.4
  - Linear regression models applied to the full data set: Figure B.6
  - Linear regression models applied to the reduced data set: Figure B.7

B.1 HO-GRM Item Parameter Estimations

Item, i	$\hat{\alpha}_i(\sigma_{\alpha_i})$	$\hat{\beta}_{i1}(\sigma_{\beta_{i1}})$	$\hat{\beta}_{i2}(\sigma_{\beta_{i2}})$	$\hat{\beta}_{i3}(\sigma_{\beta_{i3}})$	$\hat{\beta}_{i4}(\sigma_{\beta_{i4}})$
QT1	2.60 (0.23)	-2.01 (0.12)	.0.89 (0.06)	-0.16 (0.05)	1.49 (0.08)
QT2	1.31 (0.09)	-2.42 (0.20)	-0.14 (0.07)	0.83 (0.09)	3.29 (0.25)
QT3	2.35 (0.20)	-2.80 (0.20)	-1.48 (0.09)	-0.51 (0.06)	1.19 (0.08)
QT4	0.93 (0.08)	-3.30 (0.28)	-0.84 (0.11)	0.26 (0.08)	2.67 (0.22)
QF1	1.88 (0.15)	-3.22 (0.27)	-2.36 (0.16)	-1.62 (0.11)	-0.15 (0.06)
QF2	1.62 (0.13)	-3.34 (0.29)	-2.29 (0.17)	-1.33 (0.10)	0.20 (0.6)
QF3	2.01 (0.16)	-3.20 (0.27)	-2.50 (0.18)	-1.64 (0.11)	0.13 (0.05)
QF4	2.34 (0.22)	-2.96 (0.25)	-2.52 (0.19)	-1.83 (0.12)	-0.40 (0.05)
QD1	2.10 (0.18)	-3.54 (0.34)	-2.21 (0.15)	-1.47 (0.10)	-0.13 (0.05)
QD2	1.69 (0.13)	-4.13 (0.42)	-1.70 (0.12)	-0.77 (0.07)	0.60 (0.07)
QD3	2.14 (0.17)	-2.87 (0.23)	-1.60 (0.10)	-0.87 (0.07)	0.41 (0.05)
QD4	1.91 (0.16)	-3.38 (0.30)	-2.67 (0.20)	-1.83 (0.12)	-0.21 (0.06)
QB1	2.05 (0.15)	-3.02 (0.23)	-1.82 (0.11)	-0.97 (0.07)	0.67 (0.06)
QB2	3.97 (0.47)	-3.05 (0.28)	-1.81 (0.10)	-1.06 (0.06)	0.32 (0.05)
QB3	2.07 (0.15)	-3.34 (0.27)	-1.76 (0.11)	-0.73 (0.06)	0.92 (0.06)
QB4	1.69 (0.12)	-3.40 (0.27)	-1.96 (0.13)	-1.07 (0.08)	0.78 (0.07)
QP1	0.86 (0.09)	-1.81 (0.19)	0.31 (0.09)	1.59 (0.16)	3.59 (0.34)
QP2	1.58 (0.18)	-2.85 (0.25)	-1.05 (0.10)	-0.18 (0.06)	1.35 (0.12)
QP3	0.95 (0.10)	-2.88 (0.28)	-1.30 (0.14)	-0.03 (0.08)	2.40 (0.23)
QP4	1.12 (0.11)	-2.82 (0.25)	-0.82 (0.10)	0.42 (0.08)	2.28 (0.20)
QC1	2.28 (0.20)	-3.22 (0.29)	-2.40 (0.16)	-1.46 (0.09)	0.50 (0.06)
QC2	1.38 (0.11)	-4.34 (0.41)	-3.11 (0.24)	-1.85 (0.14)	0.02 (0.07)
QC3	1.60 (0.12)	-2.56 (0.19)	-1.21 (0.10)	0.37 (0.06)	1.82 (0.12)
QC4	1.91 (0.15)	-2.48 (0.18)	-1.30 (0.09)	-0.25 (0.06)	1.48 (0.09)

