LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

Person parameter estimation in the polytomous Rasch model

Thesis for attaining the grade Master of Science

Author:
Thomas Welchowski (B.Sc., Dipl.-Betriebswirt FH)

Supervisors:
Prof. Dr. Tutz and Dr. Draxler

Institution:
Ludwig-Maximilians University Munich
Faculty of Mathematics, Informatics and Statistics
Institute of Statistics

Date:

August 8, 2014

Contents

List of Figures

List of Tables

List of Abbreviations

CMLE = Conditional Maximum Likelihood Estimation
CI = Confidence Interval
CPC = Category Probability Curve
CTT = Classical Test Theory
CV = Cross Validation
EIRC = Environmental, Institutional and Regional Covariates
EPC = Educational Process Covariates
IRT = Item Response Theory
MLE = Maximum Likelihood Estimation
MCS = Minimum Chi-Square
NIRT = Nonparametric Item Response Theory
PCM = Partial Credit Model
JMLE = Joint Maximum Likelihood Estimation

MCL = Miss Classification Loss

MMLE = Marginal Maximum Likelihood Estimation

PE = Pairwise Estimation

r & c cv = Row & Column Cross Validation

RPE = Restricted Pairwise Estimation

RMSE = Root Mean Squared Error

SM = Sequential Model

WMLE = Weighted Maximum Likelihood Estimation

Acknowledgements

Special thanks to Prof. Tutz for his support, the exam data set and his inspiring advices. I also thank Dr. Draxler for the good collaboration during our work in the DFG-project *Sampling designs for the Rasch Model* and interesting discussions in which basic ideas for this thesis were derived. Additionally, thanks to Prof. Weber and Dr. Tost, for providing the Ascot pre-test data set. Last but not least my most gratitude goes to my parents, who always supported me, in all efforts.

Abstract

An important task in psychometric measurement is to identify the skills of persons, by accounting the difficulty degree of item responses in questionaries. The **Item Response Theory** (IRT) is the base for a valid measurement of the skills and the **Rasch Model** a well established probabilistic model, in the implementation of the theory. More recent extensions of the Rasch Model include the Partial Credit Model and the Sequential Model. For both models, different estimation methods are proposed. In this thesis the estimation accuracy of these methods, in the case of polytomous items of categorical variables, is examined by intensive computational simulations (about 180 000 cases).

The outcome of the **Partial Credit Model** simulations, with focus on person parameter estimations, is: The marginal maximum likelihood and joint maximum likelihood estimations have the lowest mean squared error in the majority of cases. The former approach is recommended, because it uses the theoretical well established framework of the generalized, linear, mixed models. In this framework standard errors, statistical tests and model diagnosis are easy to conduct, compared with the joint maximum likelihood method. The pairwise estimation methods show an average performance. The restricted version of the pairwise method is in most cases less effective, than the unrestricted version, in terms of mean squared error. In the case of polytomous items, the theoretic variance reduction by fixing coefficients of equal person scores is offset by the information loss, using only persons with different total scores. The conditional estimation methods performed low, but weighted maximum likelihood estimates can improve the performance in most of the cases. Choosing a higher number of items, results in lower variance of the estimates, but the ranking does not change in general.

The outcome of the **Sequential Model** simulations, with regard to the person parameter estimations, is: In all cases the marginal maximum likelihood has the lowest mean squared error, followed by the conditional maximum likelihood and last but not least by the joint maximum likelihood approach.

For the **assessment of the predictive accuracy** of IRT models a specific loss function and resampling method is proposed. The idea of the proposed resampling method (based on cross validation), is to predict the performance of new individuals in the test data, given the estimated item parameters of the training data. It was evaluated, if this procedure, combined with a marginal maximum likelihood estimation, can discriminate between the sequential (true model) and partial credit model in terms of predictive accuracy. The proposed loss function could not identify the true model. Instead of this, the 0-1 loss function is better suited for the predictive accuracy assessment. By changing the loss function, the sequential and partial credit models were better separated by the proposed resampling procedure. The reason is, that the proposed loss function (above) is only zero, in the case of a perfect deterministic model. This requirement is too strong for the intended evaluation. The goal is to find the best of bred stochastic models, which approximates the data generating process well.

The results of the simulation studies are well supported by an empirical **Data Analysis**, conducting IRT in practise. Two data sets were analysed: The **first** set consists of the examination results of statistics students at the University of Munich (LMU) for a lecture in multivariate statistics. Different estimation methods

for the sequential model were compared. In terms of predictive performance, the marginal maximum likelihood approach - on average - lead to slightly better results, than the conditional or joint maximum likelihood estimation.
The **second** set consists of closed-ended questions/questionaire for a group of vocational student in Business Administration (Munich), measuring their project performance (self-assessment). For this data, the sequential model on average show slightly better results, in terms of the predictive performance (0-1 loss function), than the partial credit model. Both models were estimated by using the marginal maximum likelihood approach.

The empirical data set results, overall lead to similar assessment patterns, with regard the predictive accuracy of the analysed estimation methods and are in-line with the simulations. For practical analysis the sequential model with marginal estimation is recommended as it is more flexible, easier to interpret and yielded better performance in data analysis than the partial credit model.

1. Introduction: Measuring the performance of minds (IRT, CTT)

On the cover page you see a picture of the famous statue called *The Thinker* created by Auguste Rodin (1840-1917) (Davies et al. 2011; see pages 924-927), one of the most influential sculptor's and symbolist's of the late nineteenth century. With his works he wanted express the depths of mind. *The Thinker* is in a typical position, laying his head on his hand, droping his arm over the knees and sitting in a bent position. This reflects a thought process and is common, when humans try to solve difficult questions. It is a symbol for the states of mind and is important in research areas like psychometrics, conducted to measure intelligence. This study focusses on statistical models in psychometrics and how to efficiently measure the performance of minds.

In practise there are a lot of different areas where psychological measurement is used as devices in decision processes (Borsboom 2005; see page 1): Teachers test pupils for dyslexia or hyperactivity, countries test pupils to exceed the level of university, educational systems are evaluated by comparing performances of populations, etc.. When measuring intelligence, there are a lot of different research subdomains, e. g. spatial, verbal, numerical, emotional, perceptual intelligence. But how should these concepts be measured? What has to be considered for constructing efficient tests? What constitutes a good test?

Bond (Bond and Fox 2001; e.g. pages 1-6) argues, that first and foremost the constructed scientific measurements should be abstractions of equal units, additive and reproducible. Good psychological test items are (Rost 1996; e. g. pages 31-38)

- objective (external circumstances, such as changing test supervisors should not influence the scientific conclusions)
- reliable (how reliable are the test measures in repeated measurements)
- and valid (the test measures what it is supposed to be).

These are basic requirements for measurement in physical sciences, as well. If psychological measurement wants to draw general inferences on a similiar level as physical sciences, then these requirements should also apply for psychometrics. For example raw data has to be abstracted to a meaningful scale, for drawing inferences about constructs rather than describing observations. On a scale with clear interpretation and a meaningful zero point any comparisons between measures can be made more reliable. Only a good calibrated measuring instrument can yield reliable estimates. In physical sciences these requirements were applied for long time. To construct instruments in psychometrics for measuring minds is still in development.

There are two major theories for psychomotric measuring: Classical Test Theory (CTT) and Item Response Theory (IRT). Both theories are briefly described and compared here. Essentially CTT is based on the following theoretical framework (Hambleton and Jones 1993; see page 40):

- Linear model of the form X = T + E with observable test Score X, true score T and error score E

- True scores and error scores are uncorrelated
- The average error score in the population of examinees is zero
- Error scores on parallel tests are uncorrelated
- Parallel forms are defined as tests, that measure the same content, for which examinees have the same true score and where the size of errors of measurement across forms are equal

It is assumed that parallel forms of a test can be constructed. Within these framework other equations can be derived. For example in empirical research based on CCT a central concept is the reliability of the test, because validity is not observable and hard to detect (Borsboom 2005; see pages 26-30). For constructing reliability measures three major proposals have been made. The most promising method is to find a worst case scenario, which gives a estimation of the lower bound of reliability. One example is the computation of Cronbach's Alpha (Cortina 1993; see page 100): It is a function of the extent to which items in a test have high communalities and thus low uniqueness. It is also a function of interrelatedness and it does not imply unidimensionality or homogenity. The main problem with CTT is the definition of the "true score", because it is prone to misinterpretation and it's definition is dubious (Borsboom 2005; see pages 44-47). For example, the expectation of the observed score for one person does not allow a frequentist view of probability, because identically, independent replications of the test questions are reasonable not applicable for one person. This implies, that the true score is invariant in time. But this is certainly not the case, because a person can learn, become tired or forget something, he previously did knew. CTT does not mention how to treat varying time intervals between tests. Borsboom concluded that CTT is not enough for an adequate treatment of psychological test scores. For further details of the philosophical debate see (Borsboom 2005).

As an alternative to CTT is IRT. Typically IRT models assume the following (Johnson 2007; see page 2):

1. Unidimensionality
2. Conditional independence
3. Monotonicity

The first assumption states that all the skills of a person for a given set of items can be expressed as a latent scalar value. Given the latent ability, the item response vectors $X_i = (X_{i1}, \ldots, X_{iJ})$ are independent. This means that no other factors besides the latent ability influences the responses to the test items. The relationship between the probability of a correct response and the latent abilities is a non-decreasing function over the interval $[0, 1]$. In practise an item characteristic curve is assumed as a structural assumption, which specifies how scores and item difficulties are transformed into the probability scale. For each person and item combination a conditional probability distribution is defined. This relationship usually is non-linear specified. Borsboom (Borsboom 2005; see page 84) argues that IRT has the following benefits in comparison with CTT:

- Placement of attribute measurement: Differences in attribute (either within or between subjects) lead to differences in the observations
- Introduction of some metaphysical concepts, e. g. latent variables, serve a clear purpose and lead to interesting research questions; CTT only uses metaphysics to get simple mathematical equations
- The latent variable view corresponds more closely to the way many researchers think about measurement

To summarize, (Hambleton and Jones 1993; see page 44) concludes the following advantages of both CTT and IRT:

IRT benefits

- Item statistics that are independent of the groups from which they were estimated
- Scores describing examinee proficiency that are not dependent on the test difficulty
- Test models that provide a basis for matching test items to ability levels
- Test models that don't require strict parallel tests for assessing reliability

CTT benefits

- Smaller sample sizes required for analysis
- Simpler mathematical analyses compared to IRT
- Model parameter estimation is conceptionally straightforward
- Analyses don't require strict goodness-of-fit studies to ensure a good fit of models to test data

For psychometric accurate measuring of the performance of groups within a given set of items, IRT seems to have fewer theoretical flaws and can be used better for practical measurement analysis. This thesis will focus on analysing IRT models. In the next chapter two of the most common IRT models for polytomous items and alternative ways for estimation are explained. After that the measures of performance for the estimation methods will be described. The performance results for all estimation methods of the computationally intensive simulation study are compared to each other. In chapter 6 the results of the simulation will be applied by analysing real world data. Finally, the results will be discussed and an outlook of future research areas is given.

2. Methods

In this chapter the theoretical methods used for the simulation and data analysis are explained and further references are given. For most of the estimation methods the functions were self-written by the author in R-Code and supplemented by C-Code (e. g. Likelihood function). Some packages are used as building blocks for the estimation methods, if available. Data restrictions were necessary to ensure identifiability of the estimates. In each optimization step, only the required restrictions for this step were applied, as to minimise information loss. For numerical optimization the classical, iterative Newton Raphson algorithm was used in many cases because it has a quadratic convergence rate in the univariate case (Monahan 2011; see page 192). The Likelihood functions for IRT models are usually well behaved and there is an efficient implementation in C-Code available in the base package. For further details of the algorithm see (Monahan 2011; pages 199-203).

2.1. Rasch model and extensions

In the dichotomous case, a well known IRT model mentioned in literature is the Rasch model. In this section a short review of the Rasch model and its polytomous extension is given.

In the Rasch model the probability of person i to answer an item j correctly is (Fischer and Molenaar 1995; e. g. page 10):

$$P(X_{ij} = x_{ij}|\theta_i, \beta_j) = \frac{\exp\left[x_{ij}\left(\theta_i - \beta_j\right)\right]}{1 + \exp\left(\theta_i - \beta_j\right)}; X_{ij} \in \{0, 1\} \tag{1}$$

$\theta_i \in \mathbb{R}$ are the individual person parameters, which reflect the skill level of each person i. Higher values indicate better performance. $\beta_j \in \mathbb{R}$ are the item difficulty parameters. Higher values correspond to more difficult items. The advantage of the Rasch model is, that it is well established. For example for both item parameters and person parameters sufficient statistics exists, represented by the column and row sums of the data matrix. One disadvantage is that the RM is only applicable for test items with binary outcomes.

There are several approaches for polytomous modelling in the Item Response Theory (Thissen 1986; e. g. pages 567-568). In this paper three main approaches are considered:

- DIFFERENCE MODELS appropriate for ordered responses
- DIVIDE BY TOTAL models for ordered or nominal responses
- LEFT-SIDE ADDED models for multiple-choice responses with guessing

This thesis focusses on the DIVIDE BY TOTAL models in which the concept of probability relates to the classical definition by Laplace: *as the ratio of the number of outcomes favourable to the event to the total number of possible outcomes, each assumed to be equally likely* (Barnett 1999; see page 74).

An extension of the Rasch model for categorial responses is the partial credit model

(PCM) (Masters 1982; see page 158). Alternatives of it are the rating scale model (Andrich 1978), but with the disadvantage to allow only an equal number of categories per item, therefore the PCM is prefered. The equation for the PCM is:

$$P(X_{ij} = x_{ij} \mid \theta_i,\ \underline{\beta_j}) = \frac{\exp\left[\sum_{k=0}^{x_{ij}} (\theta_i - \beta_{jk})\right]}{\sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l} (\theta_i - \beta_{jk})\right]};\ X_{ij} = \{0, \ldots, m_j\} \qquad (2)$$

It represents the probability to reach a specific score X_{ij}, given a latent ability θ_i for person i and a difficulty level β_{jk} of the item j for category k. For notational convience $\sum_{k=0}^{0}(\theta_i - \beta_{jk})$ is set to zero. PCM shares with RM the availability of sufficient statistics for person and item parameters and specific objectivity (Masters 1982; e. g. pages 159-161). Specific objectivity is given, if all comparisons of the subject abilities are in a certain sense independent of the items, used to determine the abilities (Irtel 1995). It is not an absolute concept, because it is restricted to the specific test situation. Therefore the results are only interpretable in the context of specific persons and specific item configurations. It is worthwhile to mention that the concept of specific objectivity is not limited to the dichotomous Rasch model, but could be adapted to the two parameter logistic model, as well.

2.2. Methods for estimation of Partial Credit Model (PCM)

There are six major estimation methods for the Partial Credit Model (Hambleton et al. 1991; see page 46). Each of them will be described in the following subchapters. The main difficulties in estimating Item Response Theory models are the person parameters, because with any new obersations new parameters have to be estimated. Therefore the equation system contains more parameters (count of persons + sum of step difficulties over all items), than the sample size.

The person parameters are in general not identifiable. A restriction for the person parameter estimates is necessary: In this thesis the restriction $\sum_i^N \theta_i = 0$ is applied. This is equivalent to centering the person parameters. It is less arbitrary, than choosing a single person as a norm. The interpretation is easier: Positive values indicate high performance, values around zero mediocre and values below zero low performance. For all response patterns from persons with raw score zero or perfect raw score the estimates tend to negative or positive infinity (Fischer and Molenaar 1995; see page 56) because there is no information in this samples. For each item j in each category, there has to be at least one observation. Therefore, all extreme cases are excluded in the estimation procedure. However after the estimation of the well conditioned person parameters, a cubic spline interpolation is used to approximate the missing person parameters. It is assumed, that persons with higher ability are expected to have higher total scores. For each person, the total score over all items on the x-axis is plotted against the person parameters on the y-axis. For the impact of the interpolation a small simulation is conducted in section 3.2.

Further it is assumed, that no covariates have an impact on the item difficulties. This is known as differential item functioning (DIF) e. g. as mentioned in (Tutz and Schauberger 2012). In each estimation method only the required restrictions are applied by step, so that a minimum of information is lost.

2.2.1. Conditional Maximum Likelihood estimation (CMLE)

The CMLE approach is based on sufficient statistic for the person parameters. A derivation as in (Masters 1982; see page 159) is applied here: It is assumed, that the performance in one item is independent of the performance of another items. In the PCM the probability of a score vector $\underline{x_i}$ over all items $j = 1, \ldots, J$ and given parameters is:

$$P(\underline{x_i}|\theta_i, \underline{\beta_j}) = \prod_{j=1}^{J} \left[\frac{\exp \sum_{k=0}^{x_{ij}} (\theta_i - \beta_{jk})}{\sum_{l=0}^{m_j} \exp \sum_{k=0}^{l} (\theta_i - \beta_{jk})} \right] \tag{3}$$

If $r_i = \sum_{j=1}^{J} x_{ij}$ is defined as the sum of scores across all items for person i and $\sum_{x_{ij}}^{r}$ as the sum over all possible response vectors to produce score r, then the probability to reach r is given by

$$P(r_i|\theta_i, \underline{\beta_j}) = \frac{\sum_{x_{ij}}^{r} \exp \sum_{j=1}^{J} \sum_{k=0}^{x_{ij}} (\theta_i - \beta_{jk})}{\prod_{j=1}^{J} \left[\sum_{l=0}^{m_j} \exp \sum_{k=0}^{l} (\theta_i - \beta_{jk}) \right]} \tag{4}$$

With this the following probability for the response vector can be derived given the total scores of persons:

$$P(\underline{x_i}|r_i, \underline{\beta_j}) = \frac{P(\underline{x_i}|\theta_i, \underline{\beta_j})}{P(r_i|\theta_i, \underline{\beta_j})} = \frac{\exp\left(- \sum_{j=1}^{J} \sum_{k=0}^{x_{ij}} \beta_{jk}\right)}{\sum_{x_{ij}}^{r} \exp\left(- \sum_{j=1}^{J} \sum_{k=0}^{x_{ij}} \beta_{jk}\right)} \tag{5}$$

The likelihood of the data matrix is obtained by multiplying the probability in equation 5 over all persons, because the observations are assumed to be independent. Then the log-likelihood conditional on the total of scores is maximised, regarding the item parameters, with numerical optimization e. g. Newton Raphson. It can be shown, that the item parameter estimates are consistent for growing sample sizes (Sijtsma and Junker 2006; e. g. page 86).

Having estimated the item parameters, the person parameters are estimated in the second step. This approach uses the estimated likelihood (Held and Bove 2014; see page 130), which means to use the joint likelihood conditional on the estimated item parameters. The joint likelihood of the PCM is given by (Masters 1982; see page 164) assuming independency between items and independency between person:

$$L(\underline{\theta}, \underline{\beta}|\underline{\underline{X}}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(X_{ij} = x_{ij} \mid \theta_i,\ \underline{\beta_j}) \tag{6}$$

After that the data is restricted in the following two steps: In the first step only the items are restricted and columns with all values being equal are removed. If some categories per column are missing, the scores are transformed such, that the lowest score is always zero. After the estimation of the item parameters, all rows which maximum or minimum scores are excluded before the iterative estimation of the person parameters.

2.2.2. Weighted Maximum Likelihood estimation (WMLE)

This estimation procedure is similiar to the CML approach in section 2.2.1. The first step is identical. In the second step for the estimation of the person parameters, the likelihood is weighted with the square root of the diagonal elements of the observed Fisher information matrix (Draxler 2014; e. g. pages 4-5). These elements are derived in (Masters 1982; see page 165):

$$\underline{\underline{\mathrm{F}}}(\underline{\theta}|\underline{\hat{\beta}})_{ii} = -\frac{\partial \log\left(L(\theta_i|\underline{\hat{\beta}})\right)}{\partial\theta_i\partial\theta_i} \tag{7}$$

$$= \sum_{j=1}^{J}\left[\sum_{k=1}^{m_j} k^2 P(X_{ij} = k \mid \theta_i,\ \underline{\beta_j}) - \left(\sum_{k=1}^{m_j} kP(X_{ij} = k \mid \theta_i,\ \underline{\beta_j})\right)^2\right] \tag{8}$$

The rationale behind this approach is an asymptotic bias correction, as described in (Warm 1985; e. g. pages 5-9). In theory this approach should be biased only in the order $o(n^{-2})$ which is one order less than the ML approach, used in section 2.2.1. In the dichotomous case, the estimate has the same asymptotic variance and normal distribution as the CMLE method.

The applied data restrictions are the same as in the CMLE approach, as explained in section 2.2.1.

2.2.3. Marginal Maximum Likelihood estimation (MMLE)

The MMLE method assumes a distribution for the person parameters (Johnson 2007; see pages 6-8). The most common assumption is $\theta \overset{\text{i.i.d.}}{\sim} N(0,1)$. The person parameters are integrated out so that the resulting marginal likelihood only depends on the item parameters.

$$P(\underline{x_i}|\underline{\beta}) = \int_{\Theta} P(\underline{x_i}|\theta,\underline{\beta})dF(\theta) \tag{9}$$

$$\Rightarrow L(\underline{\beta}|\underline{\underline{\mathrm{X}}}) = \prod_{i=1}^{N} P(\underline{x_i}|\underline{\beta}) \tag{10}$$

This approach leads in general to consistent item parameter estimates, as the number of persons grows (Sijtsma and Junker 2006; see page 85). After the estimation of the item parameters, the person parameters are estimated as usual by joint likelihood maximisation, conditional on estimated item parameters. But this approach has a practical disadvantage: It is computational time-consuming, because for every evaluation of the likelihood, N numerical integrals have to be solved, where N is the sample size. An alternative approach is to use an EM algorithm for generalized partial credit models. For further details of the algorithm see (Muraki 1992).

2.2.4. Pairwise estimation (PE)

The pairwise estimation method for PCM is a promising approach. The derivation of the required conditional probability $P(X_{nj} = a, X_{oj} = b|\theta_n, \theta_o, \underline{\beta_j}, \underline{f_j})$ is demonstrated in this section (Andrich 2010; e. g. pages 297-298). $\underline{f_j'} = \left(f_{0j}, f_{1j}, \ldots, f_{m_j j}\right)$ are the frequencies of the scores attained by solving item j. These are sufficient statistics for the item parameters (Andrich 2010; see page 296). The joint probability for person n to achieve score a and another person o to achieve score b is evaluated, first:

$$P(X_{nj} = a, X_{oj} = b | \theta_n, \theta_o, \underline{\beta_j}) = \frac{\exp\left[\sum_{k=0}^{a}(\theta_n - \beta_{jk})\right] \exp\left[\sum_{k=0}^{b}(\theta_o - \beta_{jk})\right]}{\sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l}(\theta_n - \beta_{jk})\right] \sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l}(\theta_o - \beta_{jk})\right]} \tag{11}$$

The individual's performance of the tests are assumed to be independent. Then the probability for $\underline{f_j'} = (0, 0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$, given the parameters of two persons and item parameters, has to be calculated. In the case of two persons there are only two possibilities for occurence of scores a and b: Either $X_{nj} = a, X_{oj} = b$ or $X_{nj} = b, X_{oj} = a$. Therefore this probability results in:

$$P(\underline{f_j} | \theta_n, \theta_o, \underline{\beta_j}) = P((X_{nj} = a, X_{oj} = b) \cup (X_{nj} = b, X_{oj} = a)) = \tag{12}$$

$$P(X_{nj} = a, X_{oj} = b | \theta_n, \theta_o, \underline{\beta_j}) + P(X_{nj} = b, X_{oj} = a | \theta_n, \theta_o, \underline{\beta_j}) = \tag{13}$$

$$\frac{\exp\left[\sum_{k=0}^{a}(\theta_n - \beta_{jk})\right] \exp\left[\sum_{k=0}^{b}(\theta_o - \beta_{jk})\right] + \exp\left[\sum_{k=0}^{b}(\theta_n - \beta_{jk})\right] \exp\left[\sum_{k=0}^{a}(\theta_o - \beta_{jk})\right]}{\sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l}(\theta_n - \beta_{jk})\right] \sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l}(\theta_o - \beta_{jk})\right]} \tag{14}$$

With these results one can obtain the probability conditional on the given frequencies, as derived in following equations (Andrich 2010; see page 298):

$$\gamma_{ij} = \sum_{l=0}^{m_j} \exp\left[\sum_{k=0}^{l}(\theta_i - \beta_{jk})\right] \tag{15}$$

$$\psi_{ab} = \exp\left[\sum_{k=0}^{a}(\theta_n - \beta_{jk})\right] \exp\left[\sum_{k=0}^{b}(\theta_o - \beta_{jk})\right] \tag{16}$$

$$P(X_{nj} = a, X_{oj} = b | \theta_n, \theta_o, \underline{\beta_j}, \underline{f_j}) = \frac{P(X_{nj} = a, X_{oj} = b, \underline{f_j} | \theta_n, \theta_o, \underline{\beta_j})}{P(\underline{f_j} | \theta_n, \theta_o, \underline{\beta_j})} = \tag{17}$$

$$\frac{\psi_{ab} / \left(\gamma_{nj}\gamma_{oj}\right)}{\left(\psi_{ab} + \psi_{ba}\right) / \left(\gamma_{nj}\gamma_{oj}\right)} = \ldots = \frac{\exp\left((a - b)(\theta_n - \theta_o)\right)}{1 + \exp\left((a - b)(\theta_n - \theta_o)\right)} \tag{18}$$

The pseudo likelihood of the entire data matrix in Pairwise Parameter Estimation is the product of the previous derived probability over all items and over all pairs of persons in reference to person n (Andrich 2010; see page 301):

$$L(\underline{\theta} | \underline{\underline{X}}) = \prod_{o, o \neq n} \prod_{j=1}^{J} P(X_{nj} = x_{nj}, X_{oj} = x_{oj} | \theta_n, \theta_o, \underline{f_j}) \tag{19}$$

This is called pseudo likelihood, because the pairwise observations are not independent (Andrich 2010; see page 302). The special case for $X_{nj} = X_{oj}$ is excluded from the definition of $\underline{f'_j}$ because the resulting probability would be 0.5, independent of the choosen person paramaters. This demonstrates, that if the responses of two persons are identical, they provide no information regarding the quality comparison between the two persons (Andrich 2010; see page 298). In the dichotomous Rasch model, the pairwise estimation of person parameters are consistent, that means - if the numbers of items grows - the person parameter estimates converge to their true value (Andrich 2010; see page 302). But in practise the number of test items is limited, so effectively, there remains some bias. Prior estimation persons with a maximum or a minimum score are removed. The items remain unrestricted.

2.2.5. Restricted Pairwise estimation (RPE)

The restricted version of the pairwise approach (section 2.2.4) is based on the article (Kreiner 2012; e. g. page 4). Because of different person patterns it happens, that unrestricted pairwise estimation may give persons with equal total scores, different person parameter estimates. This induces more variability in the estimates. One approach to reduce this variability, is to restrict the estimates to be equal for a given total score over all items. One way to achieve this, is to leave out persons with equal total scores. Therefore in the likelihood only persons with different total scores are compared to each other. Beside this difference, the approach is equal to the unrestricted pairwise estimation, described in section 2.2.4.

2.2.6. Joint Maximum Likelihood estimation (JMLE)

Another approach is to maximise the joint likelihood in equation (6) for all parameters together. The data is restricted, both in items and persons, because the parameters are jointly estimated. Because there are more parameters than the sample size, it is not possible to estimate all parameters in one step. Therefore an iterative algorithm is used:

1. Choose some starting values $\underline{\beta_0}$ for the item parameters e. g. all equal to zero.
2. For m=0, . . . , M do

    Maximise joint likelihood conditional on $\underline{\theta_m} \rightarrow \underline{\beta_m}$

    Maximise joint likelihood conditional on $\underline{\beta_m} \rightarrow \underline{\theta_{m+1}}$

    Stop if convergence criteria are met for a predefined ϵ and M_{max}
3. Convergence criteria:

$$\frac{\|l(\theta_m) - l(\theta_{m-1})\|}{\|l(\theta_{m-1})\|} < \epsilon;\ \text{if } \|l(\theta_{m-1})\| = 0 \text{ then } \|l(\theta_m) - l(\theta_{m-1})\| < \epsilon \tag{20}$$

$$\frac{\|\theta_m - \theta_{m-1}\|}{\|\theta_{m-1}\|} < \epsilon;\ \text{if } \|\theta_{m-1}\| = 0 \text{ then } \|\theta_m - \theta_{m-1}\| < \epsilon \tag{21}$$

$$M > M_{max} \tag{22}$$

For convergence criteria three conditions should be checked:
First if the log likelihood values or the person parameter estimates don't change im comparison to the previous iteration. This is important because there could be some local optima, where the algorithm jumps in loops. Another point is, that numbers with lots of

figures need not to be that accurate, as smaller numbers. Therefore precision is measured relatively to the values. Second the argument values should be checked, if they were different from the last iteration.
Third, there should be a maximum count for iterations. The choice of the norm is usually not important. A standard choice for the norm is $\|\underline{x}\|_2^2$, which is equal to the quadratic euclidean norm or $\|\underline{x}\|_1 = \sum |x|$. In this thesis, the focus lies on the estimation of person parameters. Therefore the convergence check is performed after the estimation of the person parameters, given the item parameters from the last iteration. This estimation method is quite simple, but has a drawback: In general the item parameter estimates are not consistent (Sijtsma and Junker 2006; e. g. page 86).

2.3. Sequential model (SM) for polytomous Item Response Theory

Some care should be taken with the interpretation of the step difficulty parameters. (Verhelst and Verstralen 2008; e. g. page 233) argued, that the step intepretation as a sequential model is not meaningful, because the step parameters in the PCM are not only dependent on the previous steps. Tutz (Tutz 1997; sees page 1-2) investigated that issue further and argues, that the item parameters in the PCM are local (Tutz 1990; see page 42):

$$P(X_{ij} = r + 1 | X_{ij} \in \{r, r + 1\}) = \tag{23}$$

$$\frac{\exp(\theta_i - \beta_{jr+1})}{1 + \exp(\theta_i - \beta_{jr+1})} = F(\theta_i - \beta_{jr+1}) \tag{24}$$

The PCM conditioned on the response, being in the actual or previous category can be represented by a dichotomous RM. In this case $\lambda_{j,r+1}$ are the item parameters for a dichotomous item j in score category r+1. Note that not all previous steps have to be completed before reaching a higher score. Therefore the PCM is better suited for rating scales, but less appropriate for sequential ordered responses. That means that every person has to complete all steps in an hierarchical order, to reach maximum score. Here is a small example for a mathematic test item. Consider the following equation:

$$\sqrt{2 * 6 + 4} = y \tag{25}$$

To give the correct answer for this item, the required steps are arithmetic multiplication, addition and at last taking the square root of the result. If one step is missing, it will be most likely, to get an incorrect answer. The underlying data is similiar, as in the case of the PCM with n persons, J items and scores ranging from $0, \ldots, k_j$ with k being the maximum score achieveable for item j. The responses should be at least on an ordinal scale. For this data Tutz (Tutz 1990; see pages 42-43) considered the general sequential model:

$$P(X_{ij} = r | \theta_i, \beta_{j1}, \ldots, \beta_{jr}) = \begin{cases} \prod_{s=0}^{r-1} (1 - F(\beta_{js} - \theta_i))\, F(\beta_{jr} - \theta_i) & \text{if} \quad r = 0, \ldots, k_j - 1 \\ \prod_{s=0}^{k_j - 1} (1 - F(\beta_{js} - \theta_i)) & \text{if} \quad r = k_j \end{cases}$$

$$F : \mathbb{R} \rightarrow [0,1] \text{ is right sided continuous and monoton increasing} \tag{26}$$
$$\lim_{x \rightarrow -\infty} F(x) = 0 \tag{27}$$
$$\lim_{x \rightarrow \infty} F(x) = 1 \tag{28}$$

Basically, this definition gives the probability that person i achieves score r on item j, given the ability and difficulty steps for that item. Because the difficulty steps have to be taken sequentially, the person i has to succeed in the steps $s = 1, \ldots, r-1$ with the probability $\prod_{s=1}^{r-1}(1 - F(\beta_{js} - \theta_i))$ and fail in the r-th step with probability $F(\beta_{jr} - \theta_i)$. $F(x)$ is a general distribution function defined by (Klenke 2006; see page 27). The SM is more general then the PCM:

$$P(X_{ij} > r | X_{ij} \geq r, \theta_i, \beta_{jr}) = F(\theta_i - \beta_{jr}) \tag{29}$$

The PCM needs to be conditioned on the r and r+1 categories to be reformulated as dichotomous Model. In contrast the SM is conditioned on the response being at least r, which is less restrictive then in the reformulation of the PCM. For practical evaluation, it is necessary to choose a specific distribution function. In this thesis, the well established symmetric logistic distribution function $F(x) = \frac{\exp(x)}{1+\exp(x)}$ will be used. The coefficients of the logistic distribution function can be better interpreted and standard software packages are available (e. g. CMLE).

2.3.1. CMLE for sequential model

In comparison with the PCM, there are no sufficient statistics available for direct estimation of the person or item parameters in the CMLE approach. Tutz proposed the following conditional estimation procedure for the sequential model (Tutz 1990; see pages 49-50): First, the response must be dichotomized and the transformed response is given by

$$Y_{ij}^{(r)} | M_r = \begin{cases} 1 & \text{if} \quad X_{ij} > r \\ 0 & \text{if} \quad X_{ij} = r \end{cases} \; ; \; M_r \in \{\underline{\underline{X}} : X_{ij} \geq r\}$$

$Y_{ij}^{(r)}$ is the response for person i and item j for achieving a score greater than r, given a set of persons $M_r = \{i_1, \ldots, i_m\}$. All persons which are considered in the set M_r with m < n (number of persons) must have completed a score of at least r. So by constructing pseudo items for the polytomous response, the model can be fitted by a dichotomous Rasch model:

$$P(Y_{ij}^{(r)} = 1 | \theta_i, \beta_{jr}) = \frac{\exp(\theta_i - \beta_{jr})}{1 + \exp(\theta_i - \beta_{jr})} \tag{30}$$

Essentially this means, that with some data transformation the CMLE (for the established dichotomous Rasch model) can be used for the estimation of the sequential item parameters. The estimated item parameters are inserted in the joint likelihood for estimating the person parameters. In the RM the CMLE approach (Fischer and Molenaar 1995; pages 44-46) works by conditioning on the elementary symmetric functions γ_{r_i}:

$$\gamma_{r_i} = \sum_{\underline{x}|t_i} \exp(-\underline{x}^T \underline{\beta}),\ \underline{x} \in \{0,1\}^J \tag{31}$$

These functions compute all combinatoric cases to observe the total score t_i of person i. Here the vector $\underline{x}$ denotes a possible response pattern with total score t_i across all items J.

2.3.2. PE for sequential model

The random variable $Y_{ij}^{(r)}$ (explained in last subsection 2.3.1) can be modeled by an dichotomous Rasch model. For dichotomous Rasch models, the conditional pairwise method for the estimation of the person parameters is available (Kreiner 2012) and the derivation will be shown here: First, the probability for the sum of two different persons equals 1, needs to be investigated. In this case there are two possibilites to observe: Either the first person has a higher score than r and the other person has a score equal to r, or the situation is reversed.

$$\begin{aligned}
&P(Y_{ij}^{(r)} + Y_{oj}^{(r)} = 1|\theta_i, \theta_o, \beta_{jr}) = &(32)\\
&P\left(\left\{Y_{ij}^{(r)} = 1, Y_{oj}^{(r)} = 0\right\} \cup \left\{Y_{ij}^{(r)} = 0, Y_{oj}^{(r)} = 1\right\} |\theta_i, \theta_o, \beta_{jr}\right) = &(33)\\
&P\left(\left\{Y_{ij}^{(r)} = 1, Y_{oj}^{(r)} = 0\right\} |\theta_i, \theta_o, \beta_{jr}\right) + P\left(\left\{Y_{ij}^{(r)} = 0, Y_{oj}^{(r)} = 1\right\} |\theta_i, \theta_o, \beta_{jr}\right) = &(34)\\
&P\left(Y_{ij}^{(r)} = 1|\theta_i, \beta_{jr}\right) P\left(Y_{oj}^{(r)} = 0|\theta_o, \beta_{jr}\right) + P\left(Y_{ij}^{(r)} = 0|\theta_i, \beta_{jr}\right) P\left(Y_{oj}^{(r)} = 1|\theta_o, \beta_{jr}\right) = &(35)\\
&\frac{\exp(\theta_i - \beta_{jr}) + \exp(\theta_o - \beta_{jr})}{(1 + \exp(\theta_i - \beta_{jr}))\,(1 + \exp(\theta_o - \beta_{jr}))} &(36)
\end{aligned}$$

After that the common probability of observing $Y_{ij}^{(r)} = 1$ and $Y_{ij}^{(r)} + Y_{oj}^{(r)} = 1$ has to be derived:

$$\begin{aligned}
&P\left(\left\{Y_{ij}^{(r)} = 1\right\} \cap \left\{Y_{ij}^{(r)} + Y_{oj}^{(r)} = 1\right\} |\theta_i, \theta_o, \beta_{jr}\right) = &(37)\\
&P\left(Y_{ij}^{(r)} = 1, Y_{oj}^{(r)} = 0|\theta_i, \theta_o, \beta_{jr}\right) = &(38)\\
&\frac{\exp(\theta_i - \beta_{jr})}{(1 + \exp(\theta_i - \beta_{jr}))\,(1 + \exp(\theta_o - \beta_{jr}))} &(39)
\end{aligned}$$

With these interim results, the conditional probability can be calculated independent of the item parameters β_{jr}.

$$P\left(\left\{Y_{ij}^{(r)}=1\right\}\mid\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\},\theta_i,\theta_o,\beta_{jr}\right)= \tag{40}$$

$$P\left(\left\{Y_{ij}^{(r)}=1\right\}\cap\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\}|\theta_i,\theta_o,\beta_{jr}\right)/P\left(Y_{ij}^{(r)}+Y_{oj}^{(r)}=1|\theta_i,\theta_o,\beta_{jr}\right)= \tag{41}$$

$$\frac{\exp(\theta_i-\beta_{jr})}{\exp(\theta_i-\beta_{jr})+\exp(\theta_o-\beta_{jr})}= \tag{42}$$

$$\frac{\exp(\theta_i)}{\exp(\theta_i)+\exp(\theta_o)}= \tag{43}$$

$$\frac{\exp(\theta_i-\theta_o)}{1+\exp(\theta_i-\theta_o)} \tag{44}$$

$$\Rightarrow P\left(\left\{Y_{ij}^{(r)}=0\right\}\mid\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\},\theta_i,\theta_o\right)= \tag{45}$$

$$1-P\left(\left\{Y_{ij}^{(r)}=1\right\}\mid\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\},\theta_i,\theta_o\right)= \tag{46}$$

$$\frac{1}{1+\exp(\theta_i-\theta_o)} \tag{47}$$

$$\Rightarrow P\left(\left\{Y_{ij}^{(r)}=y_{ij}^{(r)}\right\}\mid\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\},\theta_i,\theta_o\right)= \tag{48}$$

$$\frac{\exp\left(y_{ij}^{(r)}(\theta_i-\theta_o)\right)}{1+\exp(\theta_i-\theta_o)} \tag{49}$$

Now the pseudo likelihood for all person parameters can be constructed over all items J and different persons:

$$L(\theta_1,\ldots,\theta_n|\underline{\underline{Y}})= \tag{50}$$

$$\prod_{i\neq o}\prod_{j(r)=1}^{J}P\left(\left\{Y_{ij}^{(r)}=y_{ij}^{(r)}\right\}\mid\left\{Y_{ij}^{(r)}+Y_{oj}^{(r)}=1\right\},\theta_i,\theta_o\right) \tag{51}$$

The items are assumed to be independent, but the person combinations are not independent. Therefore the term is called pseudo likelihood. The first product ranges over all possible pairwise combinations of persons. Further the log likelihood is maximised regarding the person parameters.

2.3.3. JMLE for sequential model

For the unconditional estimation of the sequential model Tutz (Tutz 1990; see pages 47-49) proposed the following procedure: First the polytomous response with scores $r=0,\ldots,k_j$ is dichotomized as dummy variables for each person i and item j:

$$u_{ij}=\left(u_{ij0},u_{ij1},\ldots,u_{ij(k_j-1)}\right) \tag{52}$$

$$u_{ijr}=\begin{cases}1 & \text{if response of person i to item j is r}\\ 0 & \text{otherwise}\end{cases}$$

After some derivations Tutz showed, that the unconditional log likelihood with both person and item parameters is given by

$$l_{ij} = \sum_{i,j} \sum_{r} u_{ij} \ln(\pi_{ijr}(\theta_i, \beta_j r)) + \left(1 - \sum_{r=0}^{k_j - 1} u_{ijr}\right) \ln \left(1 - \sum_{r=0}^{k_j-1} \pi_{ijr}(\theta_i, \beta_{jr})\right) \quad (53)$$

$$\pi_{ijr}(\theta_i, \beta_{jr}) = P(X_{ij} = r | \theta_i, \beta_{j1}, \ldots, \beta_{jr}) \quad (54)$$

This log likelihood will be minimised similiar to the JMLE for the PCM in section 2.2.6. First, all the person parameters are set equal to zero. Second the item parameters - given the person parameters - are estimated. In the third step, the person parameters are estimated - given the estimated item parameters - and then the procedure is repeated until convergence.

2.3.4. MMLE for sequential model

It is possible to estimate the sequential model with a marginal approach (Tutz 1997; see pages 7-9). As shown in the previous estimation methods for the sequential model, at first a data transformation is applied. The transformations u_{ijr} from last section 2.3.3 are recoded in transitions:

$$v_{ijh} = 1 - (u_{ij0} + \ldots + u_{ij,h-1}), \ h = 1, \ldots, k_j \quad (55)$$

This coding gives a vector of how many transitions from score h-1 to h, person i was able to achieve on item j. For example, if person i has scored one point with a max score $k_j = 3$ on item j, then $u_{ij} = (0, 1, 0)$. This is equivalent to the transition coding $v_{ij} = (1, 0, 0)$. If a person i scored zero points then $v_{ij} = (0, 0, 0)$ and with a max score three $v_{ij} = (1, 1, 1)$. Based on transitions the marginal Likelihood for a person i is represented by (Tutz 1997; see page 8):

$$L_i = \int \prod_{j=1}^{J} \prod_{s=1}^{s_{ij}} F(\theta - \beta_{js})^{v_{ijs}} \left(1 - F(\theta - \beta_{js})\right)^{1 - v_{ijs}} g(\theta) \ d\theta \quad (56)$$

So all answers from person i over all items are taken into account and the likelihood is marginalized with priori density $g(\theta)$. Numerical integration methods have to be used to approximate the likelihood for a person i. Fortunately the equation 56 can be estimated by using standard generalized linear mixed model software. The response with v_{ijs}, the linear regression term $\theta + z_{is}^T \beta$ with the structure $P(V_{ijs} = 1) = F(\theta + \beta_{js})$ have to be specified. Additionally a normal distribution $\theta \sim N(0, \sigma^2)$ is assumed (Tutz 1997; see page 9). In this thesis an EM type algorithm is used for estimation of the parameters. In the Expectation step, the likelihood is marginalized and then the estimated likelihood is maximised. The likelihood in the E-step is not available in an analytic closed form and has to be approximated by the Gaussian Hermite quadrature. In the simulation study 10 quadrature points were adaptively choosen. For further details see (Stiratelli et al. 1984).

2.3.5. Confidence intervals for sequential parameter estimates

In data analysis where the true model is unknown, confidence intervals for parameters quantify the variability of the estimation. In the case of the data transformation of the

MMLE, it is possible to use the asymptotic Wald confidence intervals, because no external restrictions are needed and after the transformation there are enough observations available. But for the self-implemented JMLE and CMLE, with additional data restrictions and parameter interpolation, this approach does not work. An alternative to the asymptotic theory is to use bootstrap confidence intervals: Nonparametric or parametric bootstrap (Davison and Hinkley 1997; see pages 15 - 27). The nonparamtric bootstrap with block design takes samples with replacement of the observations $b = 1, \ldots, B$ times. For every bootstrap sample b the coefficients are estimated. Then the quantiles corresponding to the significance level are computed for calculation of the bootstrap percentile intervals (Davison and Hinkley 1997; see pages 202-211). In the paramatric bootstrap, the samples are taken from a parametric distribution, with parameters estimated from the data. After this, the procedure is the same as with nonparametric bootstrap confidence intervals. In this thesis parametric bootstrapping is used. One theoretical advantage of this approach is, that the search space is usually larger, which means that the amount of possible bootstrap samples is higher. With more variation in the samples, more granular estimates of the variance can be made.

2.4. Measures of performance in the simulation study

This thesis focuses on the estimation quality of the person parameters. All different estimation methods for the PCM and the sequential model will be compared with the multidimensional root mean squared error: $\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{tr}\left\{\mathrm{E}\left[\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)^T\right]\right\}}$; $\theta, \hat{\theta} \in \mathbb{R}^p$ with $\hat{\theta}$ as the estimated person parameters, θ as true person parameters and p the number of persons. In practise the empirical analoga will be used to estimate the respective covariance matrix. It is a well known result, that the MSE can be decomposed in bias and variance of the estimator: $\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$. This is the classical bias-variance tradeoff. The two components of the RMSE are analysed to investigate, what causes the deviations from true parameters. In this thesis the sum of measures across all persons is evaluated: $\mathrm{Bias}(\hat{\theta}) = \left(\mathrm{E}(\hat{\theta}) - \theta\right)^T \underline{1}$ and $\mathrm{Var}(\hat{\theta}) = \mathrm{tr}\left\{\mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))(\hat{\theta} - \mathrm{E}(\hat{\theta}))^T\right)\right\}$ Based on this information either bias corrections or variance reducing modifications of the estimator can be developed. The theoretical MSE will be approximated by the empirical counterparts, which means, that the expected value is replaced by the arithmetic mean.

Having estimated all six methods (CMLE, WMLE, PE, RPE, JMLE, MMLE), the RMSE results are supplementary analysed through the use of nonparametric Monte Carlo tests. First, in the context of root mean squared errors of the estimators, there are no reasons to assume some specific distributional assumptions, e. g. normal distribution and variance homogenity. Therefore the t-Test for two- dependent-samples is not appropriate (Sheskin 2000; e. g. page 451). Dependent samples have evaluated the same persons, but under different conditions, which is the case here. Second it is not clear, if the asymptotic assumptions hold with 1000 observations. The Monte Carlo tests dont rely on the asymptotic theory of the null hypothesis. In the general Monte Carlo test (Silva and Assuncao 2011; see page 1), the test statistic $T(X_0)$ is calculated from the available data. Then the test statistics $T(X_1), T(X_2), \ldots, T(X_B)$ are simulated under the null hypothesis of equal groups. The maximum likelihood estimator of the p-value is the mean of the number of simulated statistics greater or equal to the observed value $T(X_0)$. The paper (Silva and

Assuncao 2011) demonstrates, that the general Monte Carlo test and exact tests have similiar magnitude in power. In this thesis 10000 Monte Carlo replicates are evaluated.

In the simulation study the Friedman two-way-analysis of variance by ranks test (Sheskin 2000; e.g. pages 681-690) is used as global test. For pairwise comparisons the Wilcoxon matched pairs signed rank test will be used as implemented in the R-package (Hothorn et al. 2013). This procedure was applied in typical benchmarking experiments (Eugster et al. 2008). Both tests are nonparametric in a sense that they have less assumptions than their parametric counterparts. The Friedman test is a kind of analysis of variance by ranks for $k \geq 2$ of dependent samples and tests, if at least two samples represent different median values. The null hypothesis and the test statistic are as follows (Sheskin 2000; see pages 682, 684):

$$H_0 : \text{Median}(X_1) = \text{Median}(X_2) = \ldots = \text{Median}(X_k) \tag{57}$$

$$\chi_r^2 = \frac{12}{nk(k+1)} \left[\sum_{j=1}^{k} \left(\sum R_j \right)^2 \right] - 3n(k+1) \tag{58}$$

First for every person the corresponding ranks from the obversations over all methods $j = 1, \ldots k$ are evaluated. In this study the ranks 1-6 are allocated to the five methods. Then for each method the sum of ranks $\sum R_j$ is calculated and inserted in χ_r^2. The statistic is asymptotically χ^2 distributed with k-1 degrees of freedom. If the test is significant, then there are differences in the groups. But the Friedman test does not investigate which one. Therefore further two-sided Wilcoxon tests are used for all pairwise comparisons. The following assumptions are made by the Wilcoxon test for two-dependent-samples (Sheskin 2000; see page 484):

- Persons are randomly sampled from a given population
- Original scores have an interval scala
- The distribution of the difference-scores in the populations is symmetric around the median of the population of difference-scores

The null hypothesis states, that the medians for both samples are equal. For the Wilcoxon test statistic the absolute differences are evaluated from the two methods. After that, the sum of all absolute differences, with negative and positive signs, is computed. The smaller of the two values is used. If the populations are equal, both sums should be identical.

In this thesis a global significane level of 0.05 is used. The classical Bonferroni correction is applied to ensure a family wise type I error, less or equal than alpha (Abdi 2007; see page 6). A significant result between some pair-of-methods from the two sided Wilcoxon test indicates, that the populations are different. If that is the case, the median value for the RMSE across the Monte Carlo samples is computed. If the median of method 1 is less than the median in method 2, than the method 1 is regarded as significantly superior. If it is less, than method 1 is significantly inferior. With this information a topological order of methods with greater or less signs can be constructed, if the test is significant. Otherwise the methods are approximately equal and the null hypothesis is accepted.

2.5. Measures of performance in data analysis

In the data analysis the true data generating process is unknown. Therefore it is necessary to use other measures as in the previous section 2.5. One possibility is to take a goodness of fit measure, such as the residual sum of squares. This has the disadvantage, that it only measures data adaption on the training data, which is in many cases to optimistic. A better indicator of performance is the out-of-sample prediction error. For independent samples in regression analysis the standard approach is to use cross validation (CV) or nonparametric bootstrapping for the estimation of the prediction error. See (Efron and Tibshirani 1993; see chapter 17) for the introduction of these methods. It is important to realize, that these methods are also estimators. Of course estimators also have a bias and variance in relation to the true expected prediction error. (Molinaro et al. 2005; see page 3306) conducted simulation studies for comparison of different resampling methods in high dimensional classifcation problems. The conclusion is, that leave-one-out cross validation performed well in estimating the true prediction error. Leave-one-out CV is a computationally less efficient procedure, because as many models have to be fitted as observations are available. Fortunately the 10-fold CV approximated the leave-one-out CV very well in simulations (Molinaro et al. 2005; see page 3306). (Kuhn and Johnson 2013; see page 78) therefore recommends to use repeated 10-fold CV to reduce variance of the estimator, while still maintaining a small bias. Other authors have conducted simulations comparing bootstrap and CV methods for the estimation of prediction error such as (Kim 2009) and (Ounpraseuth et al. 2012). The result is, that 10-fold CV is a better alternative than bootstrapping for the estimation of prediction error. Therefore a 10 times repeated 10-fold CV will be used in this study, which means a 10-fold CV will be repeated 10 times and the average of the prediction error estimates will be taken as measure.

In the case of IRT the CV cannot be used in the common way. One reason is that for every row in the data set, one person parameter has to be estimated. If a random subsample from the original data is taken by CV, then the test sample has always new-persons-not-available in the training data. With the standard procedure no predictions are possible for the left-out-persons in the test data, because of missing parameters. In this thesis an alternative method for the estimation of prediction error in IRT models is presented:

1. Choose a parametric IRT model, e. g. PCM, sequential model
2. Construct randomly t times $s = 1, \dots, S$ CV test samples of the original data.
3. Estimate the item parameters of the model with the training data (all original data except, one specific test set)
4. Given the estimated item parameters of the training sample, estimate the missing person parameters in the test sample (joint maximum likelihood)
5. Calculate the estimated conditional probabilities for all possible scores of each person i and item j within the test sample
6. Transform the data values of the test sample in the vector $\underline{y}_{ij} = (y_{ij0}, y_{ij1}, \dots, y_{ijk})$ with $y_{ijk} = 1$ if the response of person i to item j falls into the category k and 0 otherwise

7. Find the probability for category k, which is the observed category in the test sample for each person i and item j

8. Compute the loss function $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = 1 - \hat{p}^*_{ijk}$ and take the arithmetic mean over all persons i and item j in the test sample

9. Repeat that procedure for all test samples and calculate the average results. This is the proposed estimator of the prediction error

The loss function can be derived, based on the following arguments: In an classification context the usual practise is to use the misclassification error rate as measure. This is the proability that the predicted class is not equal to the observed value in the test sample. For this a prediction of the categories in the test sample must be made. Normally the category with the highest conditional probability would be suggested. The observed values and the predicted values in the test samples are compared. The loss function will have the value 1, if the values are not equal, and 0 otherwise. But this loss function does not use all available information from the discrete, conditional probability distribution. One alternative would be to take the absolute deviation from the observed values and the predicted values. Then the loss function would have discrete possible outcomes of $\{0, 1, \ldots, k\}$, with k beeing the max score of a given person i and an item j. This approach is not sensitive for different probabilities in the conditional distribution, as long as the category remains the highest probability.

To illustrate that, assume that the estimated probability distribution for an item with cateogries $0, 1, 2, 3$ is $\underline{\hat{p}}_{ij} = (0.5, 0.3, , 0.15, 0.05)$ and the test set value is 0. In this case the predicted category would be 0. Now assume the estimated probabilities would be $\underline{\hat{p}}_{ij} = (0.7, 0.1, , 0.15, 0.05)$. Then the predicted category would also be 0. The previous mentioned loss function is the same in both cases. Nevertheless the probability distribution suits better the observation of the test sample, because the probability to get the observed test value is higher. Let y^*_{ij} be the value where $y_{ijk} = 1$, $\hat{p}^*_{ij}$ the matching probability and $\underline{y}^{**}_{ij}$ be the values, where $y_{ijk} = 0$. To get a more fine grained measure of the loss the following is derived:

$$L_1(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = \sum_{l=0}^{k} |y_{ijl} - \hat{p}_{ijl}| \tag{59}$$

$$= \left|y^*_{ij} - \hat{p}^*_{ij}\right| + \sum_{l=0}^{k-1} \left|y^{**}_{ijl} - \hat{p}_{ijl}\right| \tag{60}$$

$$= y^*_{ij} - \hat{p}^*_{ij} + \sum_{l=0}^{k-1} \hat{p}_{ijl} \tag{61}$$

$$= 1 - \hat{p}^*_{ij} + 1 - \hat{p}^*_{ij} \tag{62}$$

$$= 2\left(1 - \hat{p}^*_{ij}\right) \tag{63}$$

This measure has the advantage to account for even small deviations in the probability of the observed category in the test data. Further it has a range in the interval $[0, 2]$. Effectively L_1 does only use information from $\hat{p}^*_{ij}$ and all other probabilities are redundant. To normalize the measure in the interval $[0, 1]$ the used loss function is

$L(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = 1 - p^*_{ij} \propto L_1$. This measure has a clear interpretation: It is the probability for not predicting the observed class in the test sample, given the estimated probability from the choosen model for person i and item j. If $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = 0$, then the predicted class would always match the observed class in the test sample, which is locally ideal. The other extreme case $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = 1$ means, that the predicted class can not be the observed class in the test sample. Of course, that is locally the worst case possible. But a local optimum must not be a global optimum. For each test sample there are different conditional probability distributions in each cell. A similiar loss function is the predictive deviance, which is defined as $L_D(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = -\log(\hat{p}^*_{ij}) \in [0, \infty)$. This scale is not normed in an finite interval and small probabilities observed test values are punished more than in $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$. The loss function $L_D(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ increases nonlinear for decreasing probabilities. In the previous measure $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ the loss function is linear for decreasing probabilities and punishes deviations across all values in the interval $[0, 1]$ equally.

A disadvantage of the measure $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ is, that only a deterministic model can achieve a perfect loss of 0. Consider an item with 3 categories $\{0, 1, 2\}$. Suppose the person in the test data responded 1 and the true conditional distribution for the categories is $p = (0.2, 0.5, 0.3)$. In this case $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij}) = 0.5$ even if the true item parameters and person parameters are known. A deterministic model with degenerated probability distribution $p = (0, 1, 0)$ would be needed, to achieve a loss of 0. It means, that the measure is too strong to accurately discriminate between the stochastic prediction models. For comparison consider the Bayes classifier, which assigns the class to the cateogory with highest probability. Then the predicted class is compared with a 0-1 loss, which is 1, if the classes are different and 0 otherwise. With this measure the loss would be 0 in this observation case. The average over all observations yields to an estimate of the misclassification error of the test data. Most of the time $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ should be higher than the 0-1 loss.

The approach outlined in this section 2.5 may be used for another question: How well will the estimated model predict the performance of the same group of persons, if they solve new items? With some modifications this question can be investigated. Instead of subsampling the persons with cross validation, the items are subsampled and the group of persons is equal in the training and the test set. Consider a data matrix with n persons and J items as shown in table 2.5. This data set with values $x_{..}$ will be randomly split in training data columns $j = 1, \ldots, k$ and test data columns $j = l, \ldots, m$. This splitting is applied on a count of test samples S $s = 1, \ldots, S$. Analogous for repeated CV this has to be repeated t times. For every s the appropriate training sample is fitted with a choosen estimation method. Then the item parameters of the test sample are estimated with the joint likelihood, given the estimated person parameters of the training data. Given the parameters from the test sample a prediction is made for every person and item in the test sample. This prediction is evaluated with the previously specified loss function $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ of the test sample and averaged.

More generally it is possible to create test samples leaving out persons and items simultaneous from the training data. In principle it is based on a row & column cross validation (r&c cv). Additionally to the previous questions the following question can be evaluated: How well does the model predict a populuation of new persons and new items? The data looks like table 2.5 for a single split in training and test data:

$\text{Person}_i\|\text{Item}_j$	ItemTrain_1	$\cdots$	ItemTrain_k	ItemTest_l	$\cdots$	ItemTest_m
PersonTrain_1	x_{11}	$\cdots$	x_{1k}	x_{1l}	$\cdots$	x_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
PersonTrain_n	x_{n1}	$\cdots$	x_{nk}	x_{nl}	$\cdots$	x_{nm}

Table 1: Test split based on items

$\text{Person}_i\|\text{Item}_j$	ItemTrain_1	$\cdots$	ItemTrain_k	ItemTest_l	$\cdots$	ItemTest_m
PersonTrain_1	x_{11}	$\cdots$	x_{1k}	x_{1l}	$\cdots$	x_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
PersonTrain_n	x_{n1}	$\cdots$	x_{nk}	x_{nl}	$\cdots$	x_{nm}
PersonTest_o	x_{o1}	$\cdots$	x_{ok}	x_{ol}	$\cdots$	x_{om}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
PersonTest_p	x_{p1}	$\cdots$	x_{pk}	x_{pl}	$\cdots$	x_{pm}

Table 2: Test split based on persons and items

The lower right box consists of new persons and new items. At first it seems, that no prediction is possible. But with a closer look the missing person parameters can be estimated by using JMLE on the lower left test data, given item parameters from the training data. The missing items in the test sample can be again estimated by applied JMLE to the upper right test data, given the person parameters from the training sample. After these consecutive steps, all test data parameters are available for prediction. If the training and test data partially share persons and items, then it is possible to evaluate predictive performance.

