technische universität
dortmund

Master Thesis

Parsimonious Modelling of Threshold Parameters for Ordered Polytomous Response Data

Robin Grugel

March 22, 2023

Supervision:

Prof. Dr. Philipp Doebler
Chair of Statistical Methods in Social Sciences
Department of Statistics
TU Dortmund University

Contents

1. Introduction

Ordered polytomous response data arises in a variety of fields of measurement situations, ranging from social, educational, behavioural and economical to medical contexts. The scales of response categories are very commonly a Likert-type with a low number of possible ordered responses (e.g. 1-10). This enables a convenient analysis of response data using frequently used Item Response Theory (IRT) models like Rasch's Rating Scale Models (RSM), Graded Response Models (GRM) or Generalised Partial Credit Models (GPCM). Although very common, there exist situations in which it might be useful to utilize more categories to achieve higher reliability. A very common example is the Visual Analogue Scale (VAS) which was already established in the early 20th century to assess the intensity of pain on a visual scale of 10 cm ranging from "no pain" (0 cm) to "worst pain" (10 cm). The VAS leads clinical decisions of every medical profession dealing with patients' pain symptoms; physicians, nurses, physical therapists, psychotherapists, only naming a few. The response is generally evaluated by measuring the respondents mark from 0 in mm, which results in 101 possible (0-100) categories. The VAS is a tool which is typically used in daily practice of treatment evaluations for mentioned medical professions. The collected data sets commonly consist only of few individuals. The aforementioned approaches all require the specification of a large number of threshold parameters to model the probability of an outcome in a single response category. It is not entirely clear how the number of categories and therefore the number of thresholds affect the estimation of ability and item parameters in different frameworks. One possible problematic interference is the occurrence of empty/null categories in rather small sample sizes compared to the number of possible categories. Instead of the typical collapsing of response categories, the first measure in addressing difficult estimation with null categories is to choose a Bayesian framework with careful incorporation of prior information and implement it in the Bayesian programming environment `Stan`. As a guideline the work Y. Luo et al. (2018) is used and slightly modified regarding item specific hierarchical prior choice. Another possible issue could be the strong impact of response behaviour in the extremes of applied scales leading to a skewed and/or bimodal distribution of thresholds. In addition, the visual scale could show marks indicating single selected values of the scale impacting the response choice of examinees and consequently could cause null categories. This perspective of different functional shapes of thresholds can be addressed in different ways. Andrich et al. (2003) constructed a modification of the RSM using orthogonal polynomials with a lower number of coefficients (so called principal components) to model a high number of threshold parameters and to recover these by simply calculating differences of estimates and examined it further for the case of null categories (G. Luo et al., 2005). This idea of modelling a functional relation has already been proposed for the RSM and in a frequentist context is reevaluated by Tutz (2021) using B-Splines for a flexible approximation of item difficulty functions. The aim of this thesis is to use this idea of flexible B-Spline modelling on the threshold parameters in the very commonly used GRM and implement in a Bayesian framework.