TABLE B.1: Parameter estimates and SD_{post} of the HO-GRM model parameters α_i and β_{ik}

B.2 Item Characteristic Curves for all Items

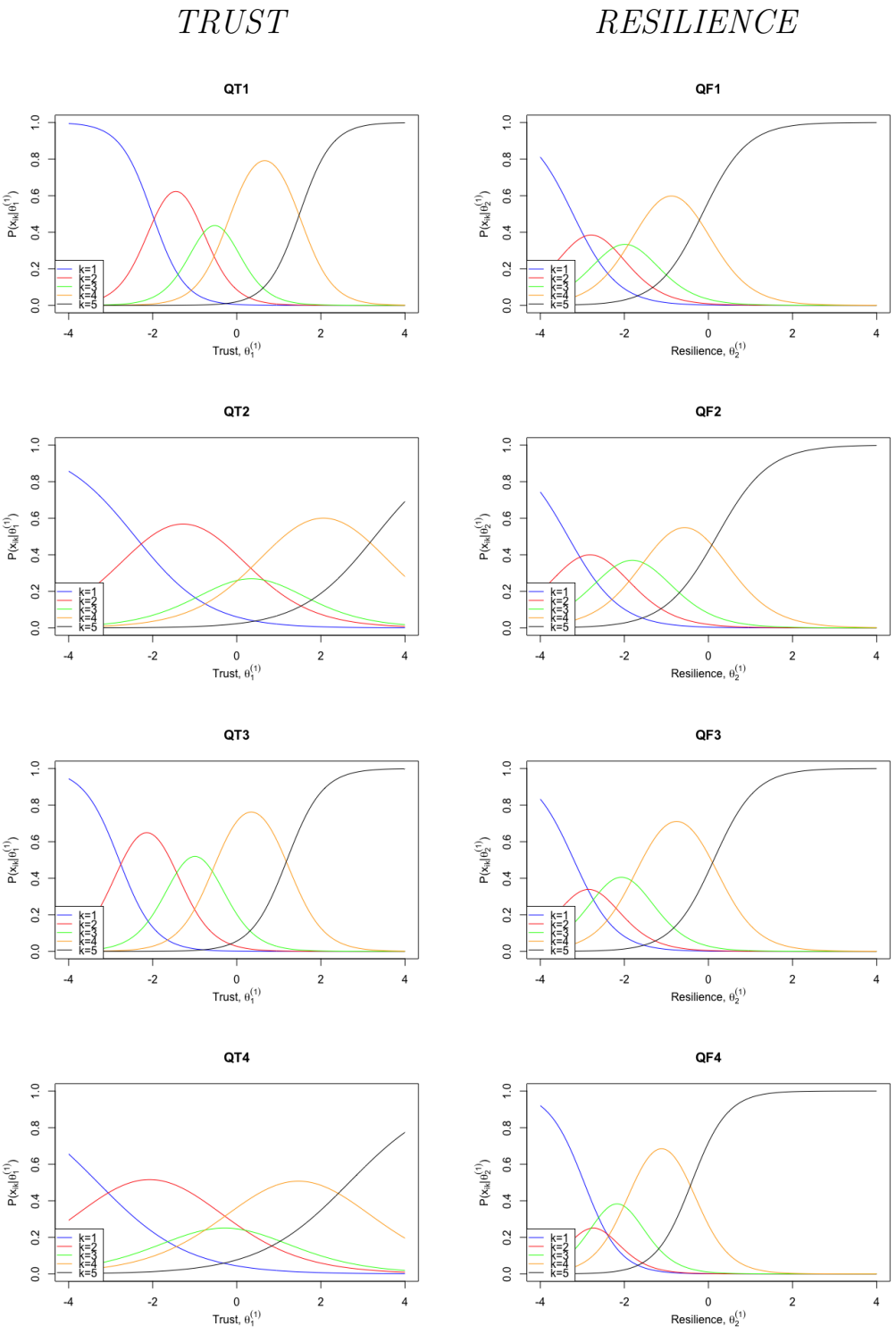


FIGURE B.1: Item Characteristic Curves for items in the domains Trust (left) and Resilience (right)