An advantage of the r&c cv approach is, that it uses all information from the training data (person parameter and item parameter estimates) and not only one parameter set to evaluate prediction. The lower left and upper right test sets will be used two times in the approach: One time for estimation of the of the unknown parameters and a second time for prediction evaluation. Ideally the test set should be only used once for prediction evaluation, because if the model partly adapts to the data before evaluation of the prediction error, the true prediction error may be underestimated. A similiar effect accurs if for example in a generalized additive model the smoothing parameter is estimated, using cross validation. If the prediction error is evaluated on the model with all data and the smoothing parameter, the information of the cross validation then already is used in the tuning process. For a discussion of this point see the reference (Hastie et al. 2011; pages 247-249). The approach with leaving out persons and items of test data produces one test set in the lower right of the table, which will only be used for assessing predictive performance. To avoid any dependency, only the test data in the lower right should be used for evaluating the prediction error. If a 10x10 design is used, repeated cross validation is not necessary, because already 100 cv samples are evaluated.

3. Simulation of PCM

3.1. Design of PCM simulation

In this chapter the details for the simulation of the data matrices are explained. In each design the number of persons is fixed to 750. This quantity is quite common in psychometric studies, high enough for good estimations of the item parameters and still computationally manageable in numbers of parameters. The item parameters are choosen deterministically from the equidistant interval $[-5, 5]$, depending on the number of items. For example with five item parameters it would be $\{-5, -2.5, 0, 2.5, 5\}$. Because of the polytomous Rasch model every item has several categories. They are fixed at five, ranging from 0 to 4 points each item. The person parameters are generated randomly.

The design covers 25 different scenarios, varying in different distributions for generation of person parameters θ and number of items. In most scenarios it is assumed, that $\theta \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with an expected value μ and a variance σ^2. An overview of the scenarios is given below:

1. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 10 items
2. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 20 items
3. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 30 items
4. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 40 items
5. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 50 items
6. $\theta \overset{\text{i.i.d.}}{\sim} N(3, 1)$ and 10 items
7. $\theta \overset{\text{i.i.d.}}{\sim} N(3, 1)$ and 20 items
8. $\theta \overset{\text{i.i.d.}}{\sim} N(3, 1)$ and 30 items
9. $\theta \overset{\text{i.i.d.}}{\sim} N(3, 1)$ and 40 items
10. $\theta \overset{\text{i.i.d.}}{\sim} N(3, 1)$ and 50 items
11. $\theta \overset{\text{i.i.d.}}{\sim} N(-3, 1)$ and 10 items
12. $\theta \overset{\text{i.i.d.}}{\sim} N(-3, 1)$ and 20 items
13. $\theta \overset{\text{i.i.d.}}{\sim} N(-3, 1)$ and 30 items
14. $\theta \overset{\text{i.i.d.}}{\sim} N(-3, 1)$ and 40 items
15. $\theta \overset{\text{i.i.d.}}{\sim} N(-3, 1)$ and 50 items
16. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 9)$ and 10 items
17. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 9)$ and 20 items

18. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 9)$ and 30 items

19. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 9)$ and 40 items

20. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 9)$ and 50 items

21. $\theta \overset{\text{i.i.d.}}{\sim} 0.5 * N(-1.5, 1) + 0.5 * N(1.5, 1)$ and 10 items

22. $\theta \overset{\text{i.i.d.}}{\sim} 0.5 * N(-1.5, 1) + 0.5 * N(1.5, 1)$ and 20 items

23. $\theta \overset{\text{i.i.d.}}{\sim} 0.5 * N(-1.5, 1) + 0.5 * N(1.5, 1)$ and 30 items

24. $\theta \overset{\text{i.i.d.}}{\sim} 0.5 * N(-1.5, 1) + 0.5 * N(1.5, 1)$ and 40 items

25. $\theta \overset{\text{i.i.d.}}{\sim} 0.5 * N(-1.5, 1) + 0.5 * N(1.5, 1)$ and 50 items

In the first five scenarios a standard normal distribution is assumed. In the cases six to 10 the group of persons is very skilled with an expected value of 3 and in the scenarios 11 to 15 a group of unskilled persons will be analysed. Then from 16 to 20 a group with high Variance is simulated. Last but not least the scenarios 21 to 25 are simulated using a symmetric, discrete mixture distribution, which has two modes at the expected values of the two normal distributions. For each scenario 1000 Monte Carlo samples are accomplished. This procedure is applied for all six estimation methods of the PCM as described in section 2.2. Altogether 150 000 models were processed to provide the results for this thesis. This is computationally costly, so part of the code, especially likelihood calculations, are programmed in C code. To ensure to have the calculations ready in reasonable time, parallel processing techniques were used.

Science work is always collaborative (Tetens 2013; e. g. pages 17-28) and a basic principle in science is reproducibility and to check other scientists works for errors. Consequently a statistical study should be reproducible for other scientists. On the other hand a simulation study like this thesis uses random numbers, which are not necesary reproducible. To overcome this downside, the following trick is applied for random sampling: Initially the random seeds for pseudo number generation are choosen randomly. Based on this seeds, the data matrices are drawn. Each cell of the matrix has another conditional probability distribution and an own seed. On this seed base the results can be reproduced by other scientists.

3.2. Sensitivity analysis: Impact of spline interpolation on performance

Before the simulation process is conducted, the impact of the spline interpolation after estimation of the PCM is to be investigated. If all Monte Carlo samples across the 25 different scenarios are evaluated the range of cases interpolated is between 0 % and 7.5 % of the observations. The mean relative proportion of interpolated cases over all samples is 0.8 %. A Sample of two scenarios is checked here:

1. $\theta \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and 10 items

2. $\theta \overset{\text{i.i.d.}}{\sim} N(3,1)$ and 10 items

For this sensitivity analysis 12000 models - without spline interpolation - were processed. In the first scenario, in 19,6 % of the 1000 Monte Carlo samples, an interpolation was necessary. In each Monte Carlo sample - which used interpolation - only 1 of 750 observations was interpolated. In the second scenario all Monte Carlo samples were interpolated, ranging from 0.1 % to 2.7 % of interpolated observations. The mean relative proportion is 1.1 %.
For comparison, the RMSE for each Monte Carlo sample - without spline interpolation - will be divided by the cases with spline interpolation for each method and each scenario. The boxplots of the relative deviations of the spline interpolated values are shown in the figure below:

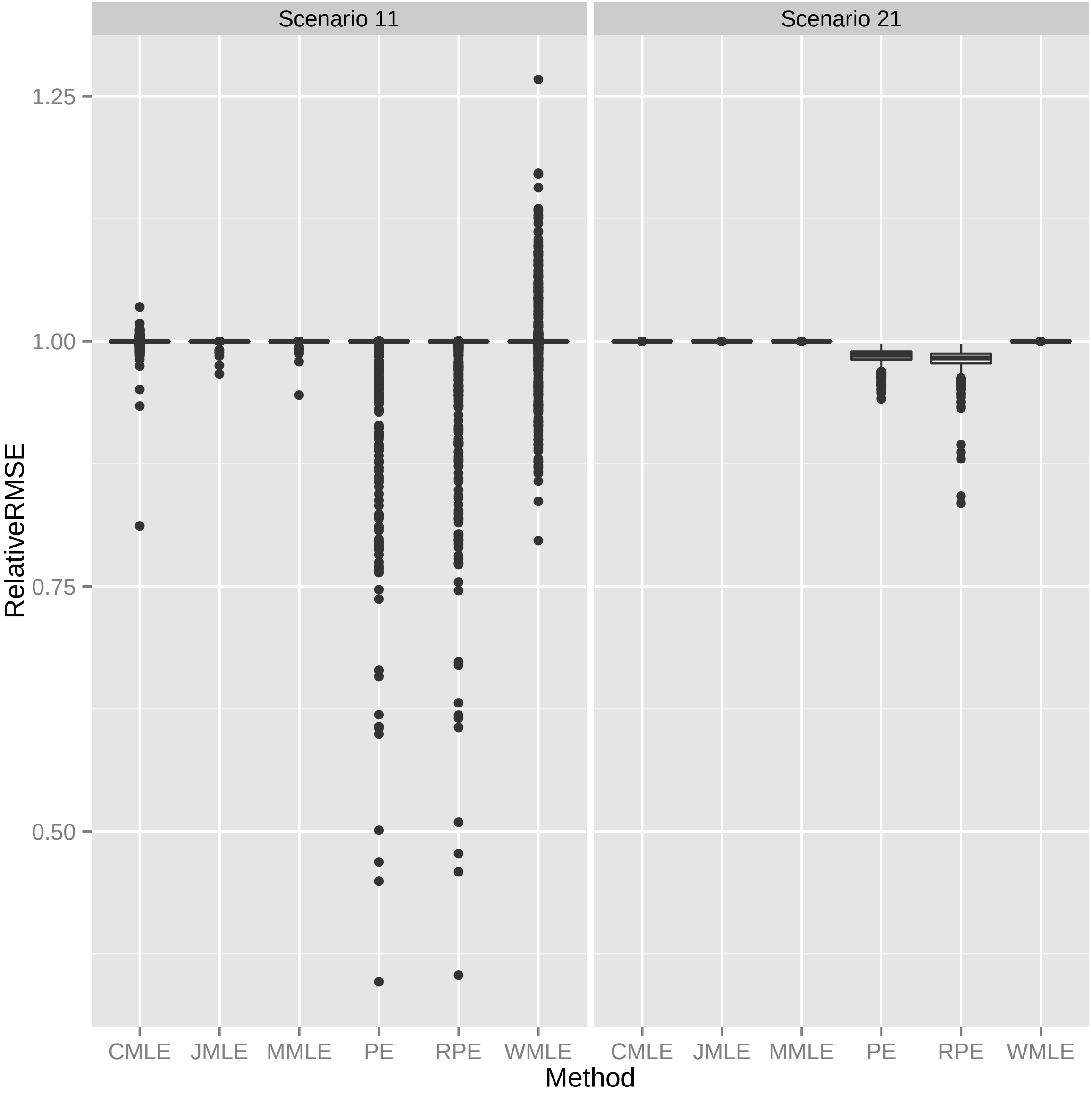

Figure 1: PCM: Relative RMSE comparison with and without interpolation

On the x-axis the different estimation methods are displayed for the two scenarios. A relative RMSE of 1 means, that the RMSE with and without interpolation is the same.

In the first scenario 11, the three methods CMLE, JMLE, MMLE have few outliers, but the majority of the 1000 Monte Carlo RMSE samples are almost equal. The inter-quartile-range is nearly 1 for the pairwise methods and there are more outliers, which are stretched further apart, but only below 1. For the WMLE the outliers are almost symmetrically distributed around 1. This is similiar in the case of CMLE with smaller outlier variation. In the second scenario 21 there is almost no difference between interpolated and not interpolated values. An exception are the pairwise methods. This time there are fewer outliers, but the interquartile range is shifted below 1. So the majority of cases is different, but the difference is small enough for beeing neglectable. Most of the other not mentioned scenarios have a similiar magnitude of interpolation. From this it is concluded, that the minor deviations in the results have no impact on topological ordering of the methods based on significance.

3.3. Results of PCM simulation

In the following section only parts of the scenarios - as defined in section 3.1 - are displayed. In every section an RMSE ist calculated for the mean of all parameters of one Monte Carlo sample. For each method a summary of the Monte Carlo samples is provided in a boxplot. All methods are set in relation to the RMSE of CMLE. With this adjustment, for every Monte Carlo sample the relative deviation to other samples of the CMLE are shown. Further the MSE over all Monte Carlo samples is investigated and the classical bias-variance split is applied, as explained in section 2.4. The sum of squared biases, and the sum of variances over all person parameters are shown in a stacked barplot. In this section only the most important graphic results are shown. For a reference to other scenarios see A.1.

3.3.1. Scenario 1 - 5: Standard normal distributed abilities

In figure 2 the boxplots based on RMSE are shown for each of the Monte Carlo samples with 10 items. On the x-axis the different estimation methods are presented. On the y-axis the relative RMSE in relation to the CMLE is shown. The first impression is, that the methods JMLE, MMLE, PE, RPE have about a 50 % smaller RMSE than CMLE and WMLE. In this scenario the WMLE has the largest box, which means, that the first and third quartiles are the biggest among the methods. The pairwise methods have both a lot of outliers (outside the 1.5 times interquartile range). The Friedman test and all pairwise Wilcoxon tests were significant with p-values smaller than 10^{-16}. On average, the standard normal distributed scenarios have the lowest RMSE in comparison with the other scenarios.

The topological order is **MMLE** $<$ **JMLE** $<$ **PE** $<$ **RPE** $<$ **CMLE** $<$ **WMLE**. So MMLE is significantly the best estimator for this scenario and WMLE is ranked last. In this case the bias correction of WMLE does not improve the CMLE.
In figure 3 the Bias-Variance decomposition of the sum of MSE over all Monte Carlo samples and parameters is ploted. As before, on the x-axis the methods and on the y-axis the values for MSE are shown. The CMLE, WMLE have an approximately equal amount of variance and bias in relation to the corresponding sum of MSE. The sum of MSE of the estimators of the other methods mainly consists of variance. MMLE has the smallest bias and PE has a little bit lower variance than RPE.

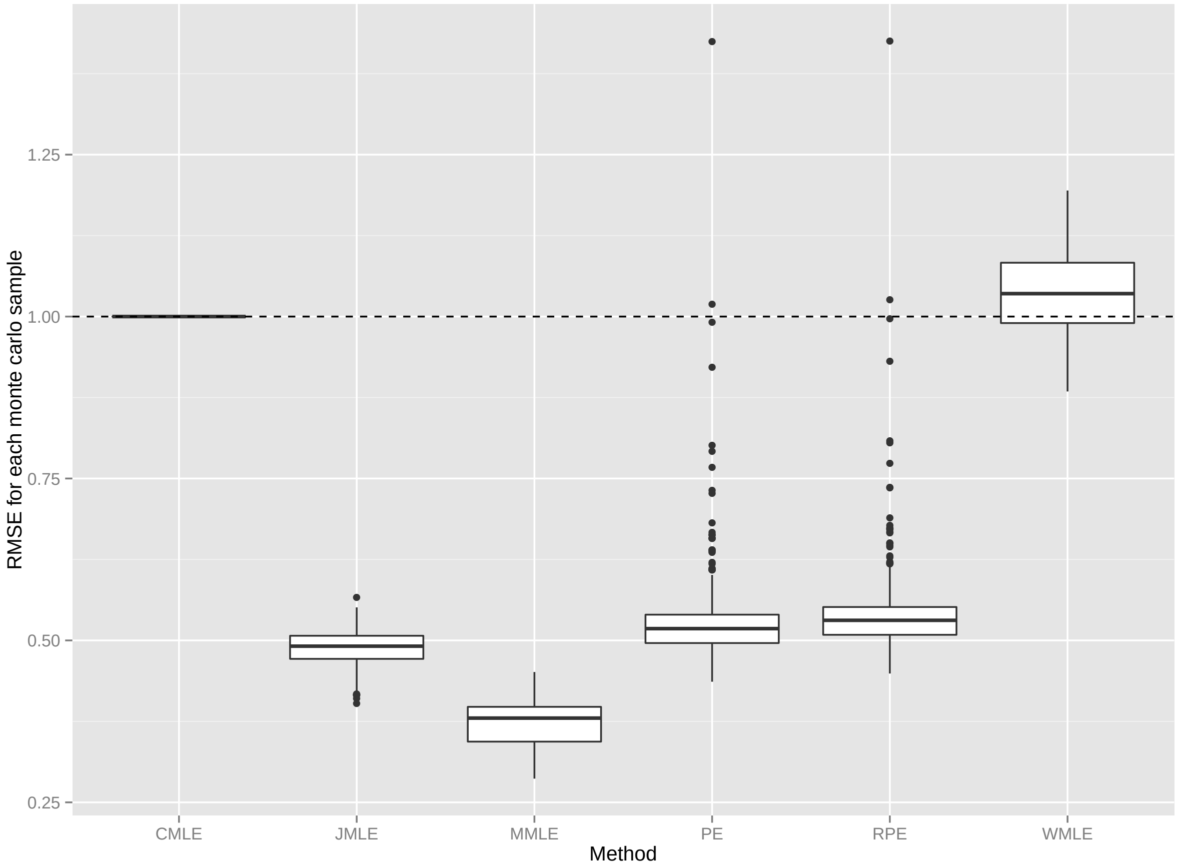

Figure 2: PCM: RMSE calculated for each Monte Carlo sample in scenario 1

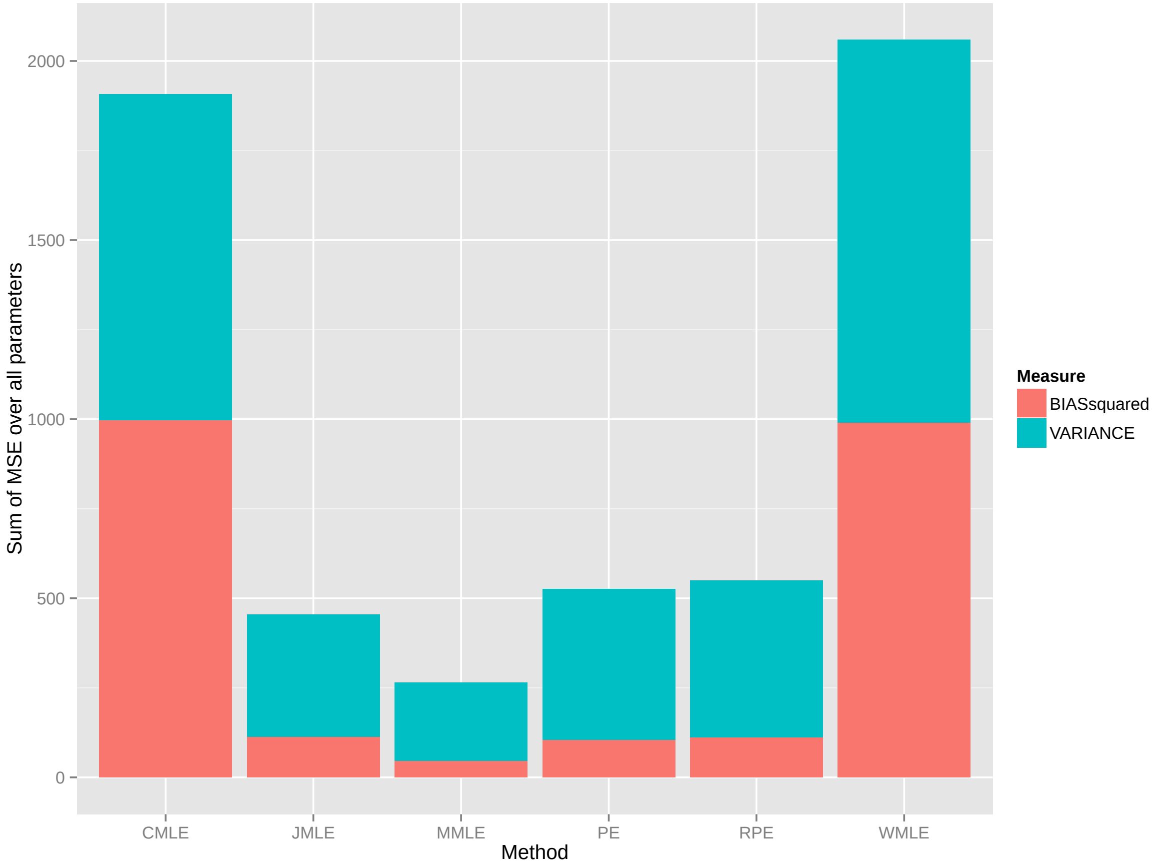

Figure 3: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 1

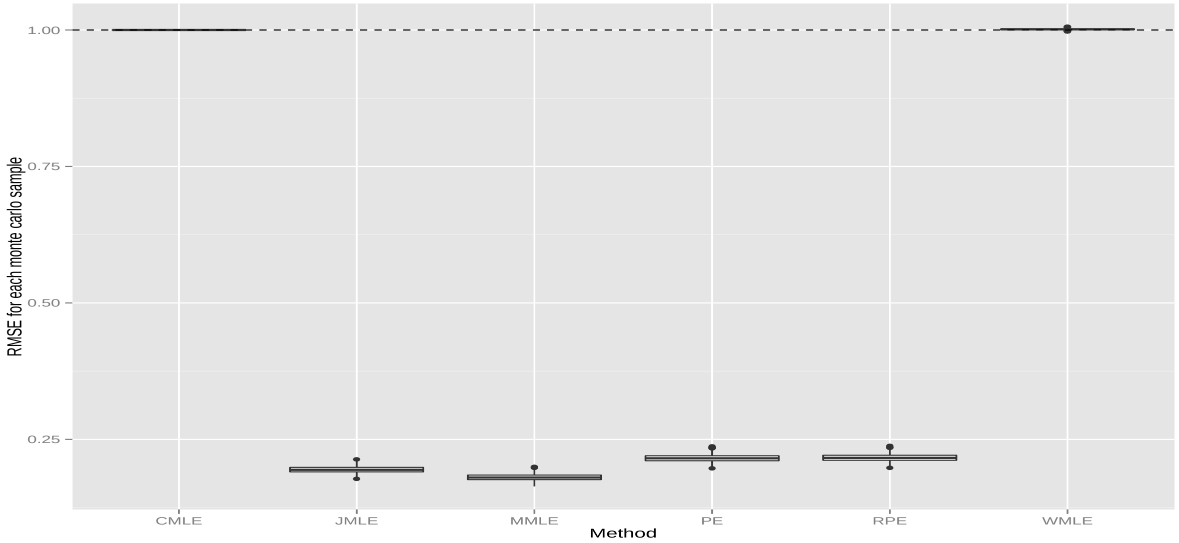

Figure 4: PCM: RMSE calculated for each Monte Carlo sample in scenario 5

For the same setting the different numbers of items do not influence the results substantially. Therefore only the fifth scenario is presented here (e. g. figure 4) and all other plots are shown in the appendix A. In scenario 5 a total of 50 items are taken into account. The gap between the methods JMLE, MMLE, PE and RPE vs. CMLE, WMLE got obviously larger and the variance of the RMSE got smaller for all methods. In figure 5 the overall sum of MSE is smaller than in scenario 1. This makes sense because with 50 items, there are more observations available and therefore more person parameters to estimate. The bias compared to variance is increased for the CMLE and WMLE, but almost vanishes for the other estimation methods. Again MMLE retains the smallest bias.

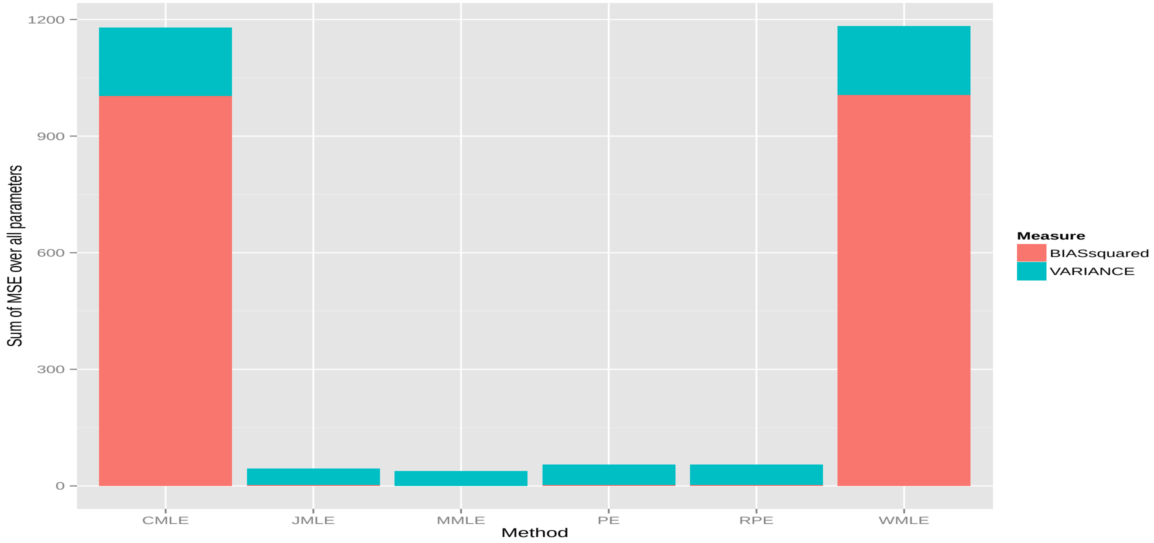

Figure 5: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 5

3.3.2. Scenario 6 - 10: Group of high performers

In the scenarios 6-10 the group consists of persons with high ability levels $\theta \sim N(3,1)$. In figure 6 the methods do not differ as much from CMLE, as in the scenarios 1-5. RPE has more and farther stretched outliers than PE. All tests in the scenarios 6-10 were significant with a p-value smaller 10^{-16}.

The resulting topological order for scenarios 6-10 is **PE** $<$ **RPE** $<$ **MMLE** $<$ **JMLE** $<$ **WMLE** $<$ **CMLE**. In comparison with the results of scenarios 1-5 the WMLE improved over the CMLE and MMLE and the Pairwise Methods changed their order. The MMLE, JMLE have higher variances than in scenario 1. In the bias-variance-decomposition (figure 7) the sum of MSE consists mainly of variance in all methods (except CMLE, WMLE). In this case a weighting of the WMLE improved the accuracy of the estimation compared to the CMLE. All other results of scenarios 6-10 are stored in appendix A.

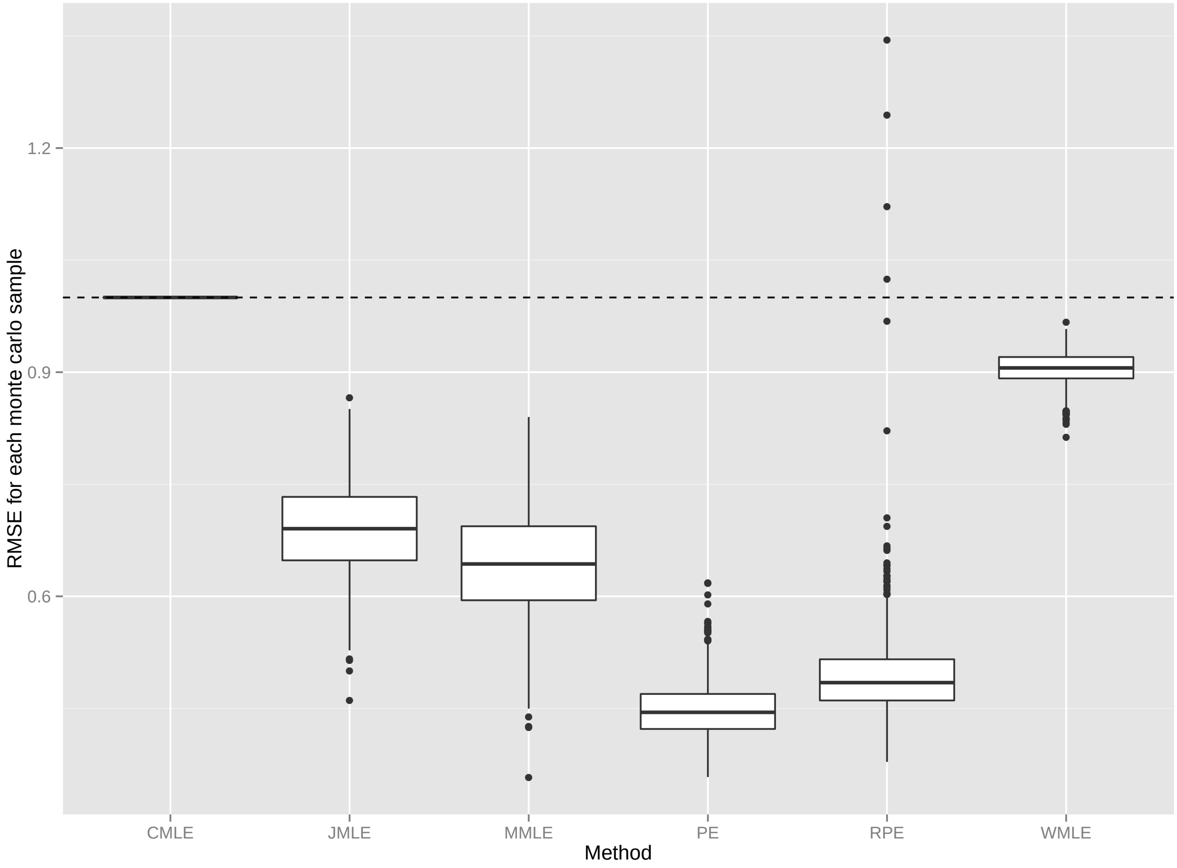

Figure 6: PCM: RMSE calculated for each Monte Carlo sample in scenario 6

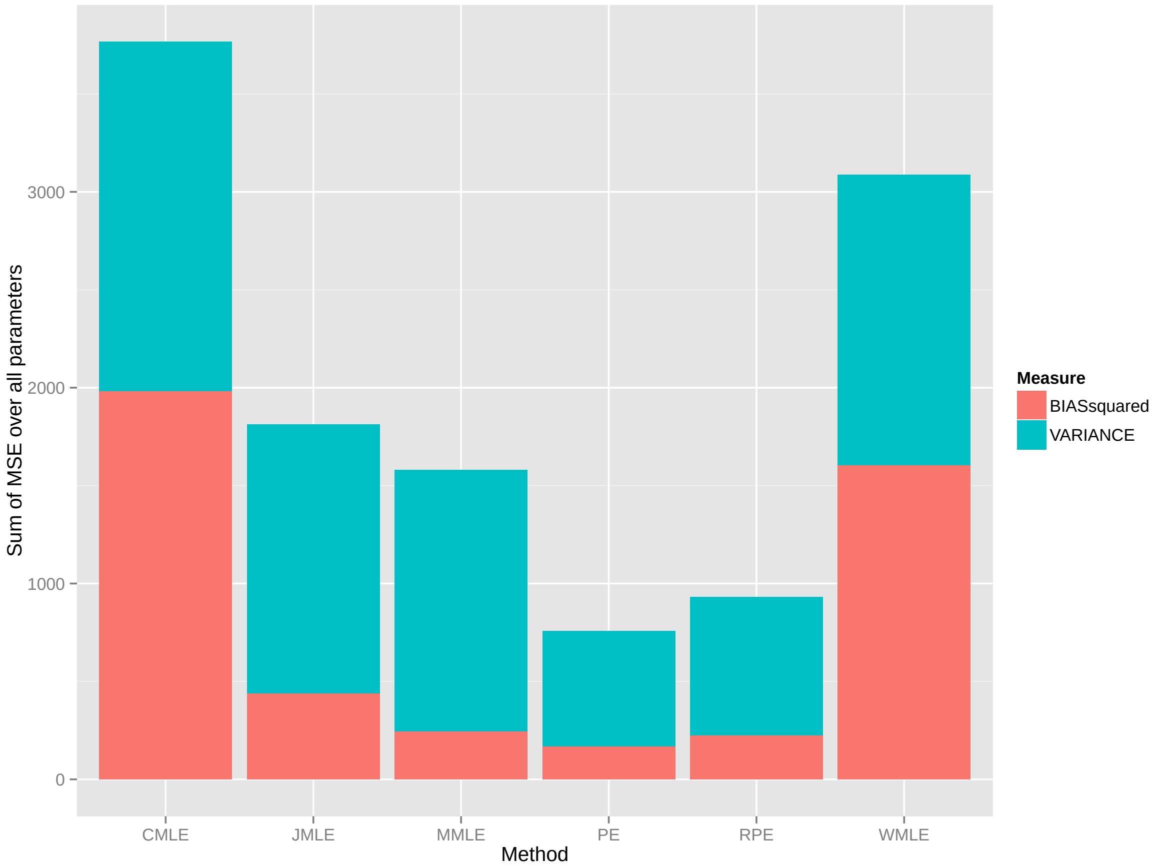

Figure 7: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 6

Another remarkable result is, that in scenario 10 the significant differences in the methods JMLE, MMLE, PE, RPE vanishes (figure 9). As the number of items increase, the differences in RMSE between the methods (except CMLE, WMLE) gets smaller. The Friedman test results are significant (p-value $< 10^{-16}$). The variability of the RMSE in the methods JMLE, MMLE across the Monte Carlo samples does not decrease with higher number of items.

The topological order is ***other methods*** $<$ **WMLE** $<$ **CMLE**. The ranking is not as clear as in the scenario 6 because **JMLE** $>$ **MMLE**, **PE** $<$ **RPE**, but **MMLE** $\sim$ **PE**, **MMLE** $\sim$ **RPE** and **JMLE** $\sim$ **PE**. With this results an ascending topological order of the estimation methods makes no sense. Similiar to section 3.3.1 the variance of the estimators decrease with more items.

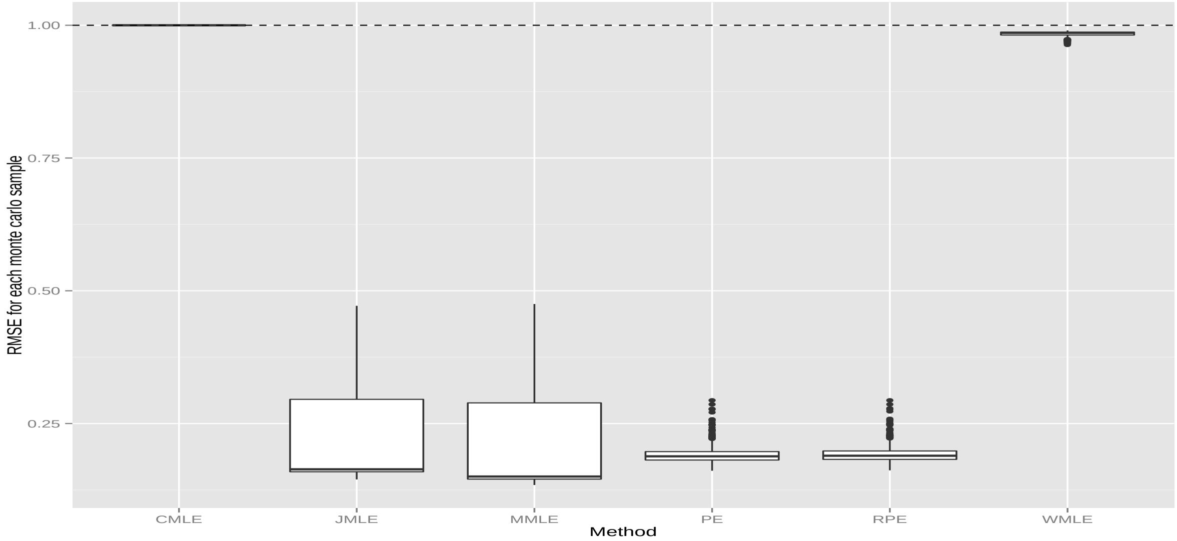

Figure 8: PCM: RMSE calculated for each Monte Carlo sample in scenario 10

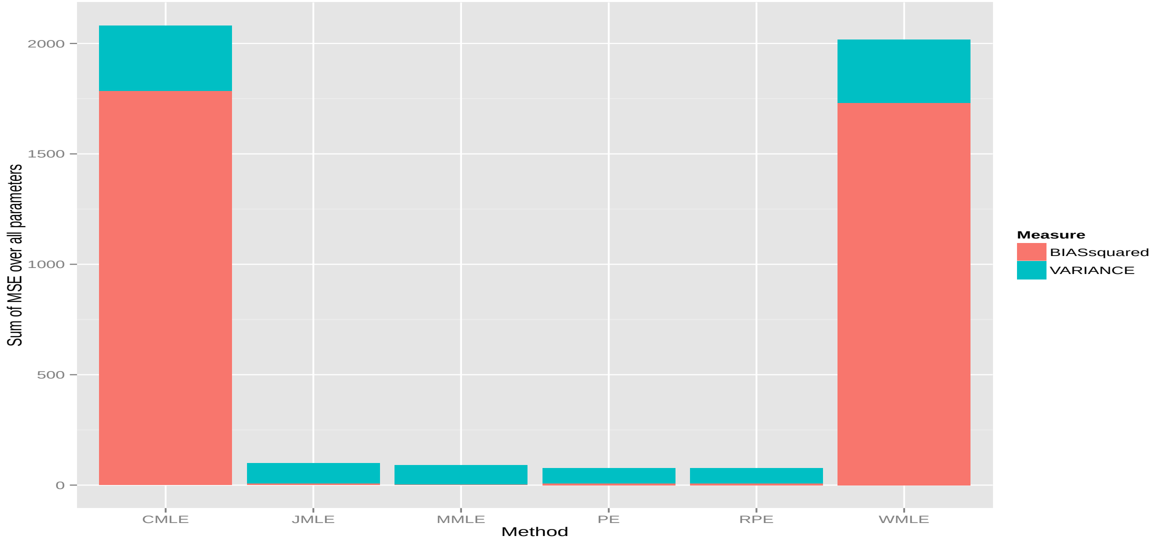

Figure 9: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 10

3.3.3. Scenario 11 - 15: Group of low performers

In this scenario only people from a group of low performers are shown $\theta \sim N(-3, 1)$. The global and pairwise tests were in the majority of comparisons significant and the results are **MMLE** $<$ **JMLE** $<$ **PE** $<$ **RPE** $<$ **WMLE** $\sim$ **CMLE**.

With a smaller number of items the tests between CMLE and WMLE showed an undetermined behaviour and could not clearly distinguish between those methods. With item counts above 40, there is a tendency to a higher rank of **WMLE** $<$ **CMLE**. Besides CMLE, WMLE rank changes the topological order remained constant across all scenarios 11-15. In contrast to scenarios 6-10 the variability across the Monte Carlo samples of the RMSE in JMLE, MMLE (figure 10) is a lot lower for low performers than for high performers. The MMLE is almost unbiased in scenario 11 with 10 items (figure 11). With an increasing number of items the relative performance of CMLE, WMLE is decreasing compared with other methods. With 50 items there are still some data sets, where the Pairwise Methods have outliers (figure 12). As shown in figure 13 the sum of MSE decreased with 50 items in comparison to 10 items and the methods JMLE, MMLE, PE and RPE are more similiar in peformance than with 10 items.

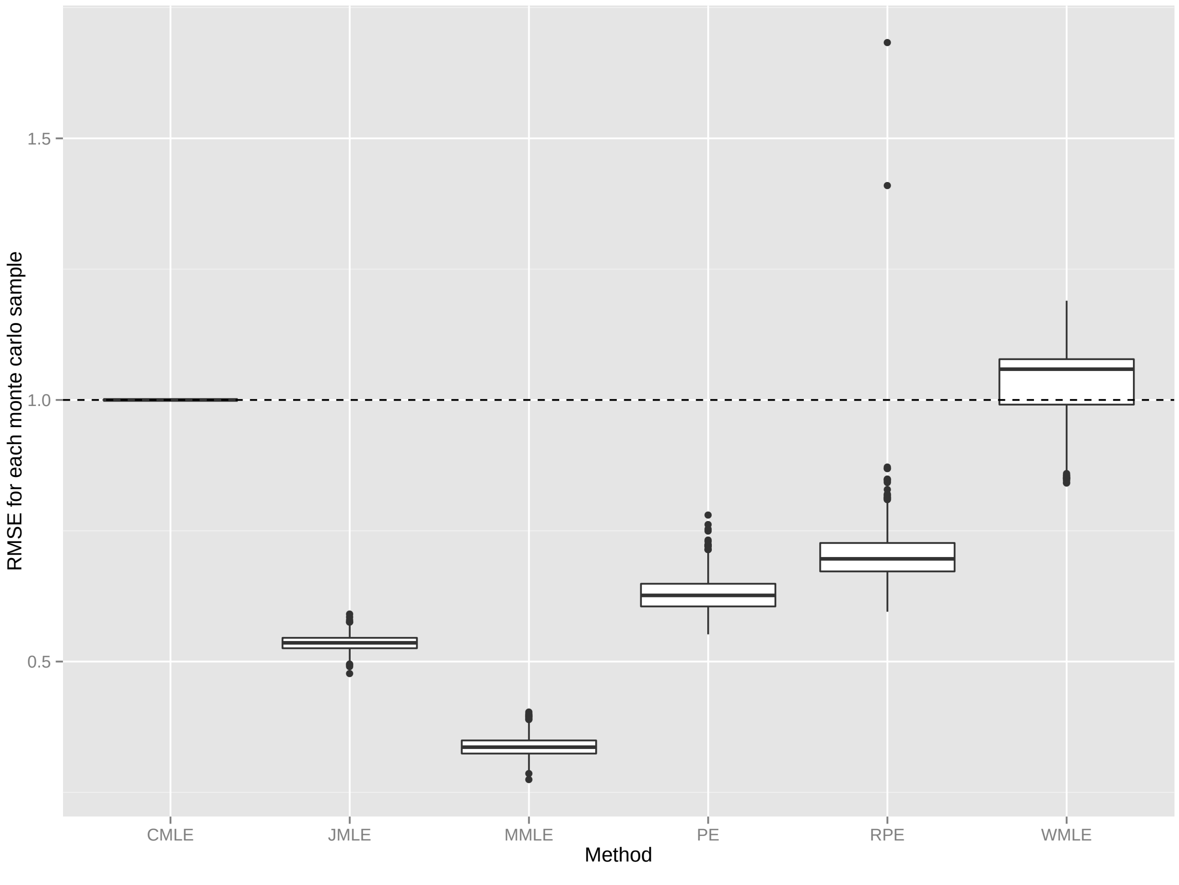

Figure 10: PCM: RMSE calculated for each Monte Carlo sample in scenario 11

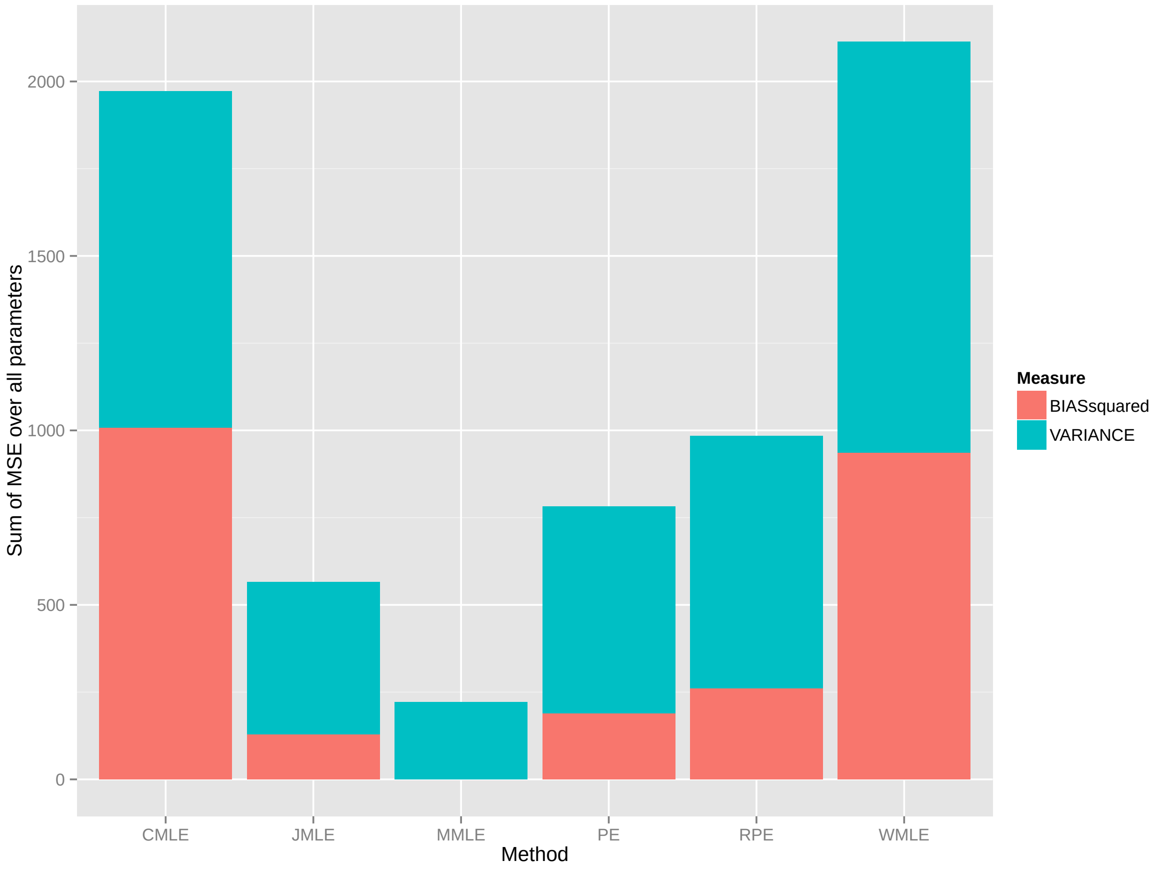

Figure 11: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 11

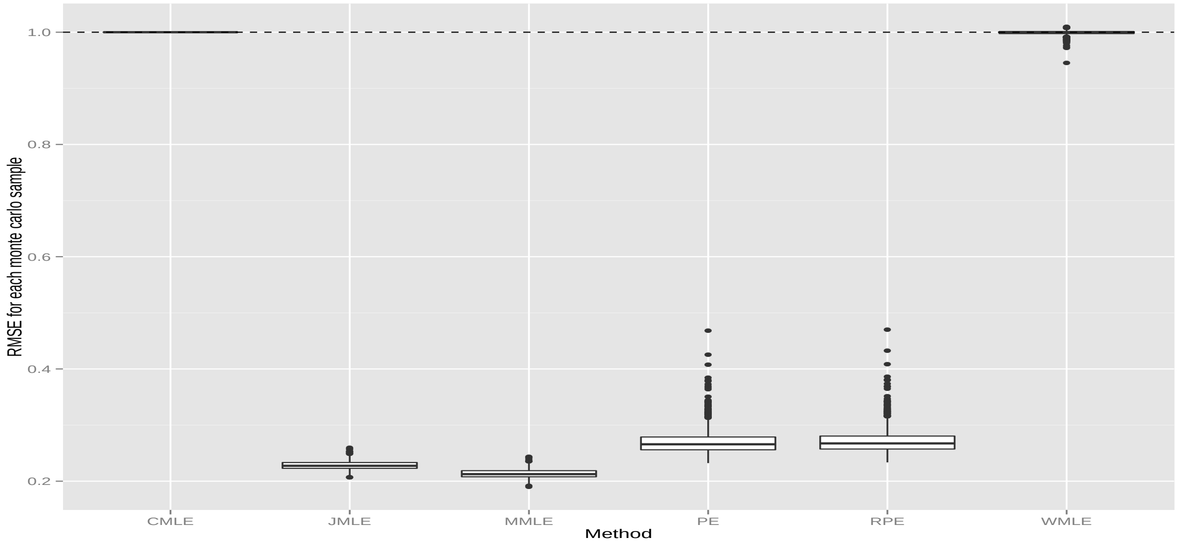

Figure 12: PCM: RMSE calculated for each Monte Carlo sample in scenario 15

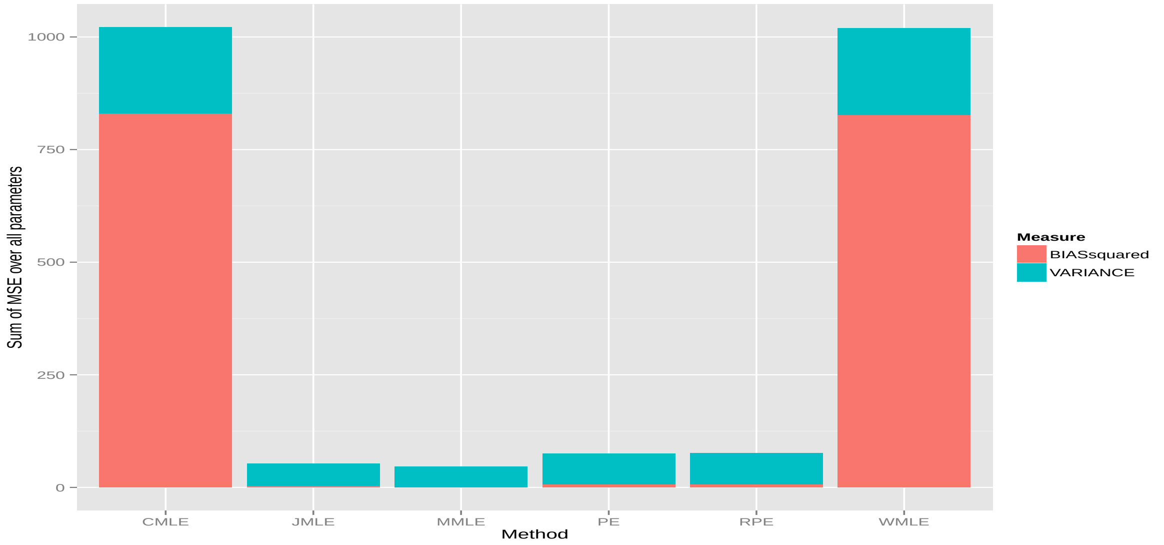

Figure 13: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 15

3.3.4. Scenario 16 - 20: Group of heterogenous abilities

In the scenarios 16-20 a normal distribution with an expectation of zero and large variance of 9 was choosen. From all scenarios in the PCM simulation, the group of heterogenous persons was the most difficult to estimate, because the RMSE was on average higher than in the other scenarios. The topological order for the scenario 16 is **JMLE $<$ PE $<$ MMLE $<$ RPE $<$ WMLE $<$ CMLE**.

The difference to the last presented scenarios is, that the JMLE is the best estimator for groups of persons with different skills. For larger number of items - greater than ten - the RPE achieves on average a better performance for identifying the true parameters than the MMLE.

Within the scenarios 17-20 the topological order stays the same **JMLE $<$ PE $<$ RPE $<$ MMLE $<$ WMLE $<$ CMLE**. The MMLE has a larger bias compared to JMLE and the Pairwise Estimators. As shown in figures 16 and 17 and other scenarios - an increasing count of available items - improves the person parameter estimates by reducing the variance.

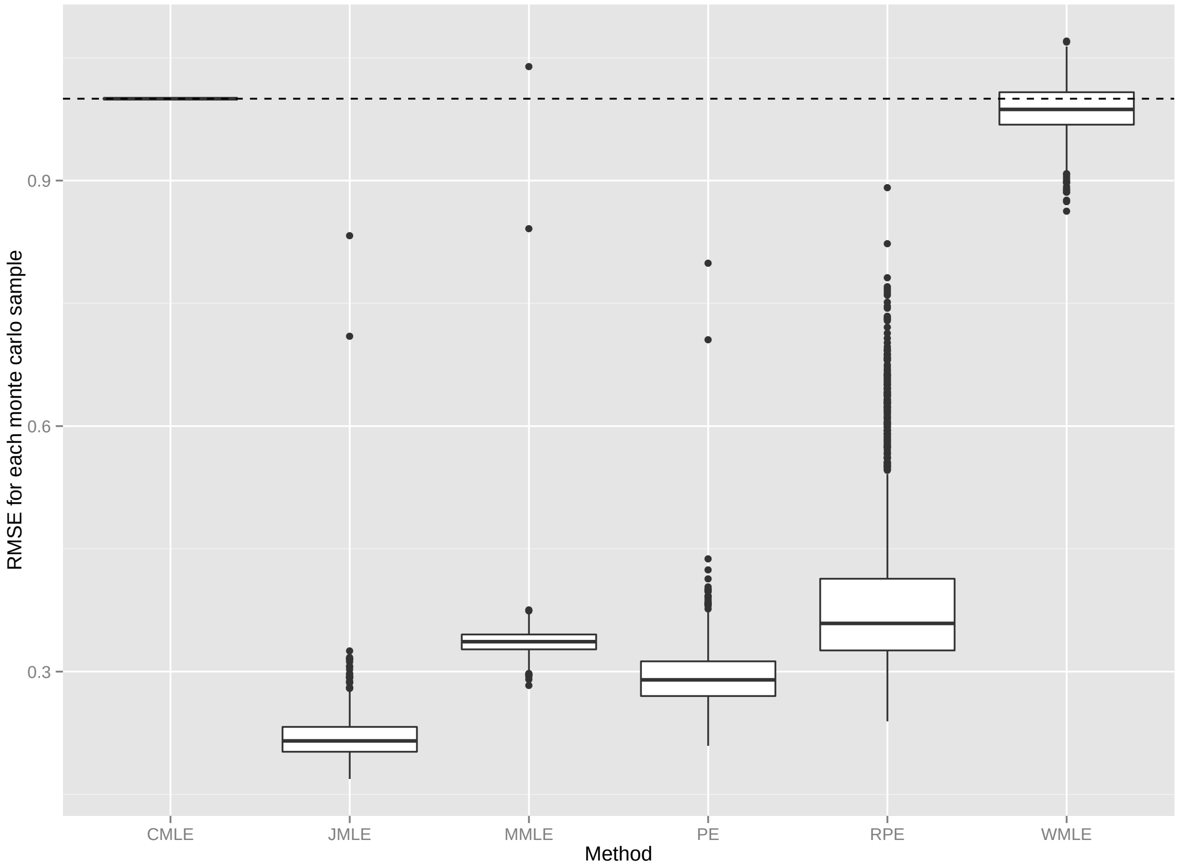

Figure 14: PCM: RMSE calculated for each Monte Carlo sample in scenario 16

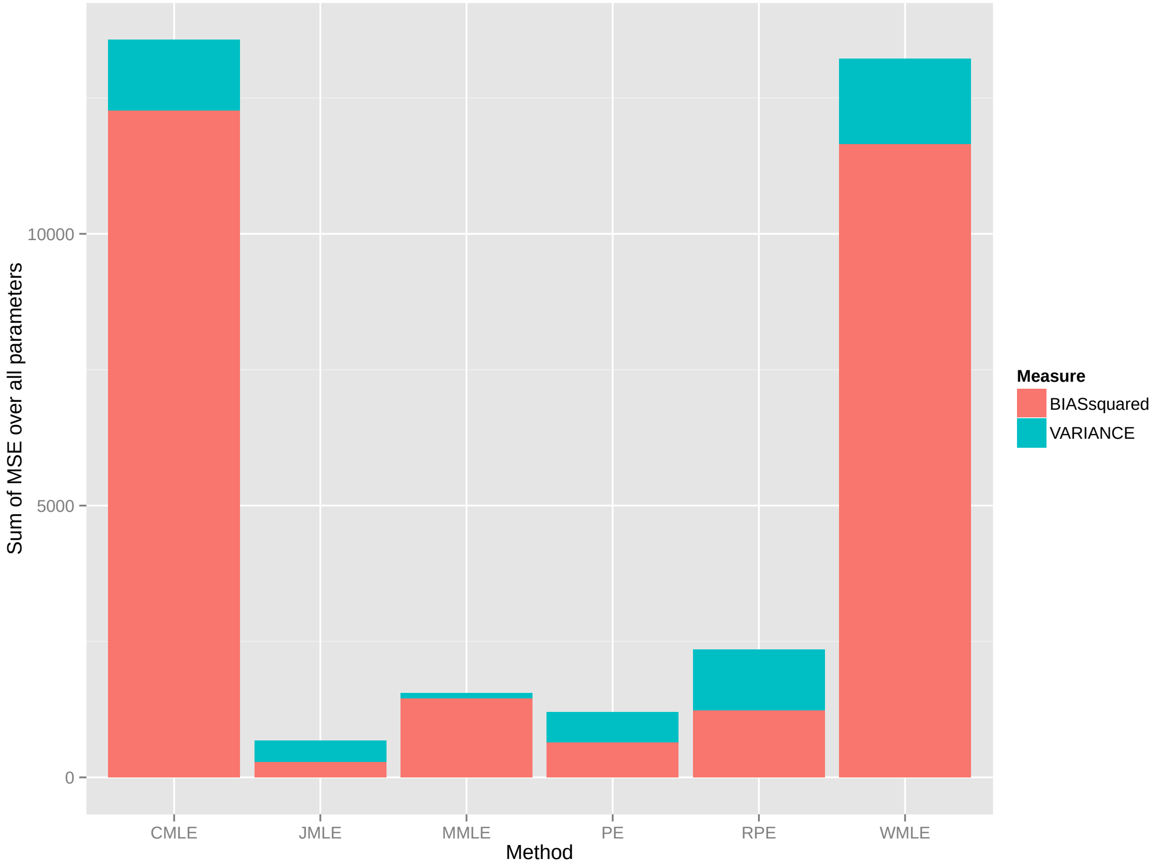

Figure 15: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 16

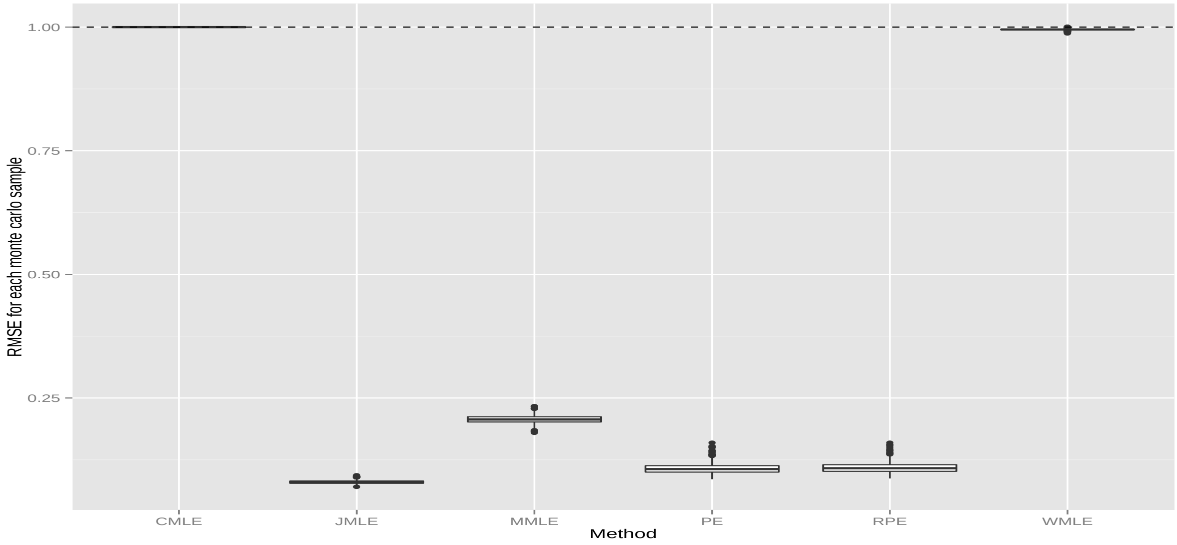

Figure 16: PCM: RMSE calculated for each Monte Carlo sample in scenario 20

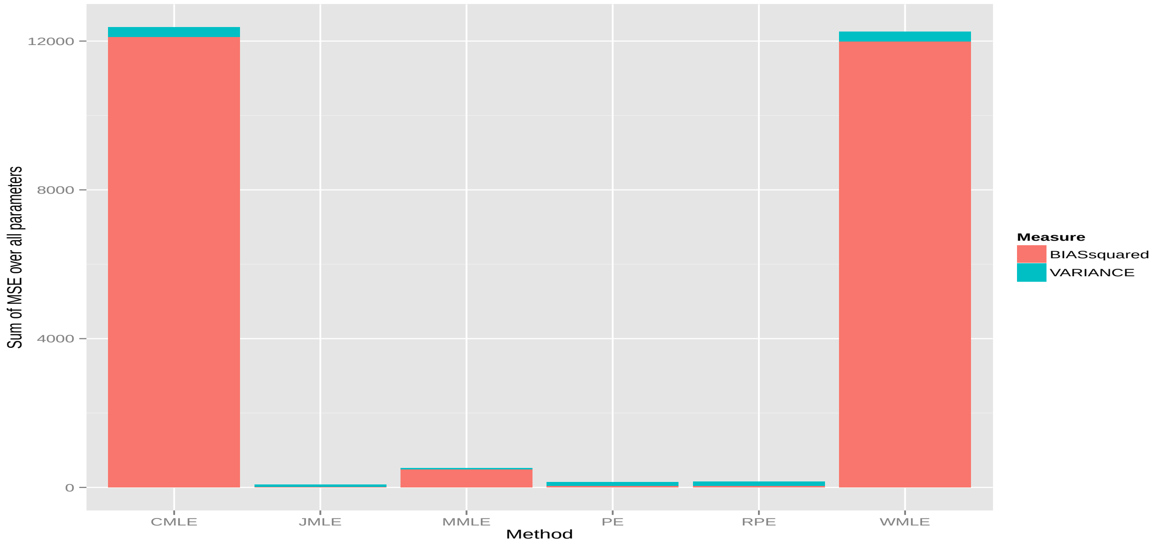

Figure 17: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 20

3.3.5. Scenario 21-25: Bivariate mixture of normal distributions

In the scenarios 21-25 a mixture of two normal distributions, subgroups with above average and below average performance expectations, were simulated. In general all estimators performed better than in the scenarios 16-20, but worser than in the scenarios 1-5. The performance of MMLE depends on the number of items. With ten items it performed best, with 20 items the JMLE and MMLE changed the order to second and first place and with 30 items or more the Pairwise Methods showed better performance than the MMLE. With 30 items or more the topological order is **JMLE** $<$ **PE** $<$ **RPE** $<$ **MMLE** $<$ **WMLE** $<$ **CMLE**.