Subsequently, the procedure is evaluated in a simulation study and is compared with the regular Bayesian GRM. In choosing B-Splines, it is expected to see a flexible approximation of unusual threshold distributions generated by uncommon measurement tools, as well as the reduction of estimation inaccuracy and bias generated by null categories. To achieve this goal, the number of B-Spline coefficients is always chosen smaller than the needed number of thresholds. Beyond that a data driven choice of knot sequence based on quantiles of a smoothed kernel density estimation is exploited. In addition, the regular GRM and B-Spline based version is applied on real data from a speeded C-test utilizing a multi-group approach for different proficiency levels of German speakers for medium many categories (0-25) and an extremely small sample case of a questionnaire capturing students' reading motivation in longitudinal design with control and treatment group for many categories (0-100). The results of the simulation study illustrate the advantages of B-spline modelling by showing a easier Bayesian implementation (less diagnostic sampling warnings regarding the MCMC draws) and more reliable point estimates with lower RMSE and bias. Both applications to the real data sets emphasize the robust behaviour in the sampling procedure of the B-spline method. However, the analyses for the speeded C-test exhibit very similar results. For the extremely small sample case in students' reading motivation, only the B-spline model could be applied, because the regular GRM shows uncontrollable divergence behaviour in the Bayesian framework. In conclusion, the B-spline approach is a promising method, that at least performs as well as the regular GRM and in some cases significantly better. The exact specification in practice and a theoretical examination though are still a necessary task to conduct. The methodical background and statistical details will be discussed in section 2 providing information about models for polytomous response data, B-Splines, the Bayesian approach for IRT models and the basic computational strategies for sampling algorithms. Section 3 will explain the simulation strategy by discussing the choice of different parameter and model setups, illustrate the evaluation technique of parameter recovery and highlight the results in a critical manner. Section 4 will present the modelling approach regarding specific prior choice, model consistency and results in detail. Finally chapter 5 will summarize the overall performance of the B-spline approach and discusses further considerations for future research.

2. Statistical Methods

2.1. Graded Response Model

Samejima (1969) proposed a family of models for response data consisting of ordered polytomous categories. Target variable of this class of models, which can be generalized to latent regression models in the structural equation modelling (Skrondal et al., 2004), is the random variable Y_{ip}, which represents the response of an examinee $p \in \{1, \dots, P\}$ of a population of size $P \in \mathbb{N}$ to the item $i \in \{1, \dots, I\}$ of a set of $I \in \mathbb{N}$ items. The actually observed value $y_{ip} \in \{1, \dots, K_i\}$ can have $K_i \in \mathbb{N}$ different values in the given categories of a possibly item-specific scale. The response behaviour of an examinee p depends on her/his abilities $\theta_p \in \mathbb{R}$, which might be a multidimensional vector $\boldsymbol{\theta}_p \in \mathbb{R}^D$ in $D \in \mathbb{N}$ different dimensions, but will for now be assumed unidimensional. Beyond this latent trait, an item specific discrimination parameter $\alpha_i \in \mathbb{R}_{>0}$ determines the discriminative ability of an item to differentiate between different ability levels of examinees. For a multidimensional ability parameter, the discrimination parameter vector $\boldsymbol{\alpha}_i \in \mathbb{R}^D_{>0}$ also subdivides its power into D dimensions. The idea of this approach is to model the probability, that an examinee with ability θ_p responds to i-th item in category k via:

$$p_{ik}(\theta_p) = P(Y_i = k|\theta_p). \tag{1}$$

This is called the category response curve (CRC), item response function (IRF) or item response category characteristic curve (IRCCC). The response random variable

$$Y_i \sim \text{Multinomial}(1, (p_{i1}(\theta_p), \dots, p_{iK_i}(\theta_p))'), \tag{2}$$

for item i is multinomially[1] distributed, as it can be seen as an urn experiment of sampling one ($n = 1$) from K_i possible categories with known frequency $p_{ik}(\theta)$ for each category (Tutz, 2011, p. 209-210) by defining the vector $\boldsymbol{U}_{ip} = (U_{ip1}, \dots, U_{ipK_i}) \in \{0,1\}^{K_i}$ of binary random variables, in which every entry is 0 except for one for the category which was drawn. In this manner Y_{ip} can be reformulated without loss of information (Agresti, 2003, p. 6-7) with the auxiliary indicator variable

$$U_{ipk} := \begin{cases} 1 & \text{for } Y_{ip} = k \\ 0 & \text{for } Y_{ip} \in \{1, \dots, K_i\} \backslash \{k\} \end{cases}, \tag{3}$$

which only is 1 for the drawn category k. The resulting special case is often called categorical distribution. As Y_i can only realize values in $\{1, \dots, K_i\}$, it is obvious that