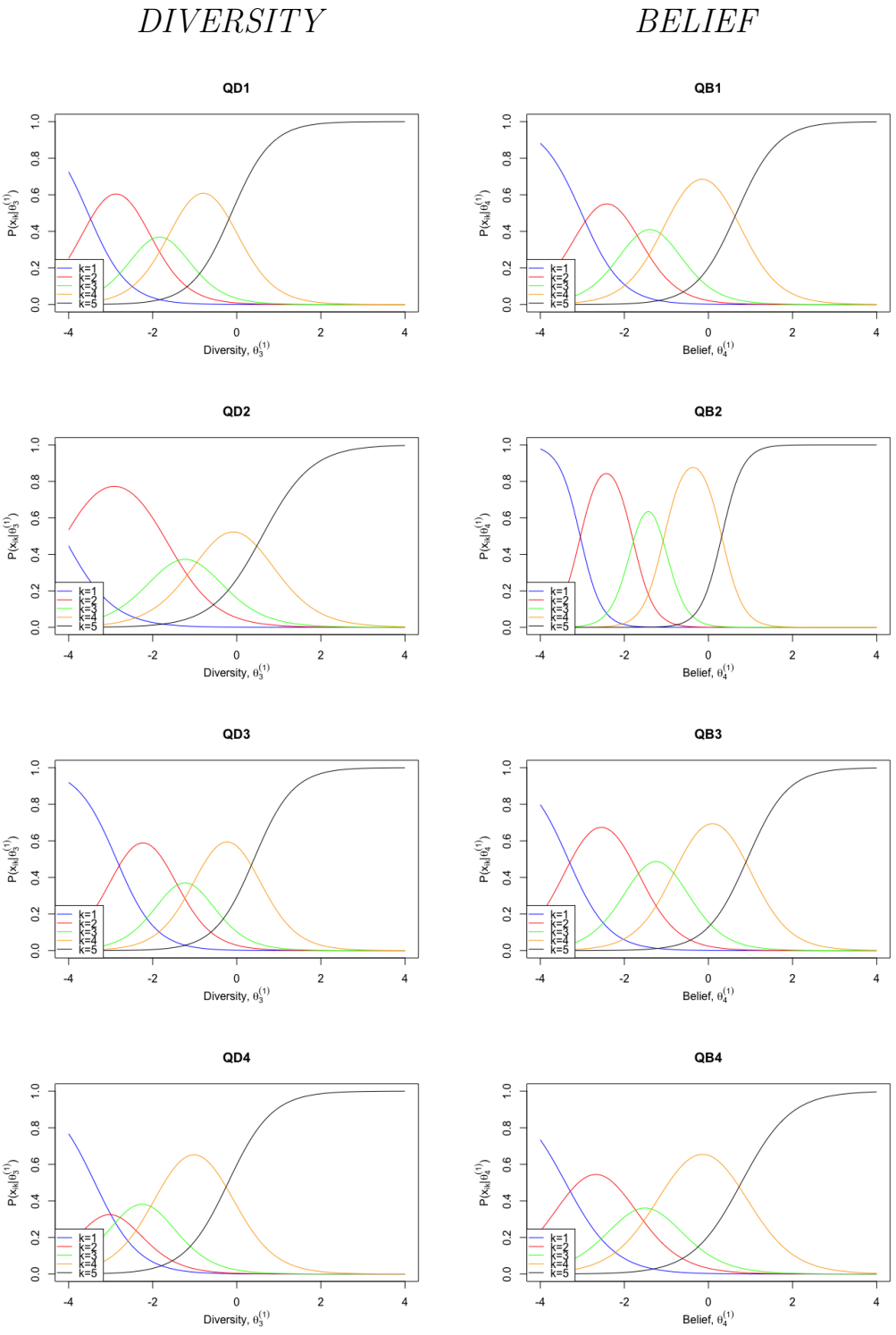


FIGURE B.2: Item Characteristic Curves for items in domains Diversity (left) and Belief (right)