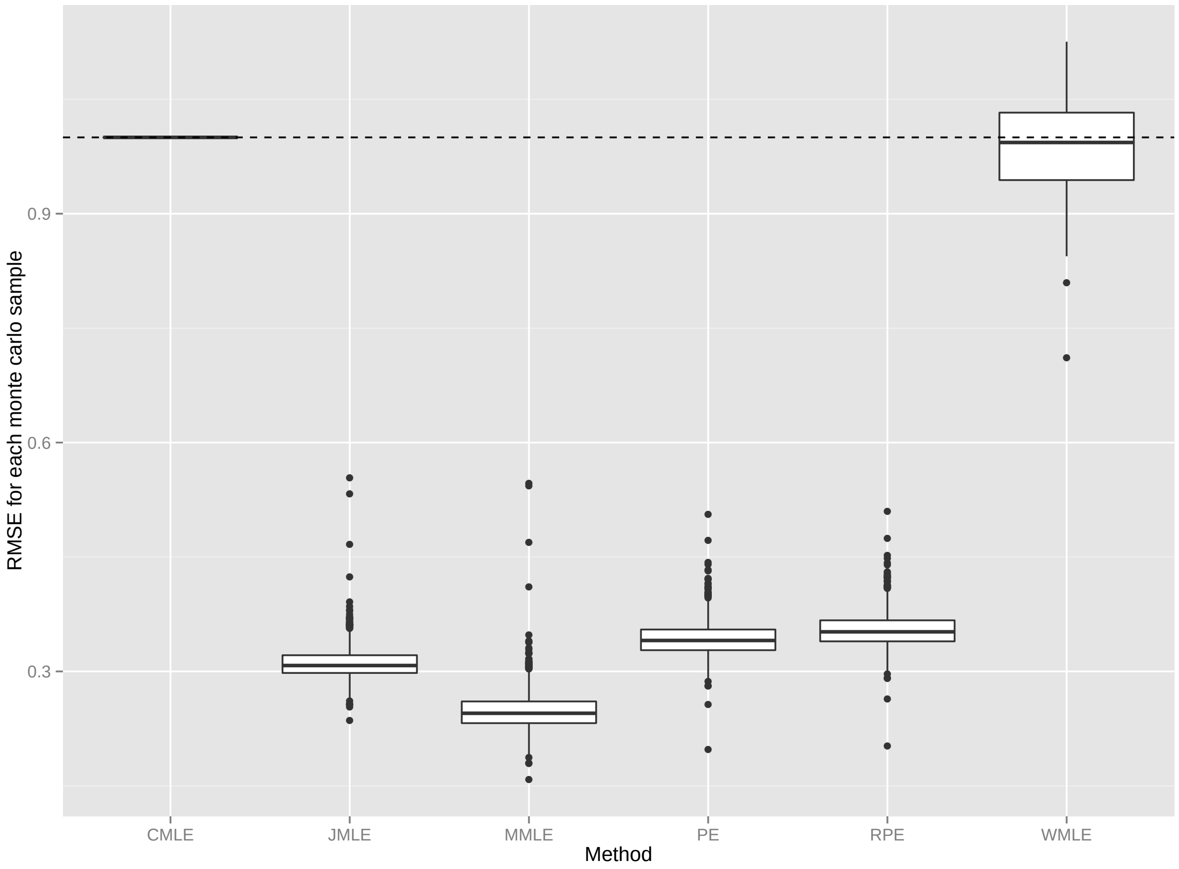

Figure 18: PCM: RMSE calculated for each Monte Carlo sample in scenario 21

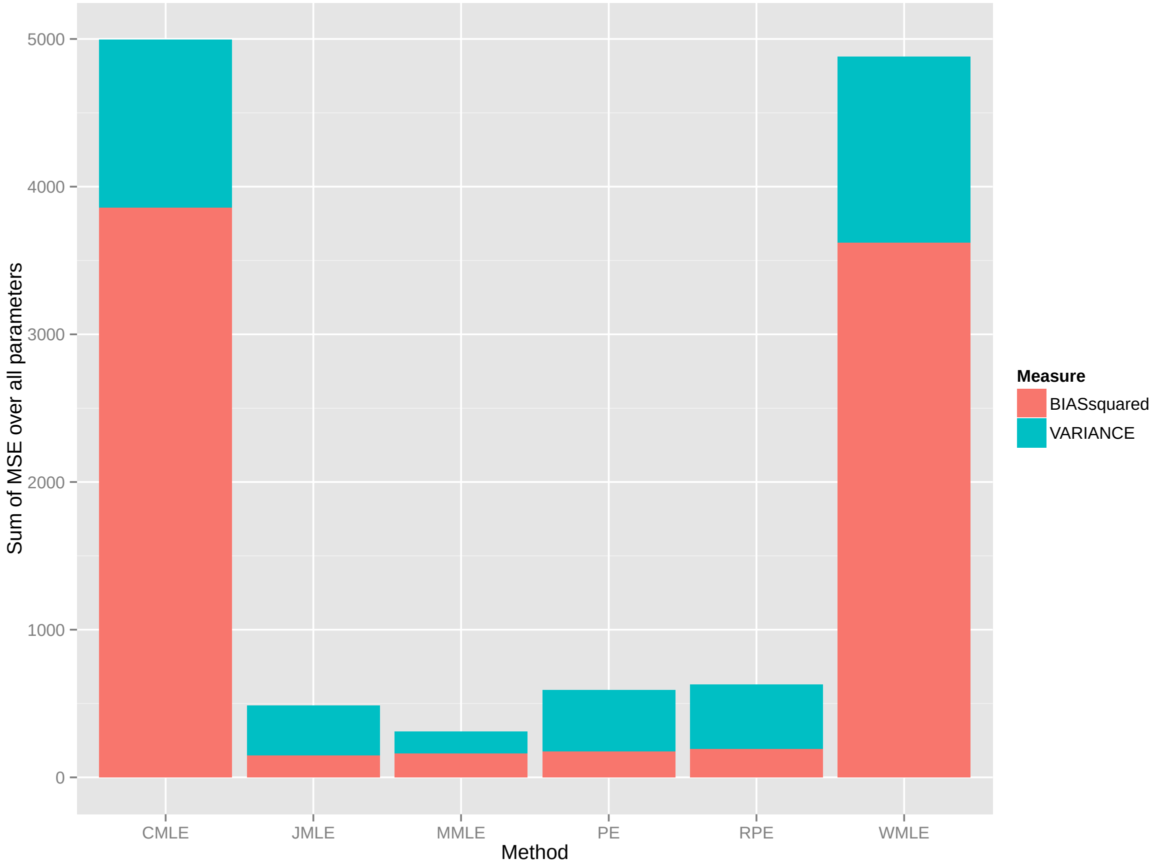

Figure 19: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 21

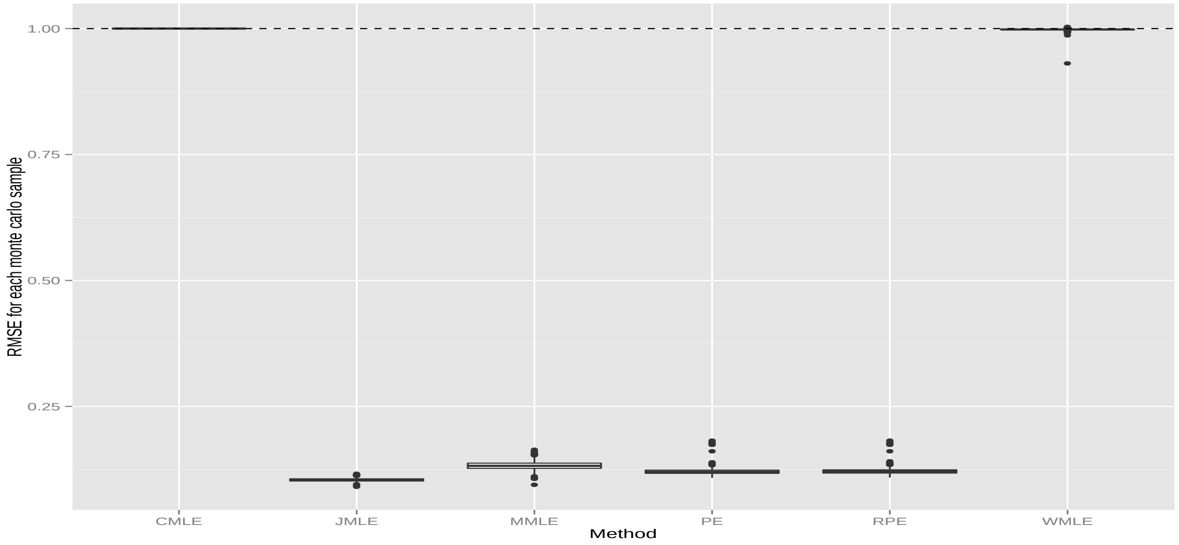

Figure 20: PCM: RMSE calculated for each Monte Carlo sample in scenario 25

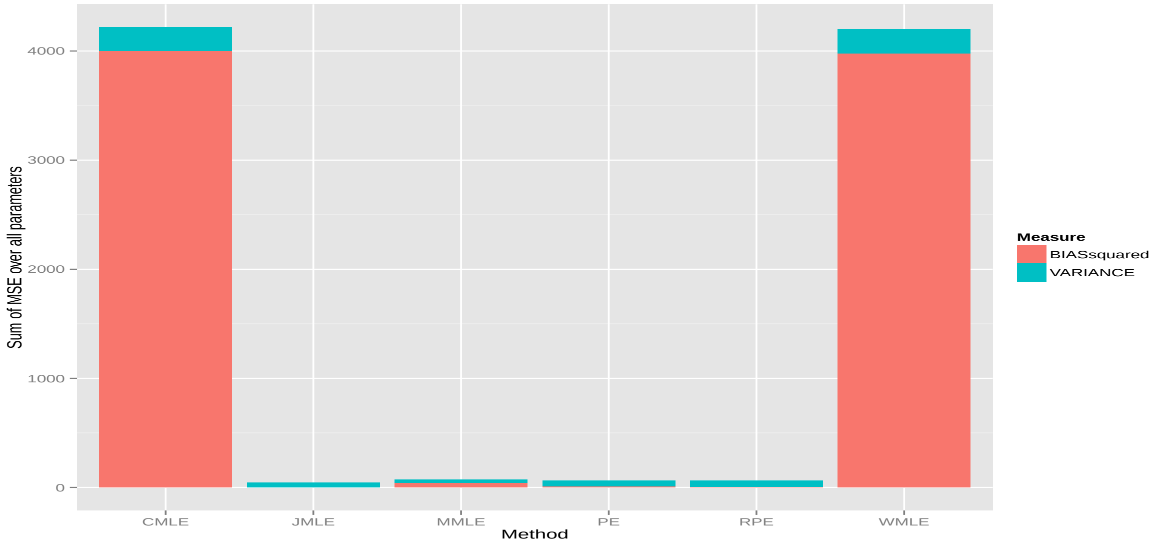

Figure 21: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 25

3.4. Summary of PCM simulation results

The overall result of the simulation study of the PCM, suggests some general advices for practical data analysis, if the person parameters are of special interest in research: In the majority of cases the best estimators for the person parameters were MMLE and JMLE. The JMLE showed good results in finite samples despite the theoretically inconsistent item parameter estimates. If there are indications that the prior distribution apparts from a normal distribution, then it is more adviseable to use JMLE or specify an appropiate prior distribution for the MMLE. The MMLE is embedded in the framework of generalized linear mixed models. This models are theoretically well understood and there are efficient algorithms available for computation. Standard errors, model diagnosis or statistical tests can be made with less efforts than for the other discussed estimation methods. An increasing number of items does decrease the variance of the estimators, but the ranking between the methods remains unchanged in most circumstances.

It is not recommended to use CMLE or WMLE approaches. Across all scenarios these estimation methods performed much poorer than all other methods. A first explaination is, that the theoretical good properties of the CMLE with consistent item parameter estimates must not hold for finite samples. Second the estimation of person parameters is only optimised once, given the item parameters. This is computationally more efficient, but it is less adaptive then the JMLE approach, which uses information of the estimates reciprocal. In both cases CMLE, WMLE the external package *eRm* was used for the conditional maximum likelihood estimation. It should be investigated if the results can be reproduced with a self made function. Despite computational issues, it would be interesting to find out, wether the CMLE would show better performance, if the person parameters are estimated conditional on the column sums of the items. In most cases (except scenarios 1-5) the WMLE approach improved the accuracy of CMLE. If conditional estimation is used, weighting with the Fisher information is meaningful. The question why these methods have a relatively high bias, remains open for further research.

With regard to the Pairwise Methods, PE was always better than RPE. The restriction to exclude cases with equal total score in pairwise comparisons is not justified for the polytomous Rasch models. In theory, by leaving out persons with equal scores, the goal in dichotomous models is to reduce the variance of the estimator. On the other hand this means, that the RPE uses less information than the PE, but less information leads to an increase in variance. The information loss outweights the reduction in variance in the polytomous case. The PE was best in five scenarios out of 25, but is computationally more demanding than the other approaches. This is because of the combinatorical quadratically growth of pairwise comparisons: There are $\frac{n(n-1)}{2}$ possible different pairings and the highest order of the input n is quadratic.

4. Simulation of SM

This section reports the results of the simulation study performed for the sequential model. The three estimation methods CMLE, JMLE and MMLE are compared with regard to their deviations from the true values. The structure of performance evaluation is the same as in chapter 3.1 but three scenarios are considered.

4.1. Design of SM simulation

For the sequential model one scenario applied with 250 persons, 10 items and three score categories $\in \{0, 1, 2\}$ for each item. In the simulation of the PCM the differences between varying counts of items were minor; therefore it is sufficient to investigate one fixed amount of items. The item parameters are generated equidistant within the closed range $[-3, 3]$. The person parameters follow the distributions:

1. $\theta \sim N(0, 1)$
2. $\theta \sim N(2, 1)$
3. $\theta \sim N(0, 4)$

For this specification 1000 Monte Carlo samples are generated. Beside the mentioned differences, the simulation procedure is the same as outlined in section 3.1.

4.2. Results of SM simulation

4.2.1. Scenario 1: Standard normal distributed abilities

In this scenario the person parameters are standard normal distributed. First the RMSE is evaluated over all items for each Monte Carlo sample. The figure 22 shows the following results for all methods.

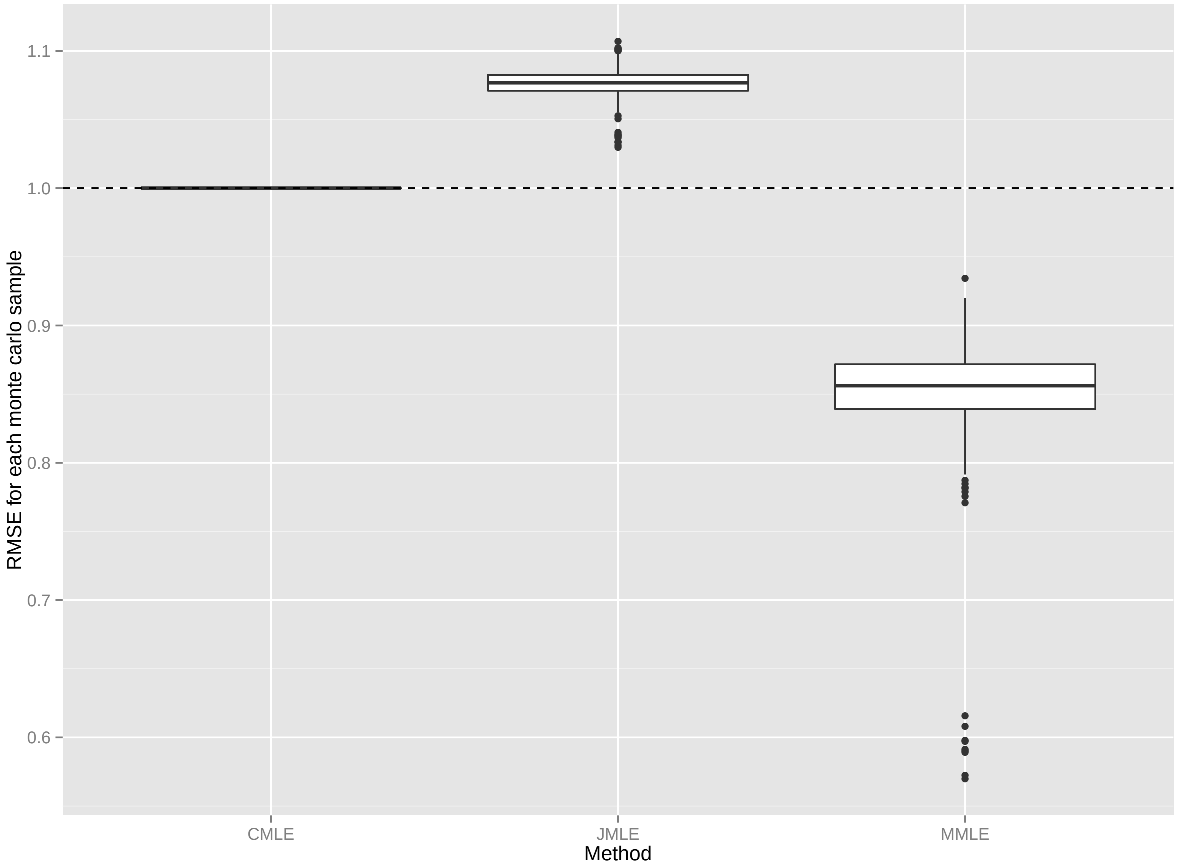

Figure 22: SM scenario 1: RMSE calculated for each Monte Carlo sample

The RMSE is computed relatively to the basis of the RMSE of the CMLE method. So the dashed line at 1 indicates all RMSE values of the CMLE method. Each value above the line is worse and each value below is better, compared to CMLE. It is obvious that the median of the JMLE in this scenario is less good than the CMLE. In contrast the median of the MMLE method is by far better than the CMLE. The variances of the relative JMLE deviations are smaller than those of the MMLE. There are a lot of RMSE outliers in the MMLE. The analysis results are supported by further formal tests. The p-value of the approximative Friedman test is below 10^{-15} and therefore significant. The pairwise comparison of all methods results in the following topological order, statistical significance is given: **MMLE** $<$ **CMLE** $<$ **JMLE**. The approximative Wilcoxon tests p-values are all below 10^{-15}.

In the next step the cause of the deviations from the true person parameters are investigated. Therefore the MSE is computed for each parameter across all Monte Carlo samples and split in squared Bias and Variance of the estimator. The MSE, squared Biases and Variances then are summed over all person parameters:

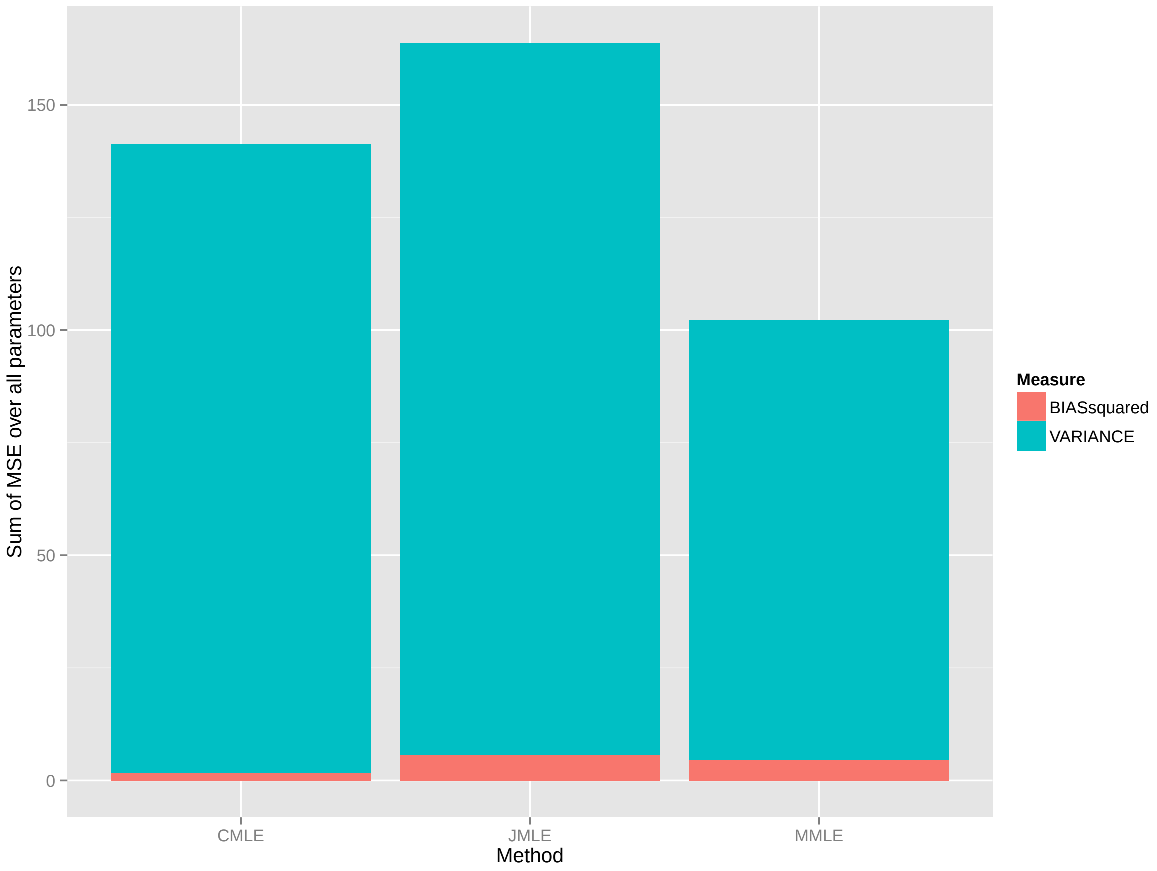

Figure 23: SM scenario 1: Sum of MSE, squared Biases, Variances over all person parameters

The result is shown in figure 23. The squared Biases of all three methods are negligible, because the variance of the estimators is dominating. The MMLE method has the smallest variance and therefore also the smallest Sum of MSE. The JMLE has a slightly higher bias than MMLE and CMLE has the least bias of all. Of course, it is to mention, that in this scenario the person parameters were simulated with a standard normal distribution, so the MMLE has an advantage because it uses almost the same underlying distribution in the fitting process. Otherwise the performance of the MMLE could be more biased, if a symmetrical normal distribution is not applicable.

Regarding computation time, the JMLE was fastest, followed by the CMLE and MMLE is quite slow. This is, because the EM algorithm switches between an expectation and maximisation step, which may be slow to convergence in practise.

4.2.2. Scenario 2: Group of high performers

In this scenario a group with person abilities above average was simulated. In figure 24 the relative deviations from the CMLE RMSE are smaller than in the scenario with the standard normal distribution. But the differences are significant between the distributions. The topological order is the same as in scenario 1.

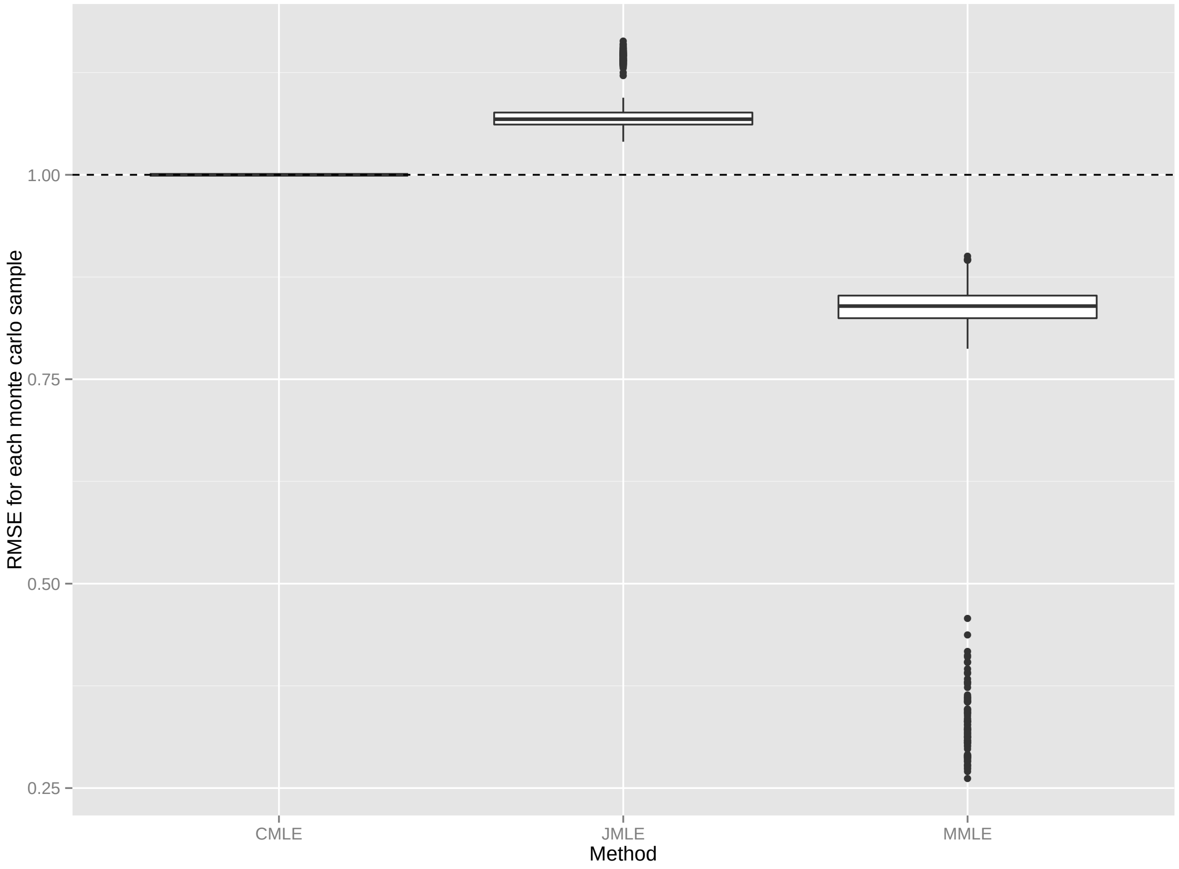

Figure 24: SM scenario 2: RMSE calculated for each Monte Carlo sample

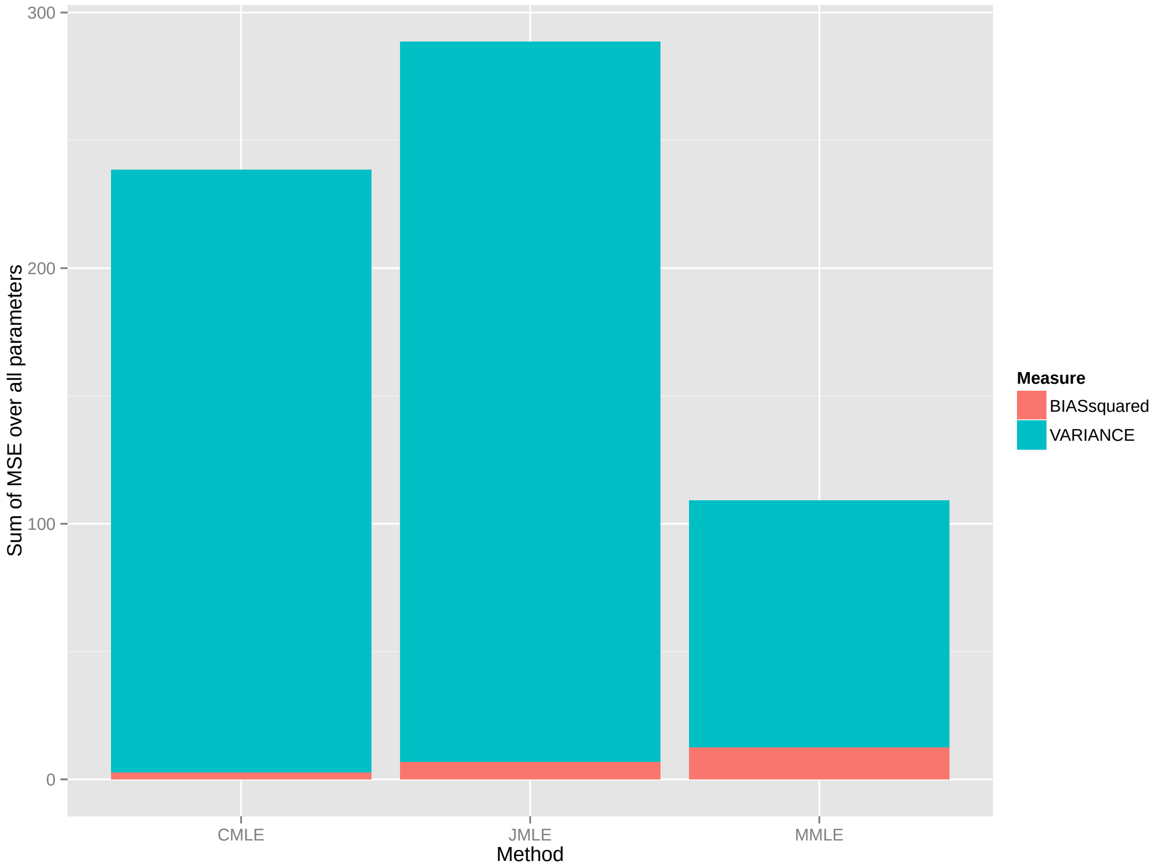

Figure 25: SM scenario 2: Sum of MSE, squared Biases, Variances over all person parameters

4.2.3. Scenario 3: Group of heterogenous abilities

In scenario 3 the results with heterogenous people skills are comparable to scenario 1. The differences are higher standard deviations of the estimates and the MMLE peformance is more close to the CMLE. Again the topological order has not changed in comparison to scenario 1. For practical analysis it is recommended to use MMLE for estimation of the SM.

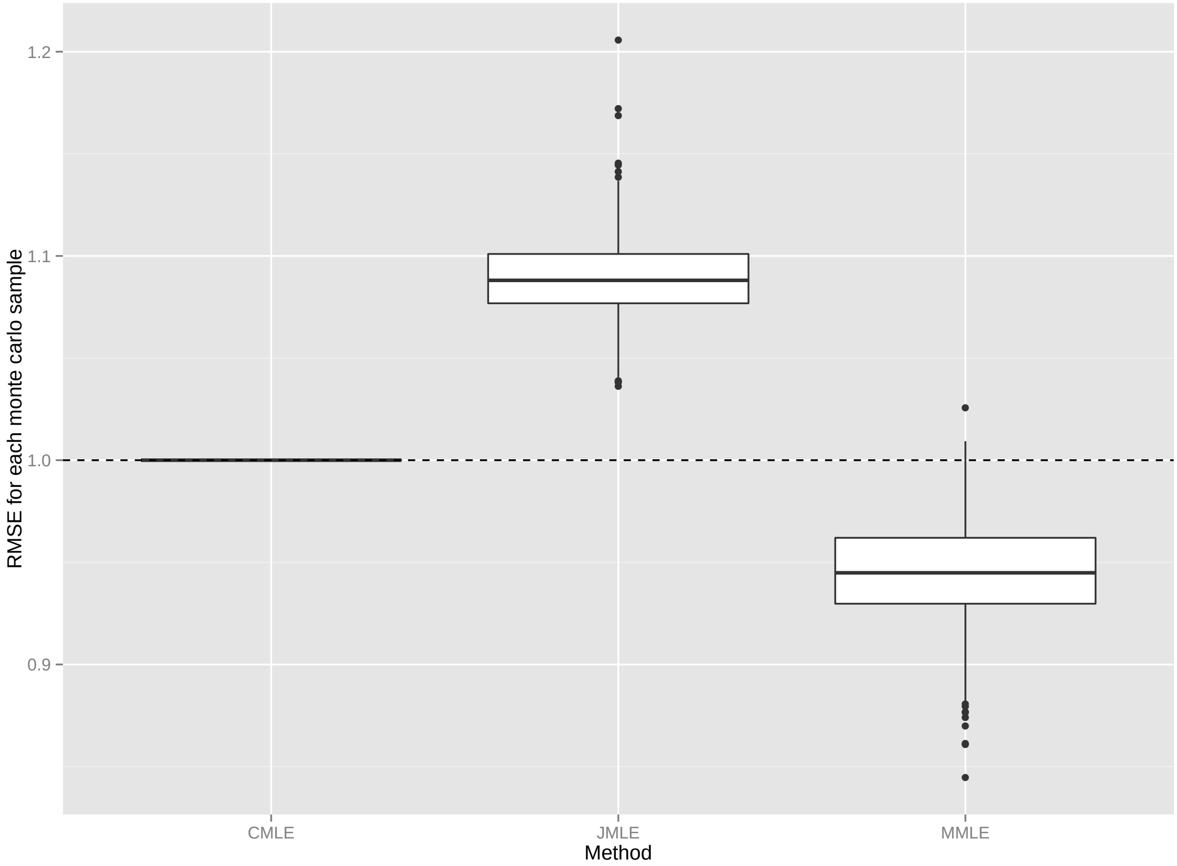

Figure 26: SM scenario 32: RMSE calculated for each Monte Carlo sample

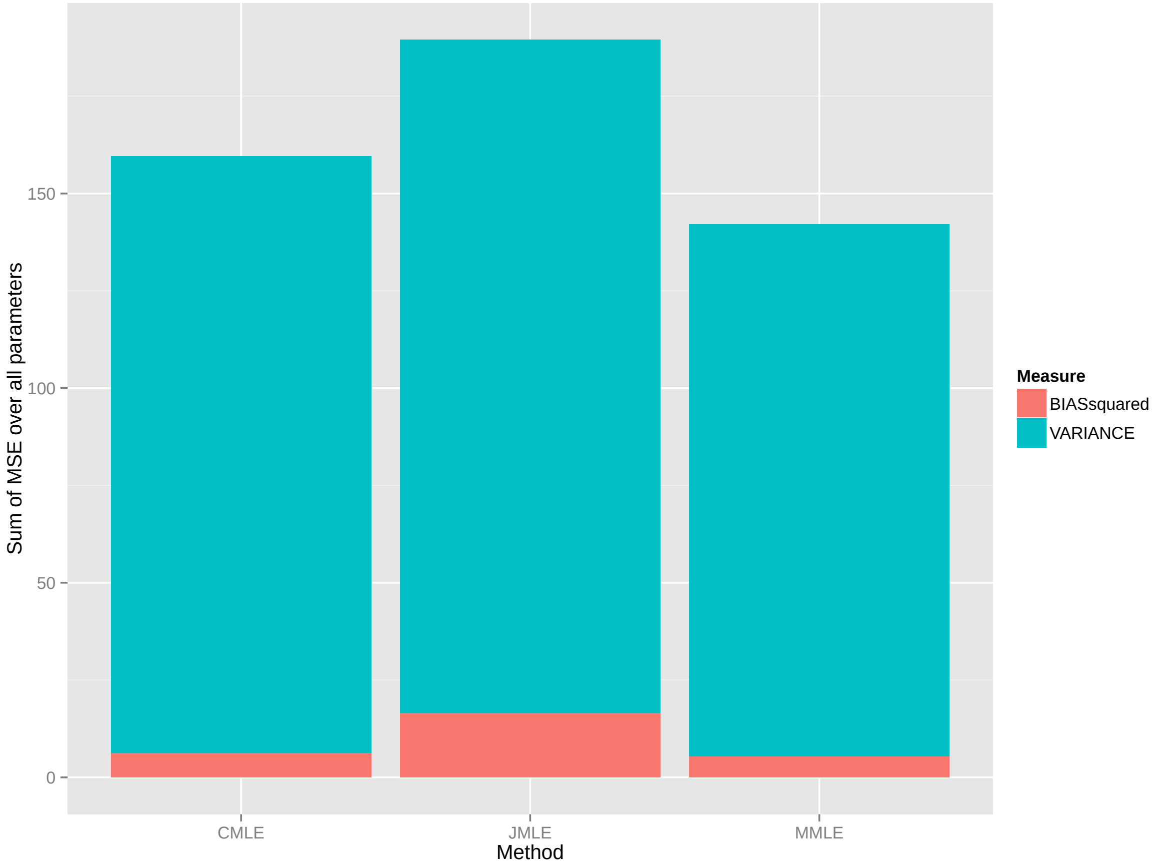

Figure 27: SM scenario 3: Sum of MSE, squared Biases, Variances over all person parameters

5. Simulation of predictive performance measures

Models which have a low average probability of misclassifcation across all persons, items and different data sets, should predict the performance of individuals better. Before applying the proposed measure the argument should be approved. Therefore a simulation is conducted how well $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ can differentiate between one true model (SM) and one false model (PCM). The expected prediction error of the true model should be smaller than the one of the wrong model. In simulations the true model is known, therefore it is not necessary to perform cross validation. Instead of this the performance will be evaluated on independent test samples generated from the true model. A good classifier for discrete outcomes should predict new observations better then random guessing. If the number of possible outcomes for item j is $k_j + 1$, then the number of misclassified observations is on average $1 - \frac{1}{k_j+1}$ by guessing randomly.

5.1. Design of predictive performance simulation

Two scenarios are considered: In the first one the PCM is the true model and in the second scenario the SM is the true model. In both scenarios the generated training and test data sets have both 250 observations. All data sets have 10 items with three categories $\{0, 1, 2\}$. The item parameters are constructed equally in the interval $[-3, 3]$ for all 20 difficulty steps. The person parameters are standard normal distributed. The training and test data use the same item parameters but with different, independently drawn person parameters. Each training and test data are replicated 1000 times using Monte Carlo sampling from the underlying true model. Then the following procedure is applied for every Monte Carlo sample of training and test data:

1. Estimate item parameters with the specified model on the training data
2. Given these item parameters, estimate the person parameters of the suitable test data
3. Evaluate predictive performance, given person parameters of the test data and item parameters from the training data, on the test data with $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ or 0-1 loss
4. Average the loss function with the arithmetic mean over all persons, items of the test sample

With these steps the expected misclassification probability for predictions of new persons is evaluated. Altogether 8000 models have been fitted in this section. After simulation the density of the mean misclassification probabilities are estimated, using bandwidth selection by pilot estimation of derivatives (Sheather and Jones 1991), with a gaussian kernel.

5.2. Results of predictive performance simulation

In figure 28 the underlying true model was the PCM. Both curves are similiar in shape, except that the PCM density has far more outliers than the SM. The densities seem to be approximately normal distributed, following the central limit theorem. Therefore, it is reasonable to use asymptotic Wald confidence intervals (CI). The grand mean misclassifcation probability is 33.96 % for the PCM and 33.47 % for the SM. The difference is

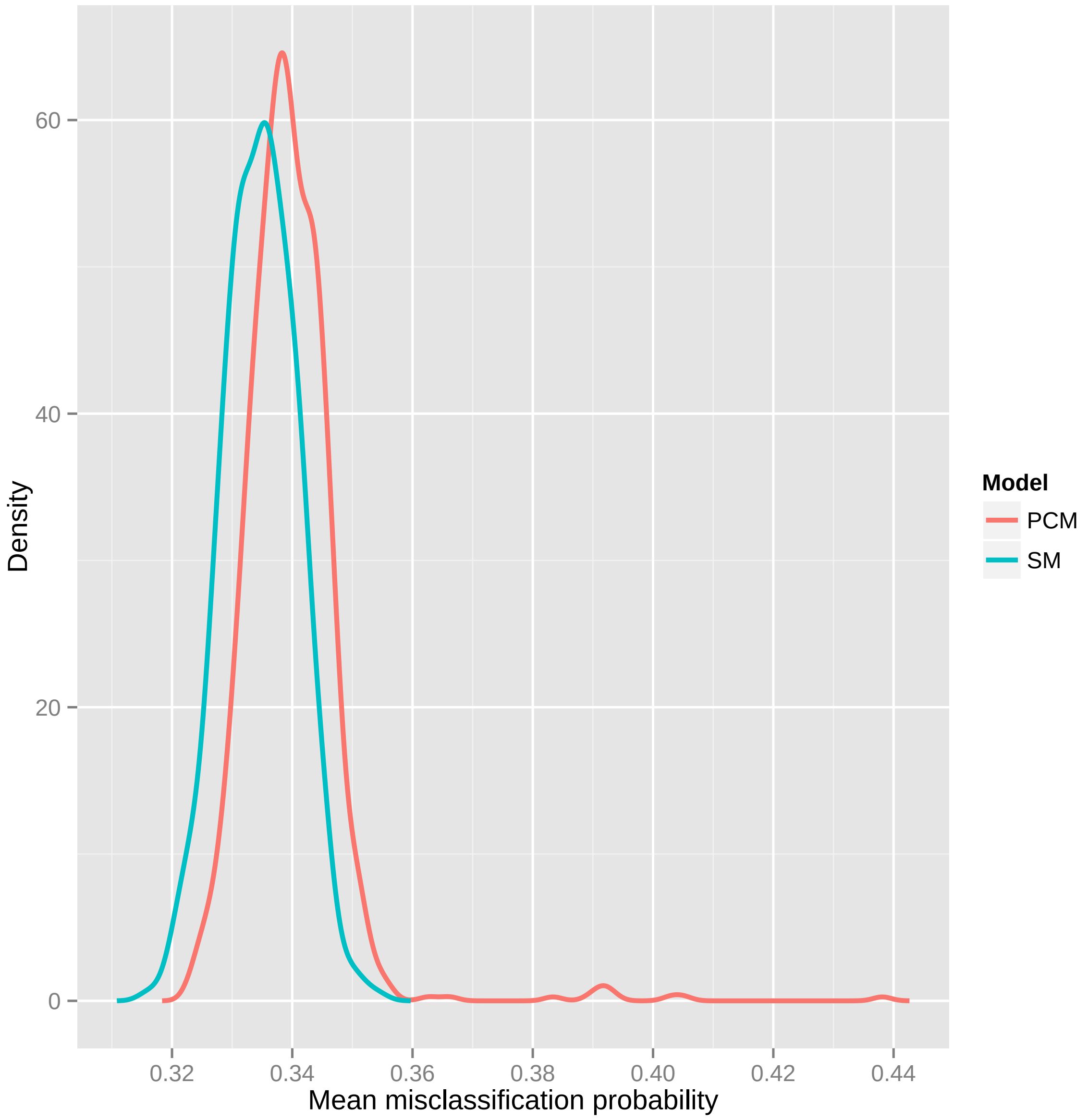

Figure 28: Density estimate of mean misclassification probability with true model PCM

significant according to a Monte Carlo Wilcoxon sign rank test, with a p-value $< 10^{-15}$. But the impact is minor because the difference is below 1 %. A 95 % confidence interval (CI) derived from the empirical quantiles for PCM is $[0.3277, 0.3515]$ and for the SM $[0.3225, 0.3457]$. Both CI's have a wide area of overlapping. This indicates that the performance measure could not discriminate well between the true model and the wrong model in this scenario.

Beside the mean misclassification probabilities, the density estimation of the mean predictive deviances are given in the appendix A.2. The results are similiar, but the outliers are stretched further appart, due to the nonlinear measure.

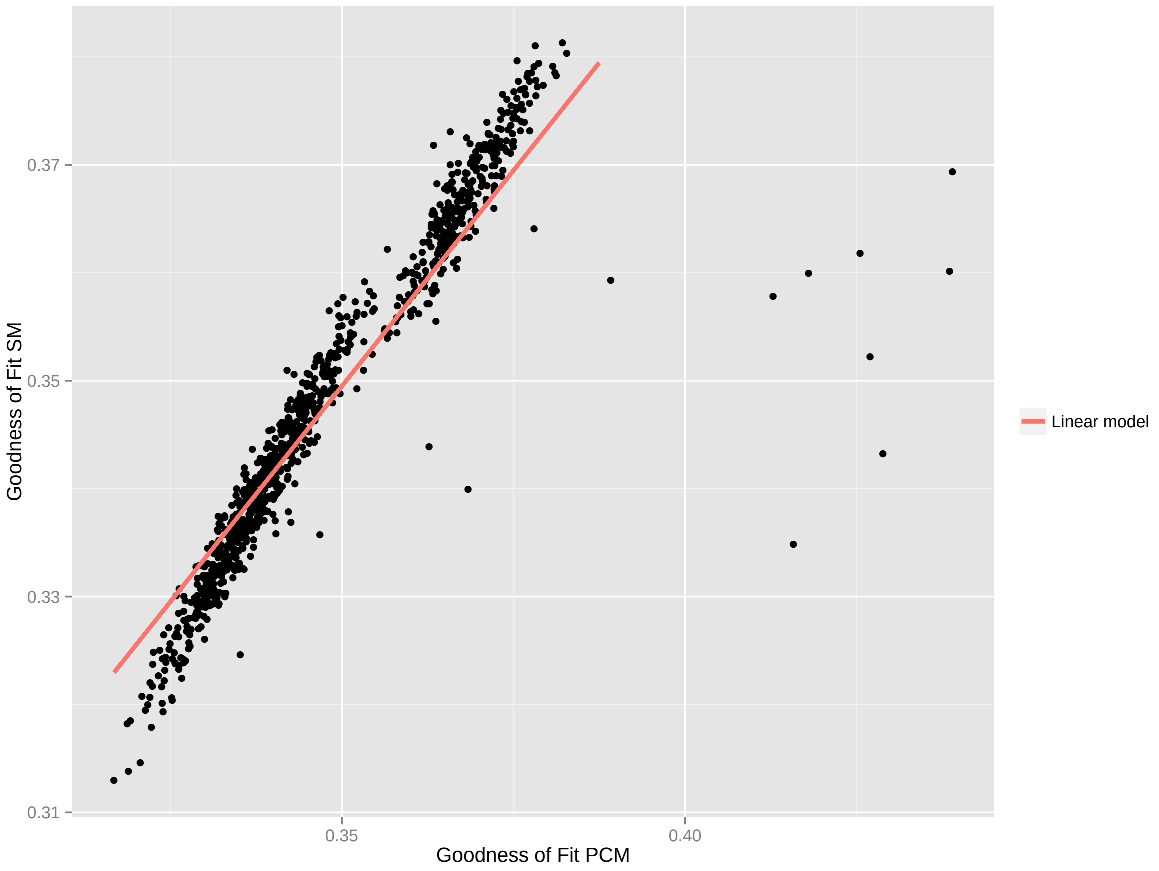

Figure 29: Pairwise comparison of goodness of fit from PCM (true model) and SM

Now the goodness of fit of both models is evaluated. Instead of the test data sets the original training data is used for evaluation of performance. In figure 29 the goodness of fits for all training data are displayed in pairwise comparison. The true model performances are on the x-axis. The majority of cases shows a linear trend (82.7 % of the variability is explained with a linear model). The slope coefficent 0.8 is significantly different from 1 by conducting a t-test with significance level 0.05. It means, that the goodness of fits are different for the PCM and SM. The SM could better adapt to the data even if the PCM is the true model.
In the results of figure 30 all design parameters were kept equal, except that the underlying true model was the SM. The situation is comparable to the last figure, except that the deviations of the grand mean between the PCM and SM are greater (0.3575 vs. 0.3664). The CI for the PCM is $[0.3451, 0.3709]$ and for the SM $[0.3530, 0.3786]$. They are nearly

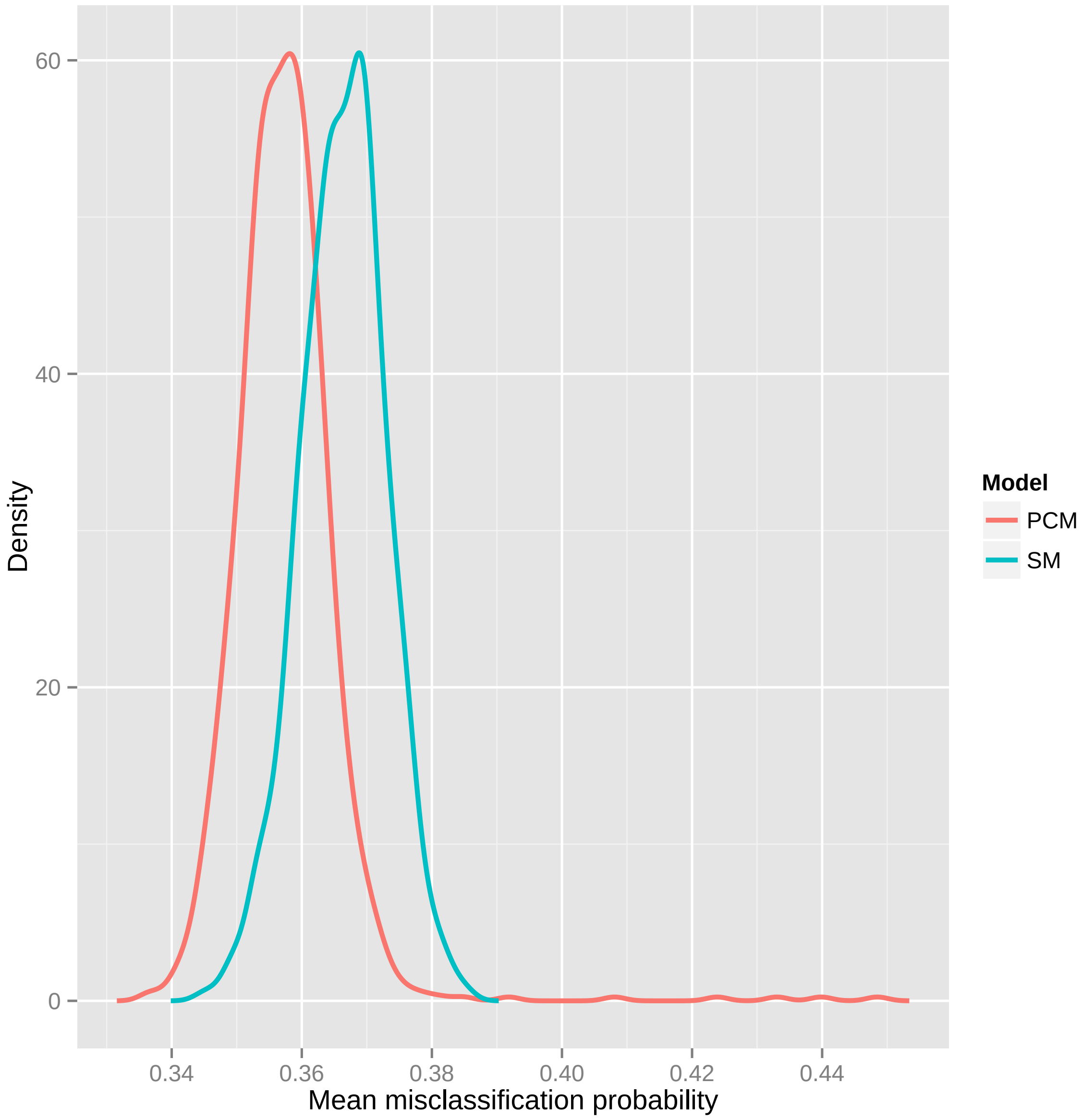

Figure 30: Density estimate of mean misclassification probability with true model SM

equal. Similiar results derive in the reverse situation, when the true model is the SM. Then the predictive measure could not discriminate the true model.

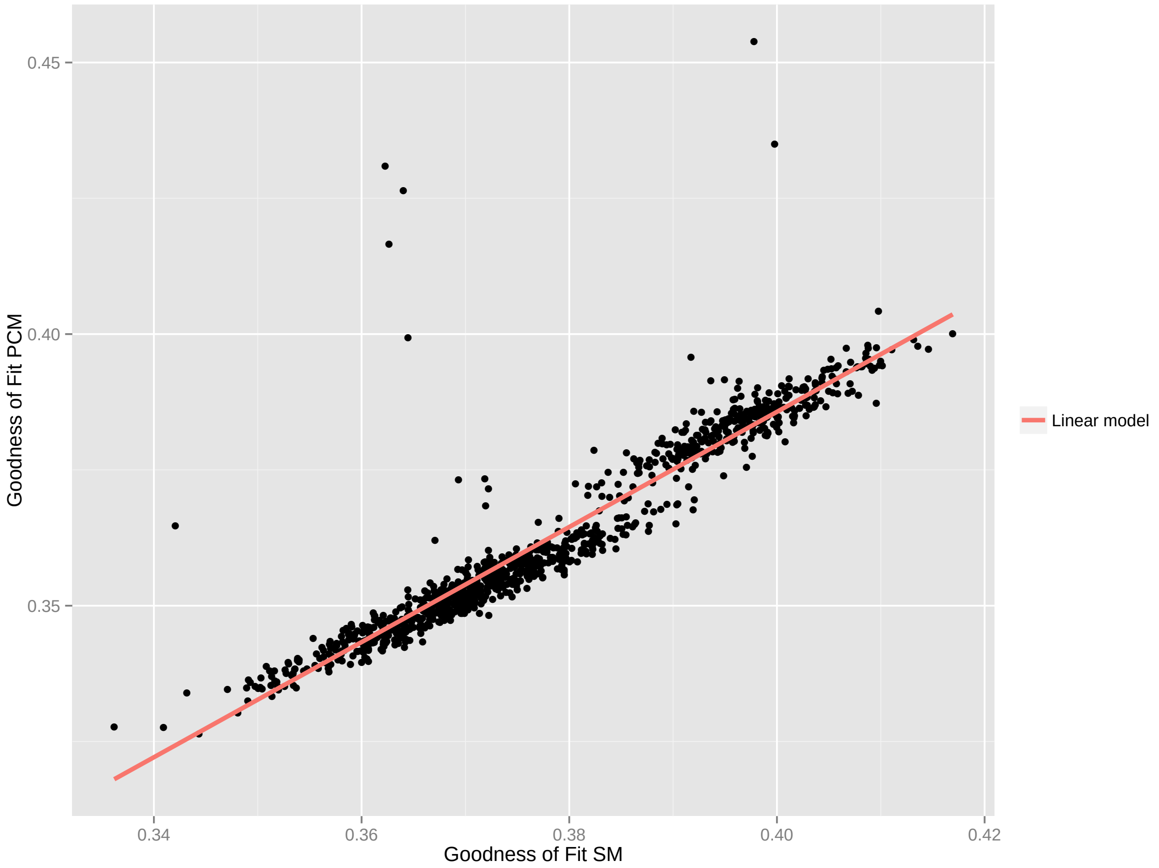

Figure 31: Pairwise comparison of goodness of fit from PCM and SM (true model)

In figure 31 the pairwise comparisons of goodness of fit are displayed in the case of SM beeing the true model. Compared to the reverse situation (with PCM as true model) the SM could better adapt to the training data. On average the PCM goodness of fits were 5 % higher then the SM goodness of fits. The coefficient is significantly different from 1 with a p-value of 4.6 %. But the difference is much smaller than in the reverse scenario.

Nevertheless any interpretation of these results should be done with care. Only one scenario with one estimation procedure was compared. Other prior distributions for the person parameters or other estimation methods could lead to different results. This outcome can be interpreted in multiple ways: Either the measure of predictive performance is not suited to distinguish the PCM and the SM in this scenario or in general the PCM, SM are both equally good models in terms of predictive accuracy. These effect is called *Rashomon* and is well known in machine learning (Breiman 2001; see page 206). The name originates from a japanese movie in which four people are telling the court what happend in an incident. Every person reported the same facts, but with quite different stories. In the case of predictive modelling the analogy is, that it is quite plausible that there exists a lot of models for a problem with given data, that yield similiar preditive performance, but assume different data generating processes.

For the comparison, a proper 0-1 loss function is computed and a design with SM as true model is investigated. By using the same simulated data, the performance of both models is largely different from the proposed loss function (figure 32). This time losses don't overlap at all. Clearly, the SM has the lower average predictive error and therefore the 0-1 loss could discriminate the true model between the SM and PCM. This simulation demonstrates that specifying a new loss function must be well thought out. The loss functions should be tested, wether it is a valid measure. Further it is recommended to compare models with regard to other criteria like context, study design and persued objectives of the research project.