[1] $\boldsymbol{Z} = (Z_1, \dots, Z_K)' \in \mathbb{N}^K$, as the multivariate generalisation of the binomial distribution, is a multinomially distributed ($\boldsymbol{Z} \sim \text{Multi}(n, (p_1, \dots, p_K)')$) vector of random variables with $z_k \in \{0, \dots, K\}$ for $K, n \in \mathbb{N}$ and $p_k \in [0,1]$ under the condition $\sum_{k=1}^K z_k = n$ and $\sum_{k=1}^K p_k = 1$, if it has the probability mass function $f(z_1, \dots, z_K|n, K) = n!(\prod_{k=1}^K z_k!)^{-1} \prod_{k=1}^K p_K^n$. Each of the Z_i has expected value $\text{E}(Z_i) = np_k$ and variance $\text{Var}(Z_i) = np_i(1-p_i)$ and covariance $\text{Cov}(Z_k, Z_{\tilde{k}}) = -np_kp_{\tilde{k}}$ for $k \neq \tilde{k}$ as it can be seen as dependently binomially distributed $Y_i \sim \text{Bin}(n, p_k)$ as described by Simonoff (2003, p. 68).

the condition $\sum_{k=1}^{K_i} p_{ik}(\theta_p) = 1$ holds true. This leads to the simple Likelihood function

$$\begin{aligned} p_{ik}(\theta_p) &= P(Y_i = k|\theta_p) && (4)\\ &= P(\boldsymbol{U}_{ip} = (u_{ip1}, \ldots, u_{ipK_i})'|\theta_p) && (5)\\ &= f(u_{ip1}, \ldots, u_{ipK_i}|n, K_i, \theta_p) && (6)\\ &\overset{n=1}{=} \left(\prod_{k=1}^{K_i} u_{ipk}!\right)^{-1} \prod_{k=1}^{K_i} p_{ik}(\theta_p)^{u_{ipk}} && (7)\\ &\overset{u_{ipk}\in\{0,1\}}{=} \prod_{k=1}^{K_i} p_{ik}(\theta_p)^{u_{ipk}}, && (8) \end{aligned}$$

using observed $\boldsymbol{u}_{ip} = (u_{ip1}, \ldots, u_{ipK_i})'$. The random vector $\boldsymbol{Y}_p = (Y_{1p}, \ldots, Y_{Ip})'$ is called response pattern of examinee p with realisations $\boldsymbol{y}_p = (y_{1p}, \ldots, y_{Ip})'$. It is desired to model the probability of a specific response pattern $\boldsymbol{y}_p$, which can be expressed by

$$p_{\boldsymbol{y}_p}(\theta_p) = P(\boldsymbol{Y}_p = \boldsymbol{y}|\theta_p) = \prod_{k\in\boldsymbol{y}_p} p_{ik}(\theta_p), \tag{9}$$

assuming local independence, which means the distribution of to item responses Y_1 and Y_2, given a specific value for θ_p are independent:

$$p_{(y_1,y_2)'}(\theta_p) = P(Y_1 = y_1, Y_2 = y_2|\theta_p) = P(Y_1 = y_1|\theta_p)P(Y_2 = y_2|\theta_p) = p_{y_1}(\theta_p)p_{y_2}(\theta_p). \tag{10}$$

Combining the the distributional model and the probability of a specific response pattern provides

$$p_{\boldsymbol{y}_p}(\theta_p) = \prod_{i=1}^{I}\prod_{k=1}^{K_i} p_{ik}(\theta_p)^{u_{ipk}}. \tag{11}$$

When additionally considering the reasonable assumption of independent response pattern of different examinees the likelihood

$$L_{\boldsymbol{Y}}(\theta_1, \ldots, \theta_P) = \prod_{p=1}^{P}\prod_{i=1}^{I}\prod_{k=1}^{K_i} p_{ik}(\theta)^{u_{ipk}} \tag{12}$$

models the complete response matrix $\boldsymbol{Y} \in \{1, \ldots, K_i\}^{P\times I}$