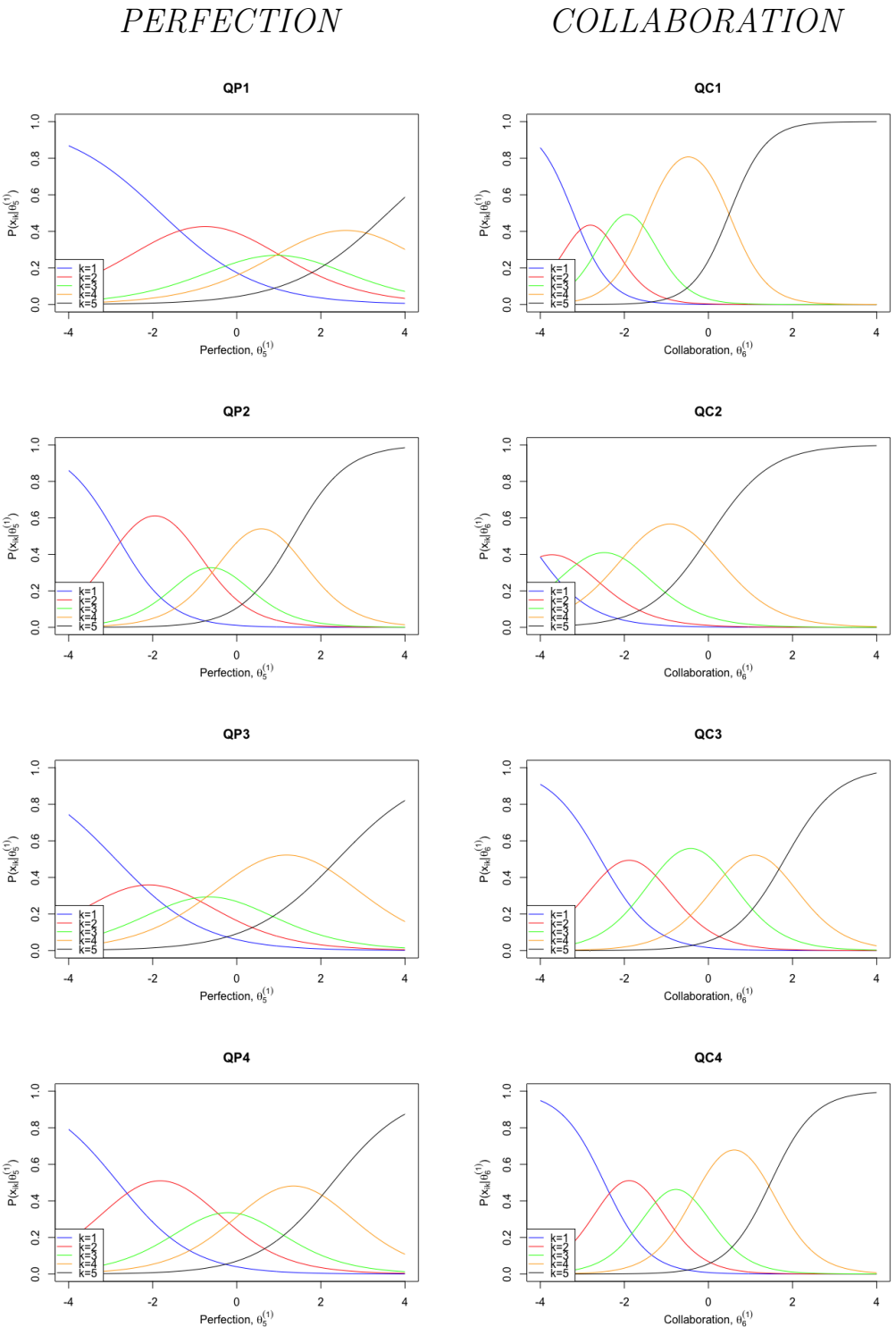


FIGURE B.3: Item Characteristic Curves for items in the domains Perfection (left) and Collaboration (right)