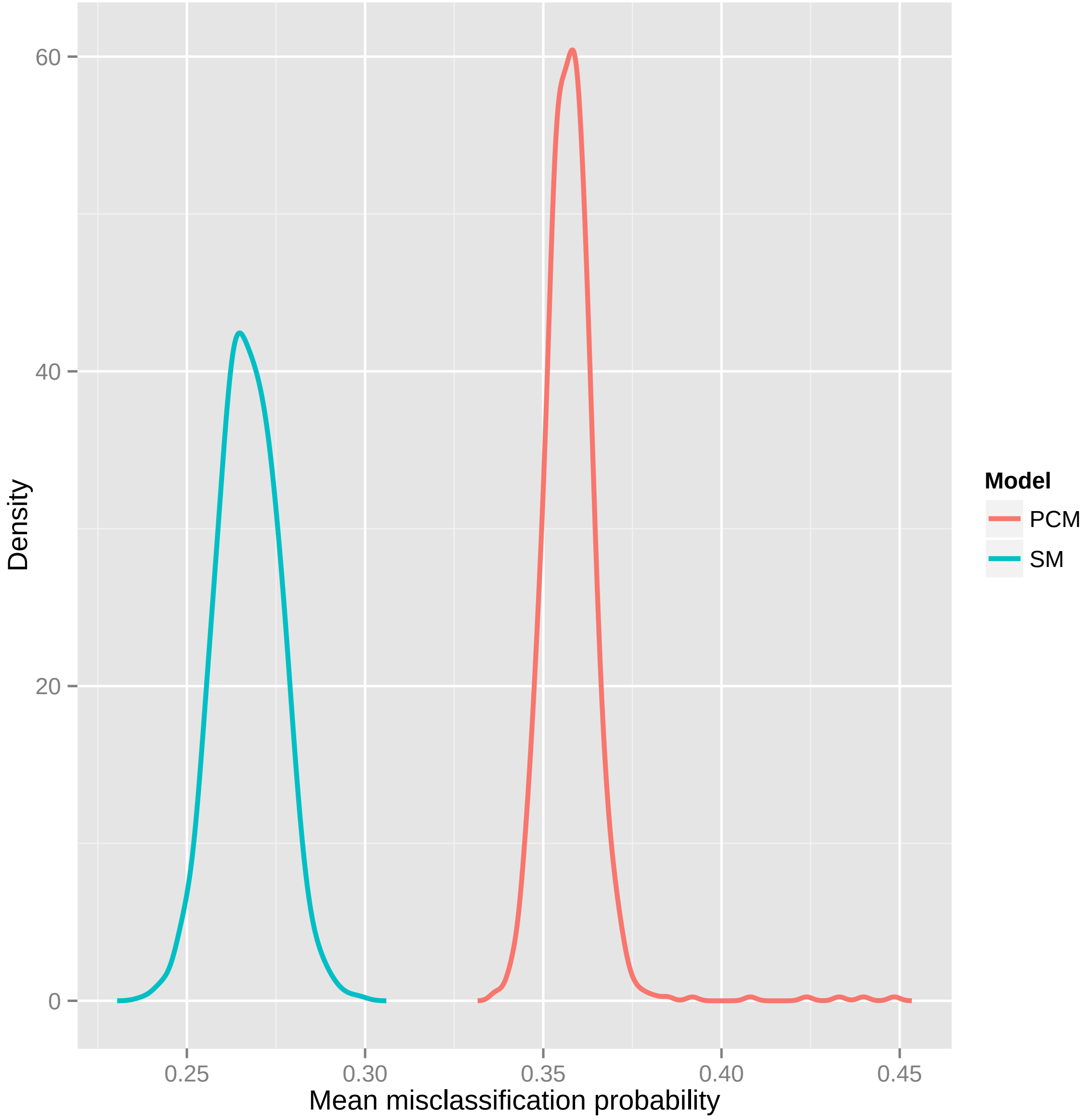

Figure 32: Predictive performance with 0-1 loss (true model SM)

6. Data analysis

6.1. Description of data sets

6.1.1. Exam data from statistics lecture

The first small data set was an exam of the lecture *multivariate statistics* for students studying at Ludwig Maximilian University in Munich. There were 18 problems to solve and a group of 57 students. Is was possible to award half points. For analysis, the half points were rounded up to the next integer. In the following two tables a short description of the variables and a summary is given:

Table 3: Description of variables of the exam data: Items (top) and additional covariates (bottom)

Variable	Description
Item1-Item18	Open questions from the subject of multivariate statistics. Most of them require analytical skills to calculate measures and give an interpretation of them.
Gender	indicator as dummy variable ($\in \{0, 1\}$) coded 1 if a person is female and 0 if a person is male
Level	indicator as dummy variable ($\in \{0, 1\}$) coded 1 if the person is a master student and 0 the person is a bachelor student in an other subject

There are some items, which were solved by the majority of students almost correctly and some items were the majority of students could not achieve a high score. The data suits the required structure of a sequential model: A person can only achieve maximum score on a specific item, if he/she is able to solve all difficulty steps in an sequential order. In inference the sequential model will be prefered over the partial credit model. As outlined in chapter 4, the MMLE method seems the most promising in this case, because it has significantly lower RMSE than the other proposed estimation methods.

Table 4: Summary of variables of the exam data

Variable	Categories/ Unit	Sample proportion/ median(range)
Item1	points	4 (0-7)
Item2	points	1 (0-3)
Item3	points	8 (0-8)
Item4	points	10 (0-12)
Item5	points	6 (0-6)
Item6	points	5 (0-6)
Item7	points	12 (0-13)
Item8	points	3 (0-5)
Item9	points	6 (0-6)
Item10	points	5 (0-5)
Item11	points	6 (0-6)
Item12	points	4 (0-8)
Item13	points	2 (0-6)
Item14	points	1 (0-3)
Item15	points	2 (0-7)
Item16	points	0 (0-4)
Item17	points	6 (0-8)
Item18	points	5 (0-7)
Gender	male	65 %
	female	35 %
Level	bachelor	84 %
	master	16 %

6.1.2. Ascot pre-test data: Industrial business management

This data was collected from a pre-test of industrial Business Management assistants in education with unipark online software (Dreßen and Trabert 2014). The empircal survey was conducted from Human Resource education & management institute (Weber and Trost 2014). It consists of 339 student observations and 55 variables. The variables consists of two parts: 25 Competence questions on project performance and 30 explanatory covariates. The individuals are anonymised. A first description of the variables is given in table 5 and a summary of items 1-5 in table 6. The items 6-25 are shown in the appendix A.6.

Table 5: Description of items of the Ascot pre-test data

Variable	Description
Item1	To start a project and continue it successfully would be an easy task for me
Item2	I know the necessary, practical details to start a project
Item3	I conduct project based activities, e. g. developing ideas, planing, implementation and motivation of participants, completely autonomous
Item4	I am able to control the processes of implementation in a new project
Item5	The skills for working in projects (e. g. developing ideas, planing, implementation and motivation of participants) should be regarded as more important in work
Item6	Project based work has more advantages than disadvantages
Item7	To work in projects gives me individual fulfillment
Item8	I act proactive on problems
Item9	I take the initiative, if others don't do it
Item10	Most of the time, I work more than required
Item11	I often suprise people with new ideas
Item12	Often I am consulted for help, if others search for inventive ideas
Item13	I prefer tasks, which need alternative solutions
Item14	At school, I tried to fulfill all required tasks
Item15	In business i tried to fulfill all required tasks
Item16	At school it was obvious for me, that I have to do it this way
Item17	In business it was obvious for me, that I have to do it this way
Item18	At school, learning and working was fun for me
Item19	In business, learning and working was fun for me
Item20	If I can achieve my goals, I will be in a higher job position during the next five years
Item21	I will do any effort to achieve a leading position in business
Item22	I am well prepared to be in a leading position
Item23	If my plans will work, I will be self-employed with my own firm in five years
Item24	I will do any effort to establish a new firm
Item25	I am prepared to do everything to be an entrepreneur

The items considered here are questions with 6 ordinal categories. It is a self-assessment of the students based on personal experience. Therefore the data could be biased, because skills will not be objectively evaluated as assumed in the simulations, where the test questions had fixed points and predefined answers. This could lead to overrating and underrating of answers. Overall impression: Most items have a median above 4.

Table 6: Summary of items 1-5 of the Ascot pre-test data

Variable	Categories	Sample proportion
Item1	0: Does not fit	0 %
	1	6 %
	2	22 %
	3	42 %
	4	23 %
	5: Fits fully	6 %
Item2	0: Does not fit	0 %
	1	7 %
	2	21 %
	3	32 %
	4	31 %
	5: Fits fully	9 %
Item3	0: Does not fit	3 %
	1	11 %
	2	25 %
	3	28 %
	4	24 %
	5: Fits fully	9 %
Item4	0: Does not fit	0 %
	1	8 %
	2	19 %
	3	32 %
	4	31 %
	5: Fits fully	10 %
Item5	0: Does not fit	1 %
	1	4 %
	2	10 %
	3	26 %
	4	42 %
	5: Fits fully	17 %

6.2. Inference of data sets

6.2.1. Analysis of the exam data with MMLE

The analysis of the exam data is carried out with the sequential model and the MMLE, with a Laplace approximation of the required integrals. The data transition coding, as outlined in section 2.3.4, leads to 6840 obersations. For this sample size it is reasonable to assume a asymptotic normal distribution for the estimated item and person parameters. On this base, Wald confidence intervals for the person $\hat{\theta}_i \pm q_{1-\alpha/2}\hat{sd}(\hat{\theta}_i)$ and item parameters $\hat{\beta}_{jk} \pm q_{1-\alpha/2}\hat{sd}(\hat{\beta}_{jk})$ can be constructed. With $\alpha = 0.05$, the corresponding normal distribution quantile is $q_{1-\alpha/2} = 1.96$ and $\hat{sd}(\hat{\beta}_{jk})$ is the estimated standard deviation for the estimated parameters. In figure 33 the person parameters for all students with confidence intervals are shown:

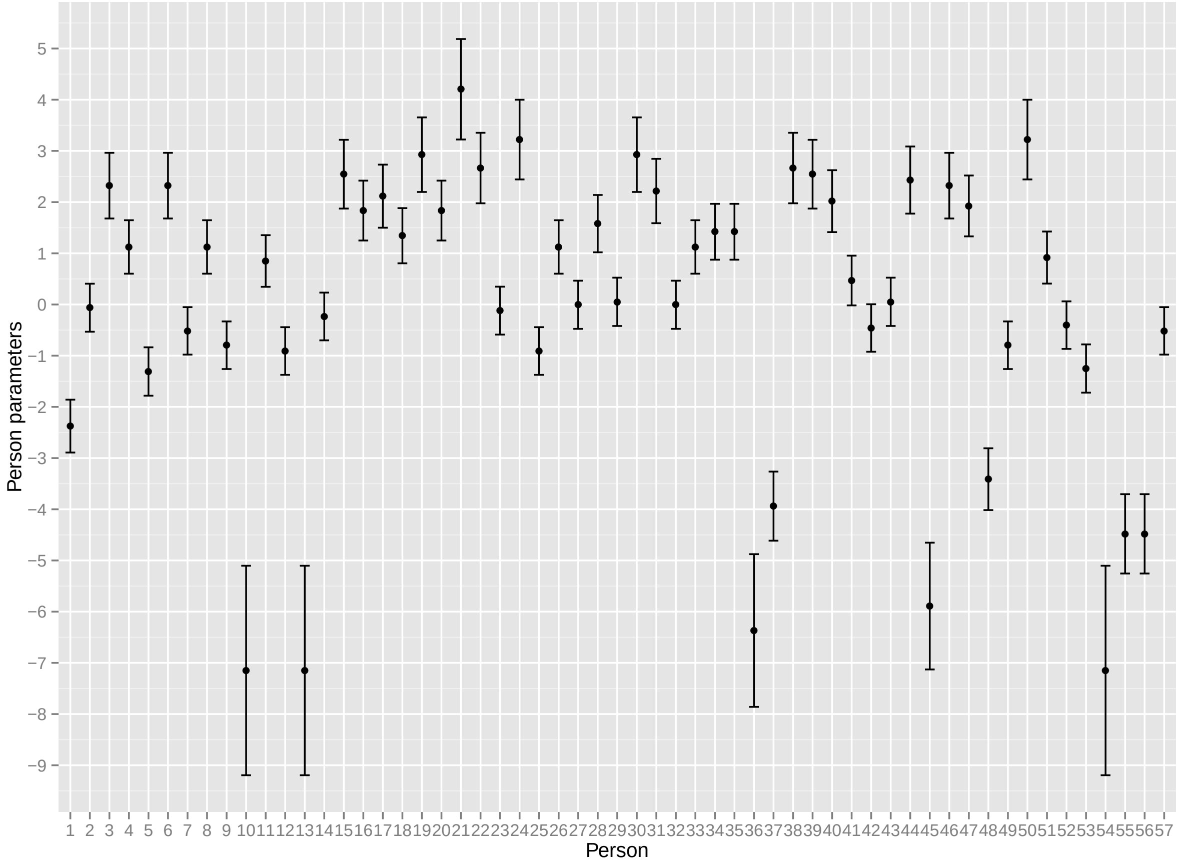

Figure 33: SM: Person parameters with 95 % confidence intervals in exam data

A group of 56 % of the persons have an ability above zero. The more the ability parameters deviate from zero, the confidence intervals get larger, because there is less data available for very low or very high scores. In this group there were no students with outstanding performances (ability > 5) but lots of mediocre and few low performers (ability < -5). The item parameters are displayed in figure 34:

For every item 1-18, the corresponding sequential difficulty steps with their confidence intervals are shown. In all items the difficulty is monoton increasing for achieving higher scores. The items 2, 4, 5, 7 have a uprising step difficulty for the last transition. In item 7 the difficulty steps - beginning with transition from 2 to 3 and onwards - have almost the

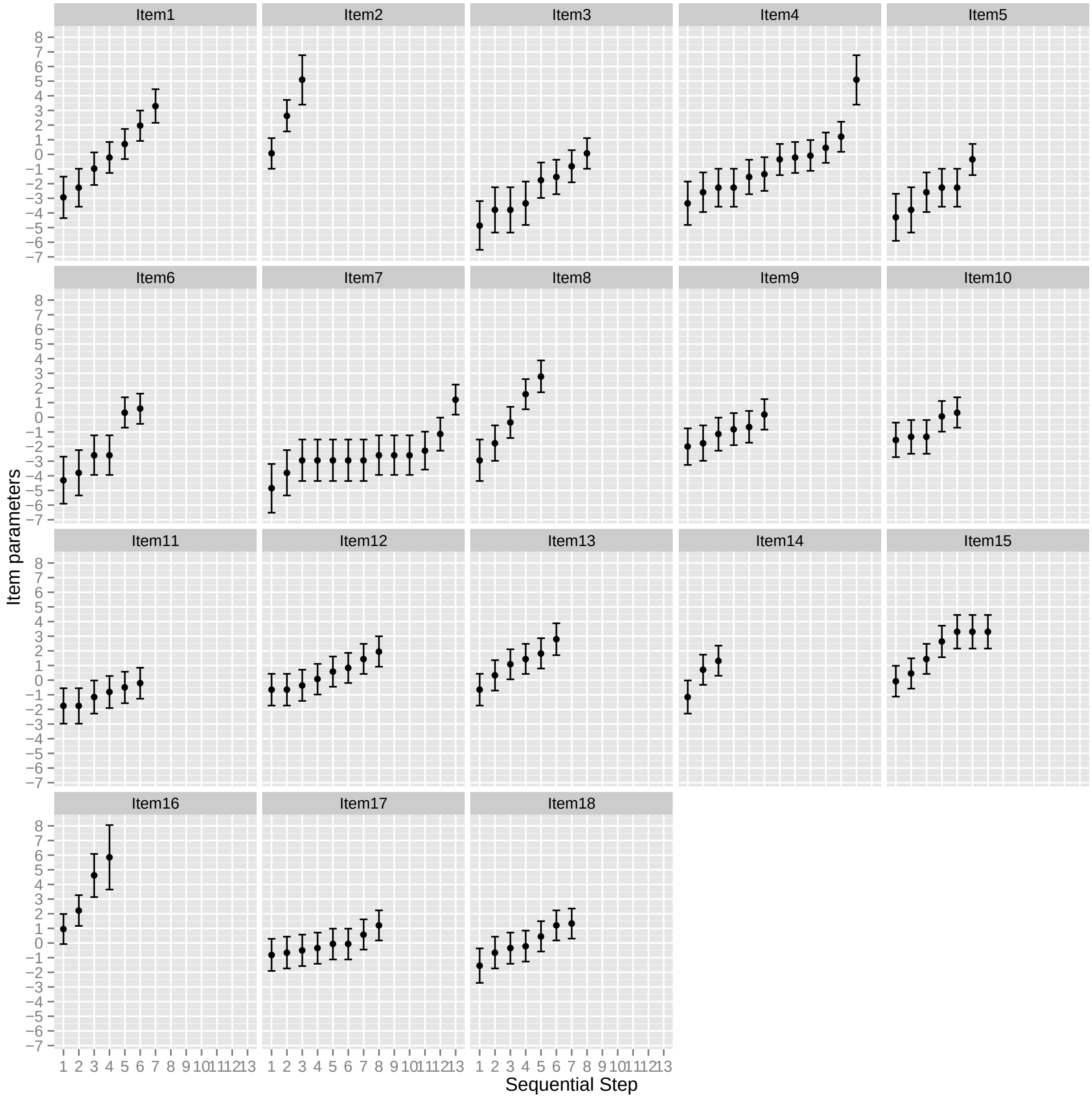

Figure 34: SM: Item parameters with 95 % confidence intervals in exam data

same requirements. In contrast the items 1, 2, 8, 16 have a steep raising. The difficulty of one item - consisting of dichotomous transitions from category k-1 to k - is not directly derived from the item parameters. To evaluate, which item is - on average - the most difficult and which is the easiest, the following derivations are made: First the expected score for an average gifted student with ability 0, conditional on the step difficulties for each item, is calculated $\sum_k prob_{jk} score_{jk}$. Then the result is divided by the maximum achievable score for the given item. This estimates what proportion of the maximum score can be expected by an average student. If the proportion is low in an item, than this item is difficult to solve. In the exam data this measure leads to a ranking, that the items 16, 15, 13 are the most difficult with expected proportions under 20 % and the easiest are items 5, 3, 7 with expected proportions above 70 %. For comparison purposes, the median is located between the items 1 and 4 with about 40 % proportion.

To get an better understanding how the person and item parameters influence the conditional probabilities of a person to achieve a specific score category, the probability curves (CPC) are plotted. An CPC represents functions for all scores depending on the ability of a person, given the appropriate item parameters. First the CPC for item two is displayed in figure 35:

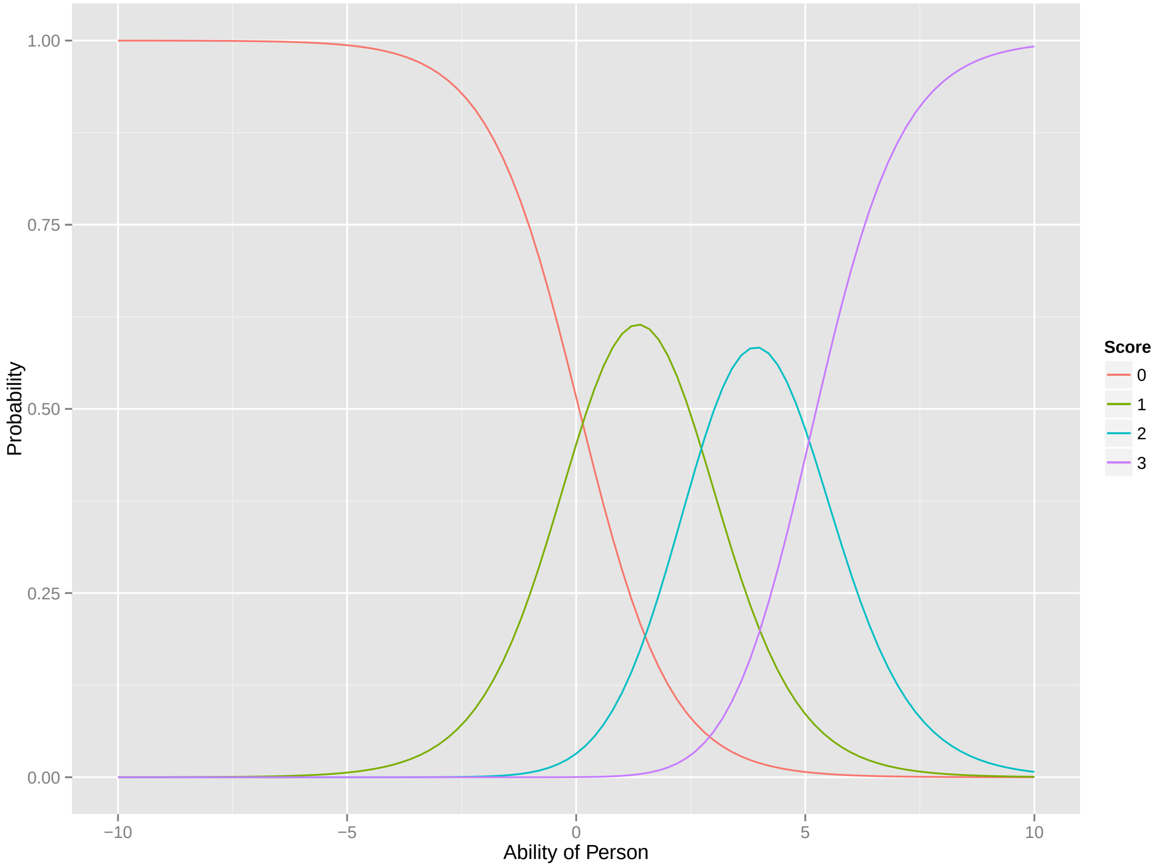

Figure 35: SM: Category probability curves for item 2 in exam data

The plot shows four possible scores ranging from zero to three. The leftmost curve is the probability to achieve a score zero depending on the ability score. It's shape is similiar to common survival curves. The more skilled a person is, the less the probability is that the person does get zero points for item two. The two curves in the middle (scores one and two) have a similiar shape, like common density curves. As the proability for score zero decreases with higher performance levels, the probability for score one increases with a higher ability up to a maximum. The rightmost curve for score three has a shape of a probability distribution function. Only persons with an ability greater than zero have a realistic chance to achieve the maximum score.

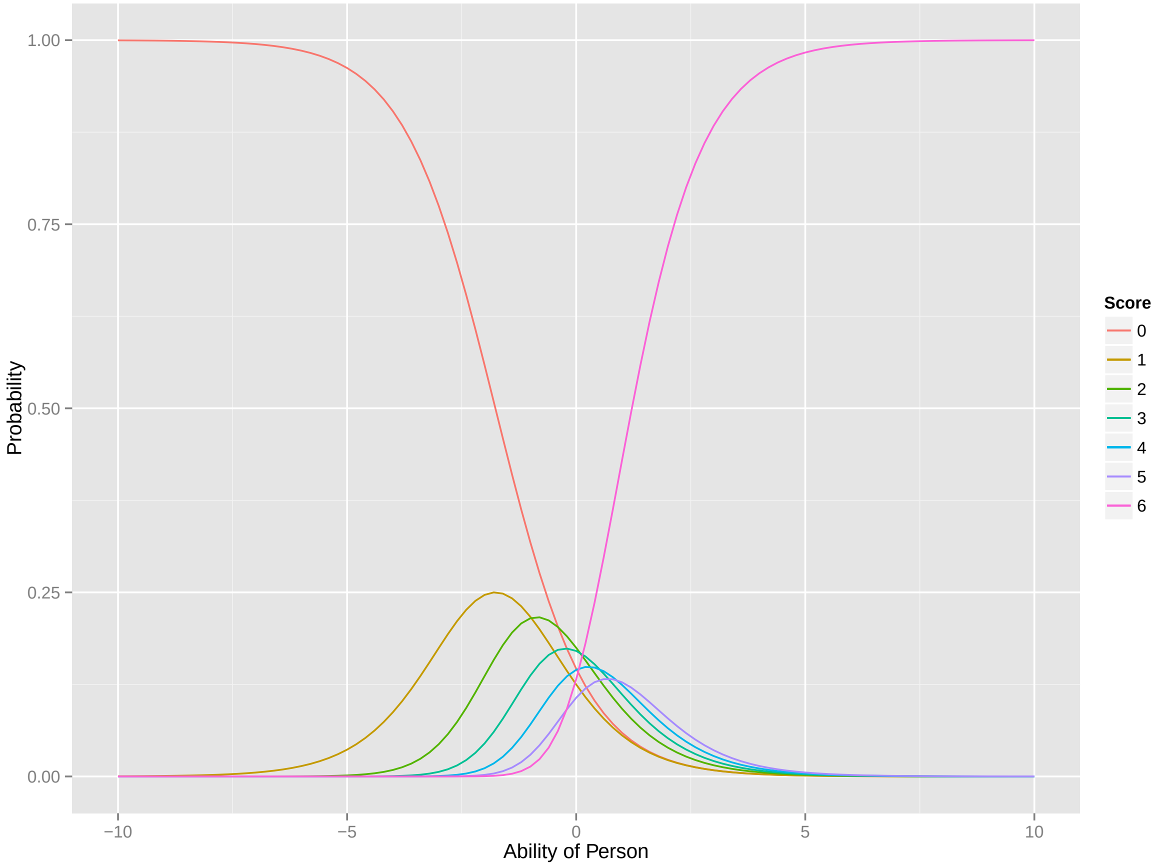

Figure 36: SM: Category probability curves for item 11 in exam data

The next figure 36 is for item 11 with a maximum score of six. In contrast to the previous CPC, there are more curves in the middle (scores one to five). A person with the ability zero has a higher probability to achieve maximum score than the one in figure 35.

Masters (Masters 1982; see pages 161-162) pointed out, that the intersection between curves of category k-1 and k of the PCM can be interpreted as the step difficulty parameter on the x-axis. But generally it does not apply for the SM. For example in figure 37 the estimated item parameters for the first four steps are ($\beta_{15;1} = -0.07, \beta_{15;2} = 0.45, \beta_{15;3} = 1.45,$ $\beta_{15;4} = 2.64$). The roots of the difference function between CPC with steps k-1 and k are the abilities $(0.82, 0.92, 1.81, 3.37)$. Obviously the intersections and item parameters are not equal. Further the CPC with scores five, four and six, five have in pairwise comparisions no intersections within the numerical accuracy. The curves with scores 1-5 converge to a probability value of zero, when the ability goes to $\pm\infty$. In these examples of CPC there are no general patterns of the observable curves evident, except that the category curves for zero have a similiar shape as a survival curve and the maximim score category has the structure of a probability distribution function.

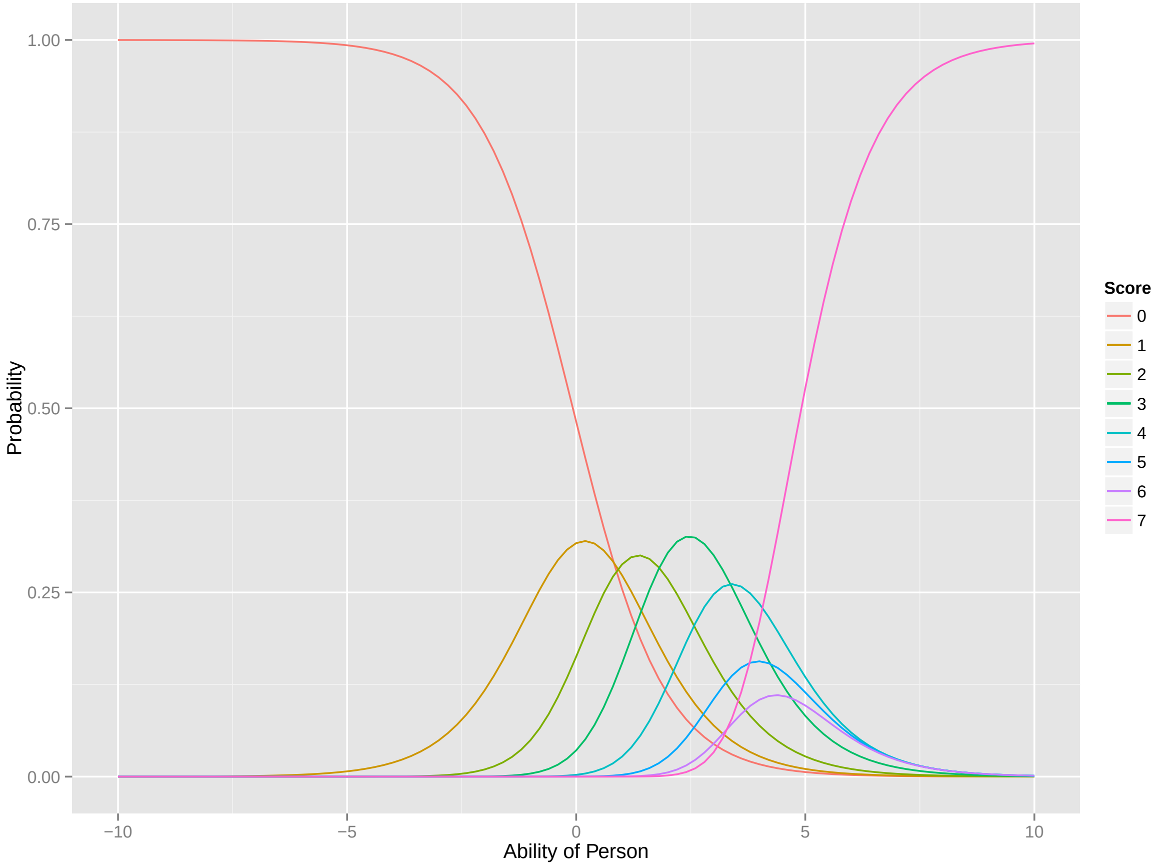

Figure 37: SM: Category probability curves for item 15 in exam data

6.2.2. Analysis of exam data with CMLE, JMLE and predictive comparison

The comparison of the estimates, confidence intervals and predictive performance of the exam data is in addition explored with JMLE, CMLE for the SM. If a bootstrap sample could not be estimated it was left out and instead of this one another bootstrap sample was estimated. Above 500 parametric bootstrap samples were estimated for each estimation method.

The person parameter estimates of JMLE are shown in the next figure 38. The structure of the person parameter estimates is similiar to the person parameter estimates from MMLE. For example, persons with low performance (person 10, 13, 36, 37, 45, 48, 54, 55, 56) are the same persons with low abilities in MMLE. The confidence intervals of JMLE are larger for persons with point estimates around zero, than those estimated in the MMLE. The upper bound for the ability estimate of person 21 is larger than in the case of MMLE.

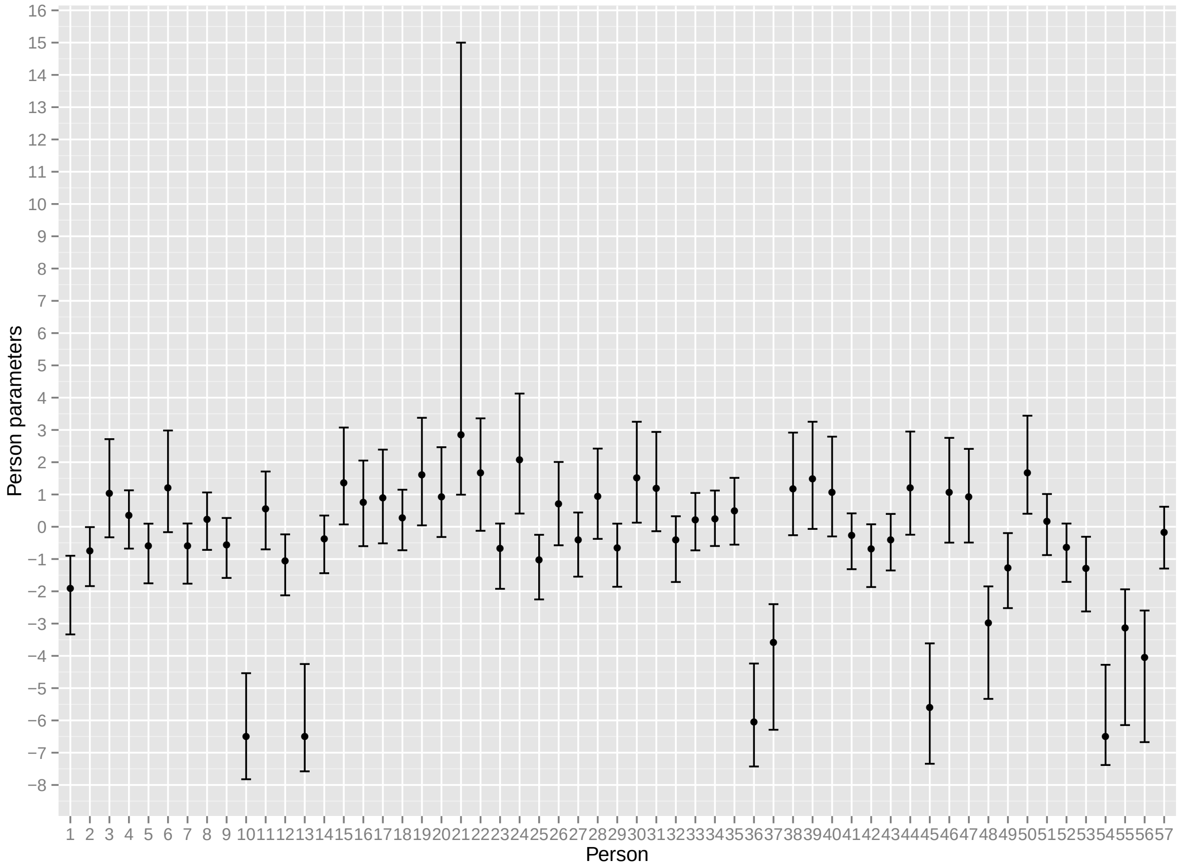

Figure 38: SM: JMLE of person parameters with parametric bootstrap percentile CI in exam data

The item parameter estimates of JMLE with parametric bootstrap confidence intervals are shown in figure 39. Especially the last two sequential steps for item 5 and item 11 are quite uncertain with higher variability than the corresponding MMLE confidence intervals. Another difference in comparison to the marginal point estimates is, that the JML step difficulties do not generally increase with increasing steps. The JML step difficulties go up and down, and are not monoton increasing as the MML point estimates. If the item difficulty is assessed, as described in section 6.2.1, then there are differences between the JMLE and MMLE. For example, the third and forth most difficult items (item 2 and item 13) changed the order, item 3 and item 5 changed place 15 and 18 and the proportion of scores of an average person is expected to achieve, on the top 5 of difficult items, is higher with JMLE than MMLE. For the most difficult item 16 both results are consistent, but in MMLE the expected proportion is 7.63 %, in JMLE it is more than twice this amount. Except for the last difficult item, the proportions are higher with the JMLE than with MMLE. So on average JML estimated, that most of the items are assessed easier. For reference the tables are shown in appendix A.4.

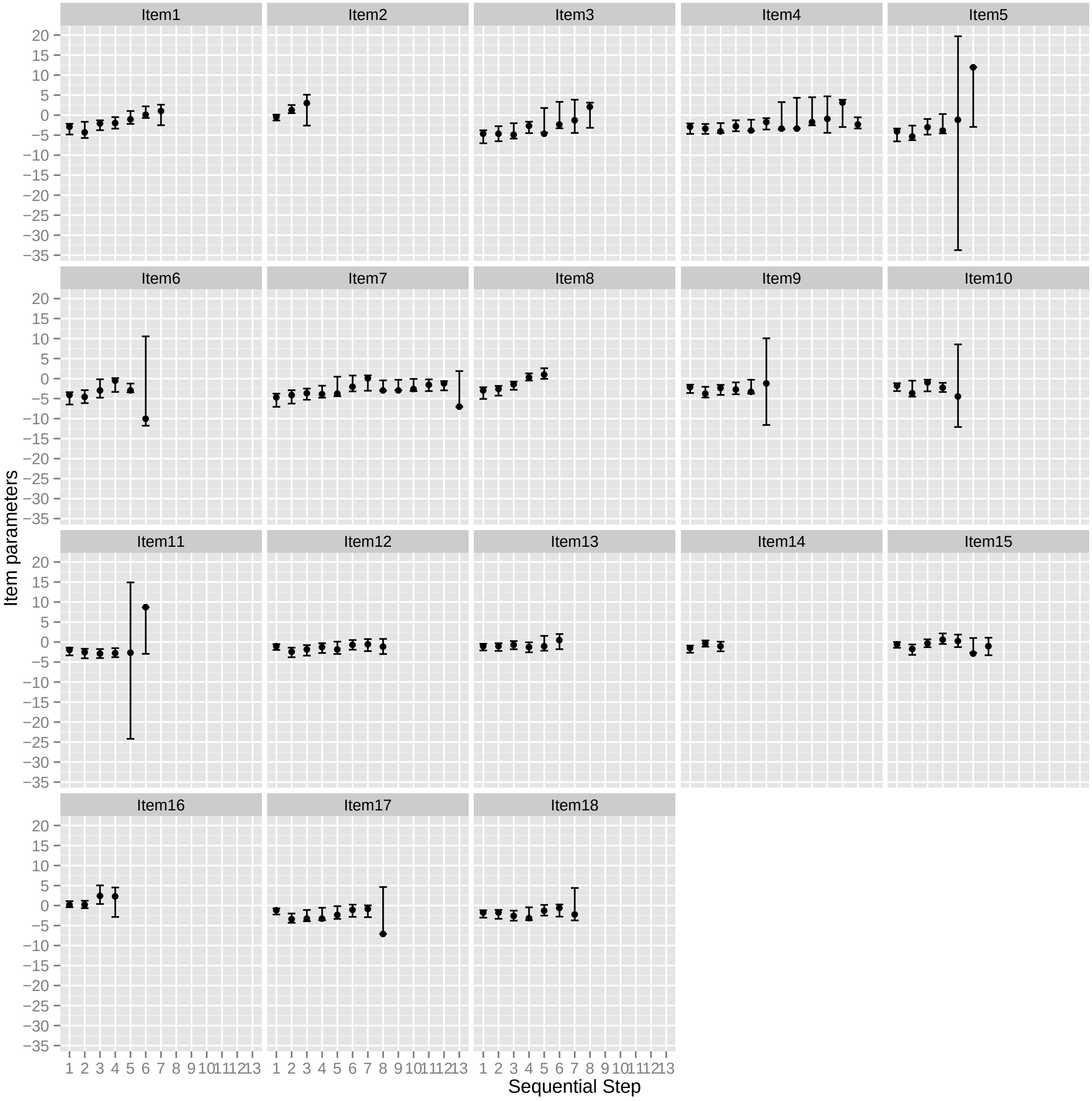

Figure 39: SM: JMLE of item parameters with parametric bootstrap percentile CI in exam data

The person parameters of the CMLE are similiar structered as in the JMLE. The bootstrap percentile confidence intervals of the estimated item parameters are in some cases wider (e. g. item 2) and in others smaller (e. g. item 5) compared to the JMLE. Because of the small changes compared to JMLE, the figures are included in the appendix A.5. In the ranking of item difficulty with CMLE item 16 is the hardest and item 5 the easiest. In comparison with JMLE the expected proportion of points is lower over all items. This implies, that the CML item parameters are higher than the JML item parameters.

Regarding the predictive performance of $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$ by leaving out persons of the three estimation methods, the MMLE performed on average better than JMLE, CMLE. From the 100 training data sets 3 cases had to be left out due to convergence problems. MMLE has an average misclassification rate of 59.1 % with 5 % and 97.5 % quantiles $[44\%, 73.4\%]$, JMLE of 61.2 % $[43.3\%, 76.5\%]$ and CMLE of 61.3 % $[42.9\%, 76.6\%]$. In some cross validation samples the peformance is lower than random guessing (misclassification rate of 66 %). In comparison with the same setting the 0-1 loss function produced the following missclassification errors: MMLE has an average misclassification rate of 48.9 % with 5 % and 97.5 % quantiles $[31.6\%, 69.4\%]$, CMLE of 50.6 % $[29.3\%, 72.4\%]$ and JMLE of 50.61 % $[29.3\%, 72.4\%]$. The 0-1 loss function is on average smaller or equal than $L(\underline{\hat{p}}_{ij}|\underline{y}_{ij})$. This is plausible, because the 0-1 loss function does penalize errors weaker, but can better discriminate between true and false statistical models, as demonstrated in section 5. The MMLE is therefore on average better than JMLE and CMLE, but the 95 % quantile range are almost completely overlapping.

The inclusion of covariates could potentially reduce the MSE. The mean estimated prediction errors are higher than those in the simulation study. In the simulation the prediction error the true data generating process is the same as in the model which is used for estimation. With real data the underlying structure is unknown and can be arbitrary complex. Therefore the performance on real data is often less effective than in simulation studies. On the other hand the data set was much smaller than the data sets used in the simulation study. With less information the variance of estimators increase and so increases the MSE, given the BIAS stays equal.

6.2.3. Analysis of Ascot pre-test data with SM vs PCM (MMLE)

In this section the Ascot pre-test data will be analysed with SM and compared to the PCM. The MMLE is the most promising estimation method (see chapters 3 and 4) for both models. In transition coding there are 42375 observations for estimation of the SM. Therefore the estimation of the SM is more computationally demanding than the PCM, which uses the original dataframe. For the evaluation of predictive performance the 0-1 loss function is used. By performing a 10 times repeated 10-fold cross validation the SM performed on average with misclassification 58.41 % error better than the PCM with 61.14 %. For this data, the theoretical flexibility of the SM results in a better average predictive performance. The stepwise interpretation is more appropiate for this data set.

In the figures 40 and 41 the empirical density estimate of the estimated person parameters is shown. In both models the person parameters are quite close to the best fitting normal distribution. In comparison between the models, the person parameters in PCM have a lower standard deviation but similiar shape as the SM estimates.
Values for the estimated item parameters are shown separately in the appendix A.8. Because of constraints and computational complexity of the algorithms no further bootstrapping of confidence intervals was performed for the Ascot data. In the SM all step

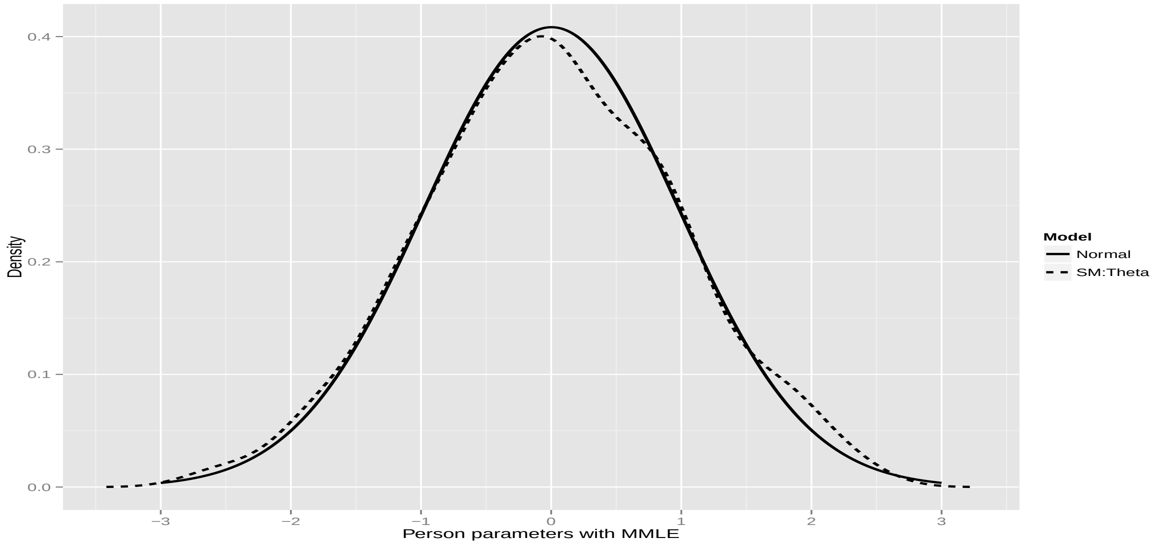

Figure 40: SM: Kernel density estimate vs normal distribution of person parameters (Ascot pre-test data)

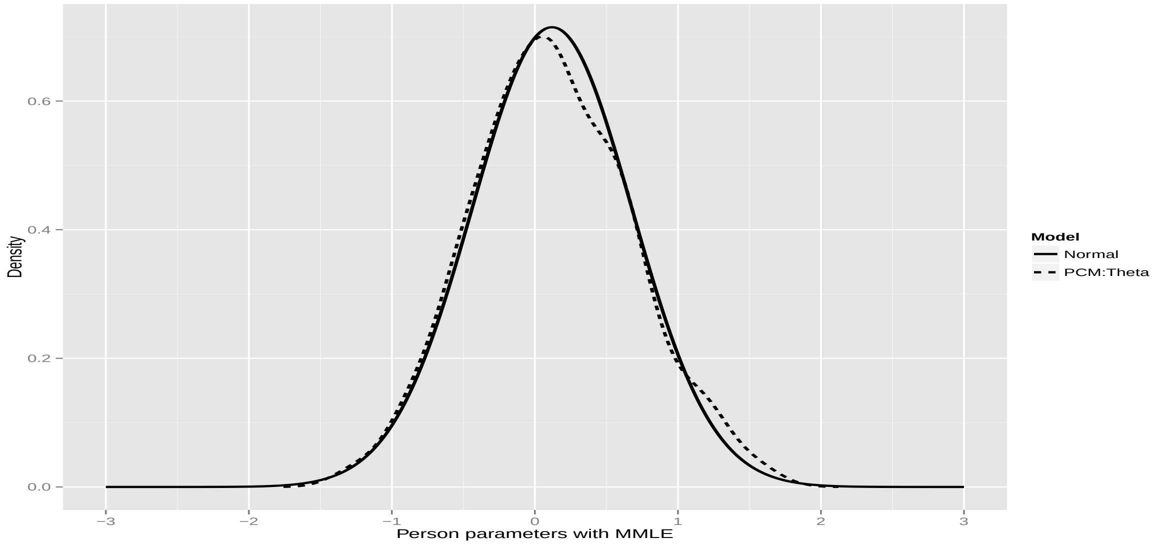

Figure 41: PCM: Kernel density estimate vs normal distribution of person parameters (Ascot pre-test data)

difficulties for each item are monoton increasing with higher scores. 63.2 % of all item parameters are below 0 and therefore most of the skill based requirements were moderate or easier. Some first step parameters, e. g. item 2, item 8 and item 9 are completed easily. The reason for this evaluation gets obvious by looking at the original data: No person anwered these questions in the first category. The estimates should be quite instabile. If one assumes that the students answers are not biased, then all students know at least a minimum of the required skills to start a project in practise (item 2). Also all students have at least some initiative for active problem solving (item 8) and engaging open problems (item 9). Most of the last difficulty steps are above 0. Noticeable are the different ranges of item parameters: No item parameter is below -4. The PCM item parameters are not monoton increasing. Instead the values of the item parameters fluctuate up and down. 65.6 % of the PCM item parameters are greater than the corresponding SM item parameters.

In figure 42 the item parameter estimates for the SM in comparison with PCM are displayed. The dotted line indicates equal coefficients. Except for the items 23-25, the first 3 category paramaters are lower for the SM than the PCM. The item parameters for the last 3 categories of the items 23-25 are higher in the SM than the PCM. There is a tendency of SM item parameters for lower categories to have lower coefficient than the PCM. The reverse situation is observed for higher categories: There the SM item parameters are usually higher than the PCM. For the median category m in each item the transition to category m+1 of the SM is in 68 % of the items the parameter with the smallest absolute deviation of the PCM. This means, that the SM and PCM item parameters have a tendency to be more similiar in the center of the response data.

Ordering the items according to difficulty, the former used approach (expected score in proportion of max score) for the exam data can not be applied here, because question answers represent ordinal categories and not absolute points. Instead the probability of average persons with ability 0 to achieve the category on item j is evaluated. Items with low values on this criteria are harder to solve fully. The complete tables for ranking of item difficulty are stored in the appendix A.7. Both SM and PCM match on the most difficult and easiest item. The most difficult item is 23 and the easiest is 15. The question 23 measures if and how ambitious students trying to become an interpreneur in five years. Obviously this should be difficult for young persons, who only studied theory but never tried it in practise. The item 15 measures if the student thinks, he has done everything what was expected from him in the firm. This is a minimum requirement, that trainees have to fulfill their duties. Some differences between the ranking of the SM and PCM of item difficulties exists, but the probabilities for achieving highest scores by average persons are always higher for the PCM than the SM. For example the most extreme deviation is observed with item 2: A rank of 9 was awarded on the difficulty item scala in the SM, but ranked 24 in the PCM. The probability of answering the last category was about 39 % different in both models. The second most exterme deviation is observed with item 14. The rank differed only by 1 (SM rank 22 and PCM rank 21), but the probability difficulty assessment differed by 11 % in absolute terms.

The CPCs for item 23 of SM (figure 43) and PCM (figure 44) show big differences in the fitted probabilities. For example, even a skilled individual with person parameter 2.5 has a probability for answering the question within category 5 of about 12.5 % in the SM

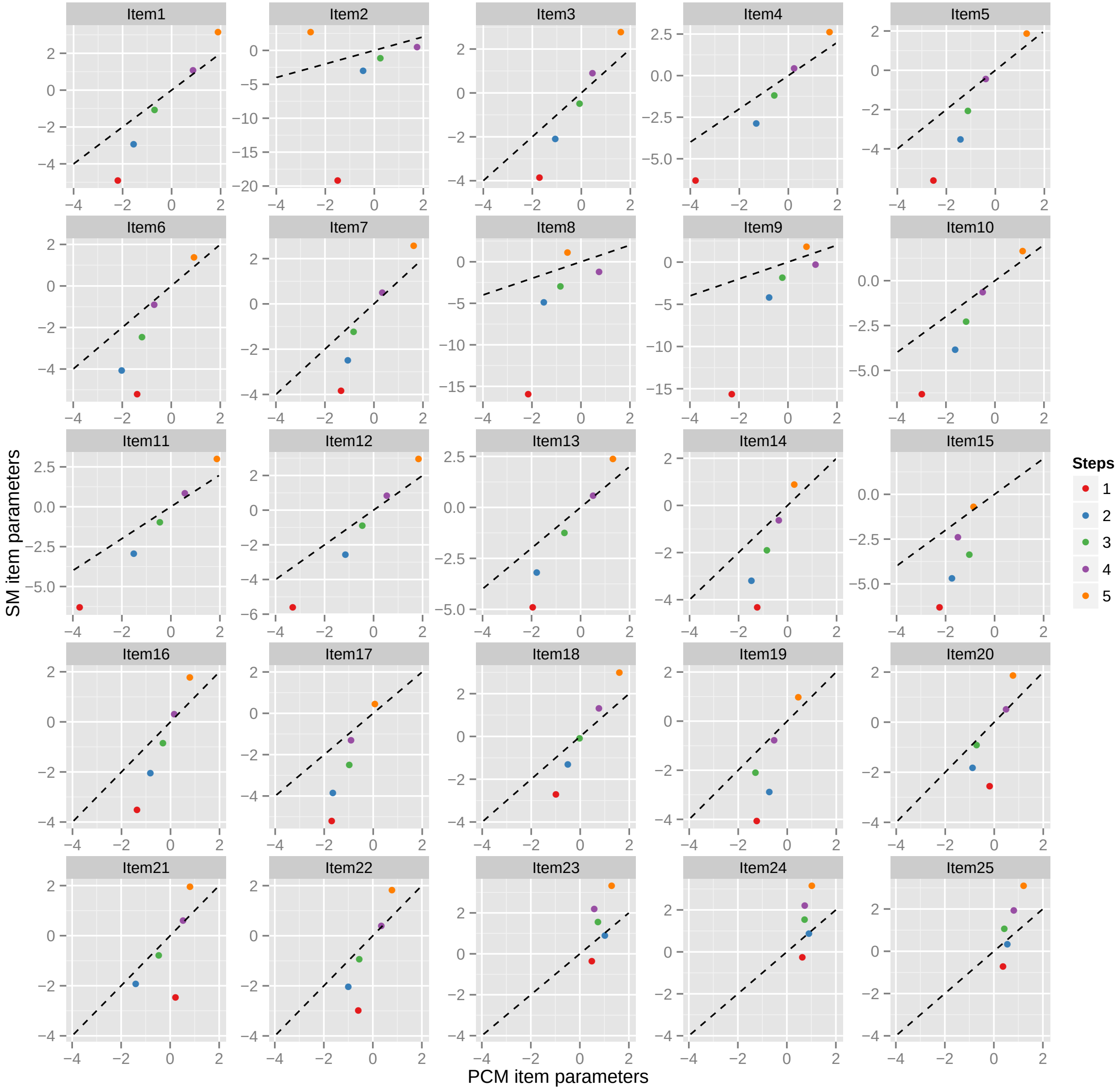

Figure 42: SM vs PCM: Estimated item parameters of Ascot pre-test data

and 75 % in the PCM. The probability of answering in the lowest category 0 is about 6 % in the SM and about 0 % in the PCM.

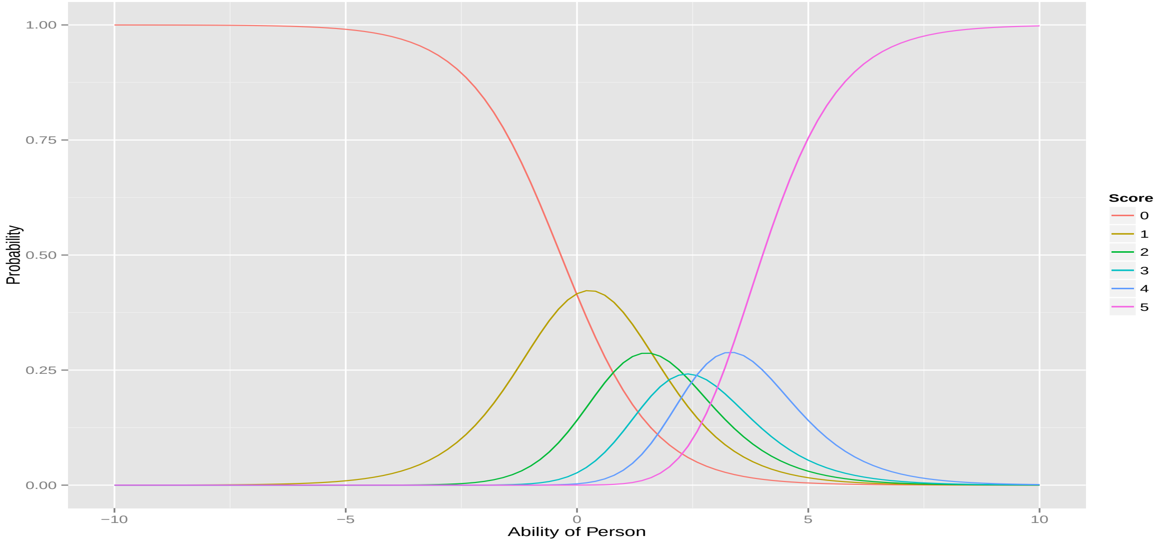

Figure 43: SM: CPCs for item 23 in Ascot pre-test data

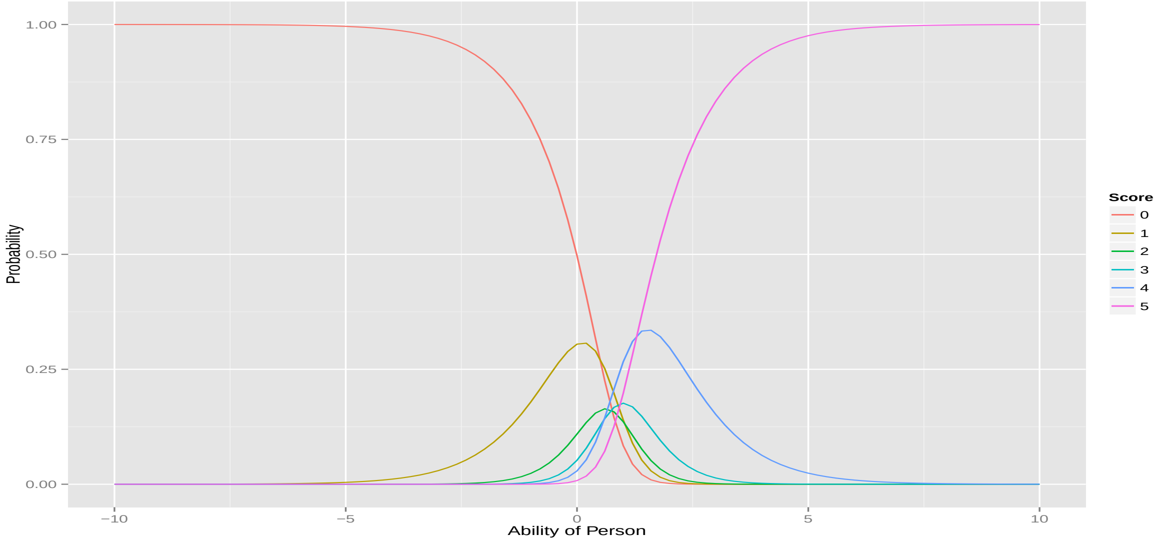

Figure 44: PCM: CPCs for item 23 in Ascot pre-test data

7. Discussion and outlook

As demonstrated in this thesis, IRT is a viable approach for measuring the performance of minds by analysing the answers to questions with multiple categories. As often in science, some answered questions generate new questions, subjet to future research.

Open issues:

An open issue is the performance of the conditional approaches in the simulation of the partial credit model. It should be checked if the conclusions can be reproduced in a similiar environment with self made functions. If this is the case, the cause of the bias should be investigated.

In this study no misspecifications of the models were tested. This means, that in most designs the data were generated with the true model and the same model was used for estimation. The new proposed subsampling of r & c cross validation could be investigated in further simulation studies. The purpose is to find out which measures are better suited to evaluate predictive performance of IRT models.

In IRT some items discriminate sharper in close ability ranges, but are less sensitive in discriminating large skill differences or reversed. The results of the simulation study indicates that loss functions could have a similiar behaviour.

Beside the questions mentioned above, some extensions of the IRT approach require further discussion. This extensions were not applied in this thesis, but could be possibly adapted and used in the SM. The following ordered issues are proposed for further studies:

1. Existence of latent variables
2. Multidimensional latent traits
3. Flexible response functions
    a) Parametric extensions
    b) Nonparametric extensions
4. Differential item functioning (DIF)
5. Application of ensemble methods for IRT
6. Other estimation methods than MLE
7. Test Equating, Scaling and Linking

From a philosophical perspective it is unclear if latent variables really exists. In probability theory Bruno De Finetti (Mura 2008; see page xviii) once argued, that *probability does not exist*. What he meant was that probability is not objective, but based on individual judgements.

An assumption in this thesis was, that the latent variable is unidimensional. That means, that the ability can be measured by one trait. If the test items are constructed properly, then this assumption may hold. In some questions the probability for a correct answer depends on different skills, because individuals can be different due to several attributes, which influence question behavior (Reckase 2009; see chapter introduction). But how to estimate the dimension of the latent variable? Even if there is a promising method to handle this, how to interpret each dimension of the latent trait?

Criticism:

The criticism of PCM or SM is, that the underlying data generating process is more complex and that the assumptions of the model could be violated. The PCM, SM are parametric models, which assume a specific form of the distribution function. In the literature there are some more flexible models available, for example 3-PL model (Harris 1989; see pages 36-37). It is called 3-PL model, because it measures three parameters for each item. In comparison with the Rasch model two additional parameters are introduced: A discrimination and a guessing parameter. The discrimination parameter measures how strong an item can discriminate between individuals with similiar latent ability. The higher the parameter, the better it discriminates small differences in abilities. But also the range of values, which can be efficitively discriminated, becomes smaller. So there is a trade-off between discrimination in short and large ranges of the linear predictor. The guessing parameter gives a lower bound for the probability to solve an item with a score, independent of the ability parameter. This would account for people who just got randomly a correct answer. Additonally there are also approaches with 5-PL models (Gottschalk and Dunn 2005). Here an additional fourth parameter for the maximal probability of solving an item with a specific score is estimated. This parameter can be interpreted as minimum error a person is able to achieve due to oversight, independent of the ability level. The fifth parameter controls the asymmetry of the distribution. A downside of this extension is, that desirable theoretic measure properties are vanishing (e.g. separability of person and item parameters). Also the optimization task may become ill-conditioned and harder to solve, because of local optima.