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1i} & \ldots & Y_{1I} \\ \vdots & \vdots & & \vdots & & \vdots \\ Y_{p1} & Y_{p2} & \ldots & Y_{pi} & \ldots & Y_{pI} \\ \vdots & \vdots & & \vdots & & \vdots \\ Y_{P1} & Y_{P2} & \ldots & Y_{Pi} & \ldots & Y_{PI} \end{pmatrix}. \tag{13}$$

In general, it is not necessary that the number of response categories is the same for every item illustrated be the response matrix. However, in practical application, it is very common to utilize the same response scale for all items. Samejima (1995) assumes a cognitive process, in which a category k can only be achieved, when all

cognitive steps up to $k-1$ were already completed. Therefore she used the term processing function $M_k(\theta_p)$ of the category response $k \in \{1, \ldots, K_i\}$ as the probability of an examinee with ability θ_p to respond with a category greater than or equal to k (the event $\{Y_i \geq k\}$) under the condition that the steps leading to $k-1$ are already completed (event $\{Y_i \geq k-1\}$):

$$M_k(\theta_p) = P(Y_i \geq k | \{Y_i \geq k-1\}, \theta_p) \tag{14}$$

$$= \frac{P(\{Y_i \geq k\} \cap \{Y_i \geq k-1\} | \theta_p)}{P(Y_i \geq k-1 | \theta_p)} \tag{15}$$

$$= \frac{P(Y_i \geq k | \theta)}{P(Y_i \geq k-1 | \theta)}, \tag{16}$$

for $k \in \{0, \ldots, K_i\}$, $M_k(\theta) = 0$ for $k = 0$ and $M_k(\theta) = 1$ for $k = K_i + 1$ (Samejima, 1996). Samejima (2016) examines a general framework for graded response models by distinguishing a homogeneous and a heterogeneous case (Samejima, 1972). The heterogeneous case allows different processing functions for each category, while the homogenous case assumes a common distribution. She derives the category response curve constructing the probability using the processing function to decompose into the probability of achieving category k or higher ($\star$) and the probability not to reach category $k+1$ ($\star\star$):

$$p_{ik}(\theta_p) = \overbrace{\left(\prod_{s \leq k} M_s(\theta_p)\right)}^{\star} \overbrace{\left(1 - M_{k+1}(\theta_p)\right)}^{\star\star} \tag{17}$$

$$= \prod_{s \leq k} M_s(\theta_p) - M_{k+1}(\theta_p) \prod_{s \leq k} M_s(\theta_p) \tag{18}$$

$$= \prod_{s \leq k} M_s(\theta_p) - \prod_{s \leq k+1} M_s(\theta_p) \tag{19}$$

$$= p^*_k(\theta_p) - p^*_{k+1}(\theta_p). \tag{20}$$

The last equation can be illustrated transforming more elaborate by expanding the product in the following manner:

$$\prod_{s \leq k} M_s(\theta_p) = \prod_{s \leq k} \frac{P(Y_i \geq s | \theta_p)}{P(Y_i \geq s-1 | \theta_p)} \tag{21}$$

$$= \frac{P(Y_i \geq 1 | \theta_p)}{P(Y_i \geq 0 | \theta_p)} \cdot \frac{P(Y_i \geq 2 | \theta_p)}{P(Y_i \geq 1 | \theta_p)} \cdot \ldots \cdot \frac{P(Y_i \geq y_i - 1 | \theta_p)}{P(Y_i \geq y_i - 2 | \theta_p)} \cdot \frac{P(Y_i \geq k | \theta_p)}{P(Y_i \geq k-1 | \theta_p)} \tag{22}$$

$$= \frac{P(Y_i \geq k | \theta_p)}{P(Y_i \geq 0 | \theta_p)} \tag{23}$$

$$= P(Y_i \geq k | \theta_p) \tag{24}$$

$$= \sum_{b=k}^{K_i} P(Y_i = b | \theta_p) \tag{25}$$

$$=: p^*_{ik}(\theta_p). \tag{26}$$

This leads to the possible representation of the CRC in means of the common representation of a given cumulative distribution function