B.3 Item and Domain Information Functions

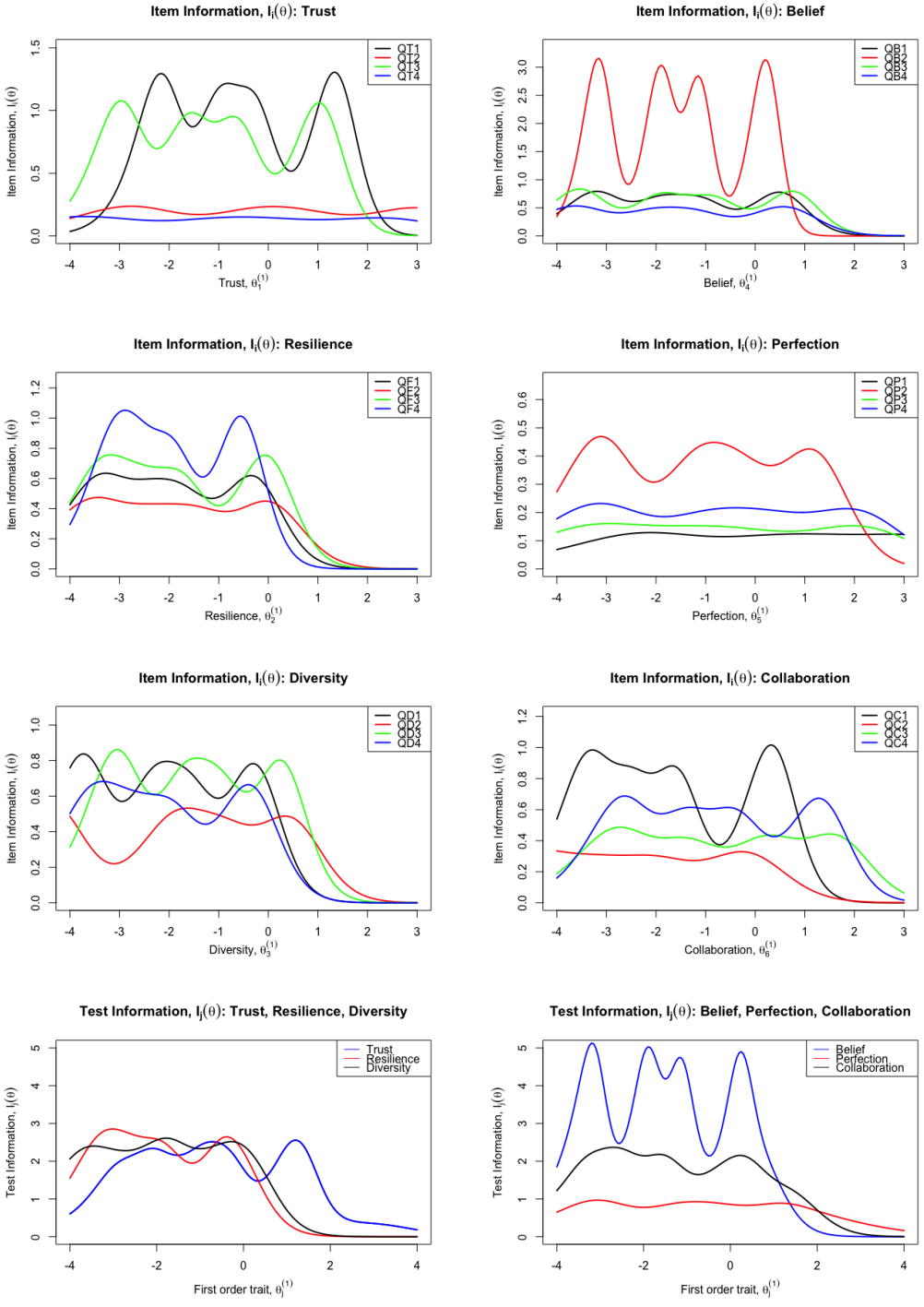


FIGURE B.4: Item and Test information functions for the HO-GRM model