Alternatives:

For explanatory data analysis and checking assumptions of parametric IRT models non-parametric IRT (NIRT) models are proposed. In contrary to the parametric RM there are less assumptions necessary on the shape of the response function. The most widely known approach in NIRT is based on kernel smoothing as implemented in (Mazza et al. 2014). The idea is to use a weighted response based on symmetric kernel function weights (e. g. gaussian kernels, Nadaraya-Watson estimator). The bandwidth can be choosen as asymptotic plug-in estimate or with cross validation. This approach is quite flexible and can detect non-monoton increasing response functions (Wang 2012).

In this study another assumption regarding the linear predictor was made: No differential item functioning (DIF) was considered. DIF is present, if the difficulty of items is different for a subgroups of persons with same skill level. Either all items are effected by a shift or only some items may be influenced by DIF. In the MMLE framework additional covariates can be added for detection of DIF. A Langrange multiplier test could be used as explained in (Glas 2001). The modelling with covariates could increase predictive performance.

In the simulation study of the sequential model the MSE mainly consisted of the variance of the estimators. In machine learning a common technique for variance reduction of adaptive estimators are the ensemble methods (Zhang and Ma 2012; see chapter 2). The most populars are bagging and boosting. The idea of bagging is to use independence of many models to construct a better lerner. Boosting uses the dependence of lerners to improve the models by refitting values with high residuals from previous iterations. Therefore bagging can be regarded as parallel ensembles and boosting as sequential ensembles. It is not guaranteed that the prediction will be better than the best classifier in the ensemble, but the likelihood decreases to choose a wrong one. However ensembles are usually less interpretable than single models.

There are other less popular alternatives for estimation of polytomous Rasch models than Maximum Likelihood. For example the Minimum Chi-Square (MCS) method for polytomous items (Linacre 2004; see pages 99-100). It is based on the idea of the chi-square statistic: Take the quadratic deviations between the observed and expected values divided by the variance. These differences are summed up over all persons or items. The expected value is calculated for every conditional distribution given person i and item j. Similiar to the JMLE approach, this loss function could also be used in a similiar manner. For a more depth discussion about the merits of MLE and MCS method see (Berkson 1980). Besides frequentist approaches bayesian techniques should be more investigated, because the recommended generalized, linear mixed model approach for estimation is a kind of empirical Bayes estimator. With the specification of prior knowledge the estimation could be more stable with noisy data and less restrictions are required.

Practical relevance:

A last point is the practical relevance for the design of questionnairies. The results in this study could be used to improve existing practise of test equating, scaling and linking (Kolen and Brennan 2014; see chapter 1) in terms of estimation accuracy. These subjects focus on comparing different test forms over time, different items and people. An example to apply the results is the Programme for International Student Assessment (PISA) (OECD 2014).

References

Herve Abdi. The bonferroni and sidak corrections for multiple comparisons. *In: Neil Salkind (Ed.) (2007). Encyclopedia of Measurement and Statistics, Thousand Oaks (CA): Sage*, 2007.

David Andrich. Application of a psychometric rating model to ordered categories wich are scored with successive integers. *Applied Psychological Measurement, Vol 2, No. 4, pages 581-594*, 1978.

David Andrich. Sufficiency and conditional estimation of person parameters in the polytomous rasch model. *Psychometrika, Vol. 75, No. 2, pages 292-308*, 2010.

Baptiste Auguie. *gridExtra: functions in Grid graphics*, 2012. URL `http://CRAN.R-project.org/package=gridExtra`. R package version 0.9.1.

Vic Barnett. *Comparative Statistical Inference.* Wiley, 1999.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-6.

Joseph Berkson. Minimum chi-square not maximum likelihood! *Annals of Statistics, Vol. 8, No.3, 457-487*, 1980.

Trevor G. Bond and Christine M. Fox. *Applying the Rasch Model : Fundamental Measurement in the Human Sciences.* Lawrence Erlbaum Associates Inc, 2001.

Denny Borsboom. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics.* Cambridge University Press, 2005.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science, Vol. 16, No. 3, pages 199-231*, 2001.

Winston Chang. *R Graphics Cookbook.* O'Reilly & Associates, 2012.

Jose M. Cortina. What is the coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology, Vol. 78, No. 1, 98-104*, 1993.

Penelope J. E. Davies, Walter B. Denny, Frima Fox Hofrichter, Joseph Jacobs, Ann M. Roberts, and David L. Simon. *Jansons's History of Art: The Western Tradition, Eigth Edition.* Pearson, 2011.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application.* Cambridge University Press, 1997.

Clemens Draxler. Comparison of maximum likelihood with conditional pairwise likelihood estimation of person parameters in the rasch model. (unpublished), 2014.

Hilarius Dreßen and Oliver Trabert. Unipark homepage, 28.06.2014, 5:45 p.m., 2014. URL `http://www.unipark.info`.

Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp, Springer, New York. ISBN 978-1-4614-6867-7.* Springer Science+Business Media, 2013.

Bradley Efron and Robert J. Tibshirani. *An introduction to the Bootstrap.* Springer Science+Business Media, 1993.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Exploratory and inferential analysis of benchmark experiments, 2008.

Gerhard H. Fischer and Ivo W. Molenaar. *Rasch Models: Foundations, Recent Developments, and Applications.* Springer Science+Business Media, 1995.

Cees A. W. Glas. Differential item functioning depending on general covariates. *In Lecture Notes in Statistics: Essays of Item Response Theory, 2001, Springer*, 2001.

Paul G. Gottschalk and John R. Dunn. The five parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry, 343, 54-65*, 2005.

Ronald K. Hambleton and Russell W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *ITEMS Instructional Topics in Educational Measurement*, 1993.

Ronald K. Hambleton, H. Swaminathan, and H. Jane Rogers. *Fundamentals of Item Response Theory.* Sage Publications, Inc., 1991.

Deborah Harris. Comparison of 1-, 2- and 3- parameter irt models. *Items, Instructional Topics in Educational Measurement*, 1989.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning, second edition.* Springer Science+Business Media, 2011.

Leonard Held and Daniel Sabanes Bove. *Applied Statistical Inference: Likelihood and Bayes.* Springer Science+Business Media, 2014.

Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software 28(8), 1-23*, 2008. URL `http://www.jstatsoft.org/v28/i08/`.

Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. *Conditional Inference Procedures in a Permutation Test Framework, Version 1.0-23*, 2013.

Hans Irtel. An extension of the concept of specific objectivity. *Psychometrika Vol.60, No.1, pages 115-118*, 1995.

Matthew S. Johnson. Marginal maximum likelihood estimation of item respone models in r. *Journal of Statistical Software, Volume 20, Issue 10*, 2007.

Karl-Heinz Jureit. Der denker (24cm), 2014. URL `http://royal-art.de/shop/kategorien/skulpturen/auguste-rodin/der-denker-24cm`.

Soetaert K. *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*, 2009. R-package version 1.6.

George Karabatsos and Stephen G. Walker. Adaptive-model bayesian nonparametric regression. *Electronic Journal of Statistics, Vol.6, pages 2038-2068*, 2012.

Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis, Volume 53, Issue 11, pages 3735-3745*, 2009.

Achim Klenke. *Wahrscheinlichkeitstheorie.* Springer Science+Business Media, 2006.

Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking.* Springer Science+Business Media New York, 2014. ISBN 978-1-4939-0317-7.

Svend Kreiner. Conditional pairwise person parameter estimates in rasch models. *Journal of Applied Measurement*, 2012.

Max Kuhn and Kjell Johnson. *Applied Predictive Modelling.* Springer Science+Business Media, 2013.

John M. Linacre. Rasch model estimation: Further topics. *Journal of applied measurement, 5(l), 95-110*, 2004.

P. Mair, R. Hatzinger, and Maier M.J. *eRm: Extended Rasch Modeling, R package version 0.15-4*, 2014. URL `http://erm.r-forge.r-project.org/`.

Geoff N. Masters. A rasch model for partial credit scoring. *Psychometrika-Vol 47 No. 2*, pages 149–174, 1982.

Angelo Mazza, Antonio Punzo, and Brian McGuire. Kernsmoothirt: An r package for kernel smoothing in item response theory. *Journal of Statistical Software, Volume 58, Issue 6.*, 2014.

Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics, Vol. 21, No. 15, pages 3301-3307*, 2005.

John F. Monahan. *Numerical Methods of Statistics, second edition.* Cambridge University Press, 2011.

Alberto Mura. *Philosophical Lectures in Probability.* Springer Science+Business Media, 2008.

Eiji Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement Vol. 16, No. 2, pp. 159-176*, 1992.

Erich Neuwirth. *RColorBrewer: ColorBrewer palettes*, 2011. URL `http://CRAN.R-project.org/package=RColorBrewer`. R package version 1.0-5.

OECD. Oecd homepage, 14 p.m., 31.07., 2014. URL `http://www.oecd.org/`.

Songthip Ounpraseuth, Shelly Y. Lensing, Horace J. Spencer, and Ralph L Kodell. Estimating misclassification error: a closer look at cross-validation based methods. *Ounpraseuth et al. BMC Research Notes*, 2012.

Mair P. and Hatzinger R. CML based estimation of extended rasch models with the eRm package in r. *Psychology Science, 49, 26-43*, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Mark D. Reckase. *Multidimensional Item Response Theory.* Springer Science+Business Media, 2009.

Dimitris Rizopoulos. ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25, 2006. URL `http://www.jstatsoft.org/v17/i05/`.

Auguste Rodin. The thinker, 2014. URL `http://www.artble.com/artists/auguste_rodin/sculpture/the_thinker`.

Jürgen Rost. *Lehrbuch Testtheorie Testkonstruktion.* Verlag Hans Huber, 1996.

S. J. Sheather and M. C. Jones. A reliable data-based bandwitdth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological), Volume 53, Issue 3, pages 683-690*, 1991.

David J. Sheskin. *Handbook of parametric & nonparametric statistical procedures, second edition.* Chapman & Hall/CRC, 2000.

Klaas Sijtsma and Brian W. Junker. Item response theory: Past performance, present developments and future expectations. *Behaviormetrika, Vol.33, No.1, pages 75-102*, 2006.

I. R. Silva and R. M. Assuncao. Monte carlo test under general conditions: Power and number of simulations. *Journal of Statistical Planning and Inference*, 2011.

R. Stiratelli, N. Laird, and J. H. Ware. Random effects models for serial observations with binary response. *Biometrics, 40, 961-971*, 1984.

Holm Tetens. *Wissenschaftstheorie.* C.H.Beck, 2013.

David Thissen. A taxonomy of item response models. *Psychometrika, Vol.51, No. 4, pages 567-577*, 1986.

Gerhard Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43, pages 49-55*, 1990.

Gerhard Tutz. *Sequential models for ordered responses, pages 139-152 in Handbook for Modern item response theory.* Springer Science & Business Media, 1997.

Gerhard Tutz and Gunther Schauberger. A penalty approach to differential item functioning in rasch models, 2012.

N.D. Verhelst and H.H.F.M. Verstralen. Some considerations on the partial credit model. *Psicologica 29, pages 229-254*, 2008.

Wenhao Wang. Are all item response functions monotonically increasing? *Dissertation in department of psychology and Research in Education and the Graduate Faculty of the University of Kansas*, 2012.

Thomas Albert Warm. Weighted likelihood estimation of ability in item response theory with tests of finite length, 1985.

PD Dr. Christian Heumann (web interface). Institute of statistics at lmu, 2014. URL `http://www.stat.uni-muenchen.de`.

Susanne Weber and Sandra Trost. Ludwig-maximilians-universität München; Fakultät für Betriebswirtschaft; Institut für Wirtschaftspädagogik; Geschwister-Scholl-Platz 1; 80539 München; Germany, 2014.

Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

Hadley Wickham. *scales: Scale functions for graphics.*, 2014. URL `http://CRAN.R-project.org/package=scales`. R package version 0.2.4.

Cha Zhang and Yunqian Ma. *Ensemble Machine Learning*. Springer Science+Business Media, 2012.