$$\begin{aligned} p_{ik}(\theta_p) &= p^*_{ik}(\theta_p) - p^*_{ik+1}(\theta_p) && (27)\\ &= P(Y_i \geq k|\theta_p) - P(Y_i \geq k+1|\theta_p) && (28)\\ &= 1 - P(Y_i < k|\theta_p) - (1 - P(Y_i < k+1|\theta_p)) && (29)\\ &= 1 - P(Y_i \geq k-1|\theta_p) - (1 - P(Y_i \geq k|\theta_p)) && (30)\\ &= P(Y_i \leq k|\theta_p) - P(Y_i \leq k-1|\theta_p) && (31)\\ &= F_{Y_i|\theta_p}(k) - F_{Y_i|\theta_p}(k-1). && (32) \end{aligned}$$

Baker et al. (2004, p. 7f., 203f.) referring to Lord et al. (2008, chapter 16) describe a continuous latent item random variable $\Gamma_i \in \mathbb{R}$, which characterizes a examinees tendency to respond in a specific range of categories of the item i. In theory this relation describes a latent regression of Γ_i on θ_p. For every examinee's ability there exists a conditional distribution for the item variable with mean and variance. In which category the examinee's response will be, is specified by using a threshold parameter $\gamma_{ik} \in \mathbb{R}$, for which the strict order condition

$$-\infty = \gamma_{i0} < \gamma_{i1} \leq \gamma_{i2} \leq \ldots \leq \gamma_{iK_i-1} < \gamma_{iK_i} = \infty, \tag{33}$$

holds. With help of this formulation the response can be reformulated by the expression

$$Y_i = k \text{ for } k \in \{1, \ldots, K_i\} \iff \gamma_{ik-1} < \Gamma_i < \gamma_{ik}. \tag{34}$$

Albert et al. (1993) utilized this augmentation to implement an early approach for Bayesian analysis of ordered polytomous response models based on Gibbs sampling. If we assume $\Gamma_i \sim N(\eta, 1)$ (normally distributed[2]) or $\Gamma_i \sim L(\eta, 1)$ (logistically distributed[3], with the linear predictor $\eta \in \mathbb{R}$ allows the equivalent reformulation by centering the item variable

$$\begin{aligned} p_{ik}(\theta_p) &= P(\gamma_{ik-1} < \Gamma_i < \gamma_{ik}) && (35)\\ &= P(\gamma_{ik-1} - \eta < \Gamma_i - \eta < \gamma_{ik} - \eta) && (36)\\ &= F_{\Gamma_i - \eta}(\gamma_k - \eta_i) - F_{\Gamma_i - \eta}(\gamma_{k-1} - \eta) && (37) \end{aligned}$$

[2] A random variable $Y \in \mathbb{R}$ is normally distributed ($Z \sim N(\mu, \sigma^2)$) with mean (location parameter) $\mathrm{E}(Z) = \mu \in \mathbb{R}$ and variance (scale parameter) $\mathrm{Var}(Z) = \sigma^2 \in \mathbb{R}_{>0}$, if the probability mass function (pmf) is of the form $f_Z(z|\mu, \sigma^2) = (\sqrt{2\pi\sigma^2})^{-1} \exp(-(y-\mu)^2/(2\sigma^2))$. The cumulative distribution function (cdf) has no closed form expression, but can be calculated by $F_Z(z|\mu, \sigma^2) = \int_{-\infty}^{y} f_Z(z|\mu, \sigma^2) dz$. In case of $N(0, 1)$, it is called standard normal distributed with probability pmf ϕ and cdf Φ

[3] A random variable Z is logistically distributed ($Z \sim L(\mu, \sigma)$) with mean (location parameter) $\mathrm{E}(Z) = \mu \in \mathbb{R}$, if the probability mass function is defined as $\psi(z|\mu, \sigma) = \exp((z-\mu)/\sigma)(\sigma(1 + \exp((z-\mu)\sigma))^2)^{-1}$. The cumulative distribution is therefore defined in a closed form $\Psi(z|\mu, \sigma) = (1 + \exp(-(z-\mu)/\sigma))^{-1}$ which explains the traditional preference.