B.4 Distributions of First Order Latent Traits

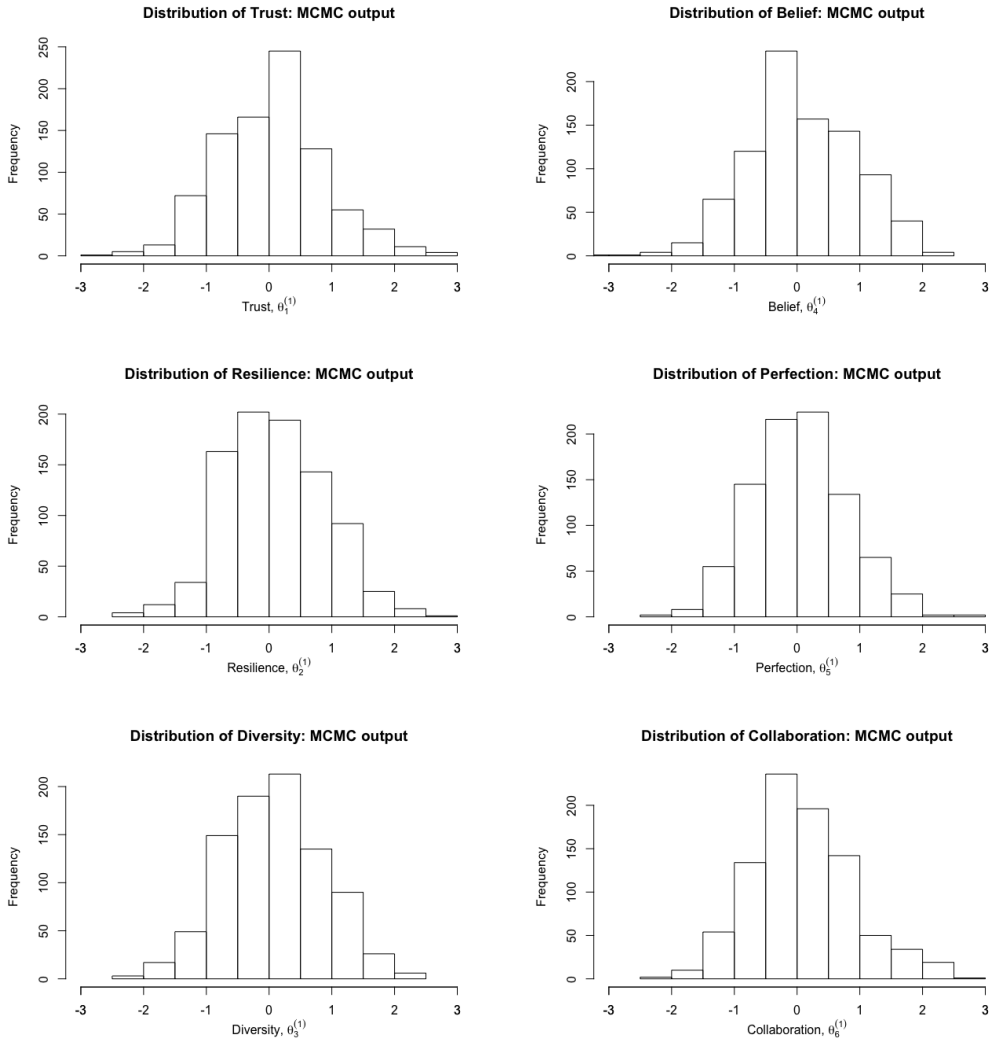


FIGURE B.5: Distribution of the first order latent traits given by the MCMC parameter estimates