A. Appendix

A.1. Simulation PCM

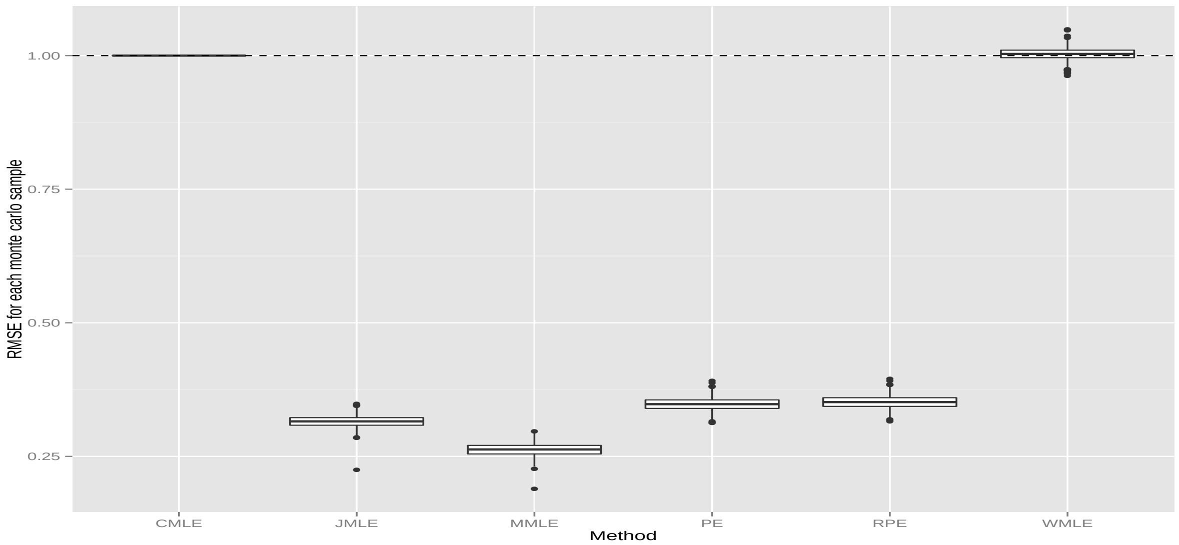

Figure 45: PCM: RMSE calculated for each Monte Carlo sample in scenario 2

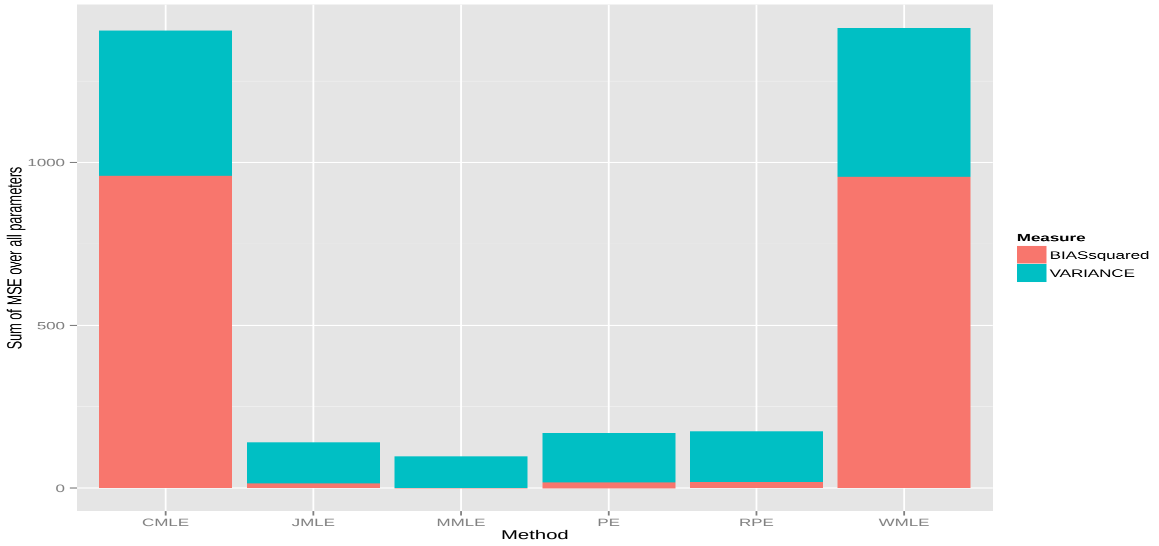

Figure 46: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 2

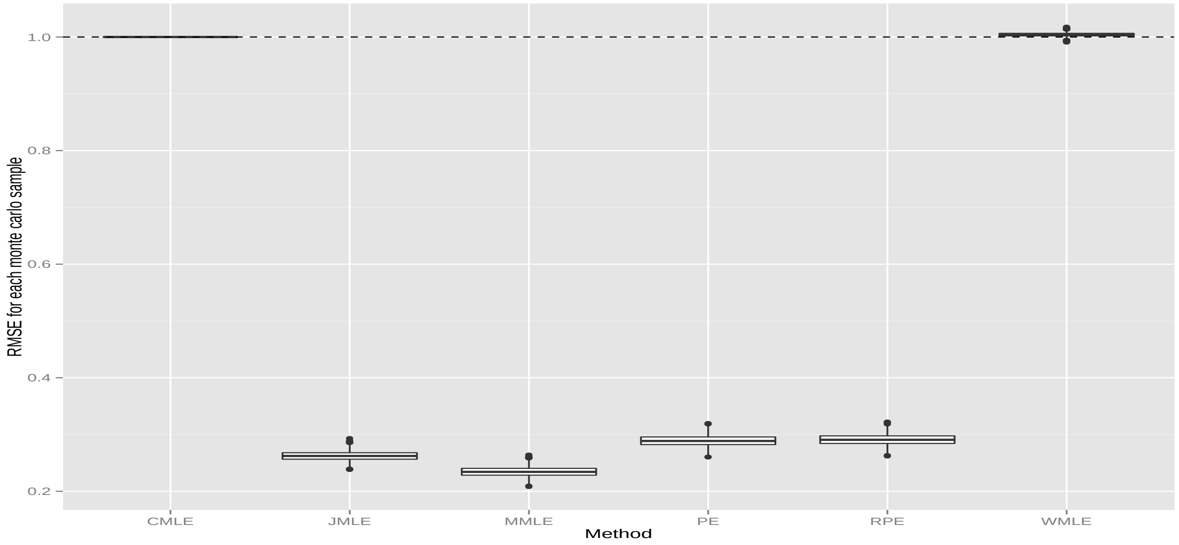

Figure 47: PCM: RMSE calculated for each Monte Carlo sample in scenario 3

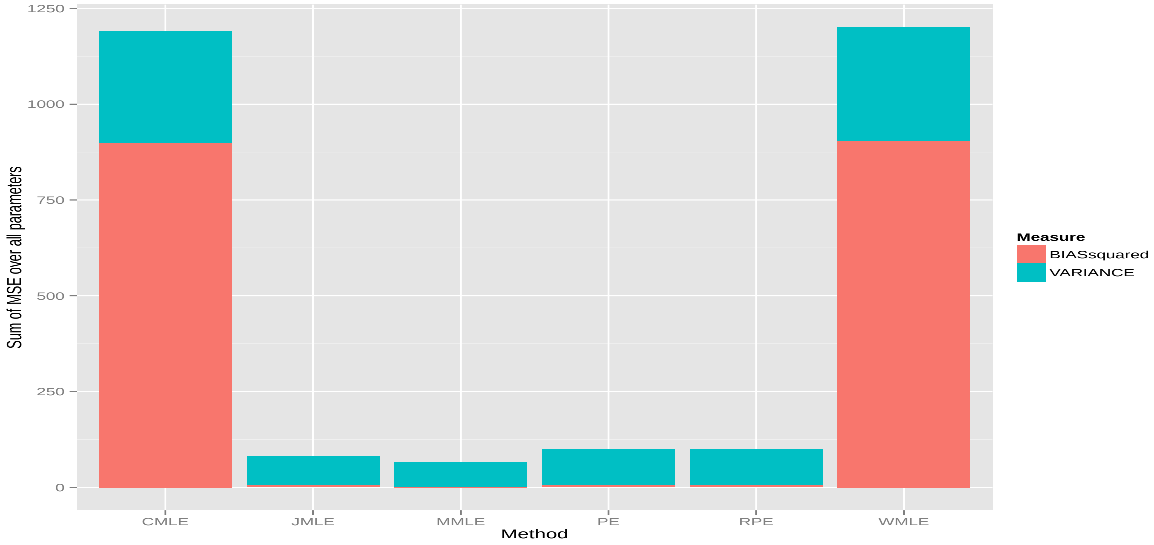

Figure 48: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 3

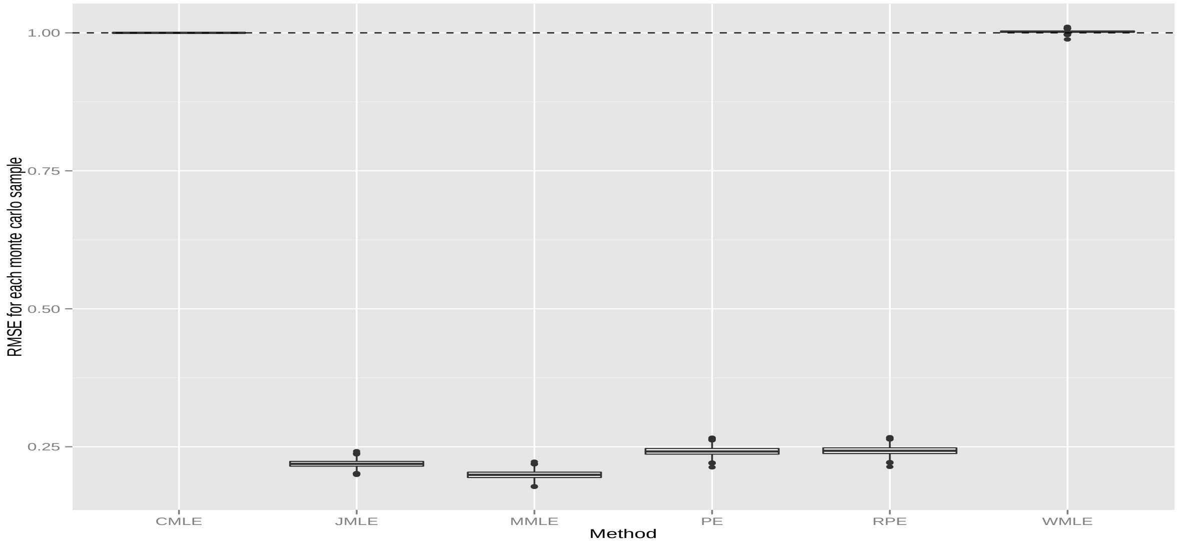

Figure 49: PCM: RMSE calculated for each Monte Carlo sample in scenario 4

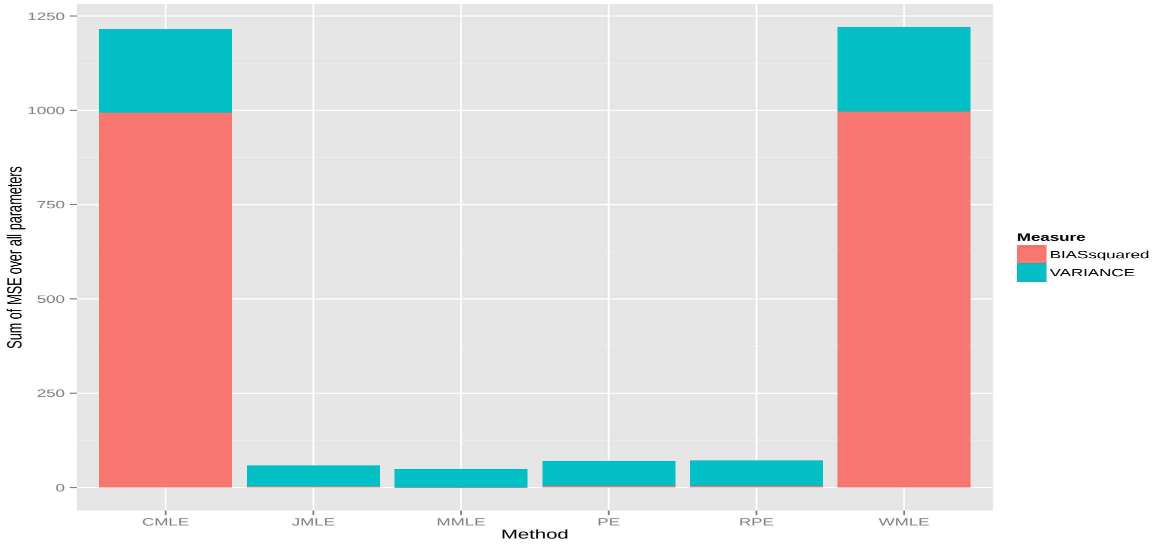

Figure 50: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 4

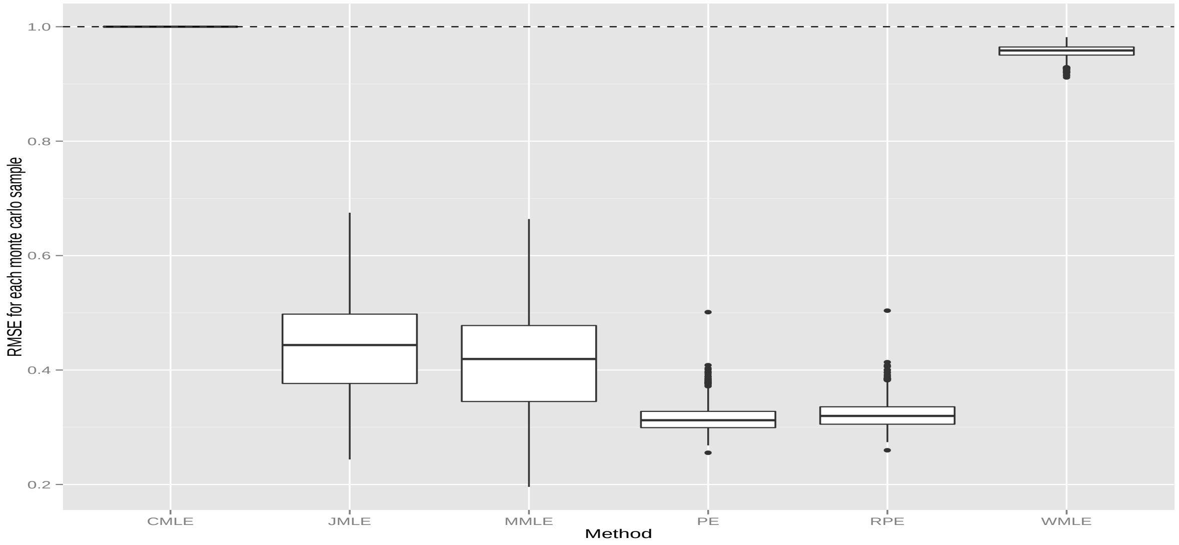

Figure 51: PCM: RMSE calculated for each Monte Carlo sample in scenario 7

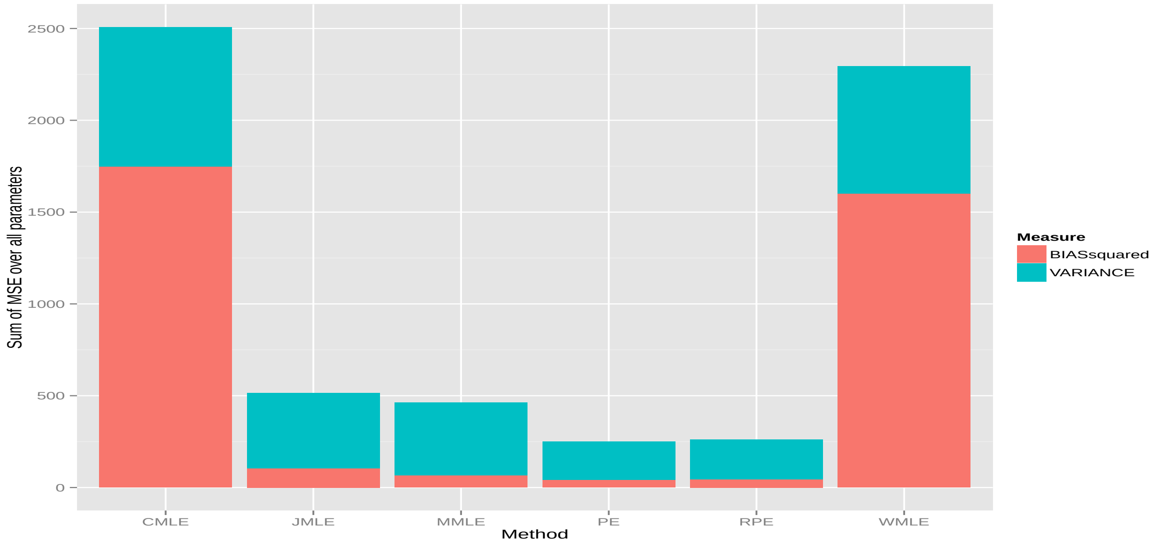

Figure 52: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 7

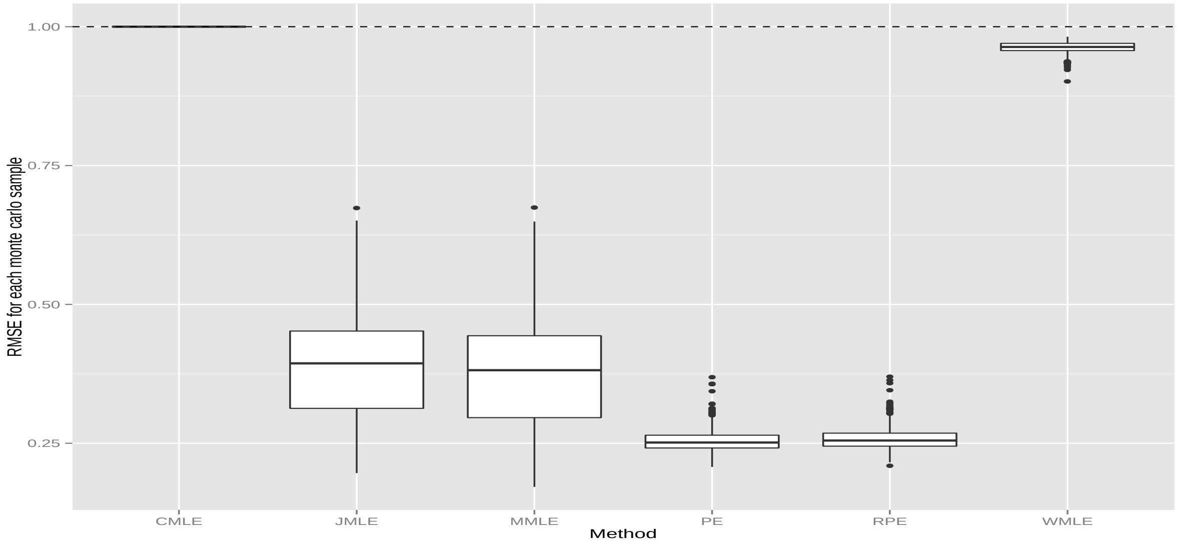

Figure 53: PCM: RMSE calculated for each Monte Carlo sample in scenario 8

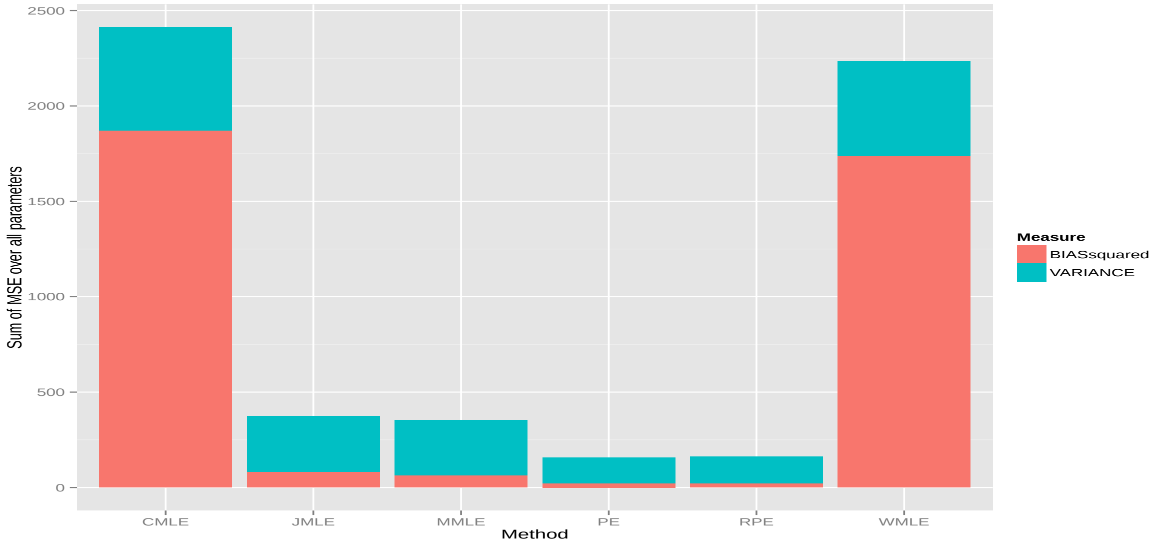

Figure 54: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 8

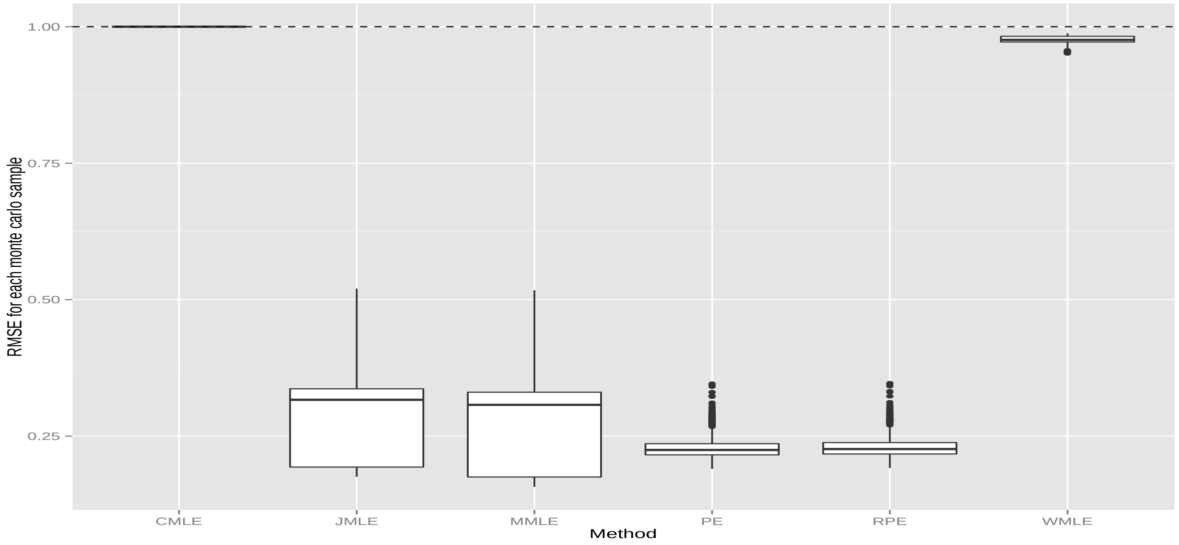

Figure 55: PCM: RMSE calculated for each Monte Carlo sample in scenario 9

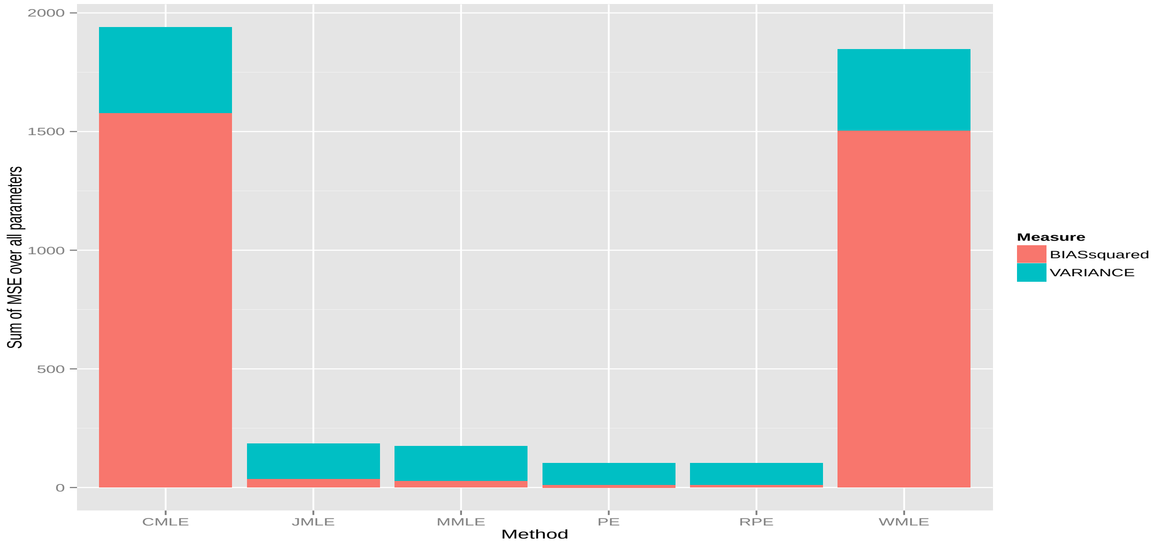

Figure 56: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 9

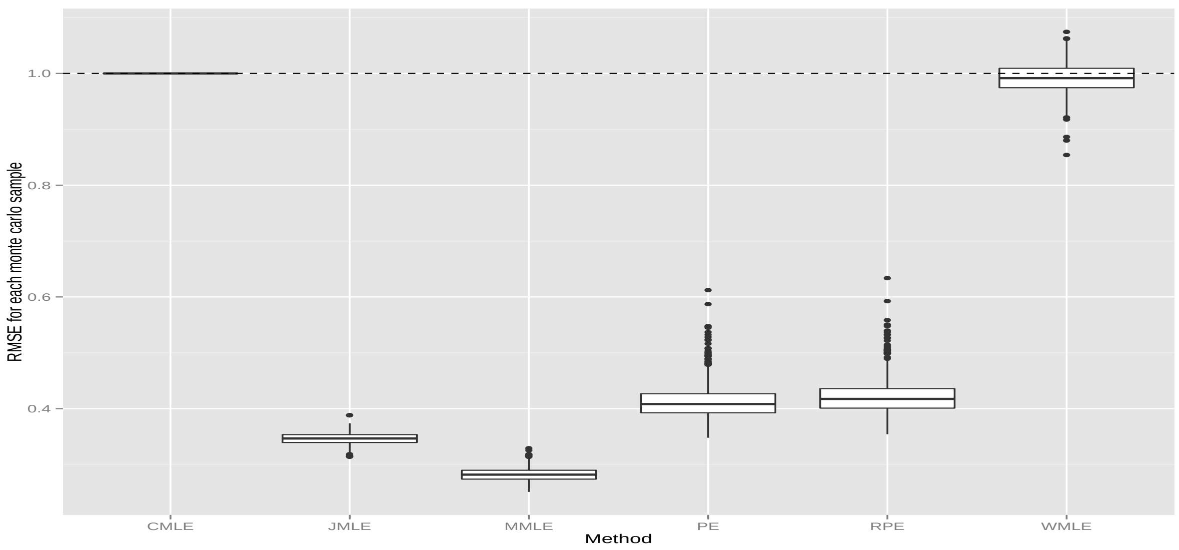

Figure 57: PCM: RMSE calculated for each Monte Carlo sample in scenario 12

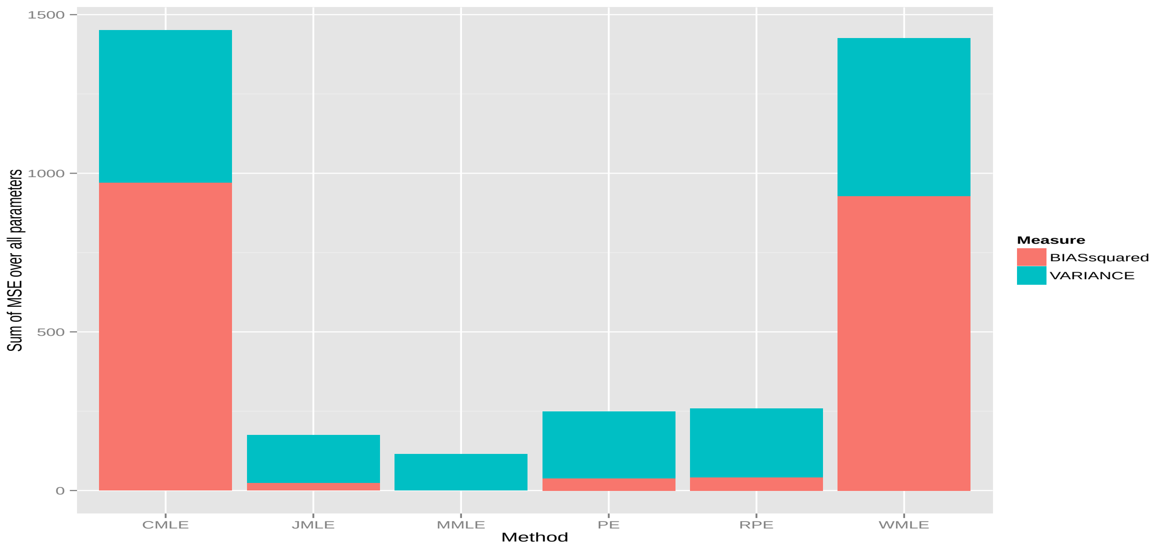

Figure 58: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 12

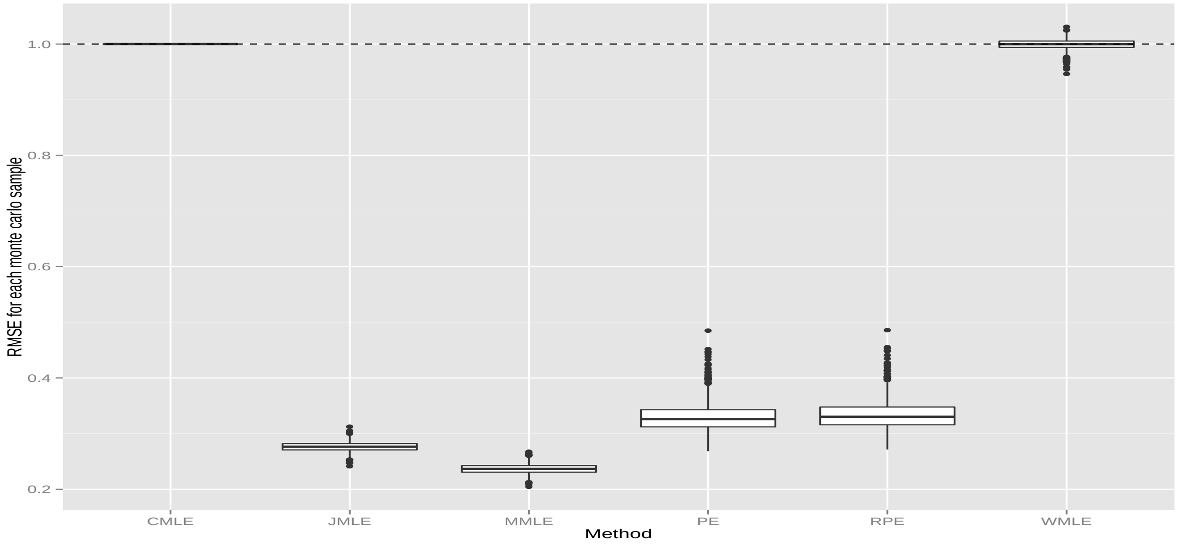

Figure 59: PCM: RMSE calculated for each Monte Carlo sample in scenario 13

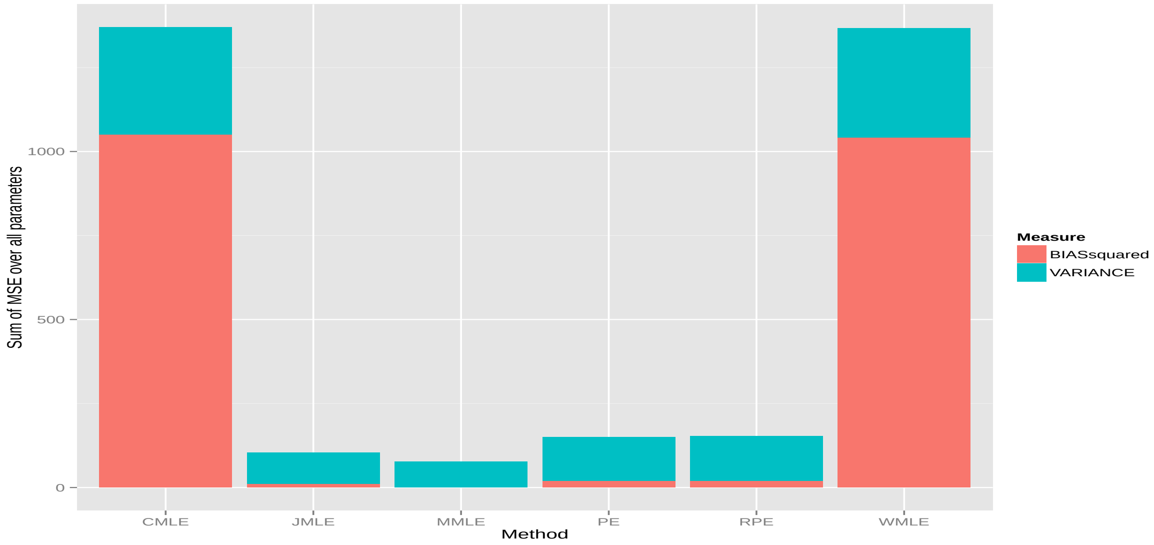

Figure 60: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 13

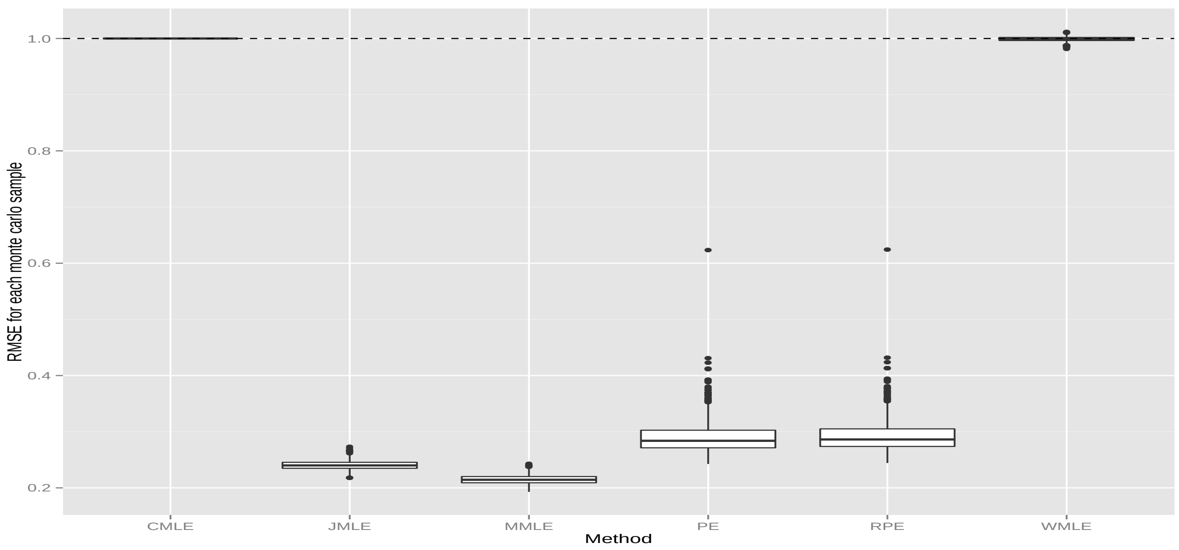

Figure 61: PCM: RMSE calculated for each Monte Carlo sample in scenario 14

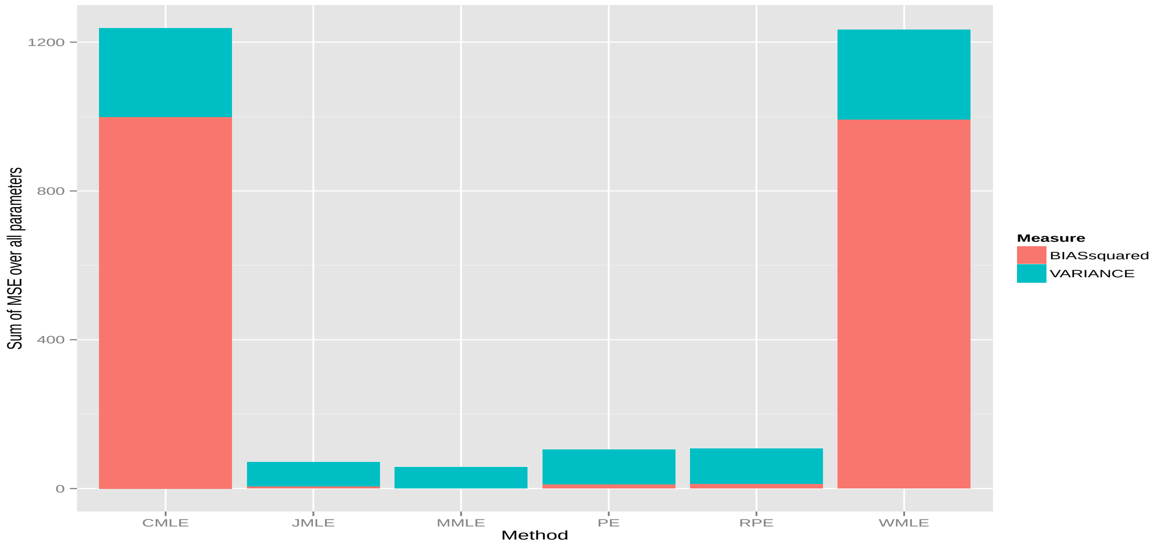

Figure 62: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 14

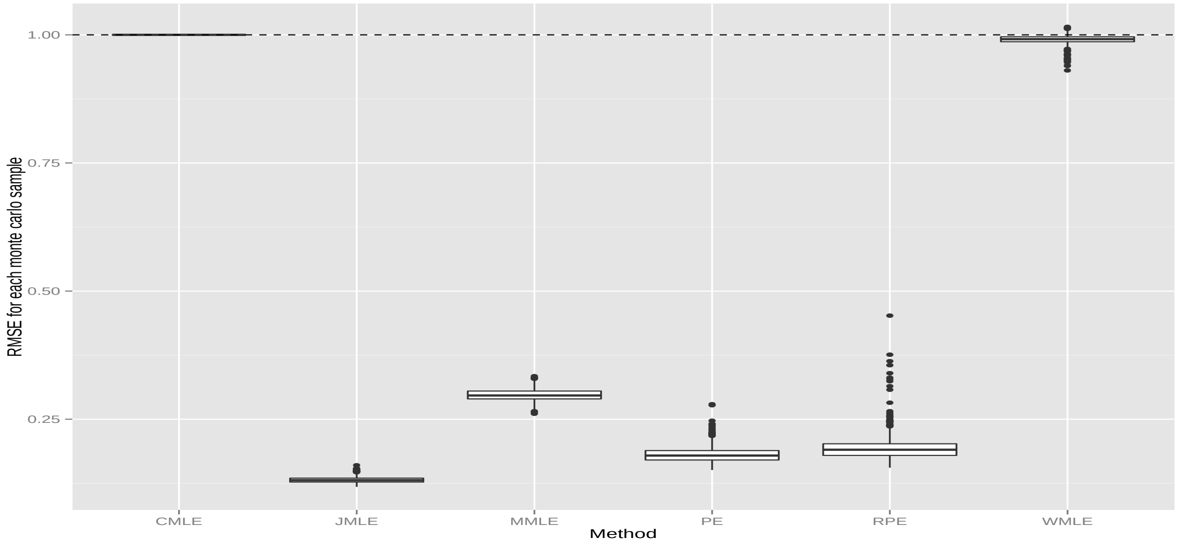

Figure 63: PCM: RMSE calculated for each Monte Carlo sample in scenario 17

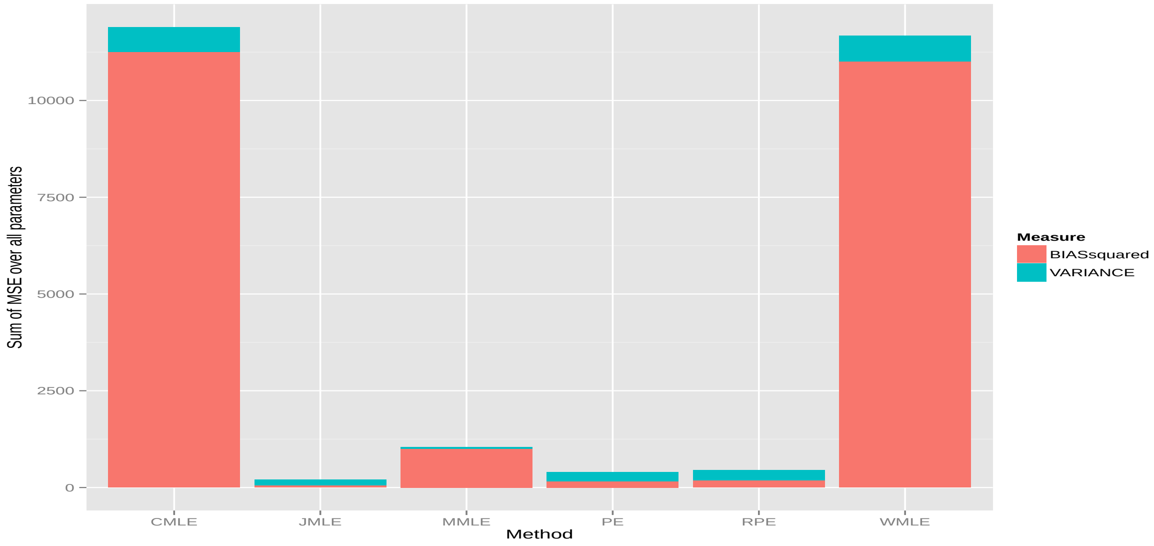

Figure 64: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 17

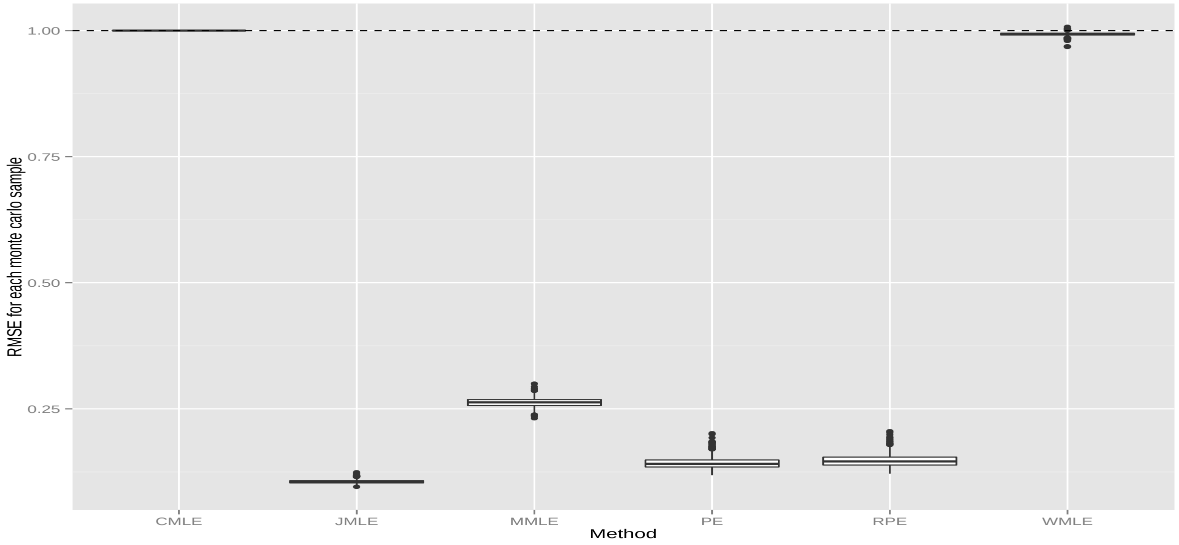

Figure 65: PCM: RMSE calculated for each Monte Carlo sample in scenario 18

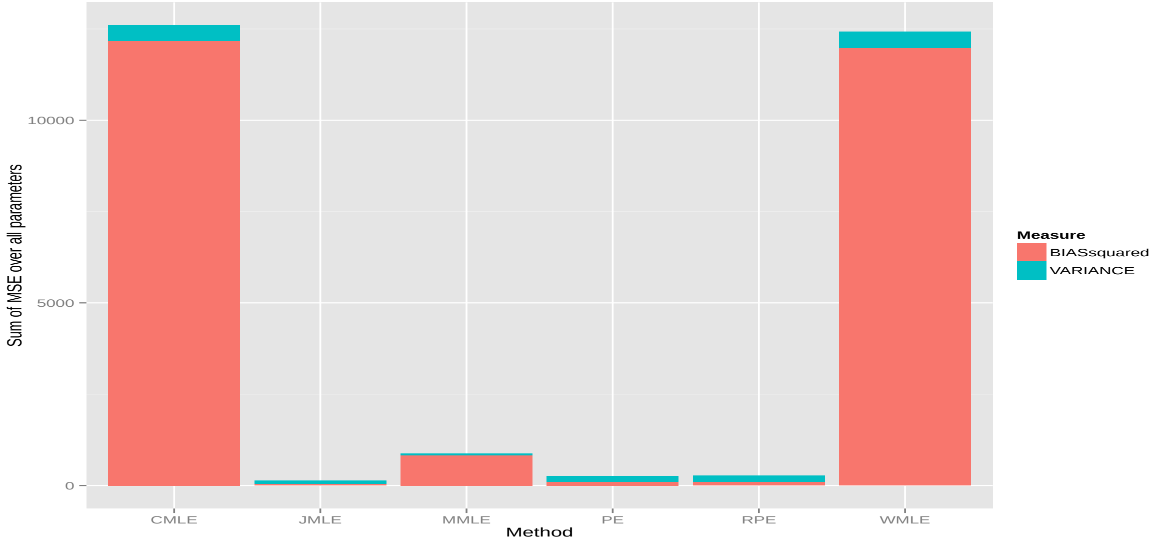

Figure 66: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 18

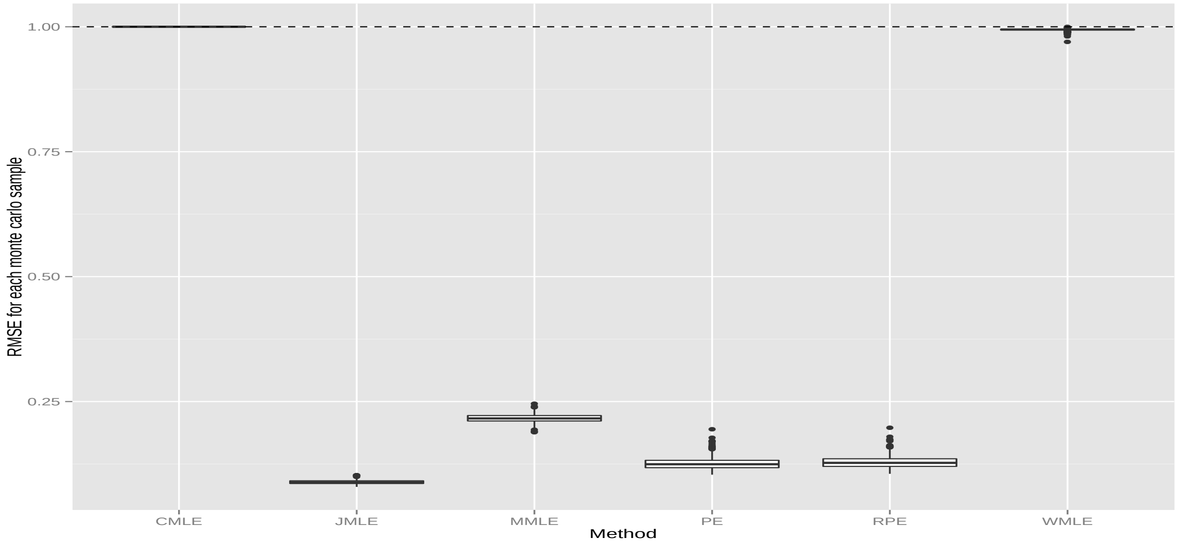

Figure 67: PCM: RMSE calculated for each Monte Carlo sample in scenario 19

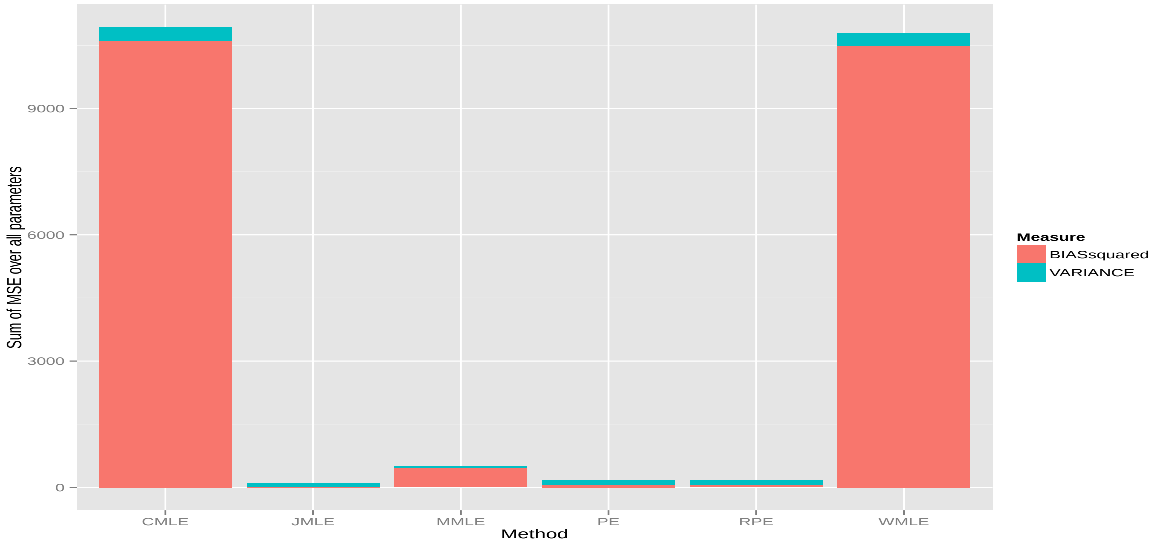

Figure 68: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 19

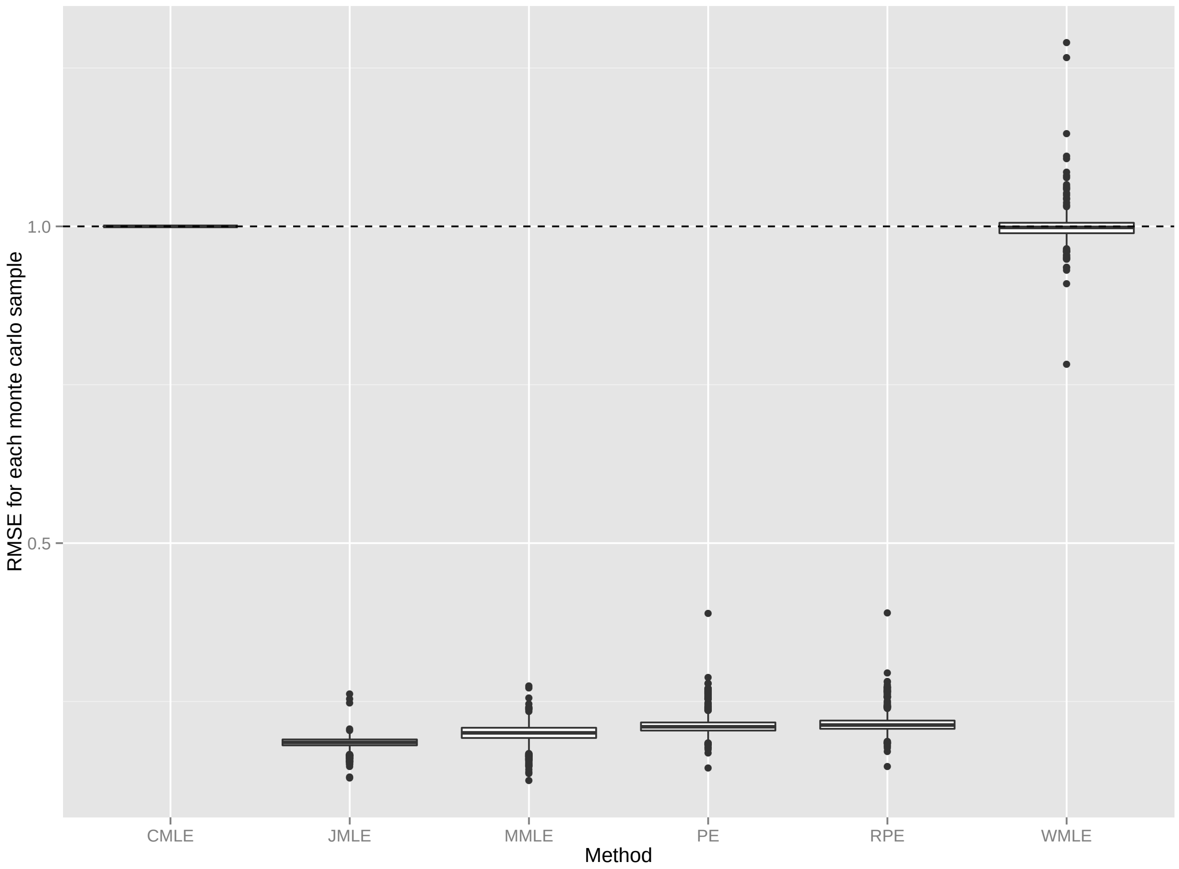

Figure 69: PCM: RMSE calculated for each Monte Carlo sample in scenario 22

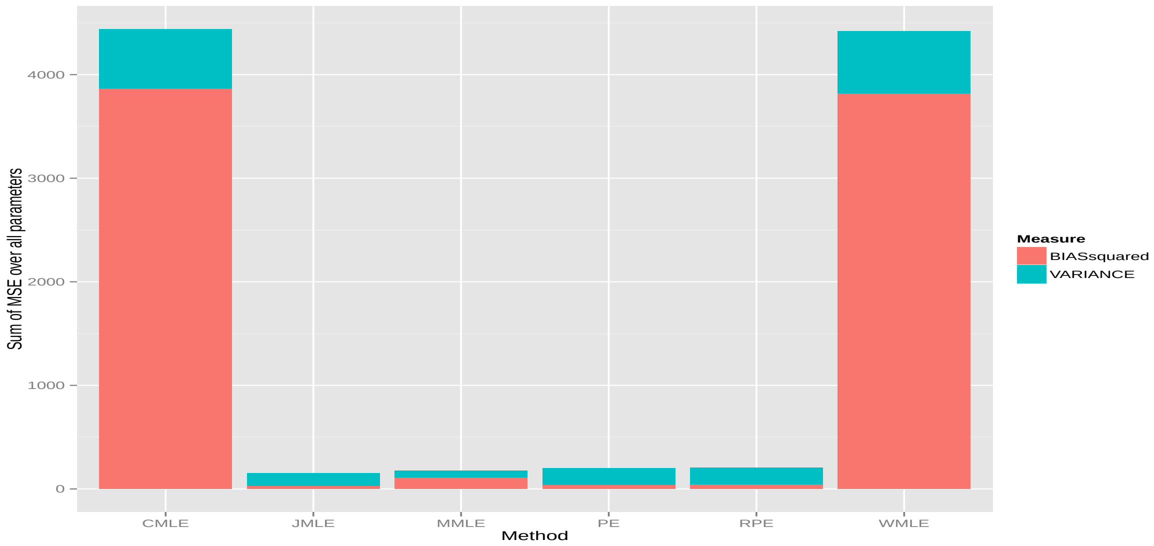

Figure 70: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 22

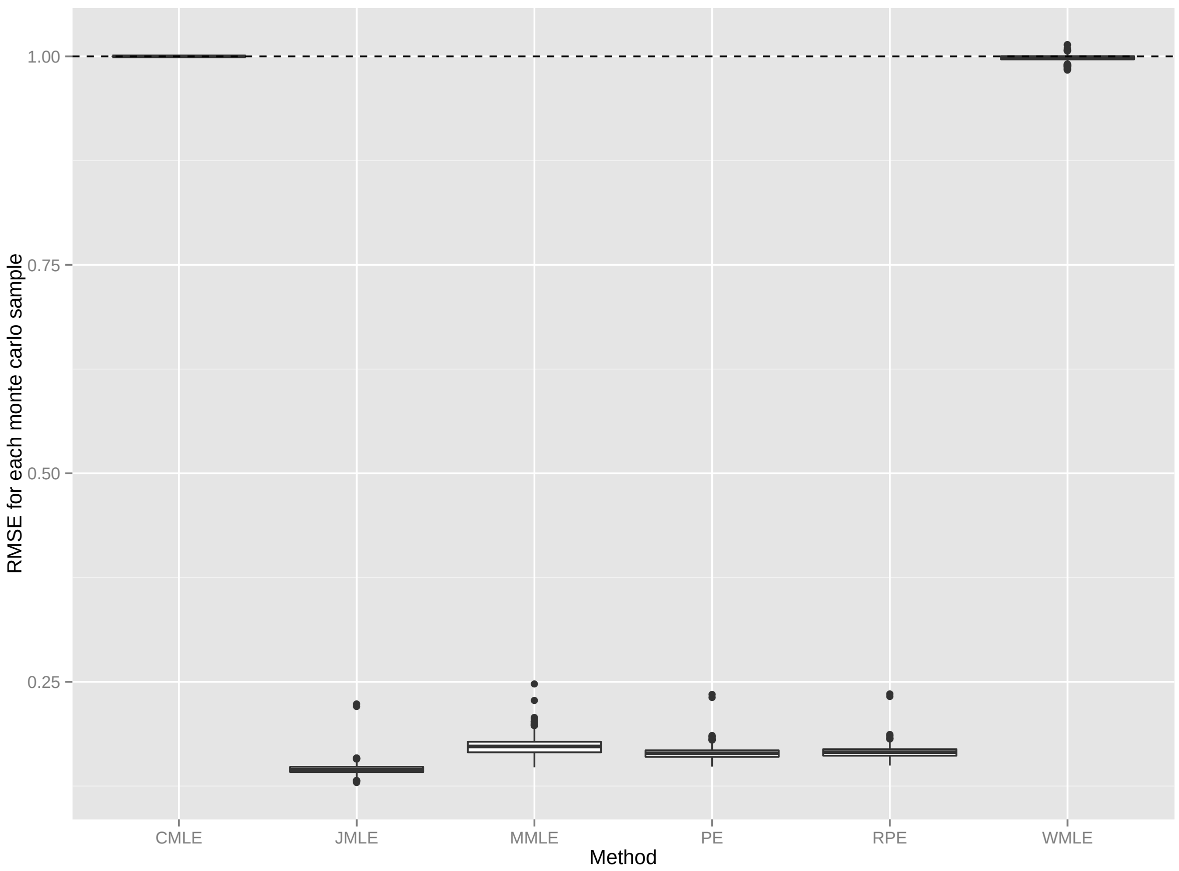

Figure 71: PCM: RMSE calculated for each Monte Carlo sample in scenario 23

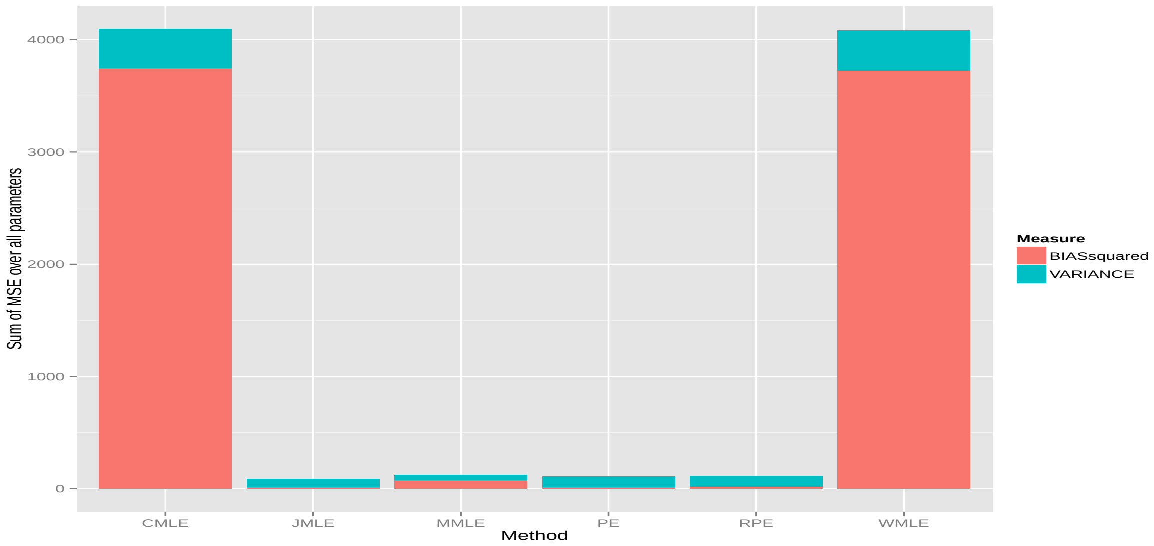

Figure 72: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 23

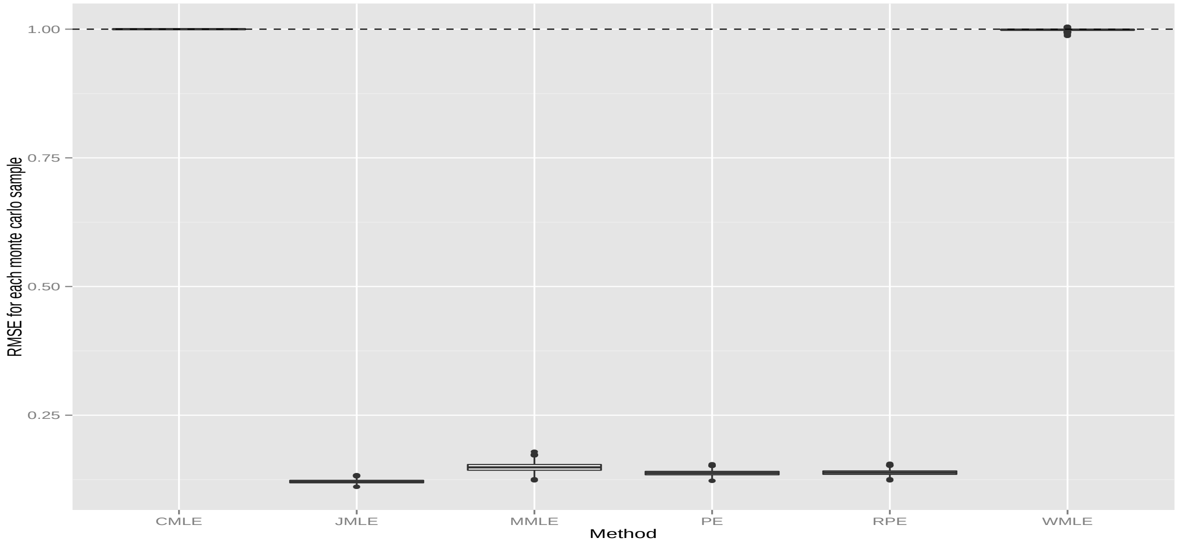

Figure 73: PCM: RMSE calculated for each Monte Carlo sample in scenario 24

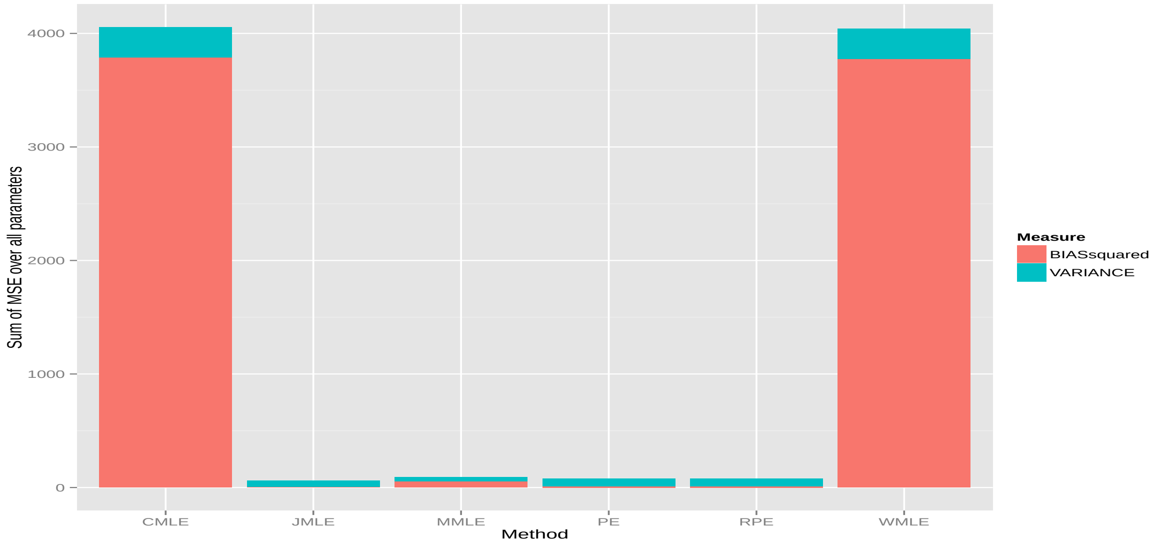

Figure 74: PCM: Sum of MSE, squared Biases, Variances over all person parameters in scenario 24

A.2. Simulation of predictive deviance density curves

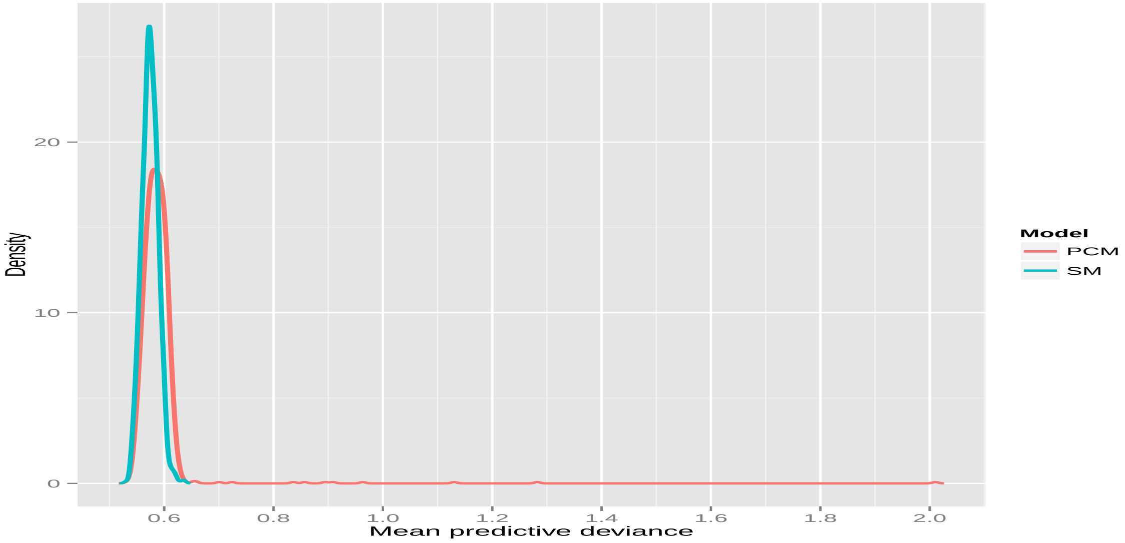

Figure 75: Density estimate of mean predictive deviance within true model PCM

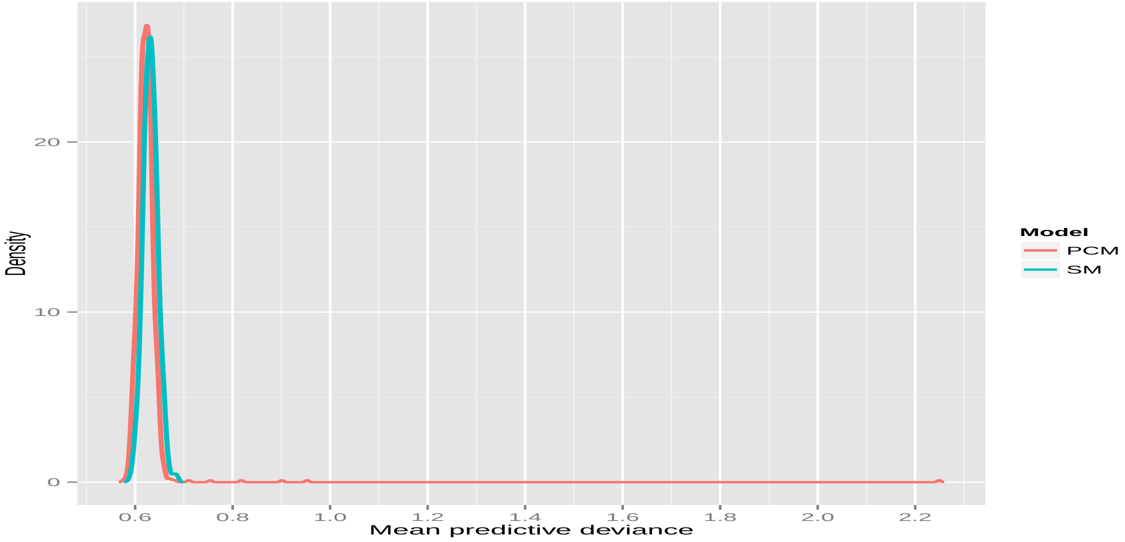

Figure 76: Density estimate of mean predictive deviance within true model SM

A.3. Estimated category probability curves of exam data for other items than 2, 11, 15

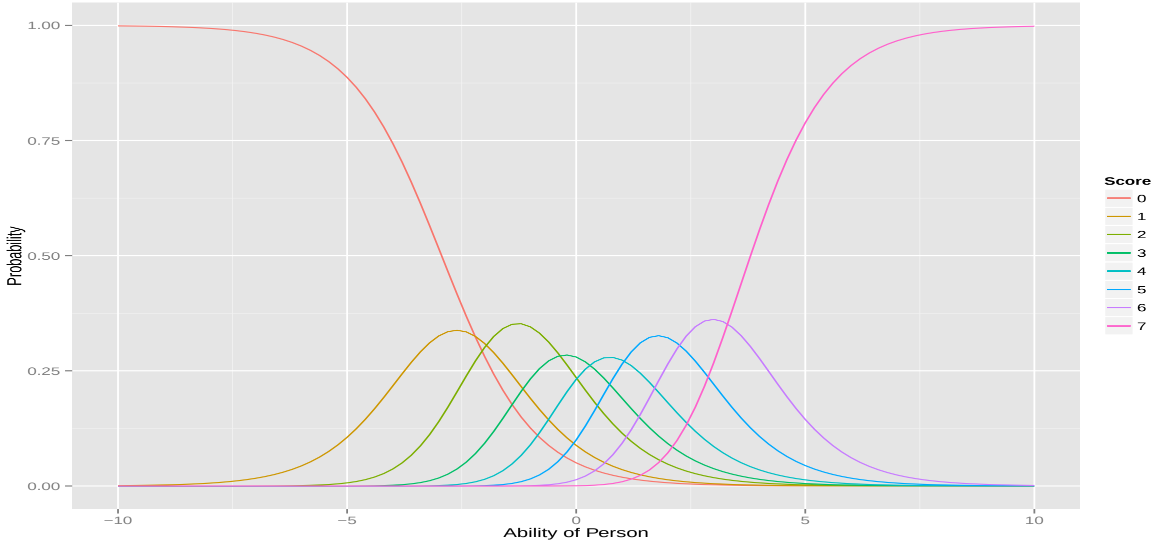

Figure 77: SM: Category probability curves for item 1 in exam data

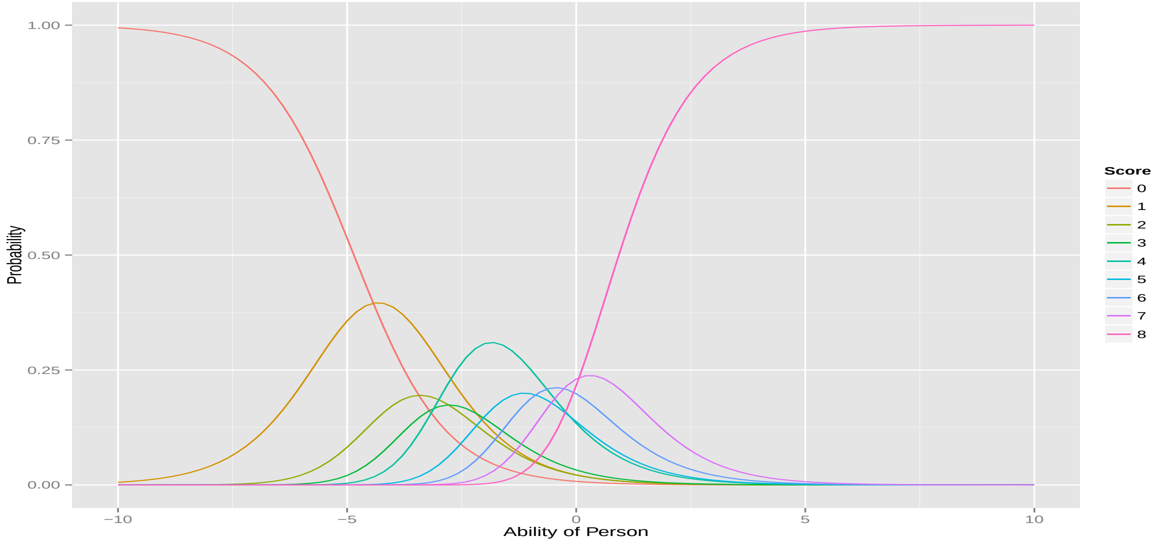

Figure 78: SM: Category probability curves for item 3 in exam data

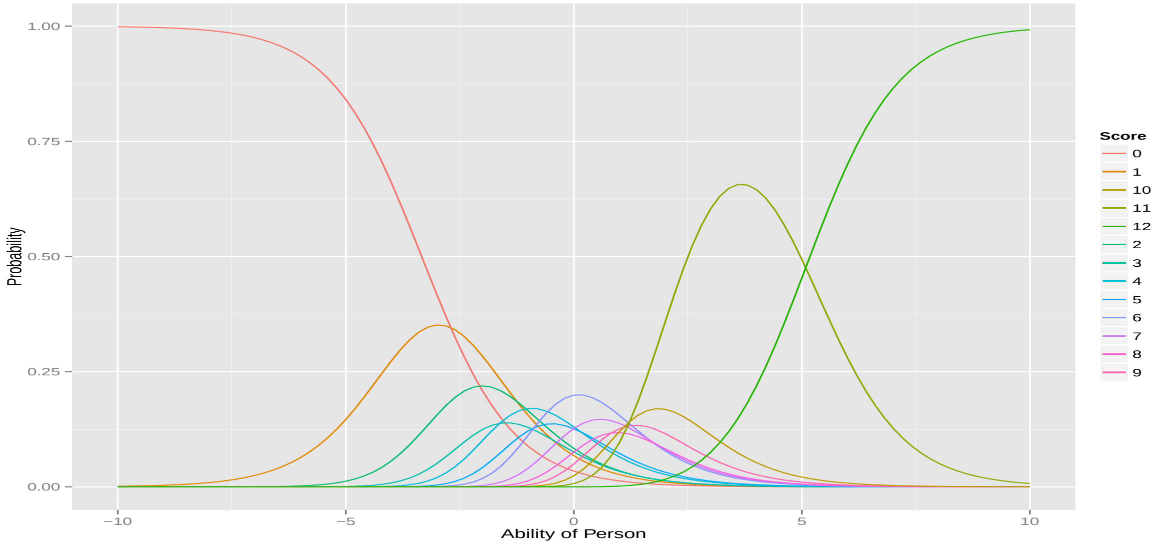

Figure 79: SM: Category probability curves for item 4 in exam data

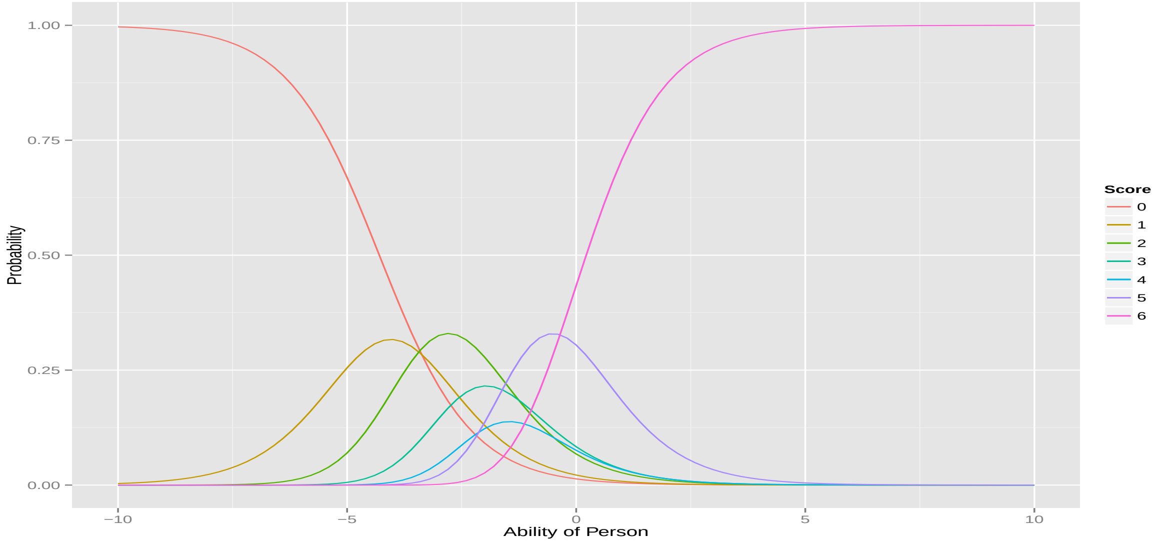

Figure 80: SM: Category probability curves for item 5 in exam data

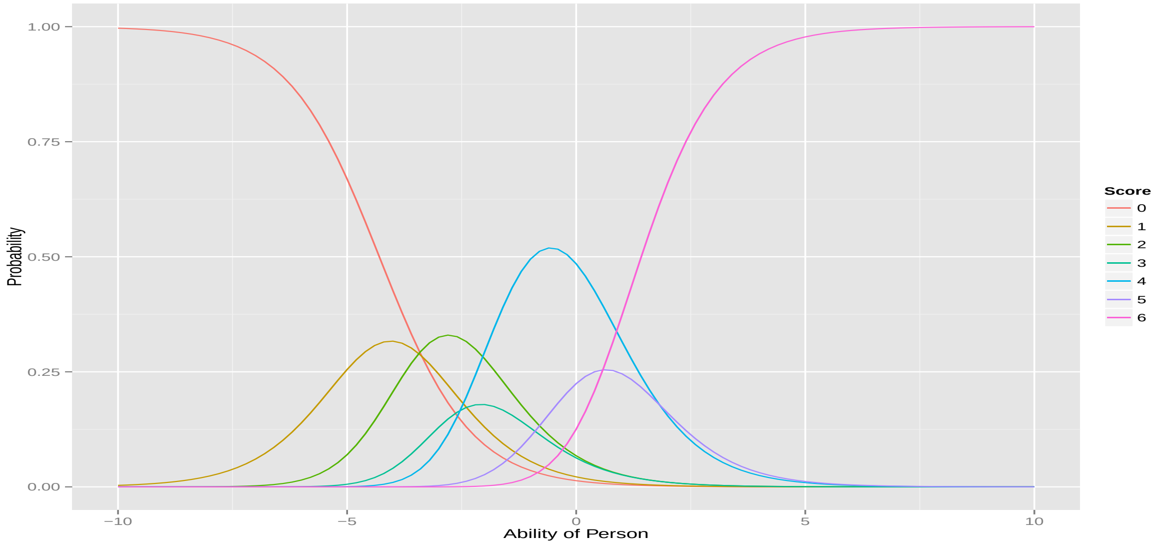

Figure 81: SM: Category probability curves for item 6 in exam data

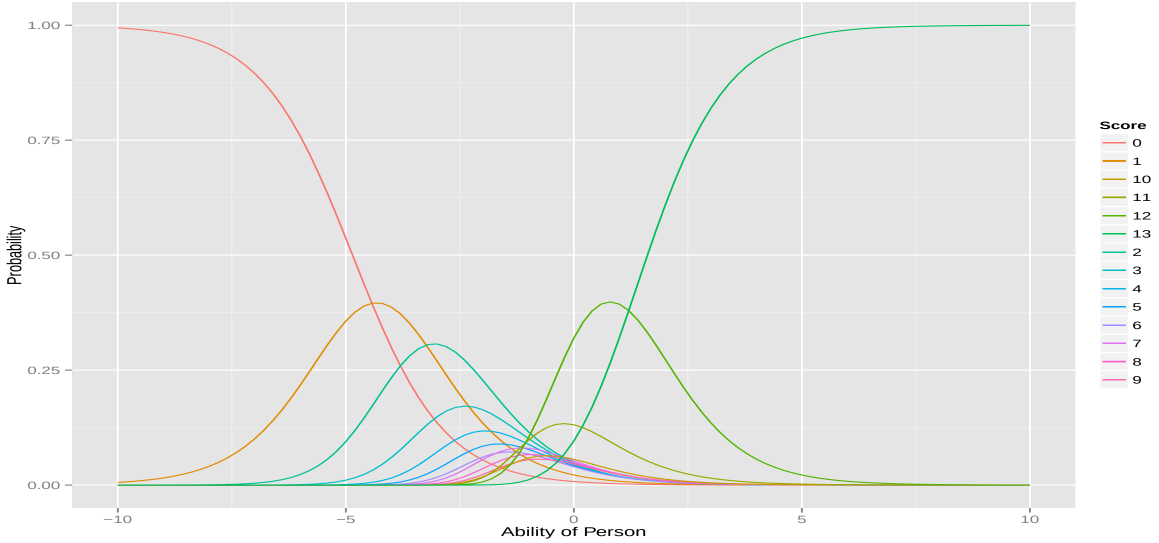

Figure 82: SM: Category probability curves for item 7 in exam data

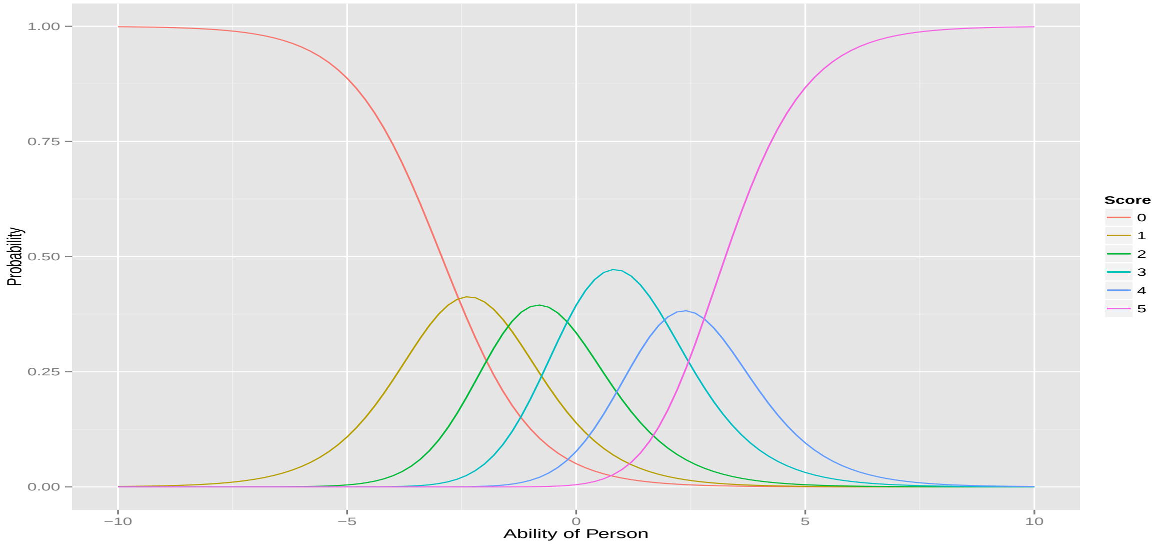

Figure 83: SM: Category probability curves for item 8 in exam data

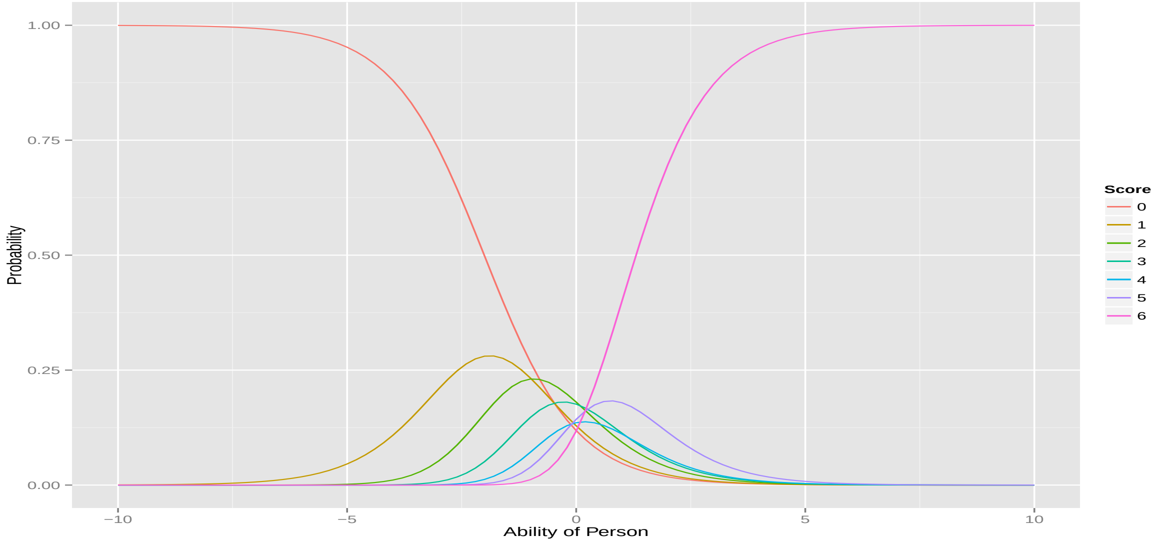

Figure 84: SM: Category probability curves for item 9 in exam data

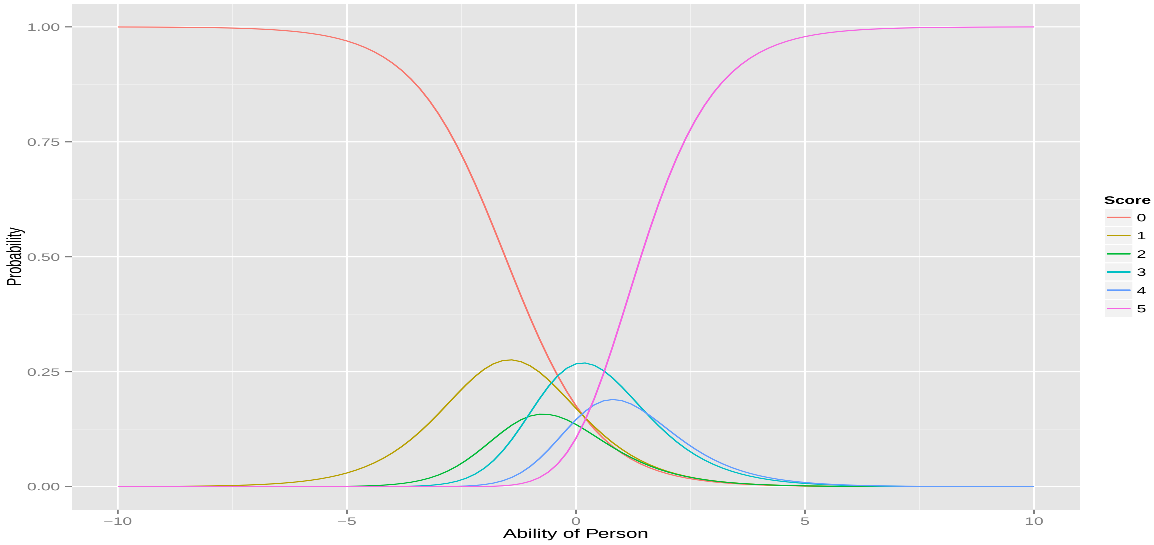

Figure 85: SM: Category probability curves for item 10 in exam data

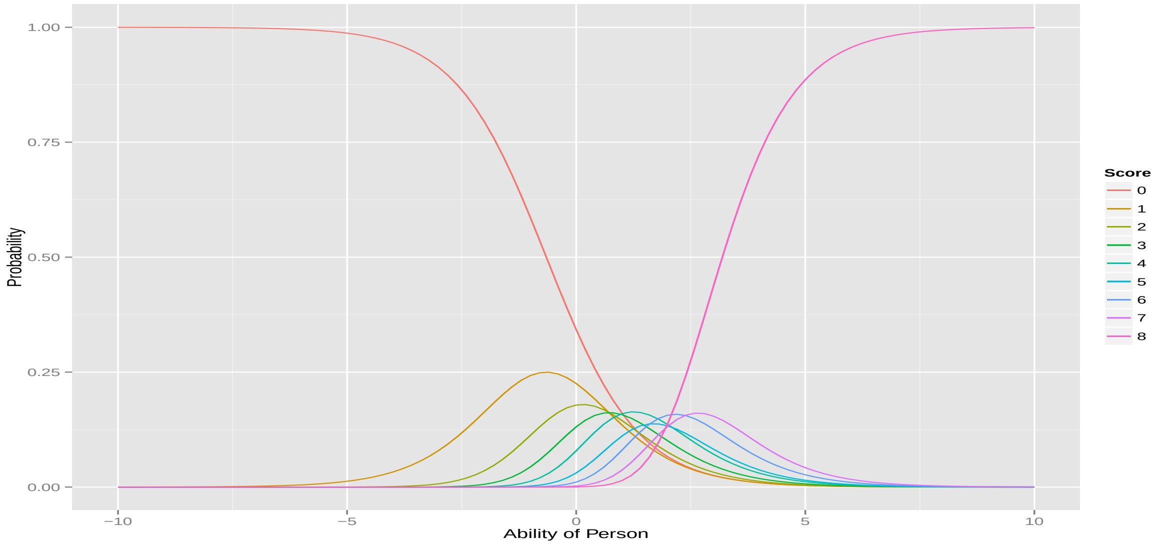

Figure 86: SM: Category probability curves for item 12 in exam data

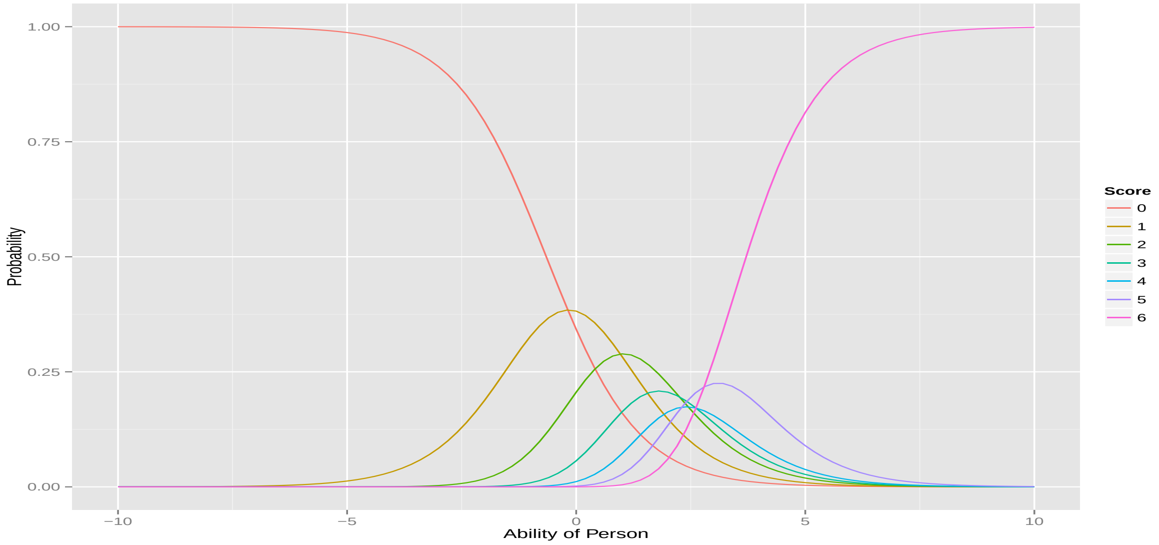

Figure 87: SM: Category probability curves for item 13 in exam data

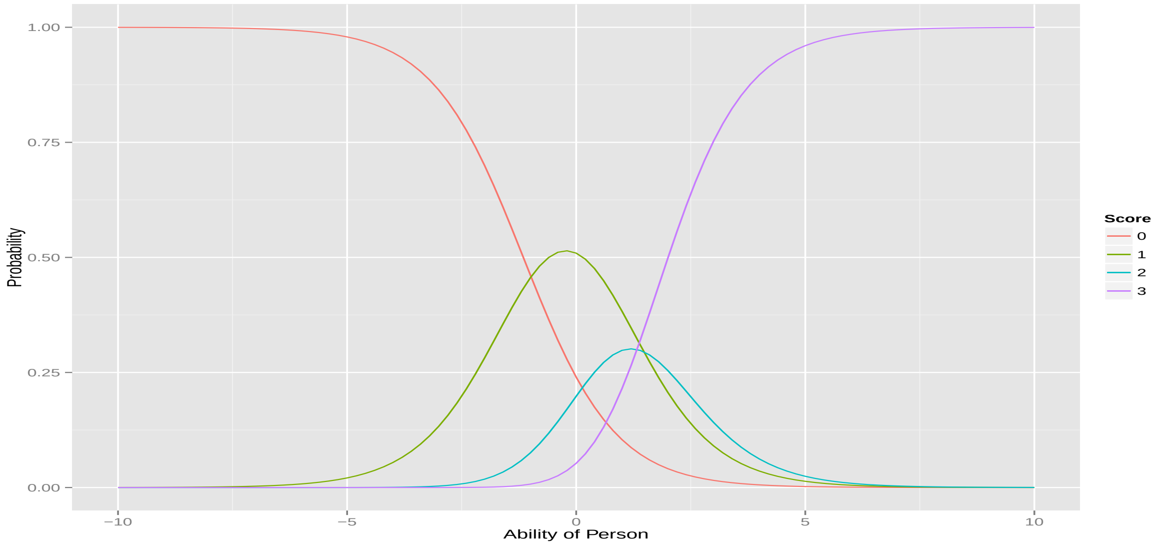

Figure 88: SM: Category probability curves for item 14 in exam data

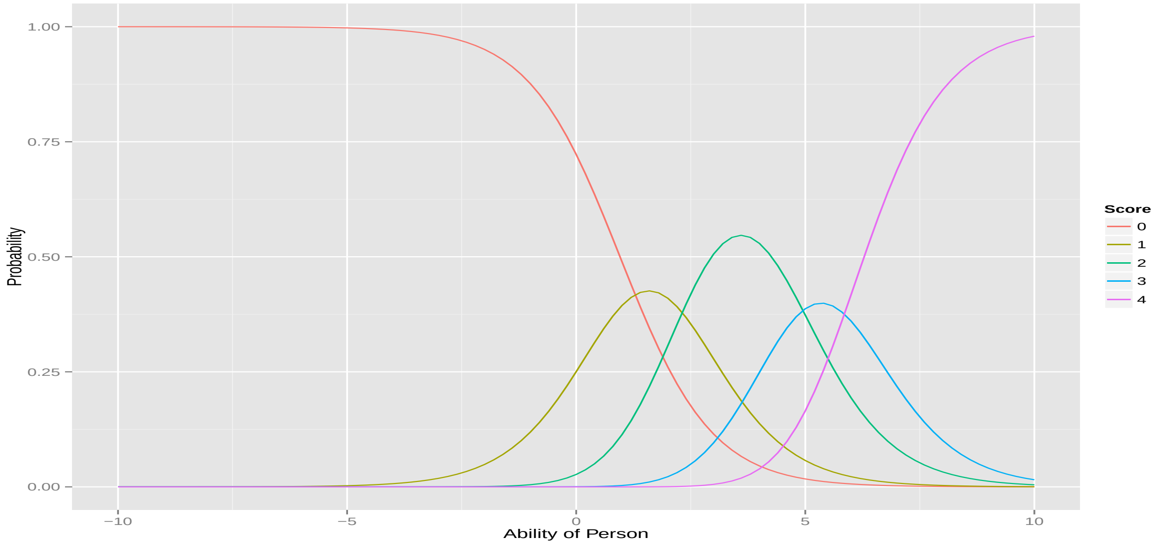

Figure 89: SM: Category probability curves for item 16 in exam data

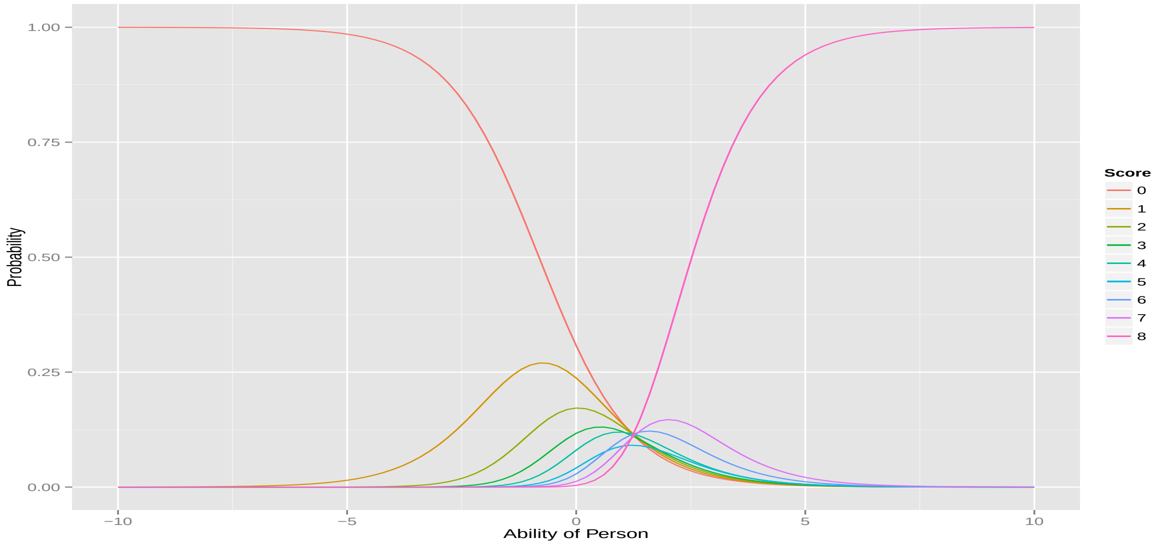

Figure 90: SM: Category probability curves for item 17 in exam data

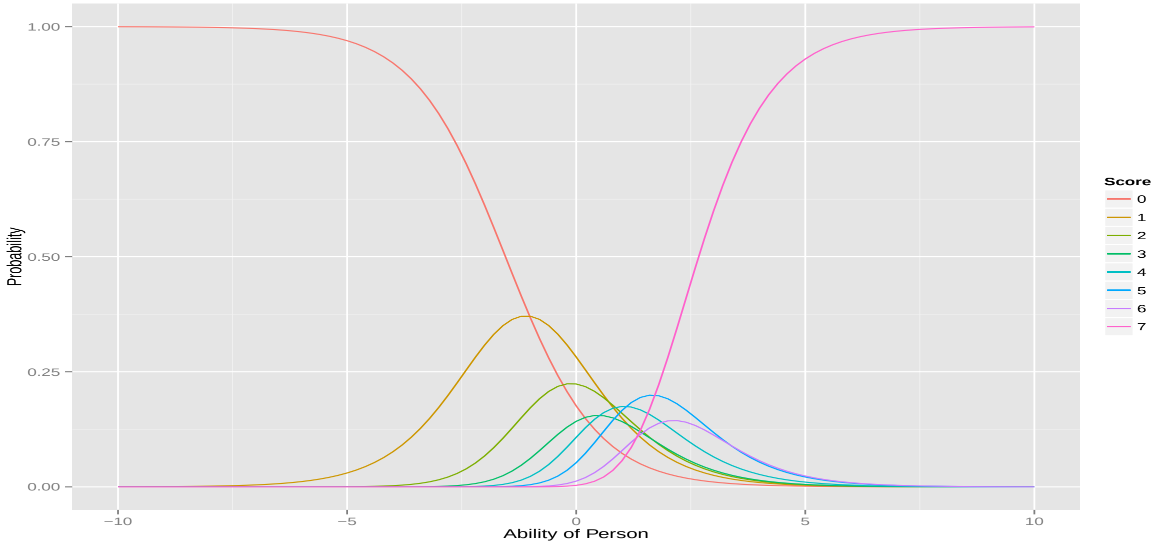

Figure 91: SM: Category probability curves for item 18 in exam data

A.4. Tables: Expected proportions of max score for exam data with decreasing order of difficulty

Table 7: MMLE: Expected proportions of max score for each item of exam data

Rank	Item	Expected Proportion
1	16	7.63 %
2	15	10.86 %
3	13	16.97 %
4	2	17.22 %
5	12	19.09 %
6	17	21.85 %
7	18	27.82 %
8	14	35.46 %
9	1	41.59 %
10	4	42.24 %
11	8	46.46 %
12	10	47.02 %
13	11	48.18 %
14	9	49.67 %
15	6	69.30 %
16	7	70.27 %
17	3	73.97 %
18	5	80.57 %

Table 8: JMLE: Expected proportions of max score for each item of exam data

Rank	Item	Expected Proportion
1	16	16.28 %
2	15	25.03 %
3	2	25.49 %
4	13	37.70 %
5	12	43.07 %
6	14	54.53 %
7	7	58.51 %
8	17	58.58 %
9	18	58.63 %
10	8	59.06 %
11	1	61.76 %
12	4	62.65 %
13	11	64.38 %
14	10	68.00 %
15	5	75.13 %
16	6	75.92 %
17	9	77.38 %
18	3	78.71 %

Table 9: CMLE: Expected proportions of max score for each item of exam data

Rank	Item	Expected Proportion
1	16	3.02 %
2	15	4.73 %
3	2	6.94 %
4	13	7.22 %
5	12	8.09 %
6	17	14.29 %
7	18	14.75 %
8	14	16.35 %
9	10	22.71 %
10	11	23.07 %
11	4	27.74 %
12	8	28.80 %
13	1	29.09 %
14	9	30.23 %
15	7	30.85 %
16	6	44.36 %
17	3	52.08 %
18	5	52.80 %

A.5. SM: CMLE of exam data

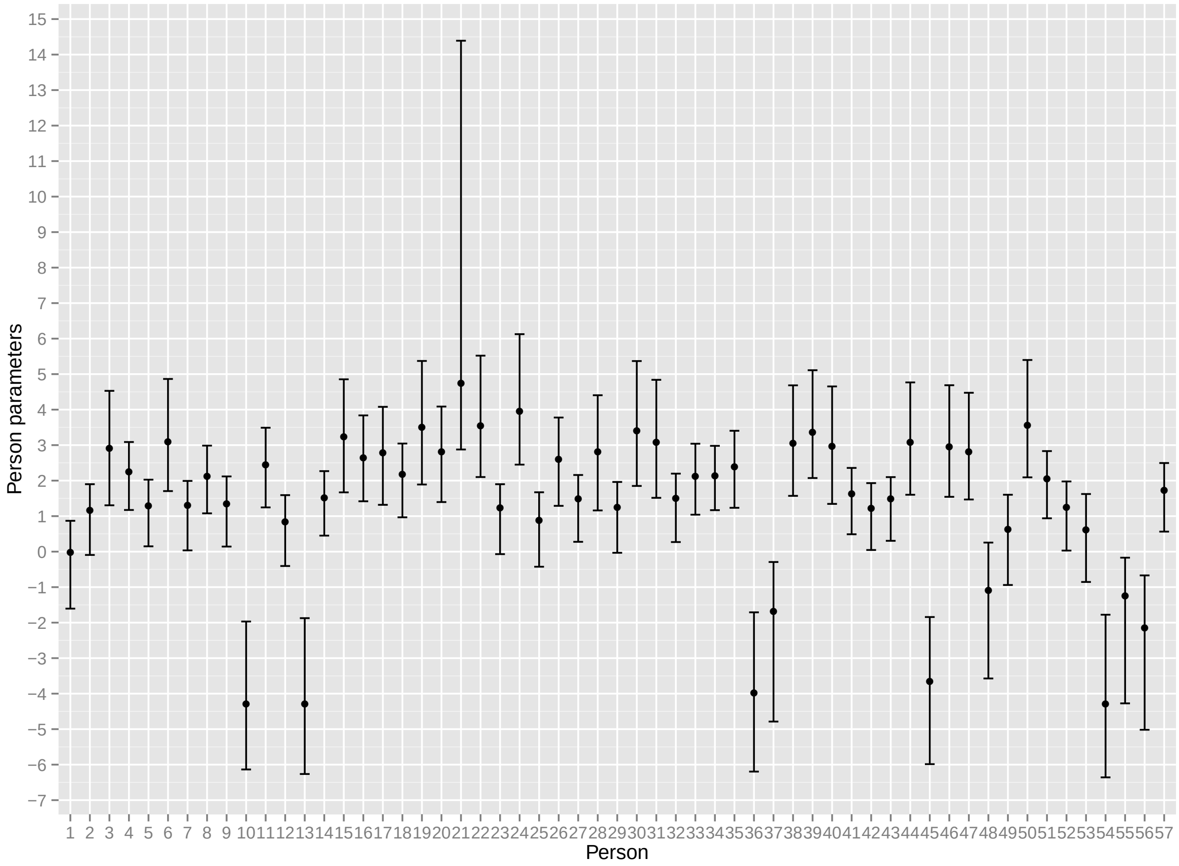

Figure 92: SM: CMLE of person parameters with parametric bootstrap percentile CI of exam data