and leads with$F_{L(0,1)}(\gamma_k - \eta) = \Psi(\gamma_k - \eta)$ (or alternative traditionally less common normal distribution: $F_{N(0,1)}(\gamma_k - \eta) = \Phi(\gamma_k - \eta)$) to

$$p_{ik}(\theta_p) = \begin{cases} \Psi(\gamma_{i1} - \eta) & , \text{ for } k = 0 \\ \Psi(\gamma_{ik} - \eta) - \Psi(\gamma_{ik-1} - \eta) & , \text{ for } k \in \{1, \dots, K_i - 1\} \\ 1 - \Psi(\gamma_{iK_i-1} - \eta) & , \text{ for } k = K_i \end{cases}. \tag{38}$$

In the classical GRM the linear predictor is defined as $\eta = \alpha_i \theta_p$ and incorporates item and examinee information and uncertainty; and can easily extended by covariates. All these relationships and the especially the latent regression aspect can be recognized in figure 1. One can see the general regression of the ability in the item variable and the distribution of every single ability parameter which determines the probability of a response in the provided categories. It is apparent, that examinees with lower ability parameters have higher probabilities to respond to the item with a lower category than to a higher one.

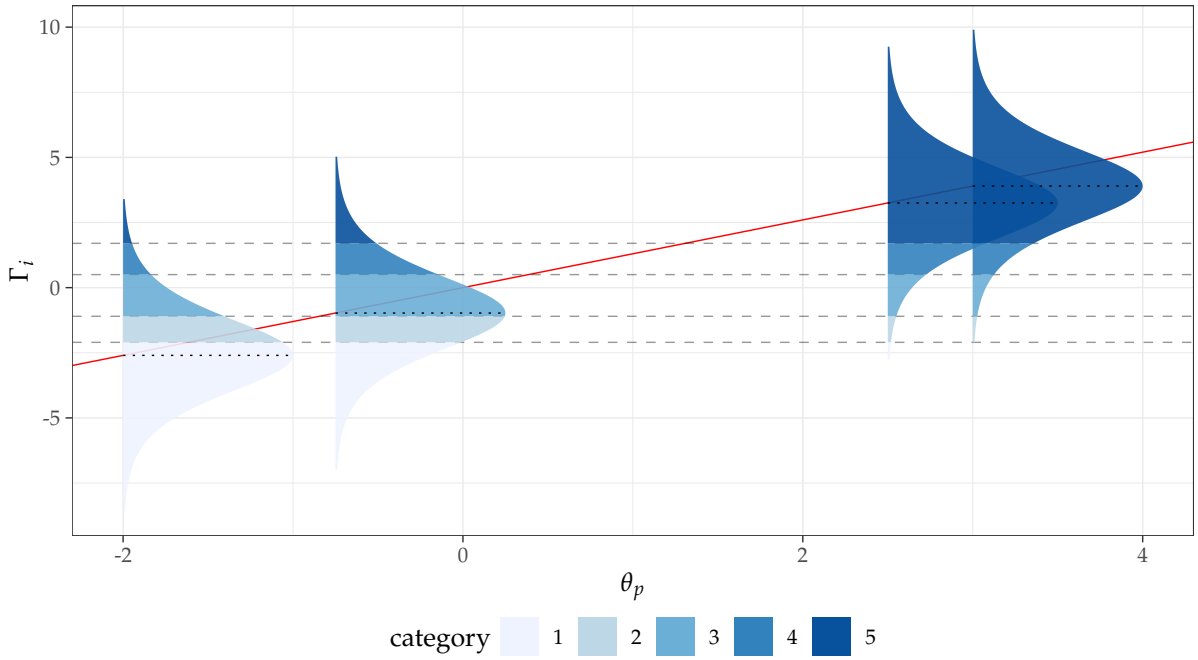

Figure 1: Latent regression of a variable Γ_i of an item i with threshold parameter vector $\gamma_i = (-2.1, -1.1, 0.5, 1.7)'$ and discrimination parameter $\alpha_i = 1.3$ showing the ability distribution for values $\theta_p \in \{-2, -0.75, 2.5, 3\}$

That behaves in the opposite way for examinees with higher ability parameters. The slope of the regression line is determined by the discrimination parameter α_i. Higher values for α_i lead visually to a wider dispersion of distribution means and therefore simplifies to differentiate between the individual response probabilities and therefore examinees response behaviour.