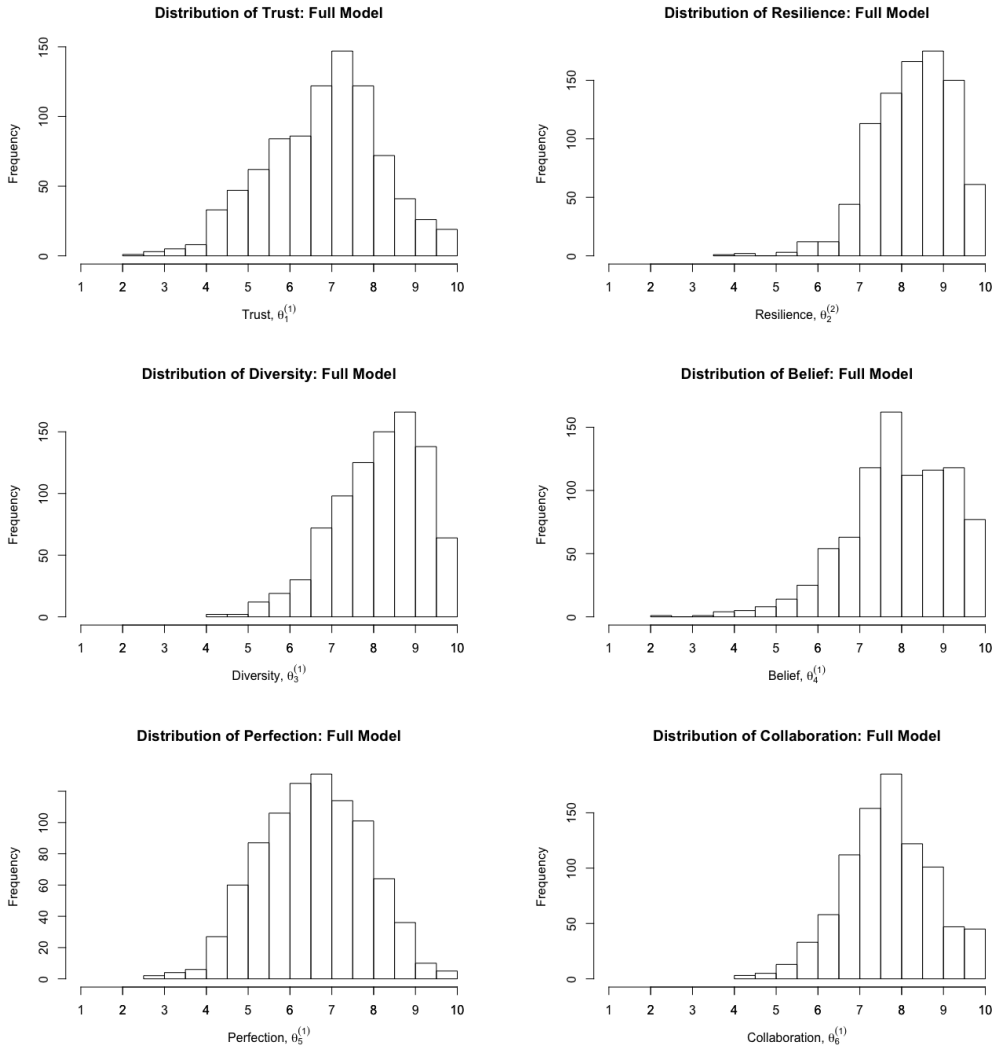


FIGURE B.6: Distribution of first order latent trait estimates obtained from the linear regression models given the full data set