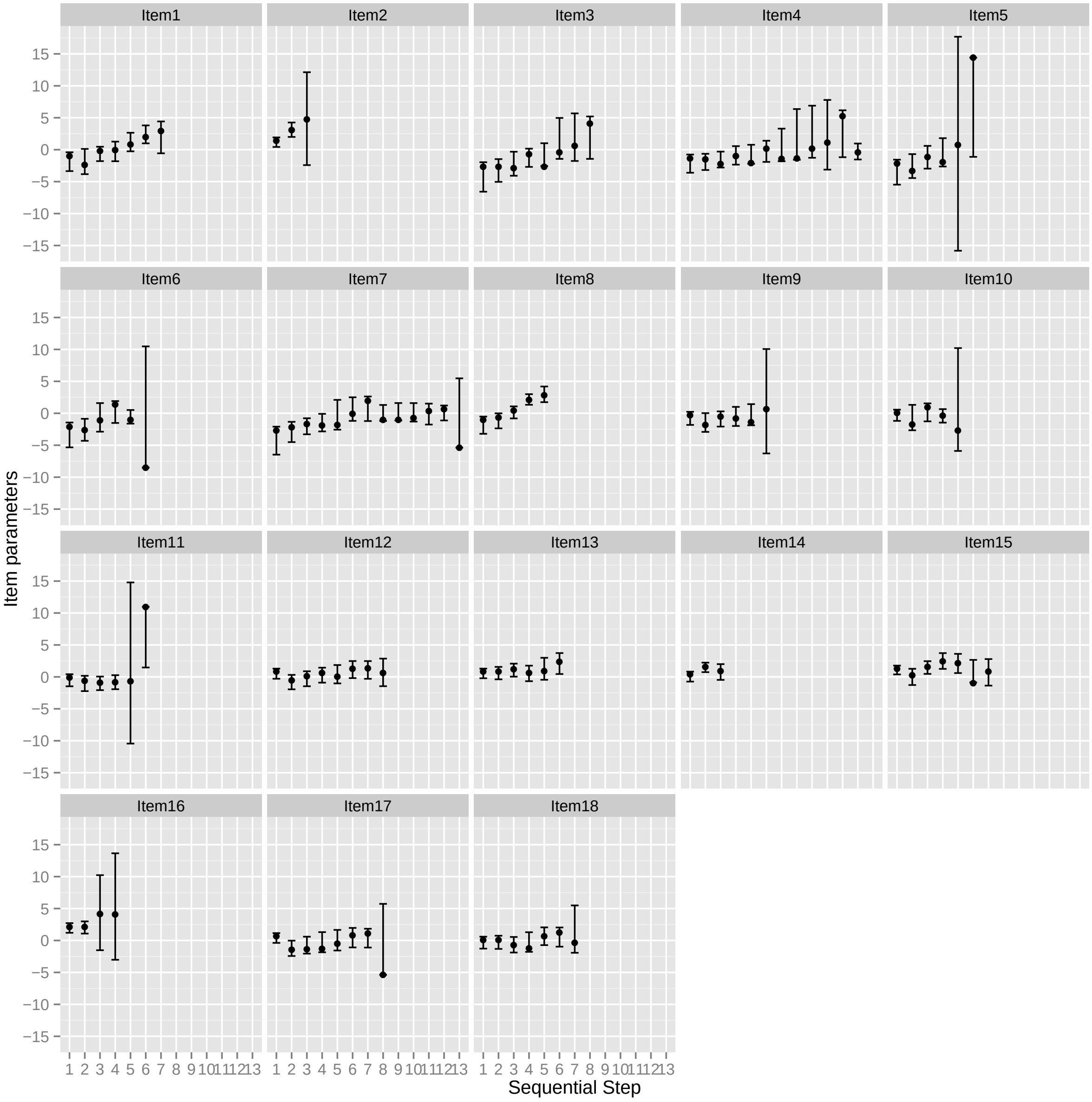

Figure 93: SM: CMLE of item parameters with parametric bootstrap percentile CI of exam data

A.6. Additional tables for description of items 6-25 of Ascot pre-test data

Table 10: Summary of items 6-10 of the Ascot pre-test data

Variable	Categories	Sample proportion
Item6	0: Does not fit	1 %
	1	2 %
	2	8 %
	3	21 %
	4	44 %
	5: Fits fully	24 %
Item7	0: Does not fit	3 %
	1	7 %
	2	16 %
	3	34 %
	4	30 %
	5: Fits fully	10 %
Item8	0: Does not fit	0 %
	1	1 %
	2	6 %
	3	20 %
	4	44 %
	5: Fits fully	29 %
Item9	0: Does not fit	0 %
	1	3 %
	2	15 %
	3	27 %
	4	37 %
	5: Fits fully	18 %
Item10	0: Does not fit	0 %
	1	3 %
	2	10 %
	3	24 %
	4	43 %
	5: Fits fully	20 %

Table 11: Summary of items 11-15 of the Ascot pre-test data

Variable	Categories	Sample proportion
Item11	0: Does not fit	0 %
	1	7 %
	2	24 %
	3	36 %
	4	26 %
	5: Fits fully	7 %
Item12	0: Does not fit	1 %
	1	10 %
	2	22 %
	3	34 %
	4	26 %
	5: Fits fully	7 %
Item13	0: Does not fit	1 %
	1	5 %
	2	20 %
	3	36 %
	4	26 %
	5: Fits fully	12 %
Item14	0: Does not fit	2 %
	1	4 %
	2	11 %
	3	21 %
	4	30 %
	5: Fits fully	32 %
Item15	0: Does not fit	1 %
	1	1 %
	2	4 %
	3	6 %
	4	24 %
	5: Fits fully	64 %

Table 12: Summary of items 16-20 of the Ascot pre-test data

Variable	Categories	Sample proportion
Item16	0: Does not fit	4 %
	1	11 %
	2	18 %
	3	23 %
	4	25 %
	5: Fits fully	19 %
Item17	0: Does not fit	1 %
	1	3 %
	2	7 %
	3	14 %
	4	34 %
	5: Fits fully	41 %
Item18	0: Does not fit	9 %
	1	16 %
	2	23 %
	3	27 %
	4	18 %
	5: Fits fully	7 %
Item19	0: Does not fit	2 %
	1	5 %
	2	7 %
	3	20 %
	4	35 %
	5: Fits fully	31 %
Item20	0: Does not fit	10 %
	1	8 %
	2	14 %
	3	28 %
	4	23 %
	5: Fits fully	17 %

Table 13: Summary of items 21-25 of the Ascot pre-test data

Variable	Categories	Sample proportion
Item21	0: Does not fit	11 %
	1	6 %
	2	18 %
	3	28 %
	4	21 %
	5: Fits fully	16 %
Item22	0: Does not fit	7 %
	1	8 %
	2	16 %
	3	27 %
	4	24 %
	5: Fits fully	18 %
Item23	0: Does not fit	43 %
	1	25 %
	2	11 %
	3	8 %
	4	8 %
	5: Fits fully	5 %
Item24	0: Does not fit	45 %
	1	23 %
	2	11 %
	3	8 %
	4	7 %
	5: Fits fully	6 %
Item25	0: Does not fit	36 %
	1	21 %
	2	14 %
	3	13 %
	4	10 %
	5: Fits fully	6 %

A.7. Tables: Probability of max score (average person) for Ascot pre-test data with decreasing order of difficulty

Table 14: SM: Probabilities of max score (average person) for Ascot pre-test data

Rank	Item	Probability
1	23	0.01 %
2	24	0.01 %
3	25	0.04 %
4	18	0.39 %
5	1	0.74 %
6	3	0.93 %
7	11	0.99 %
8	12	1.00 %
9	2	1.74 %
10	7	1.88 %
11	4	1.93 %
12	13	2.27 %
13	21	2.41 %
14	20	2.90 %
15	22	3.42 %
16	16	3.71 %
17	9	6.92 %
18	5	7.08 %
19	10	9.26 %
20	6	12.89 %
21	19	15.64 %
22	14	15.66 %
23	8	18.04 %
24	17	27.66 %
25	15	58.73 %

Table 15: PCM: Probabilities of max score (average person) for Ascot pre-test data

Rank	Item	Probability
1	23	0.79 %
2	24	0.94 %
3	25	1.35 %
4	18	2.82 %
5	1	2.98 %
6	11	3.54 %
7	12	3.69 %
8	3	4.09 %
9	4	5.48 %
10	9	5.63 %
11	7	5.66 %
12	13	6.69 %
13	21	9.27 %
14	20	10.36 %
15	22	11.00 %
16	16	11.47 %
17	5	12.42 %
18	10	14.92 %
19	6	19.02 %
20	19	25.59 %
21	14	26.61 %
22	8	28.93 %
23	17	36.91 %
24	2	40.41 %
25	15	64.04 %

A.8. Estimated item parameters for the Sequential and Partial Credit Model of Ascot pre-test data

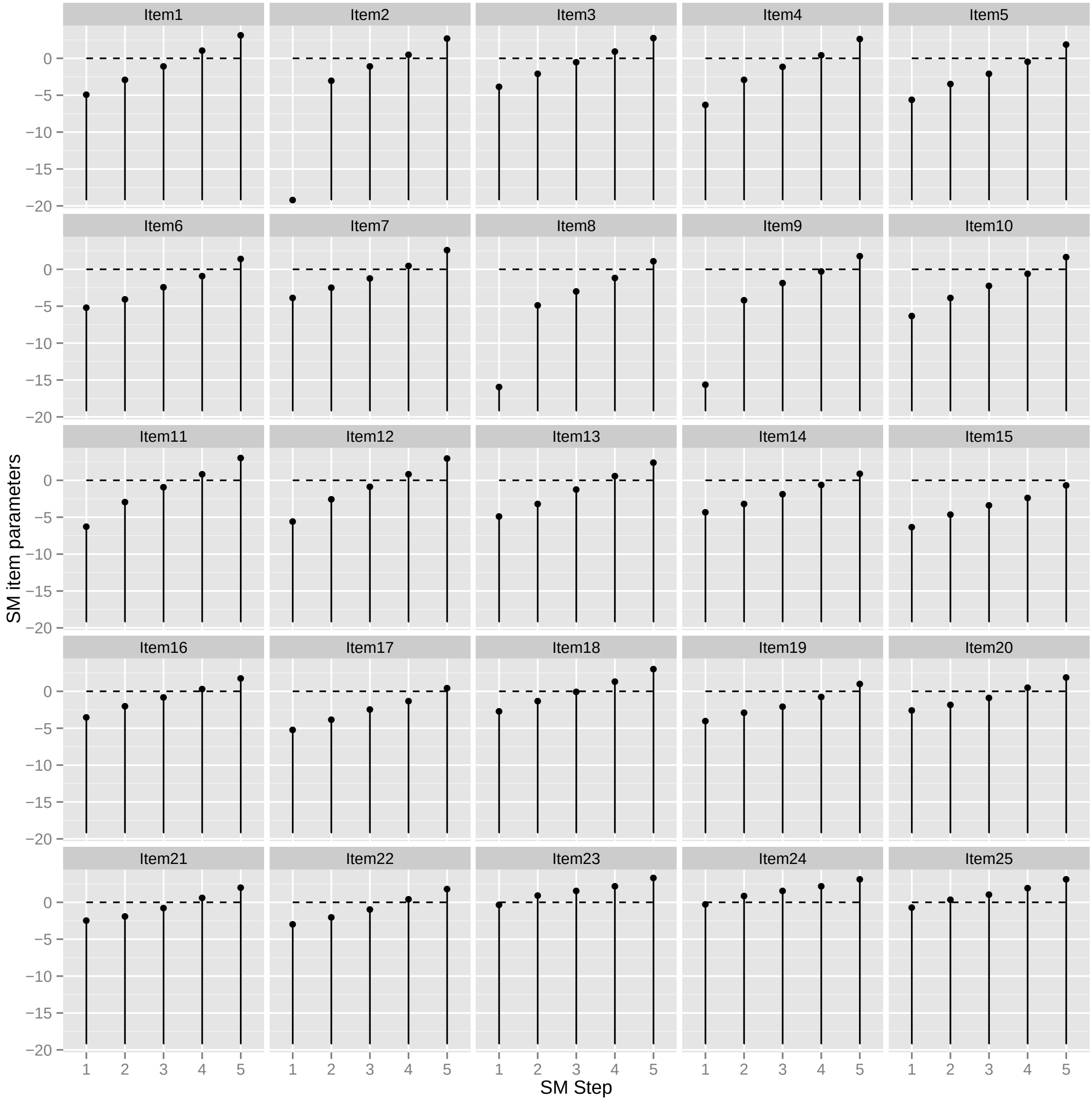

Figure 94: SM: Estimated item parameters of Ascot pre-test data

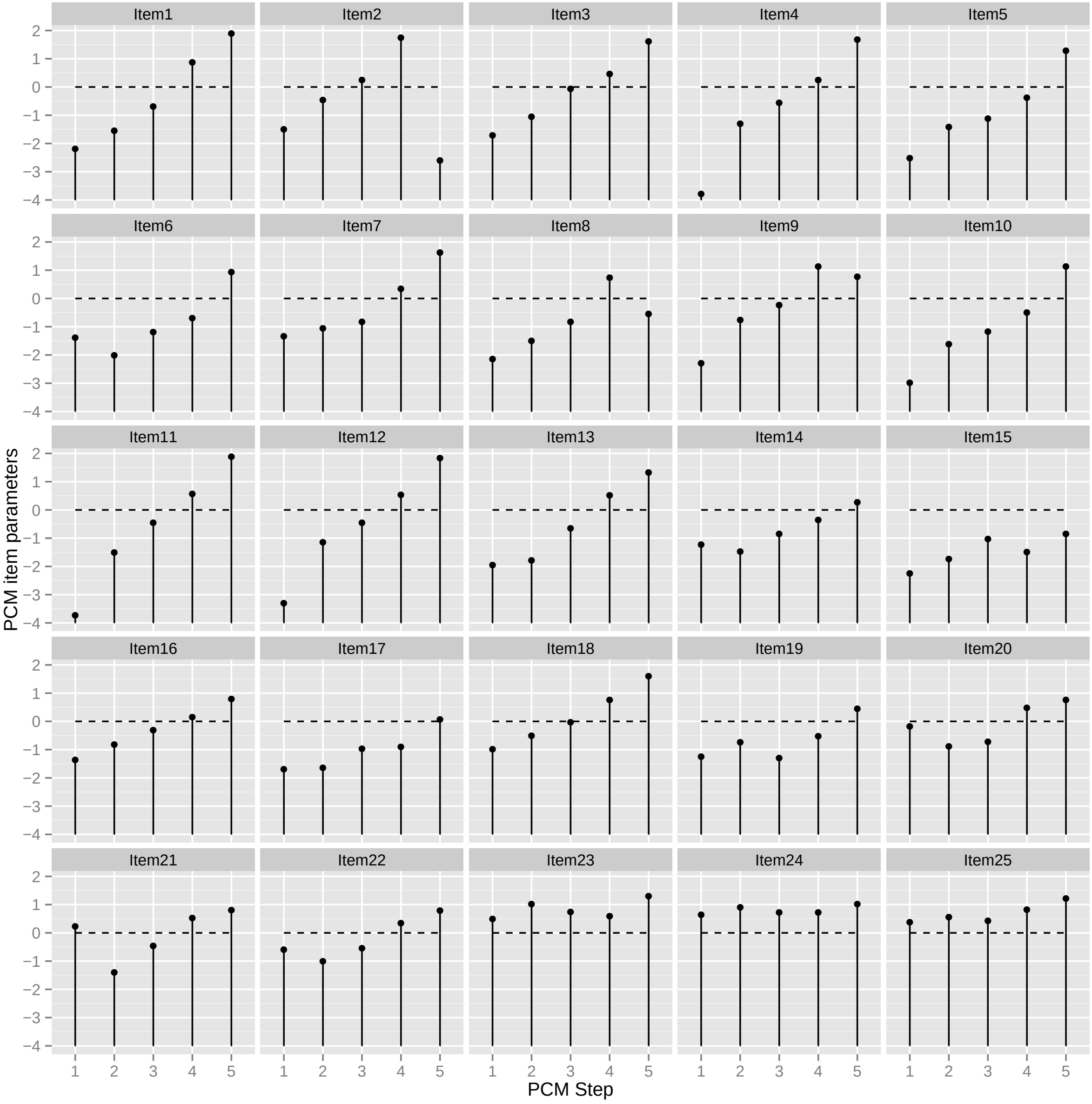

Figure 95: PCM: Estimated item parameters of Ascot pre-test data

A.9. Category probability curves for rest of items of Ascot pre-test data

A.9.1. Category probability curves for Sequential Model

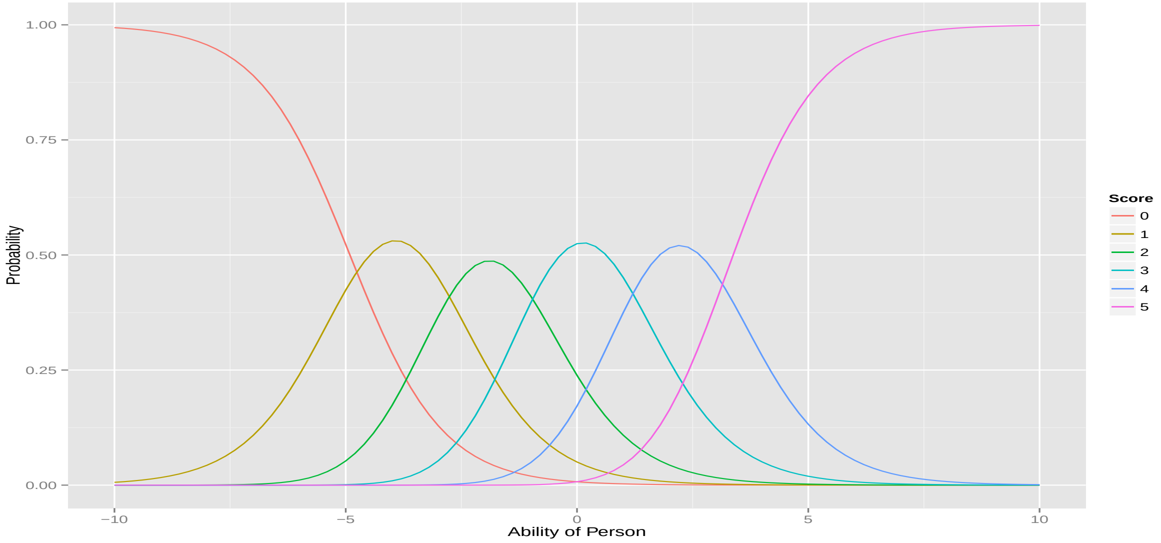

Figure 96: SM: CPCs for item 1 in Ascot pre-test data

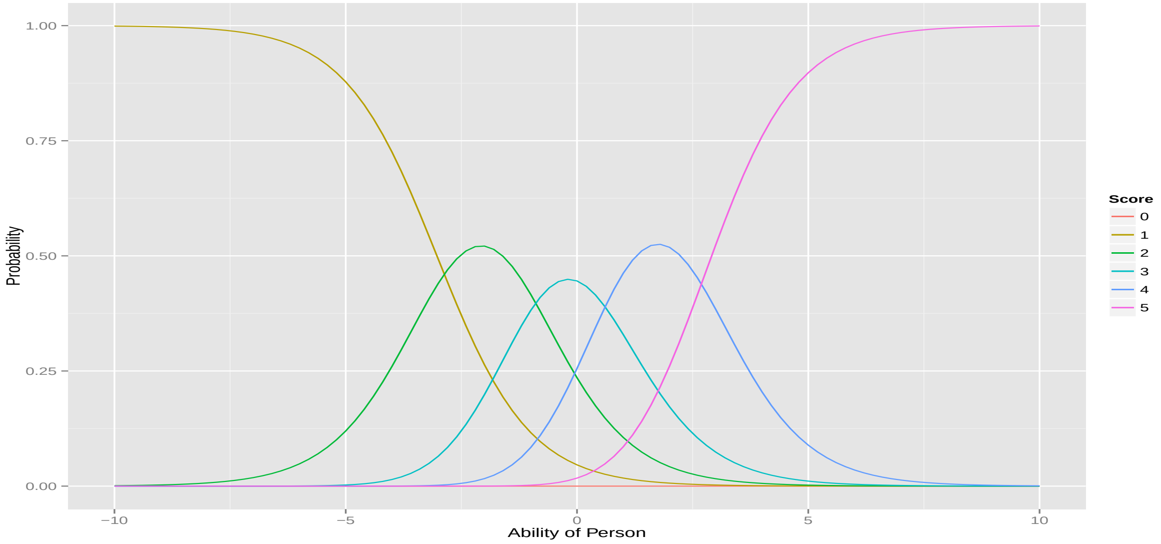

Figure 97: SM: CPCs for item 2 in Ascot pre-test data

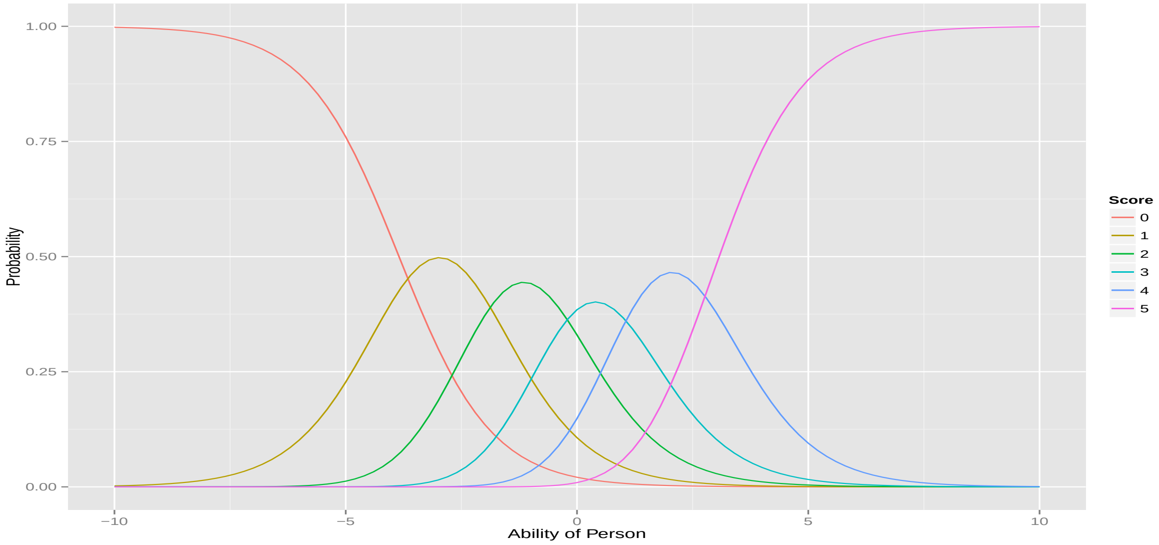

Figure 98: SM: CPCs for item 3 in Ascot pre-test data

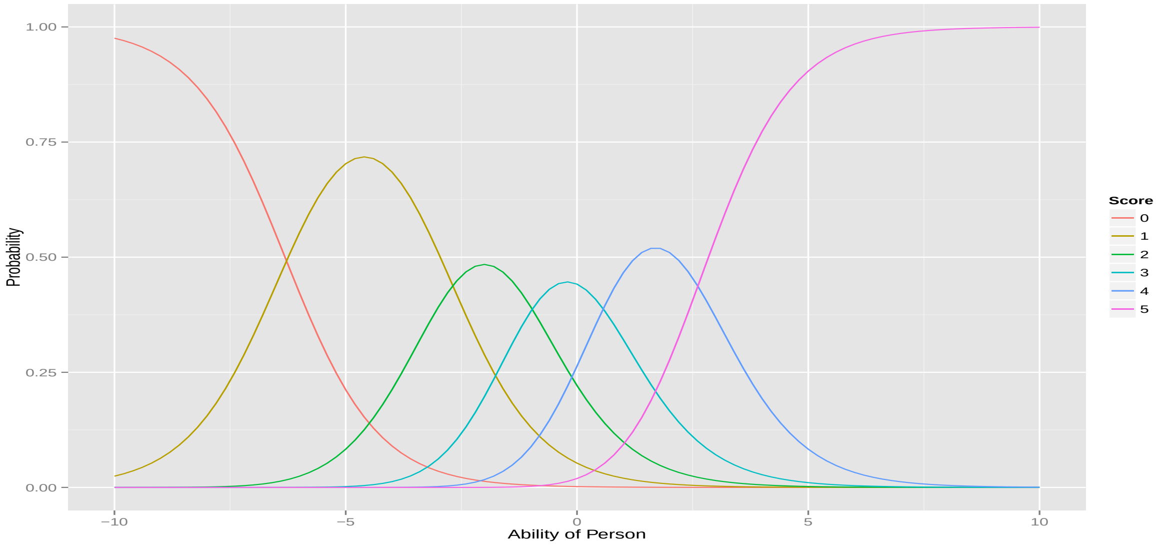

Figure 99: SM: CPCs for item 4 in Ascot pre-test data

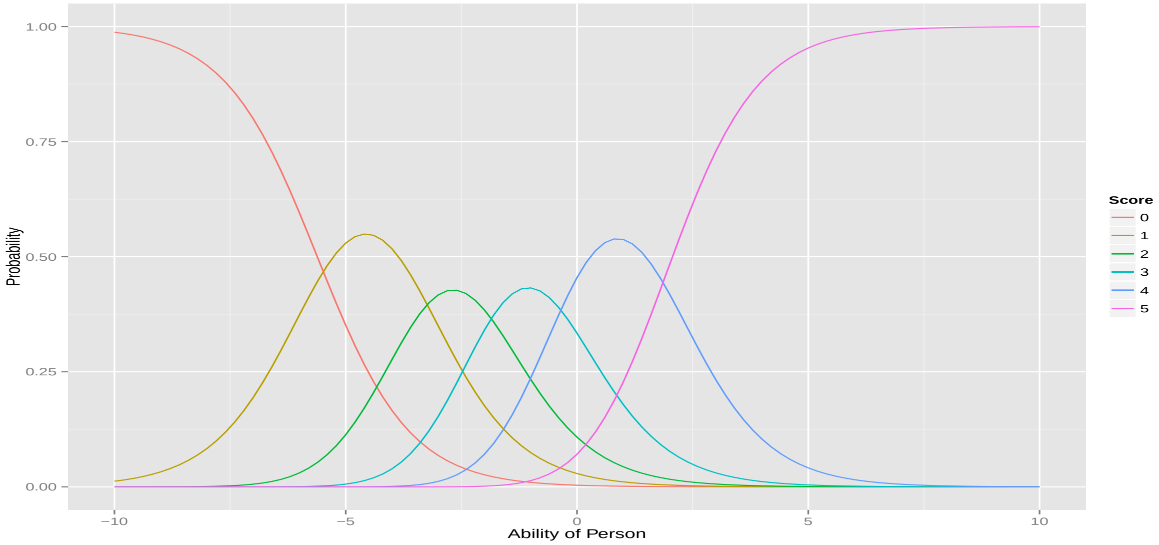

Figure 100: SM: CPCs for item 5 in Ascot pre-test data

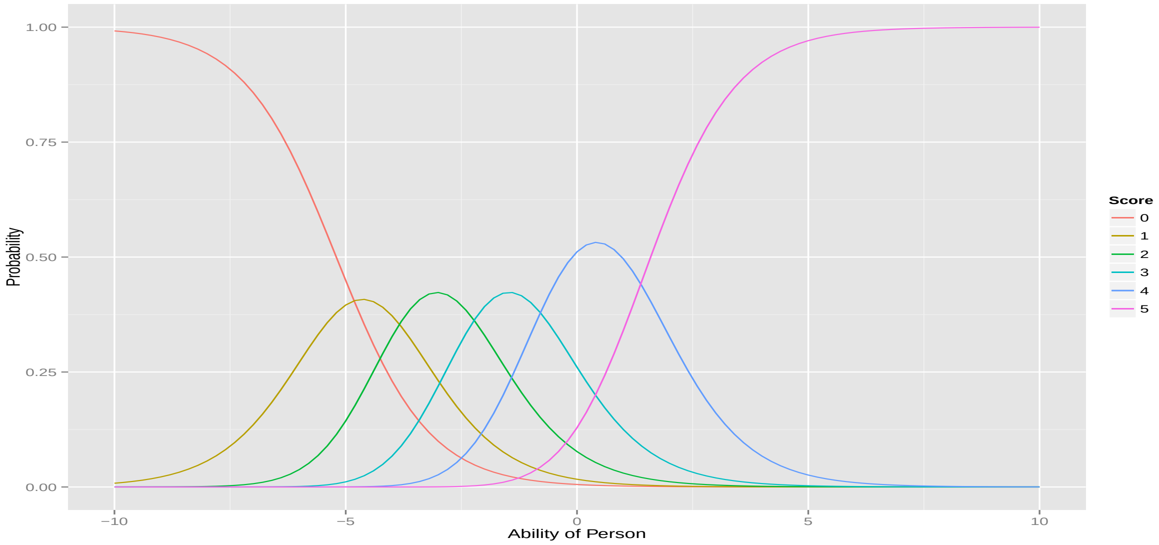

Figure 101: SM: CPCs for item 6 in Ascot pre-test data

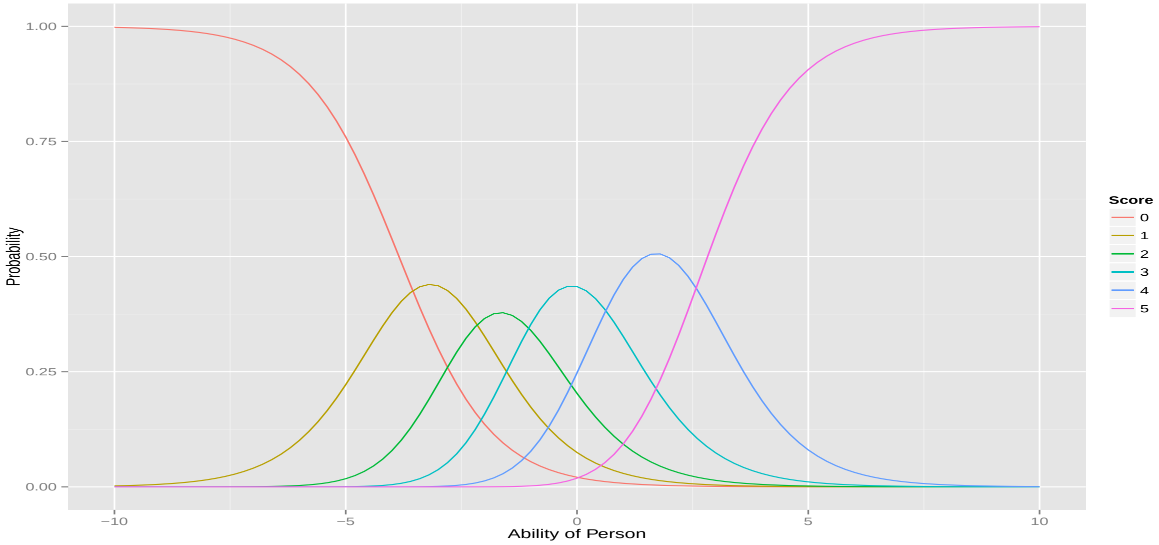

Figure 102: SM: CPCs for item 7 in Ascot pre-test data

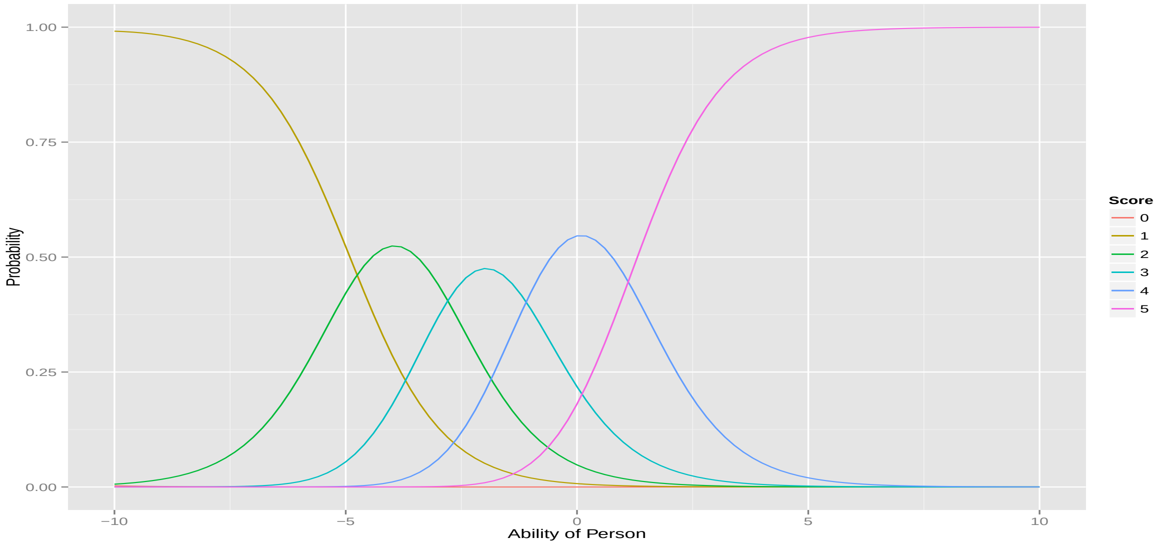

Figure 103: SM: CPCs for item 8 in Ascot pre-test data

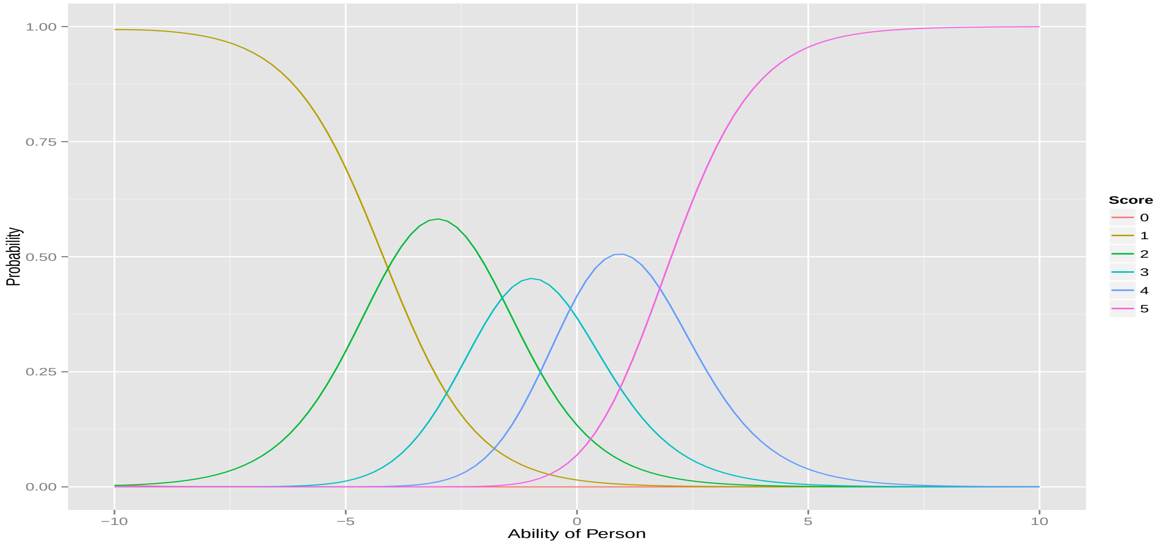

Figure 104: SM: CPCs for item 9 in Ascot pre-test data

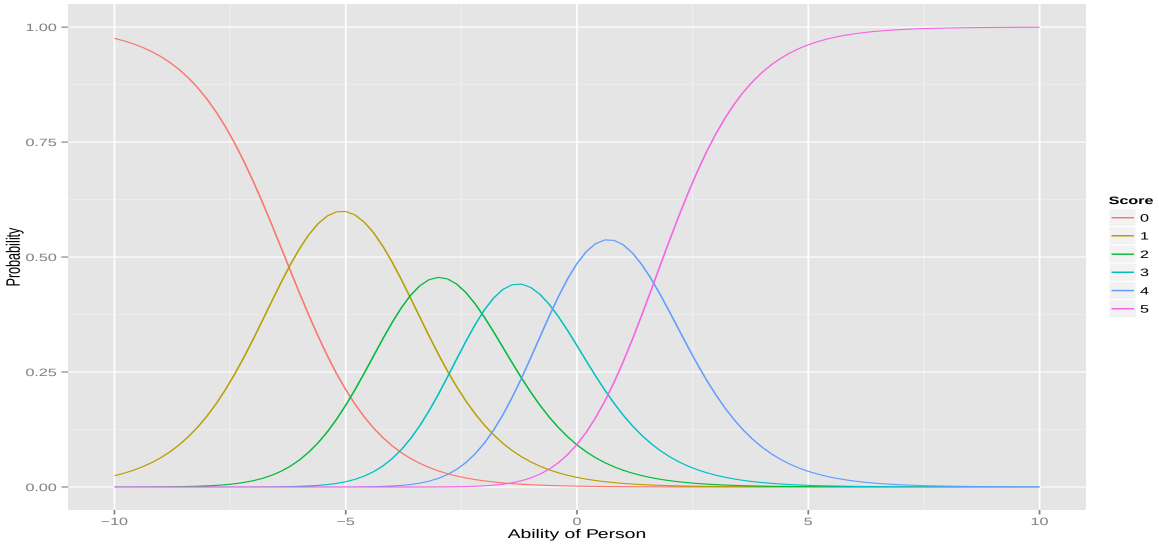

Figure 105: SM: CPCs for item 10 in Ascot pre-test data

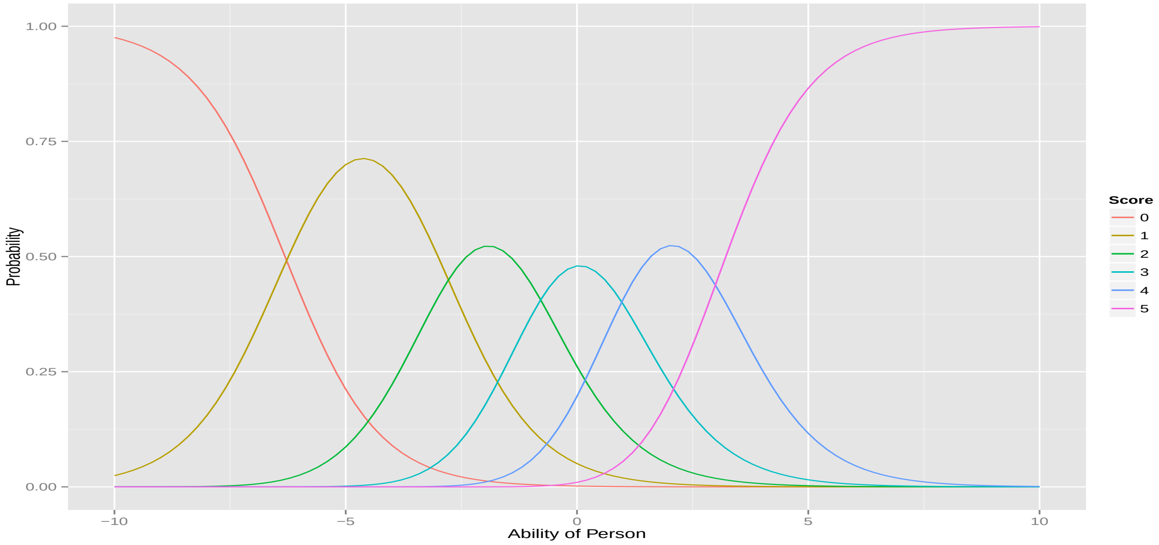

Figure 106: SM: CPCs for item 11 in Ascot pre-test data

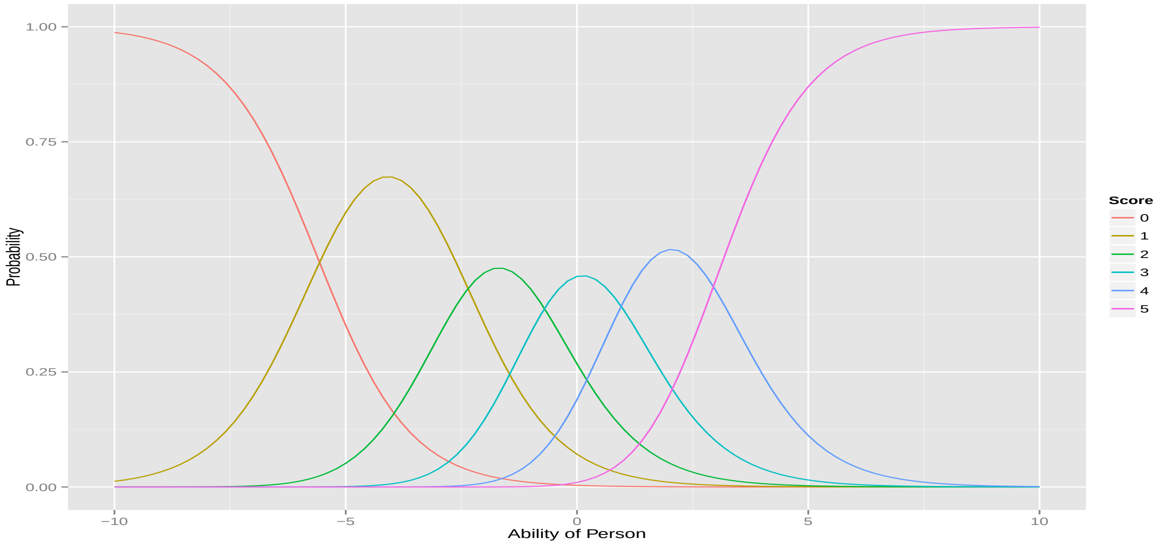

Figure 107: SM: CPCs for item 12 in Ascot pre-test data

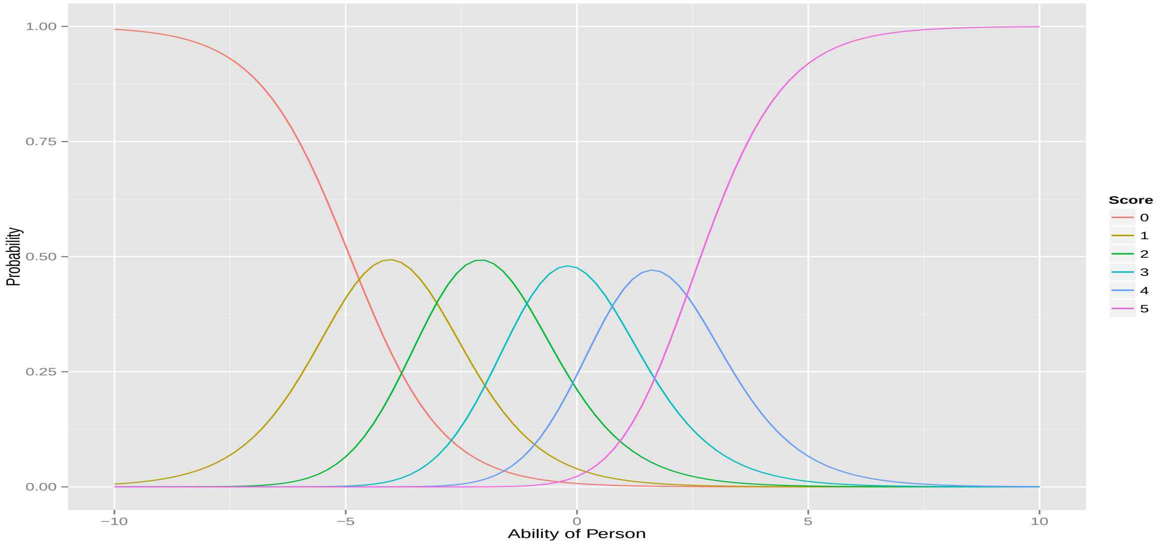

Figure 108: SM: CPCs for item 13 in Ascot pre-test data

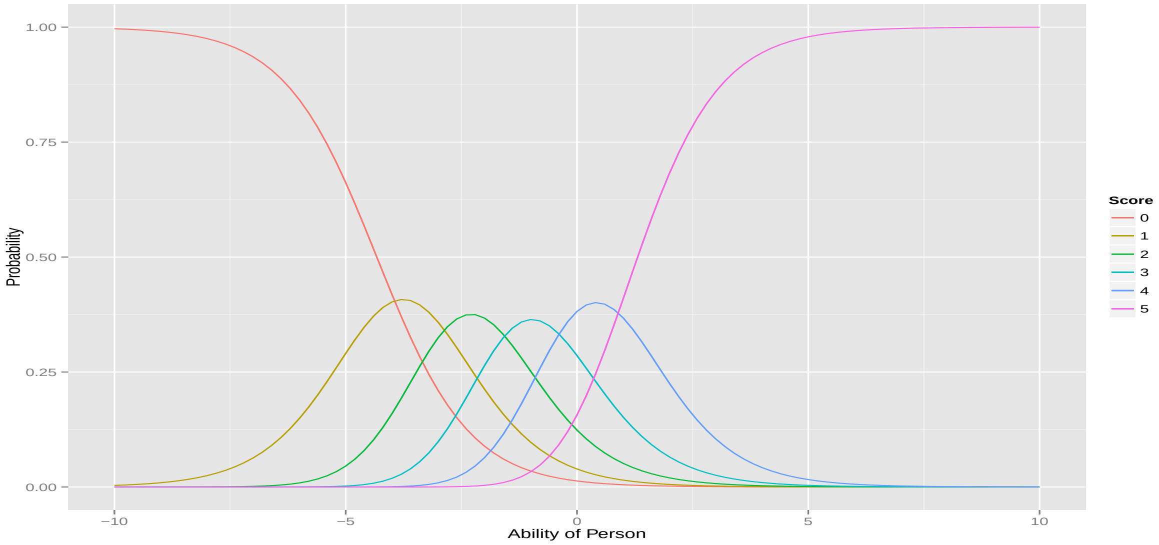

Figure 109: SM: CPCs for item 14 in Ascot pre-test data

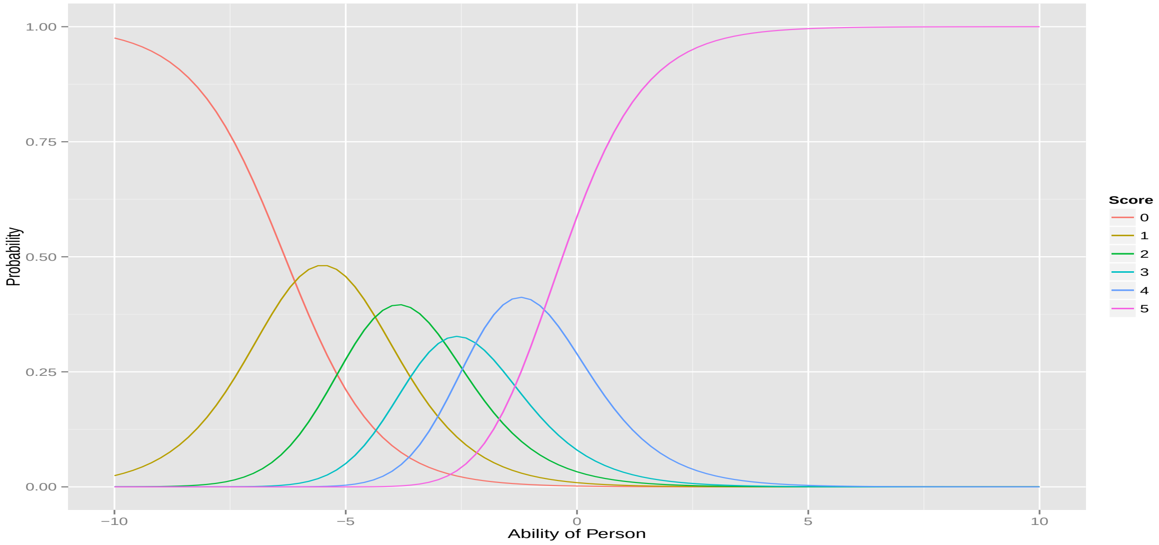

Figure 110: SM: CPCs for item 15 in Ascot pre-test data

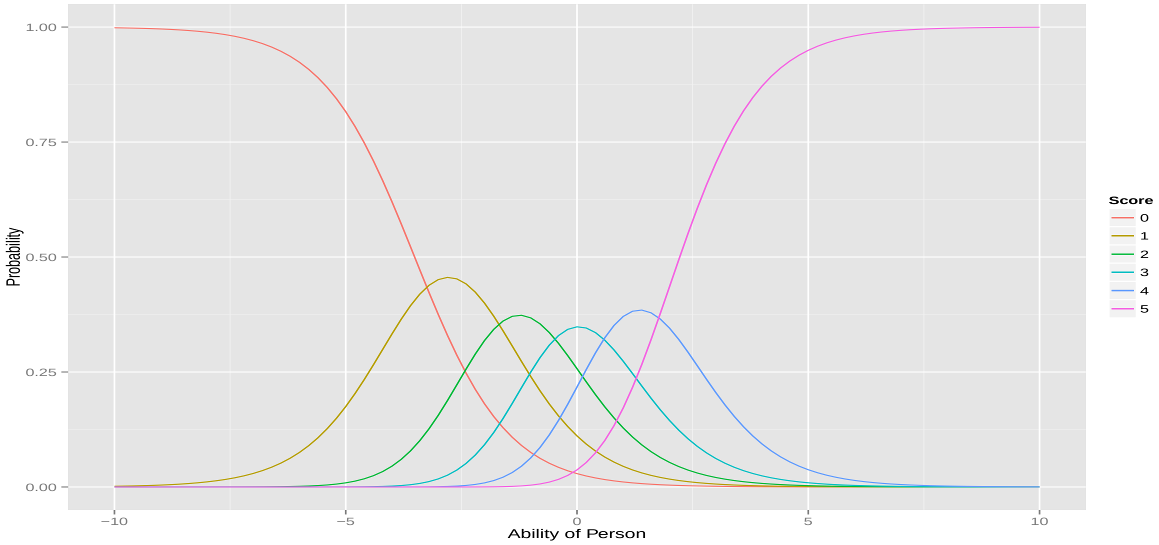

Figure 111: SM: CPCs for item 16 in Ascot pre-test data

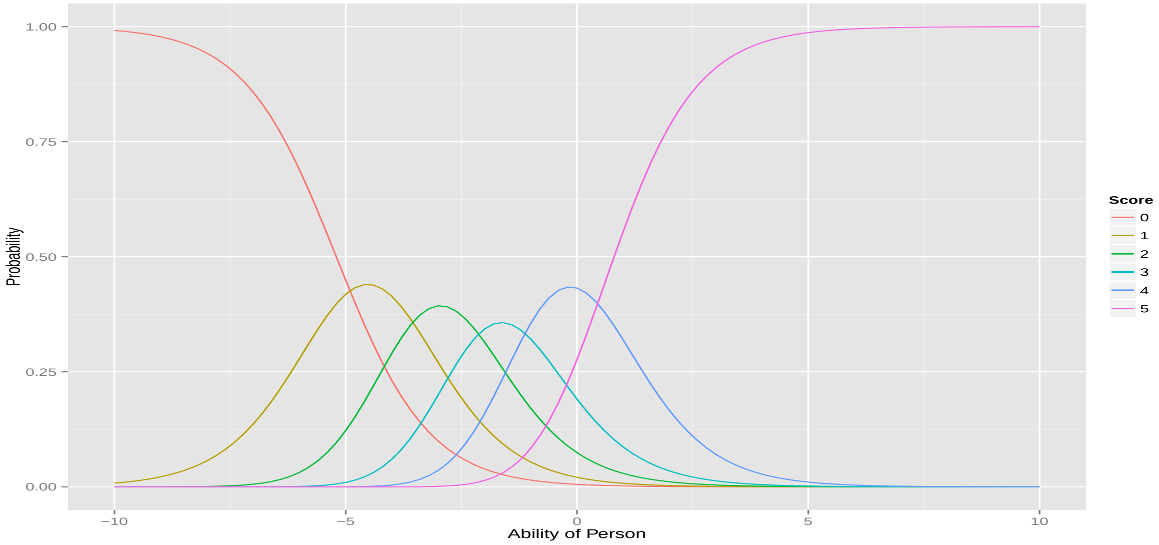

Figure 112: SM: CPCs for item 17 in Ascot pre-test data

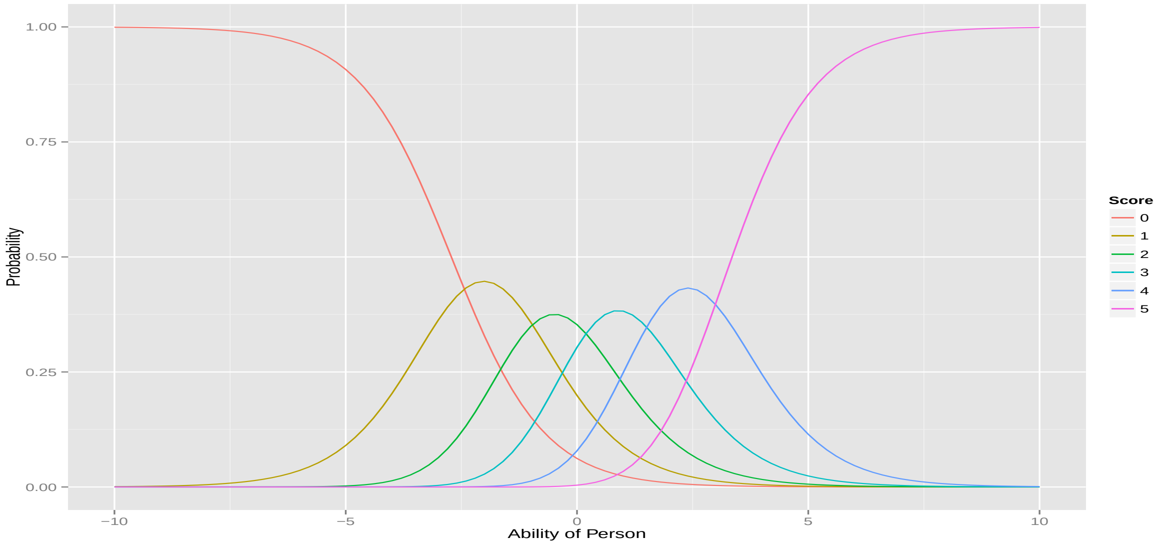

Figure 113: SM: CPCs for item 18 in Ascot pre-test data

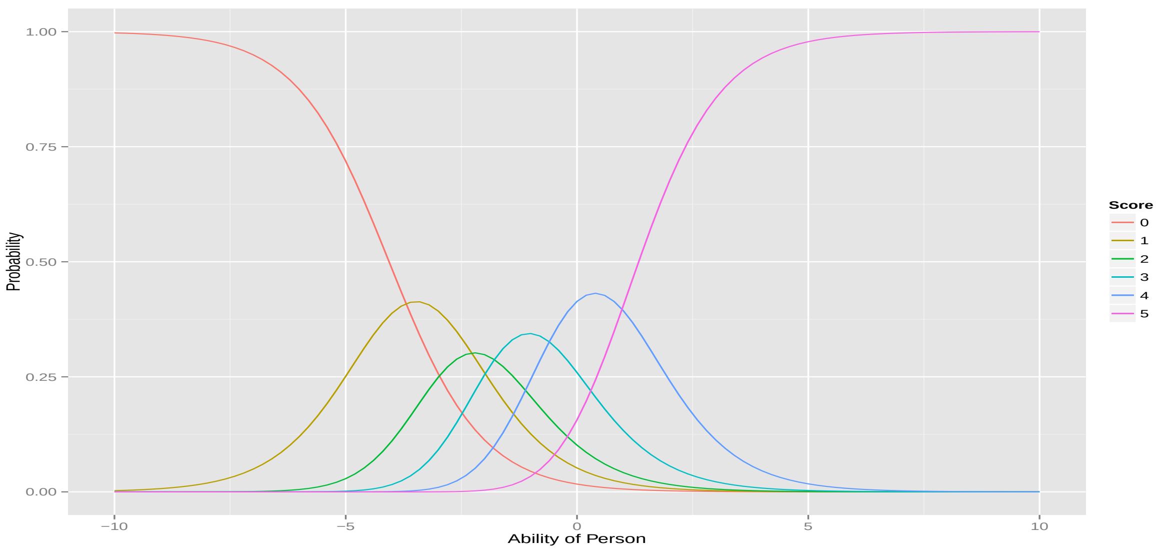

Figure 114: SM: CPCs for item 19 in Ascot pre-test data

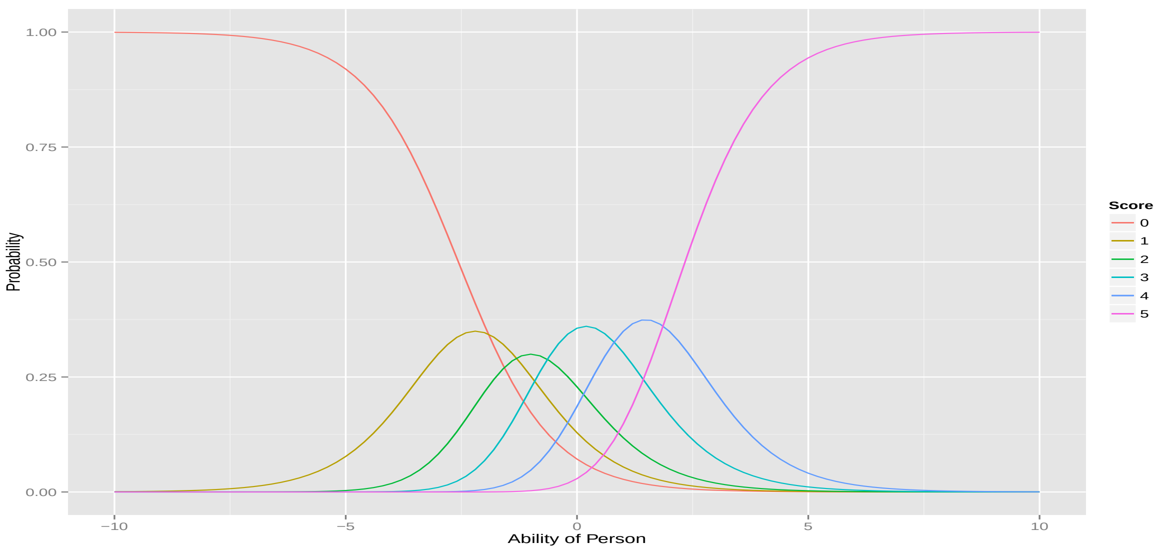

Figure 115: SM: CPCs for item 20 in Ascot pre-test data

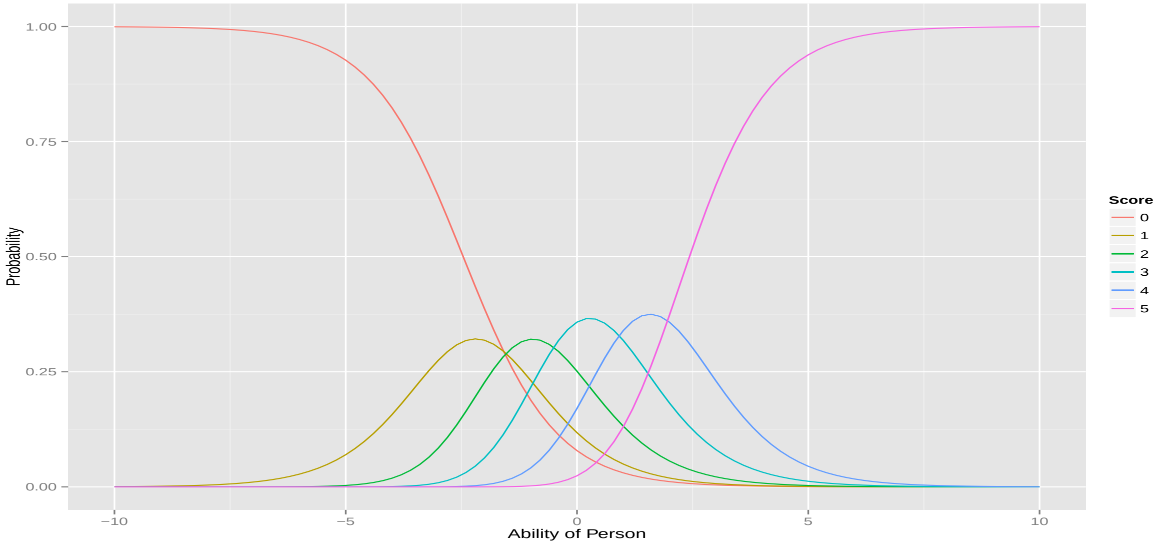

Figure 116: SM: CPCs for item 21 in Ascot pre-test data

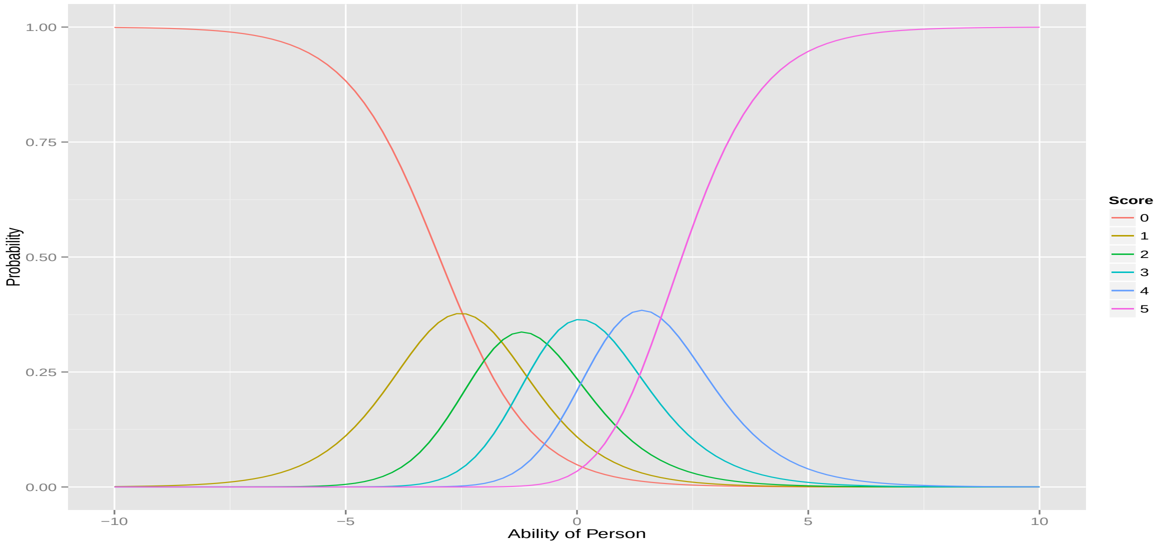

Figure 117: SM: CPCs for item 22 in Ascot pre-test data

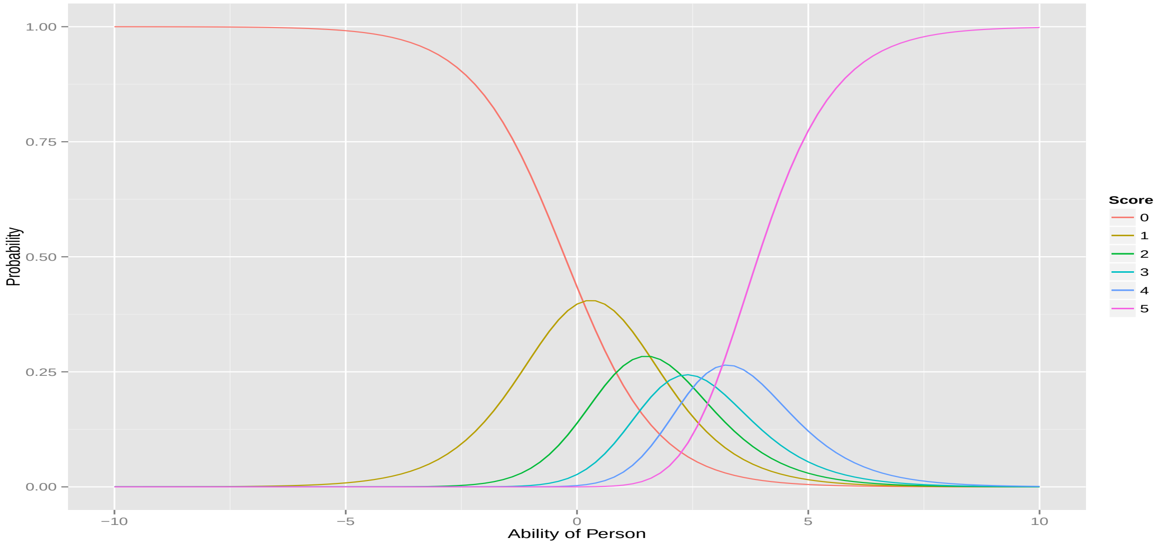

Figure 118: SM: CPCs for item 24 in Ascot pre-test data

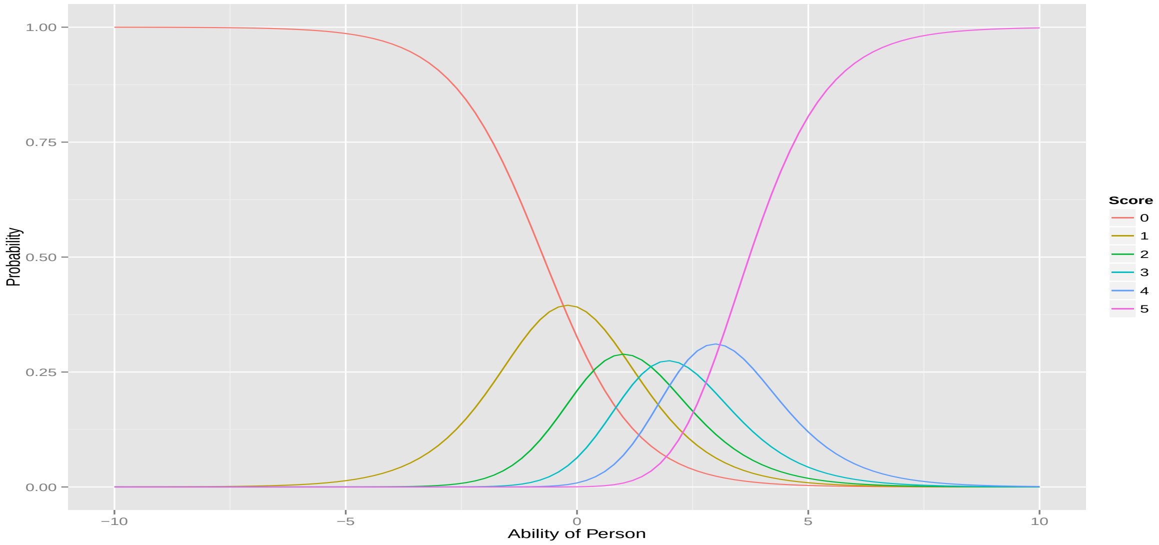

Figure 119: SM: CPCs for item 25 in Ascot pre-test data

A.9.2. Category probability curves for Partial Credit Model

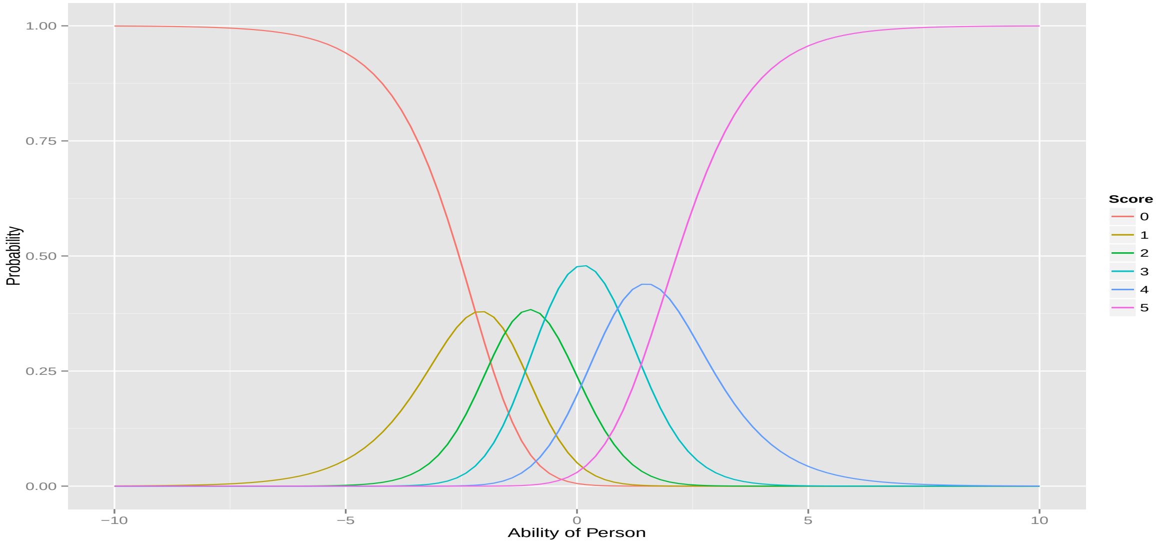

Figure 120: PCM: CPCs for item 1 in Ascot pre-test data

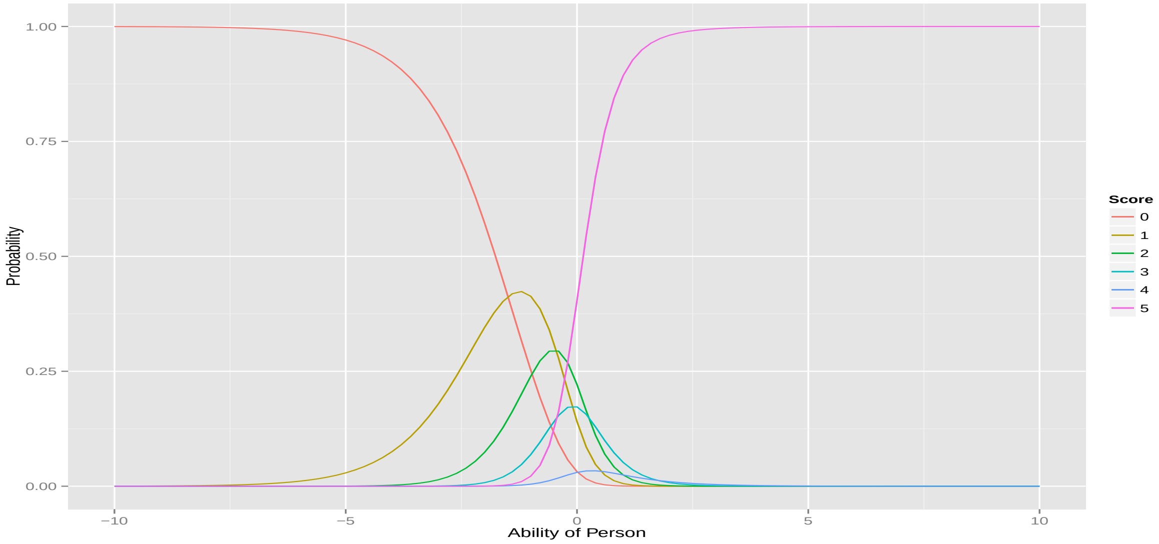

Figure 121: PCM: CPCs for item 2 in Ascot pre-test data

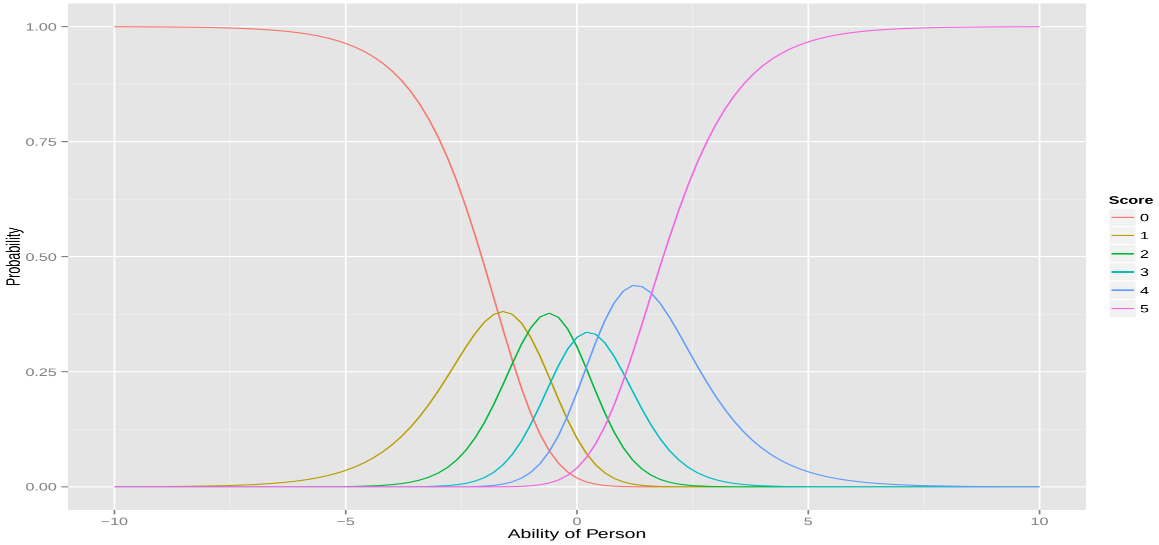

Figure 122: PCM: CPCs for item 3 in Ascot pre-test data

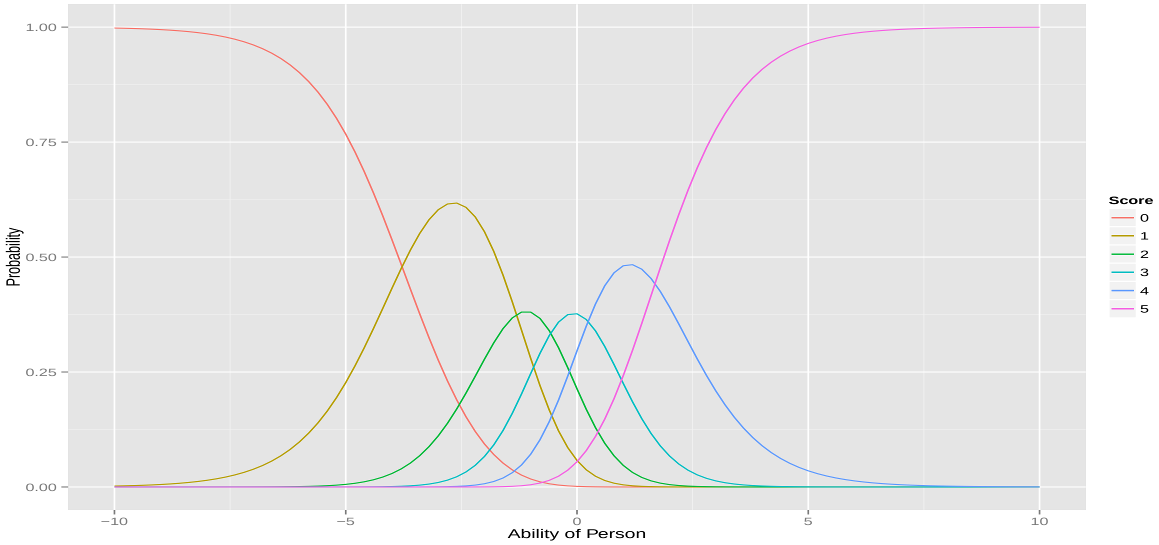

Figure 123: PCM: CPCs for item 4 in Ascot pre-test data

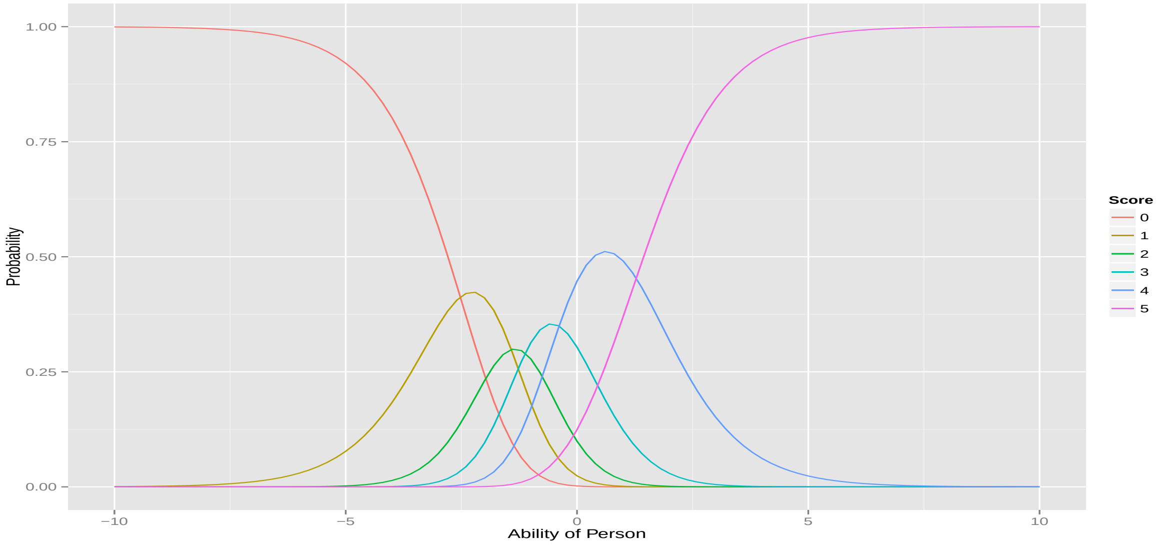

Figure 124: PCM: CPCs for item 5 in Ascot pre-test data

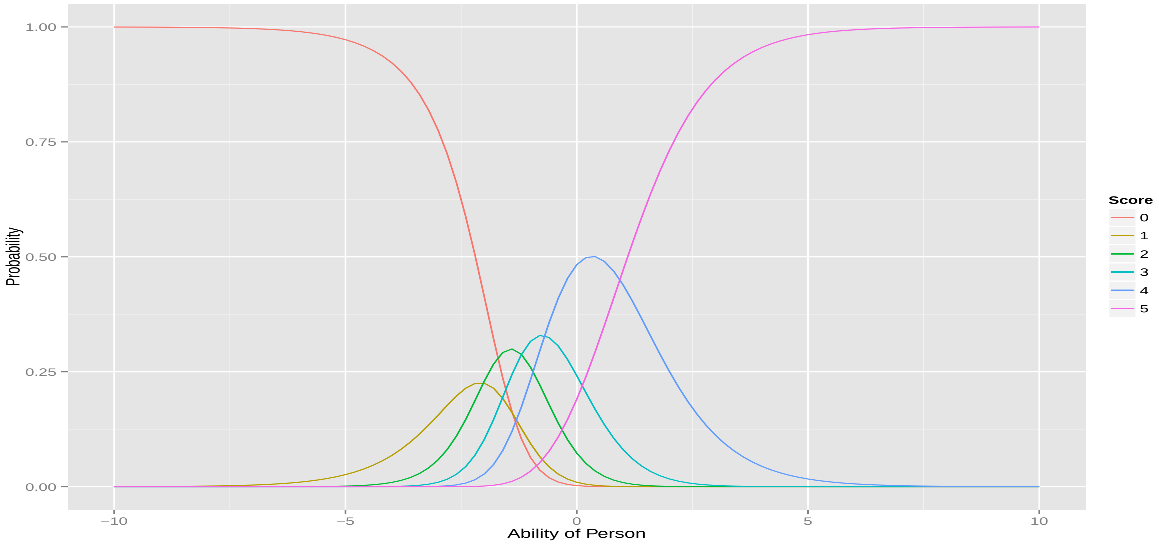

Figure 125: PCM: CPCs for item 6 in Ascot pre-test data

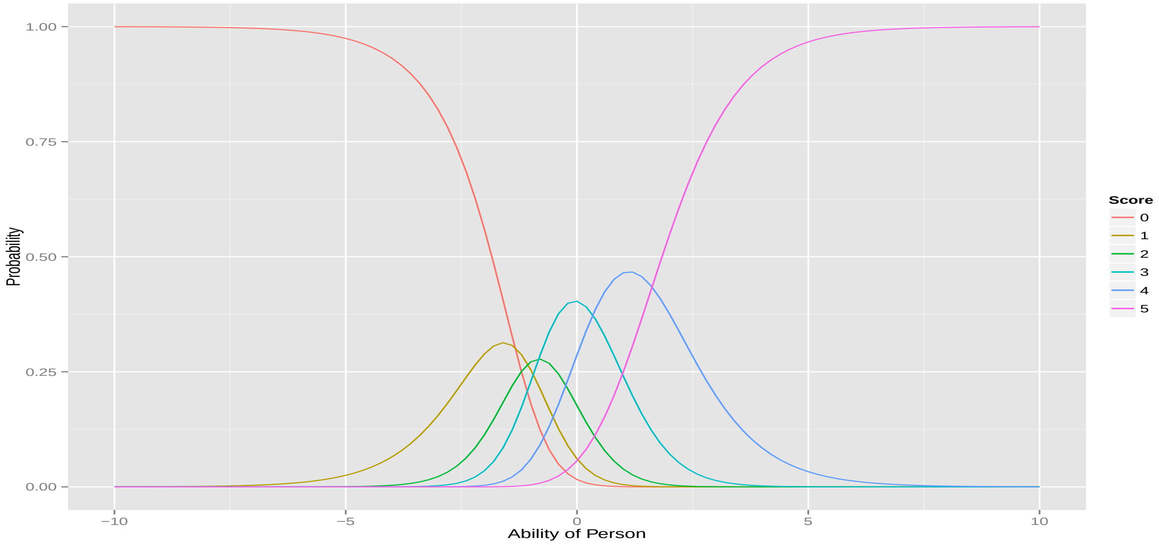

Figure 126: PCM: CPCs for item 7 in Ascot pre-test data

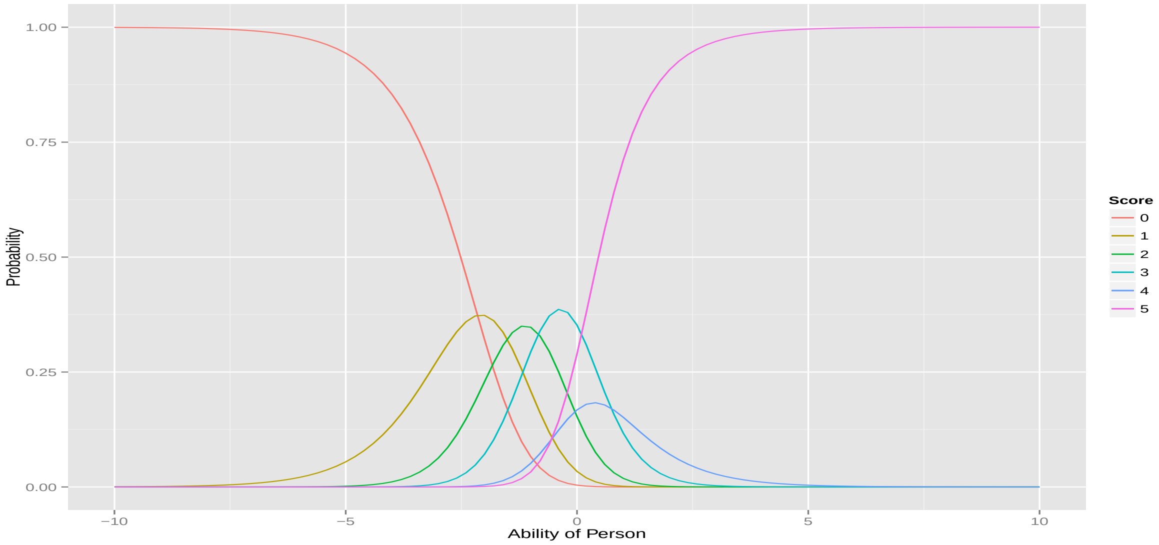

Figure 127: PCM: CPCs for item 8 in Ascot pre-test data

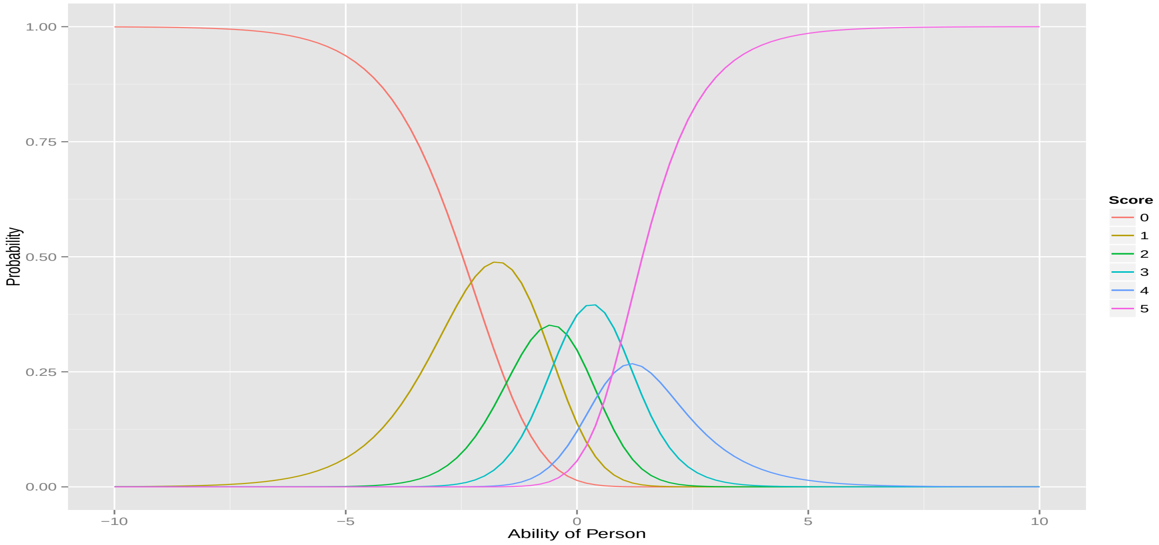

Figure 128: PCM: CPCs for item 9 in Ascot pre-test data

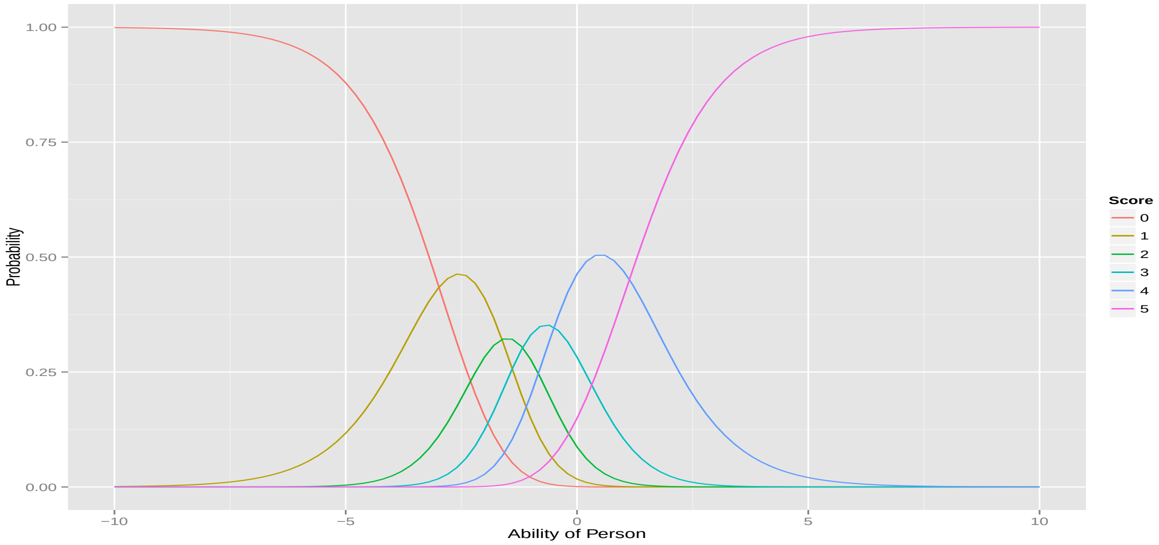

Figure 129: PCM: CPCs for item 10 in Ascot pre-test data

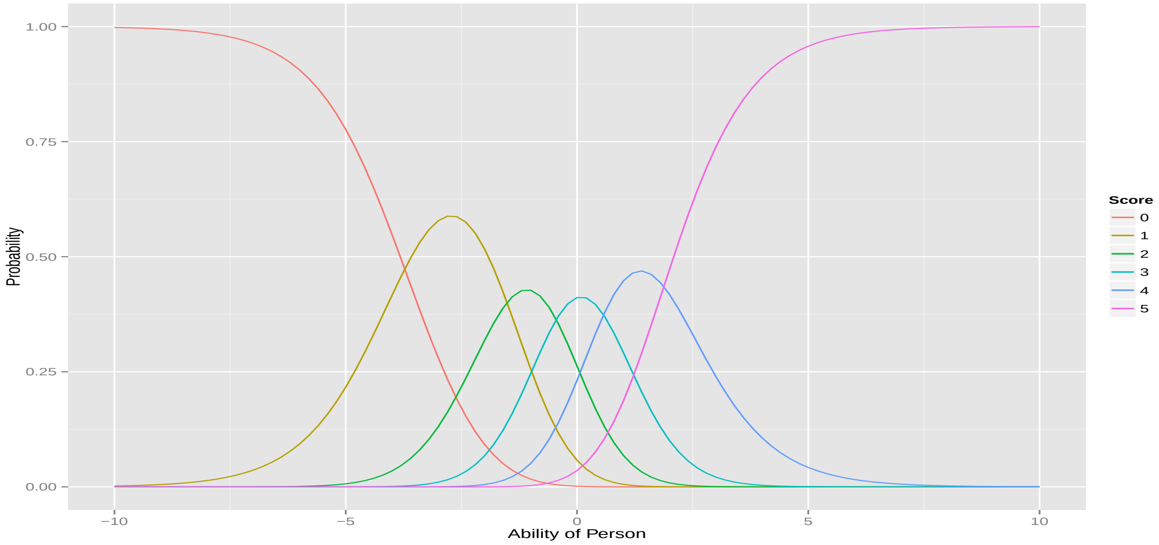

Figure 130: PCM: CPCs for item 11 in Ascot pre-test data

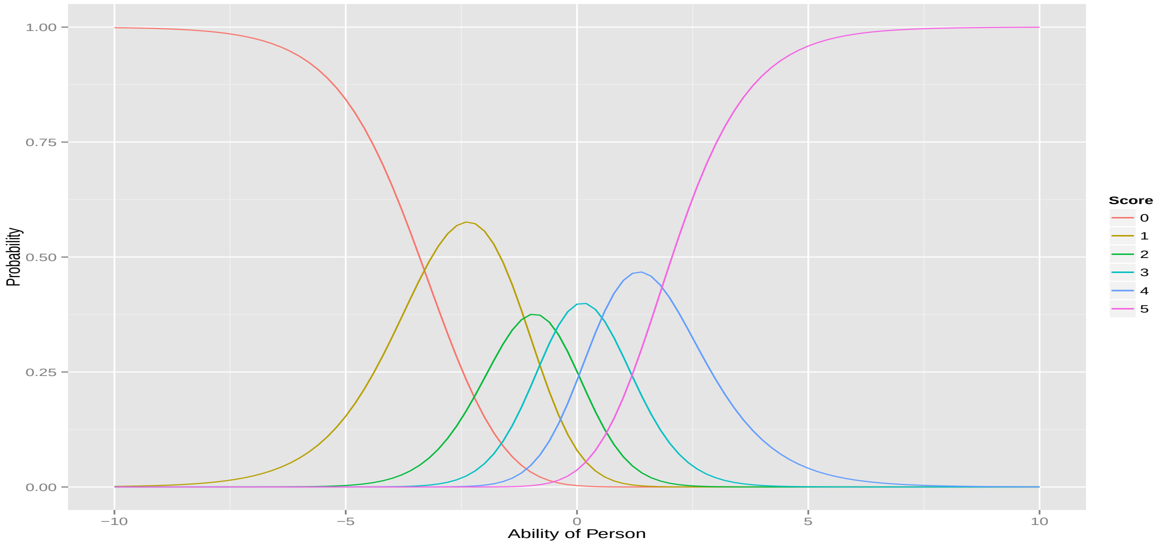

Figure 131: PCM: CPCs for item 12 in Ascot pre-test data

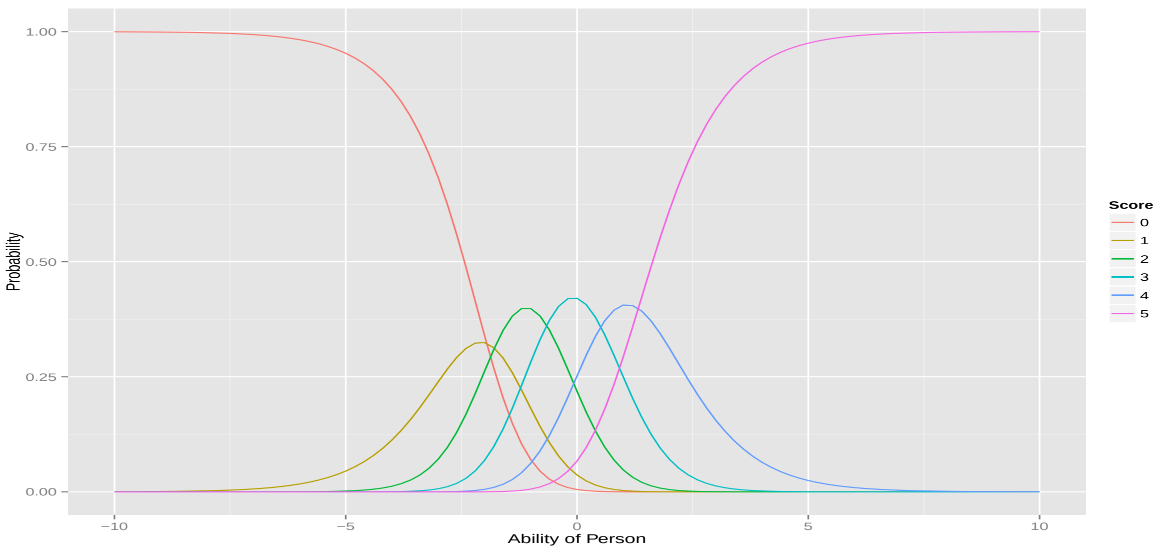

Figure 132: PCM: CPCs for item 13 in Ascot pre-test data

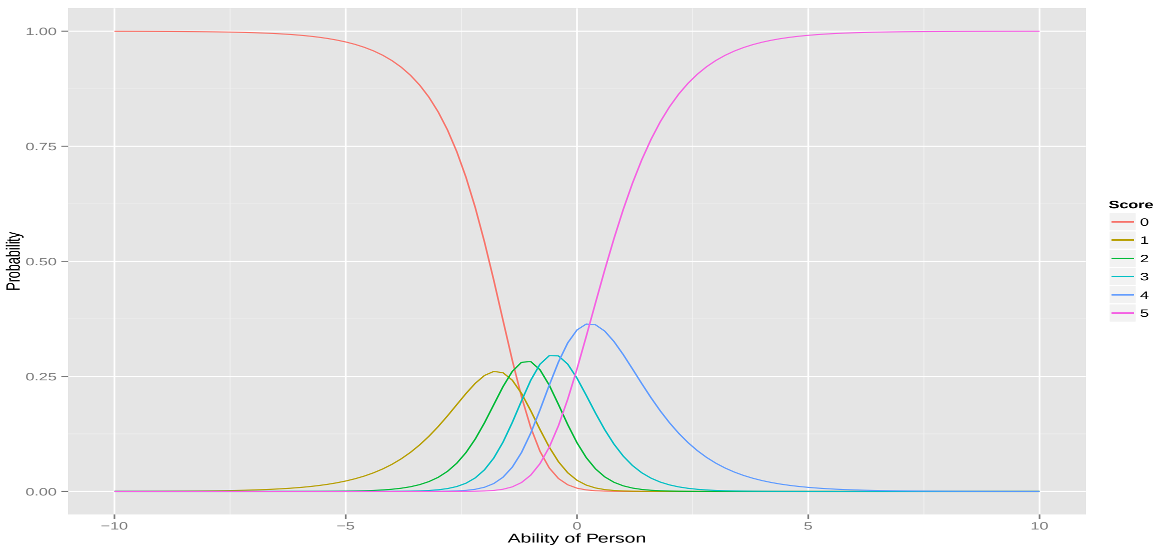

Figure 133: PCM: CPCs for item 14 in Ascot pre-test data

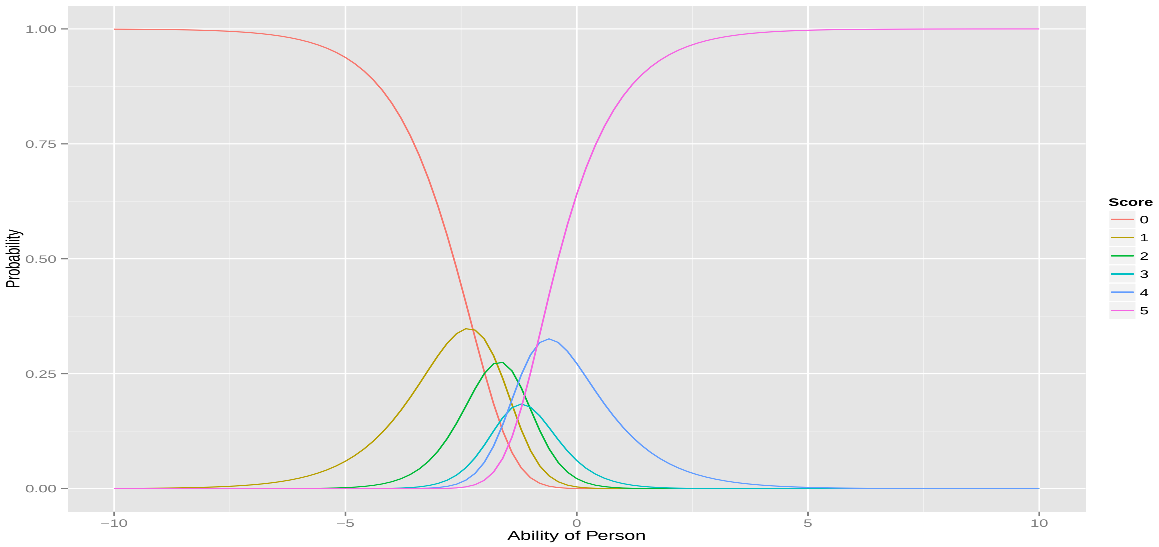

Figure 134: PCM: CPCs for item 15 in Ascot pre-test data

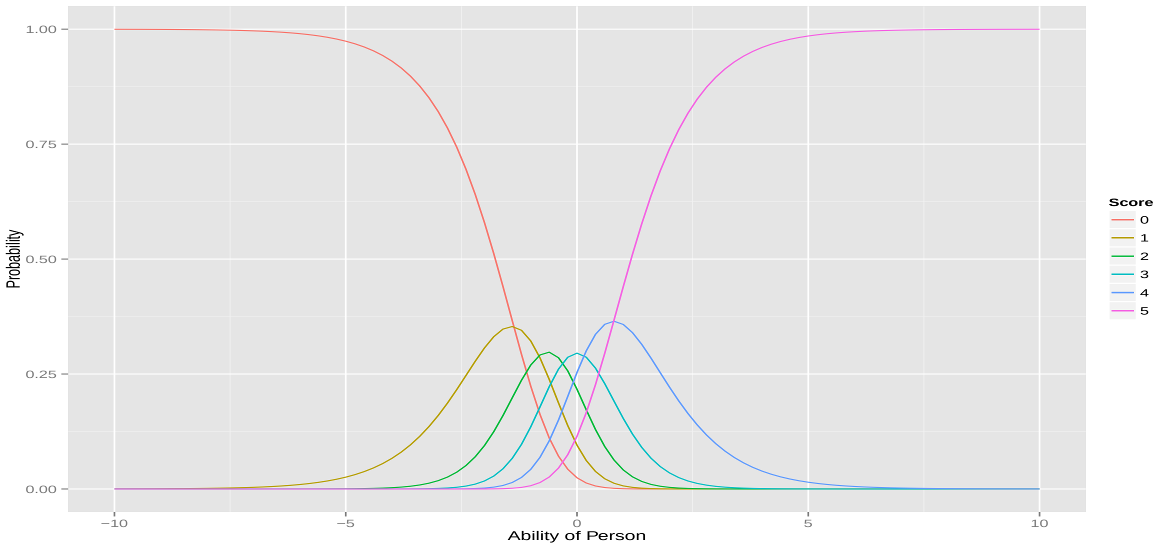

Figure 135: PCM: CPCs for item 16 in Ascot pre-test data

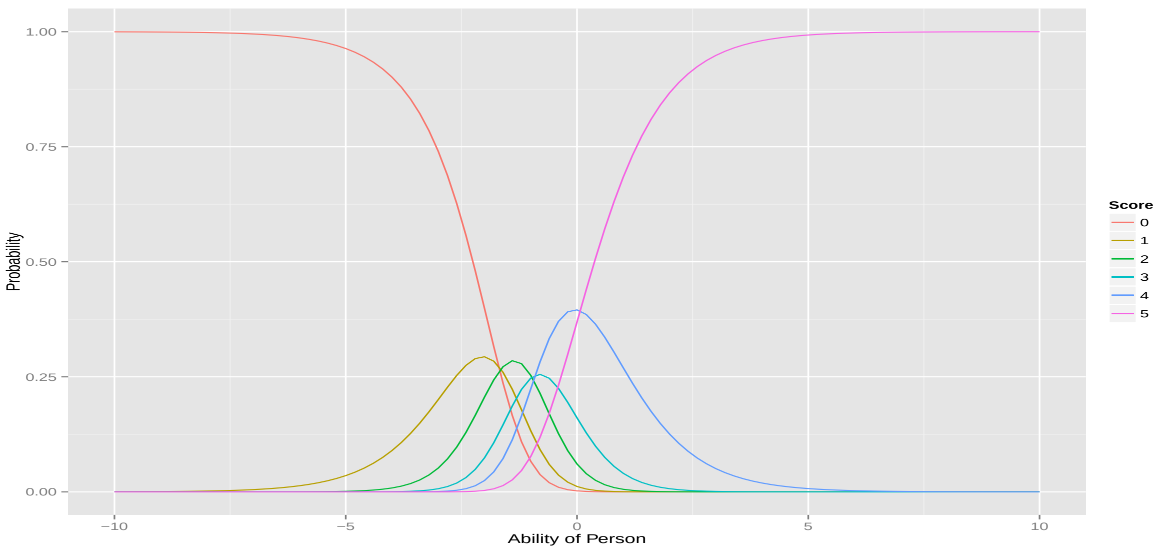

Figure 136: PCM: CPCs for item 17 in Ascot pre-test data

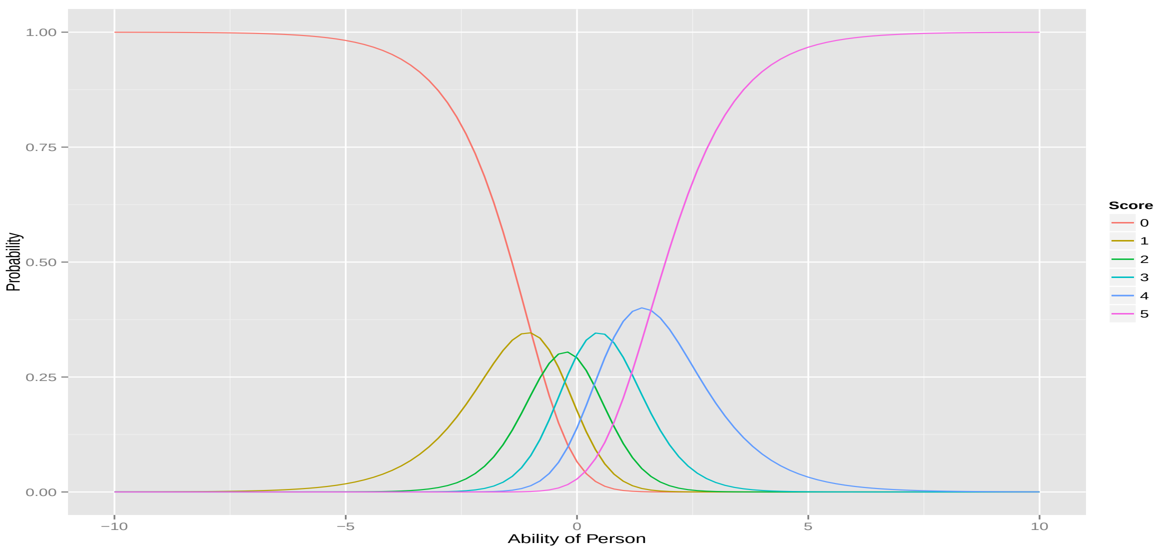

Figure 137: PCM: CPCs for item 18 in Ascot pre-test data

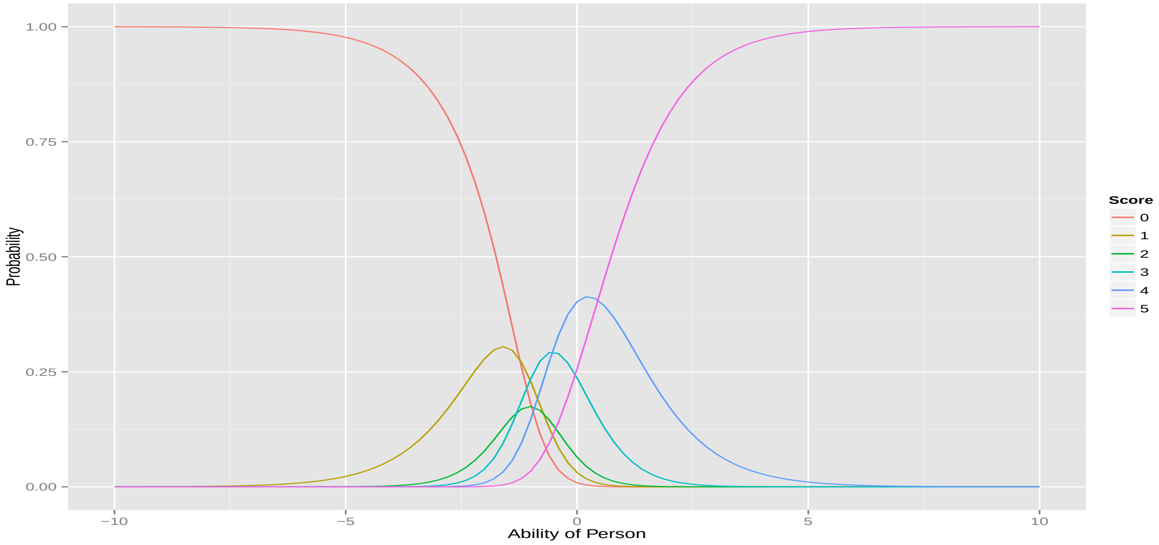

Figure 138: PCM: CPCs for item 19 in Ascot pre-test data

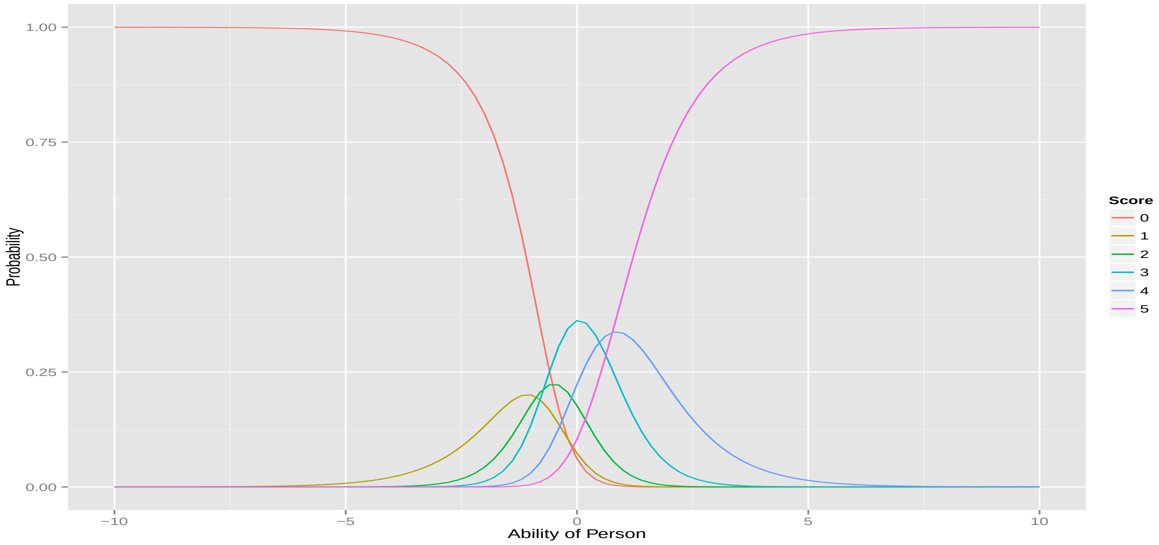

Figure 139: PCM: CPCs for item 20 in Ascot pre-test data

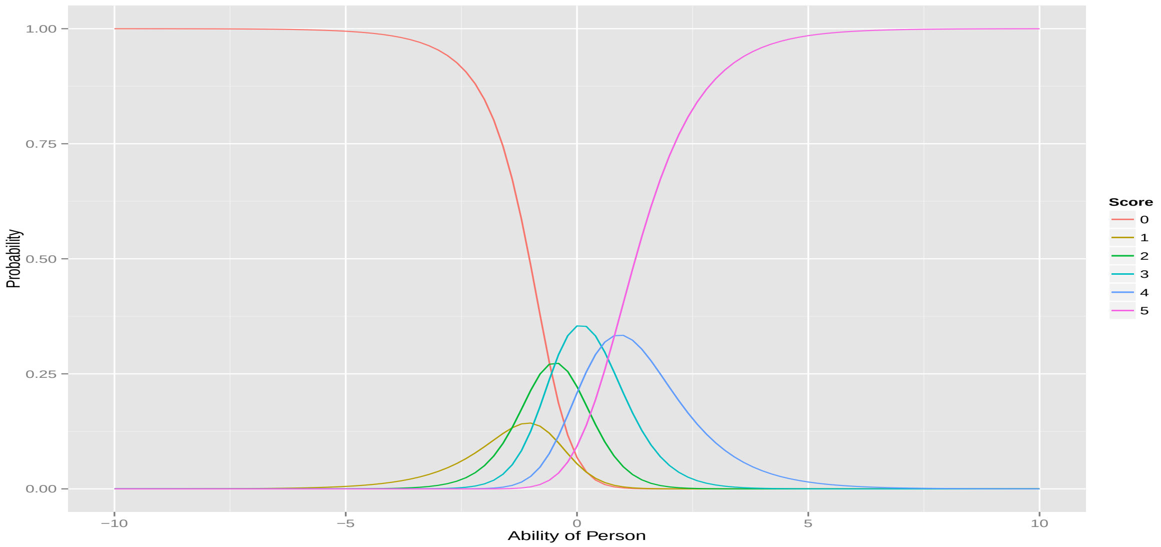

Figure 140: PCM: CPCs for item 21 in Ascot pre-test data

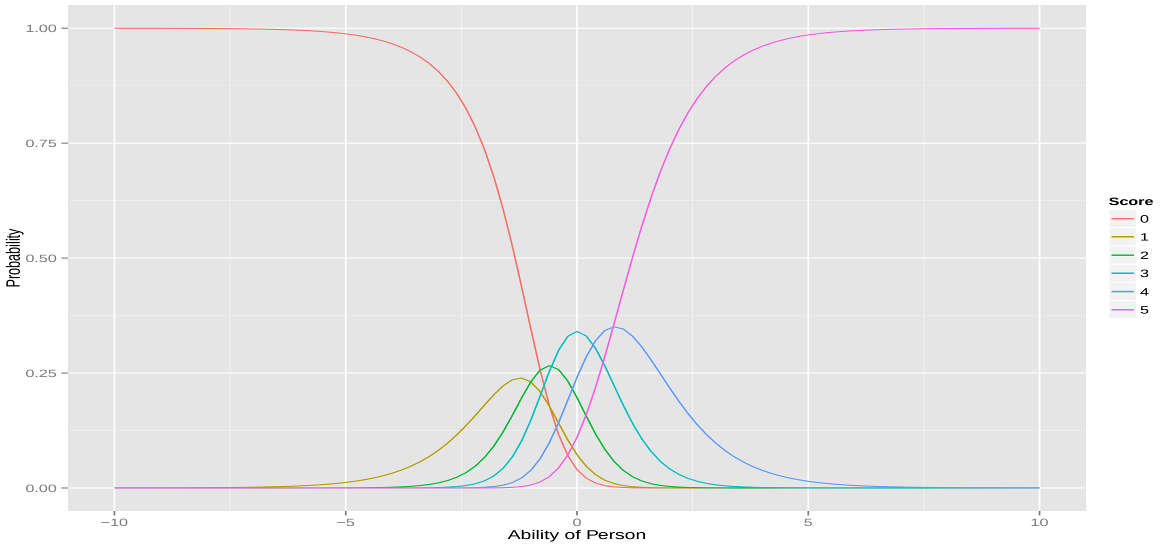

Figure 141: PCM: CPCs for item 22 in Ascot pre-test data

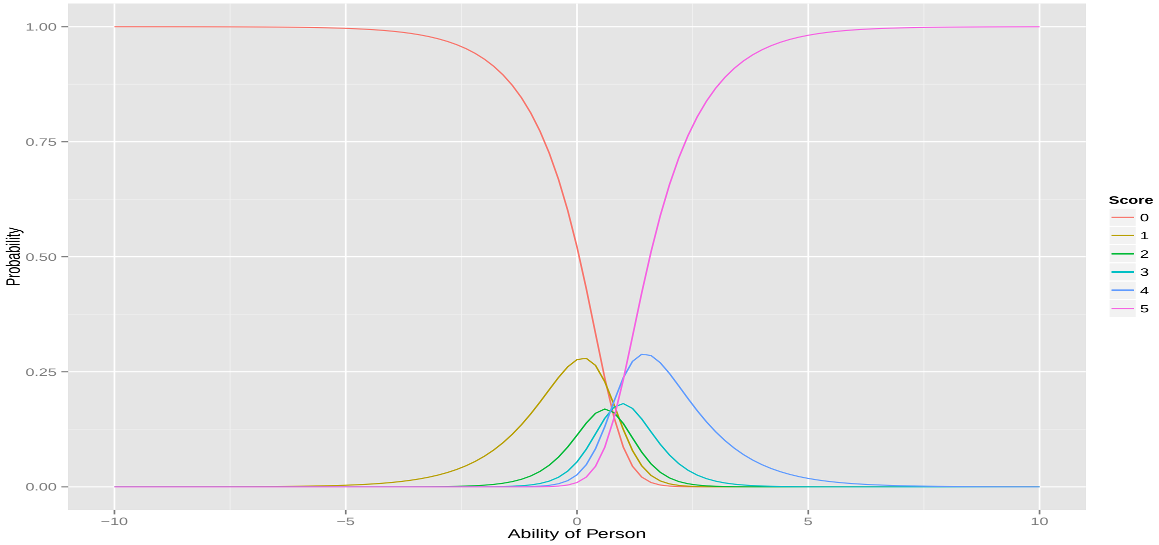

Figure 142: PCM: CPCs for item 24 in Ascot pre-test data

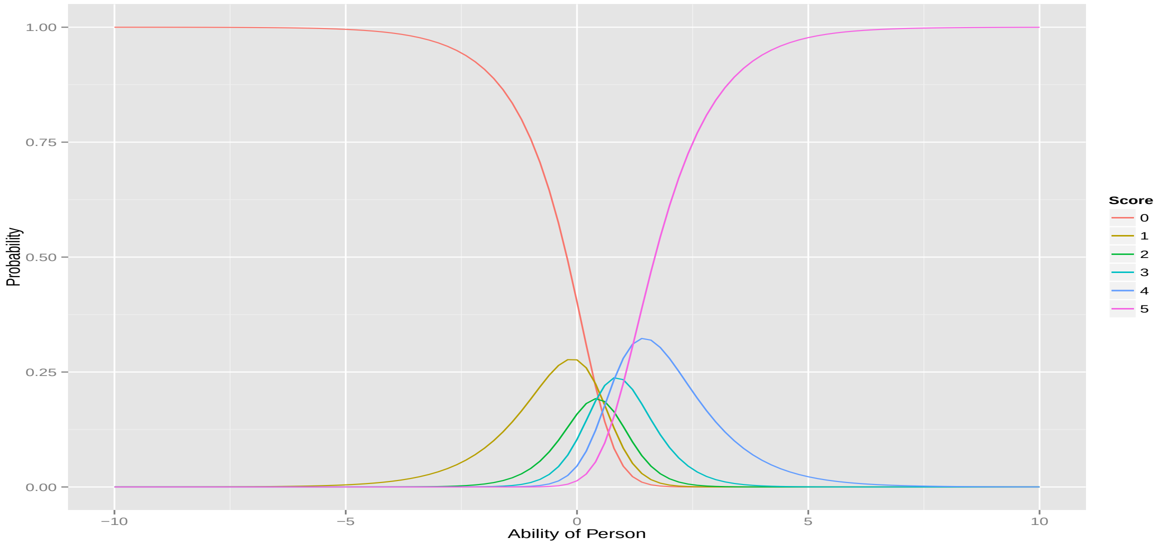

Figure 143: PCM: CPCs for item 25 in Ascot pre-test data

A.10. Statutory declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Location, date Signature

Munich; August 8, 2014 Thomas Welchowski