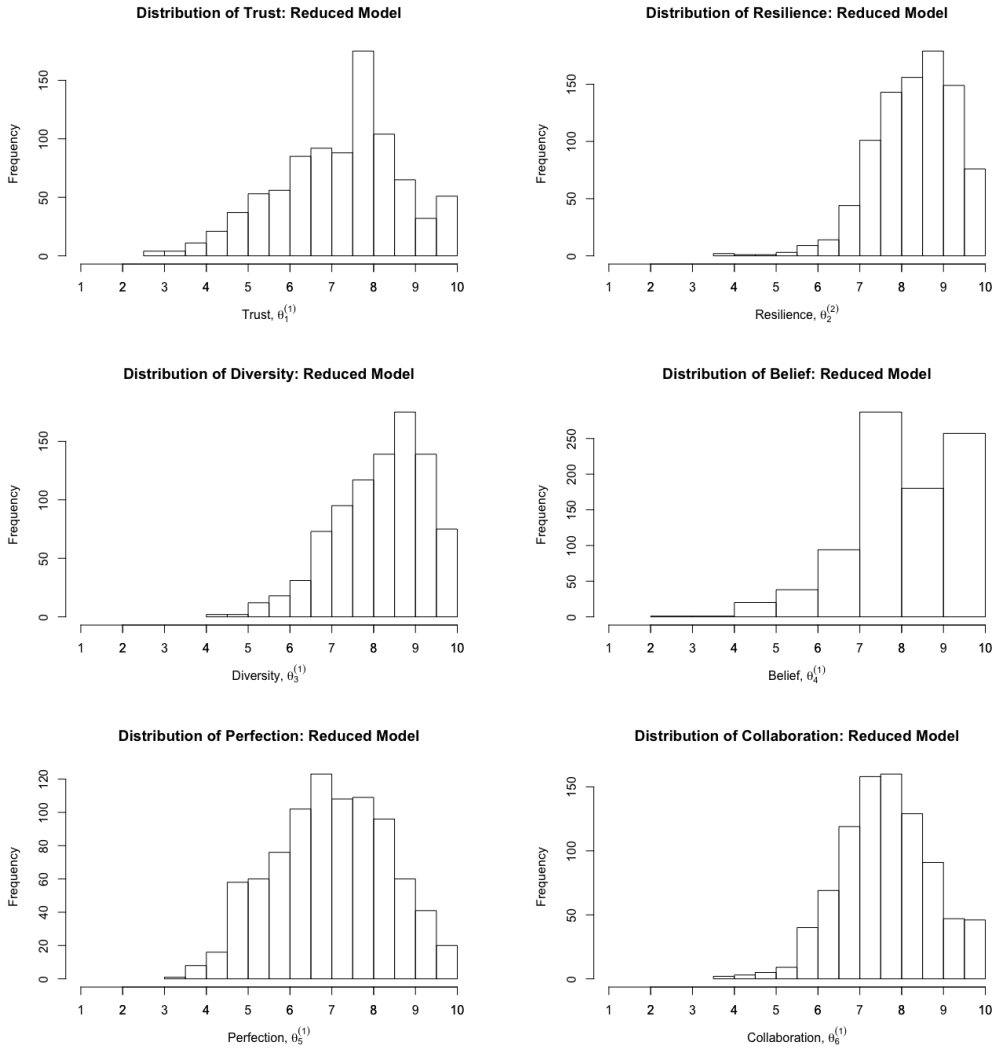


FIGURE B.7: Distribution of first order latent trait estimates obtained from the linear regression models given the reduced data set

Master's Theses in Mathematical Sciences 2016:E48
ISSN 1404-6342
LUTFMS-3309-2016
Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden
http://www.maths.lth.se/