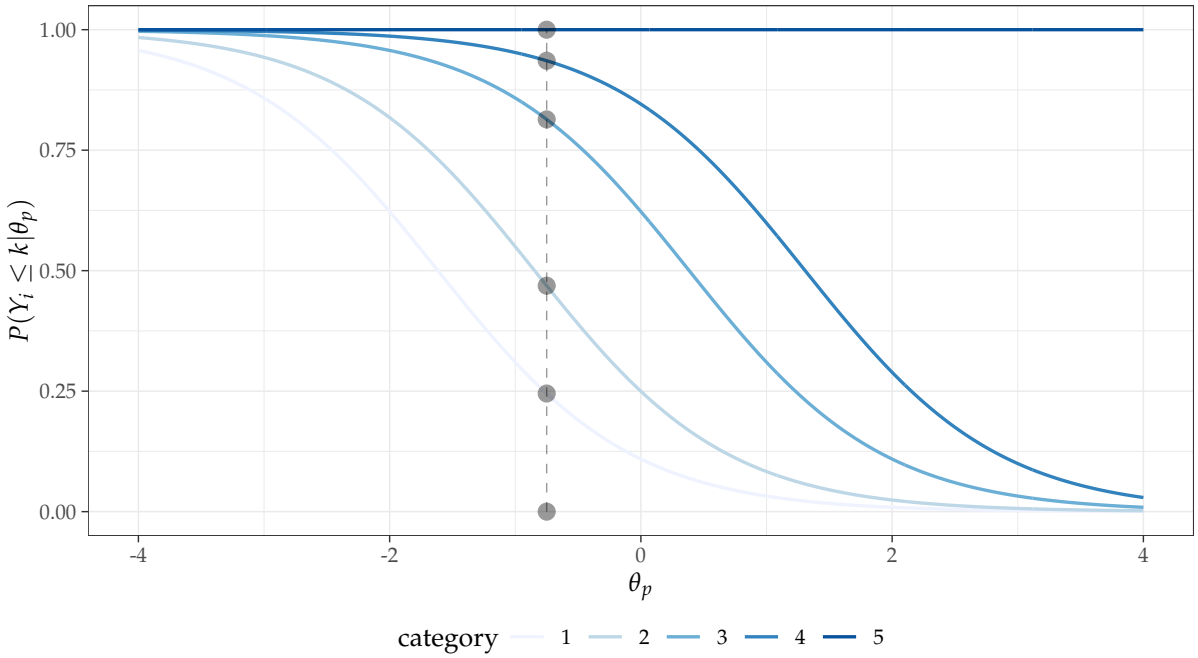

Figure 2: Category response boundary curve (CRBC) $P(Y_i \leq k|\theta_p)$ of an item i with threshold parameter vector $\gamma_i = (-2.1, -1.1, 0.5, 1.7)'$ and discrimination parameter $\alpha_i = 1.3$ evaluated once in $\theta_p = -0.75$

To calculate and display these distributions it is necessary to utilize the category boundary curve $P(Y_i \leq k|\theta_p)$, seen in figure 2. For one specific ability θ_p, one can calculate the differences of $P(Y_i \leq k|\theta_p)$ for adjacent categories, which leads to the response probabilities $P(Y_i \leq k|\theta_p)$. This vertical differences determine the CRC for every θ_p. Figure 3 shows the shape of CRCs for the complete space of ability parameters.

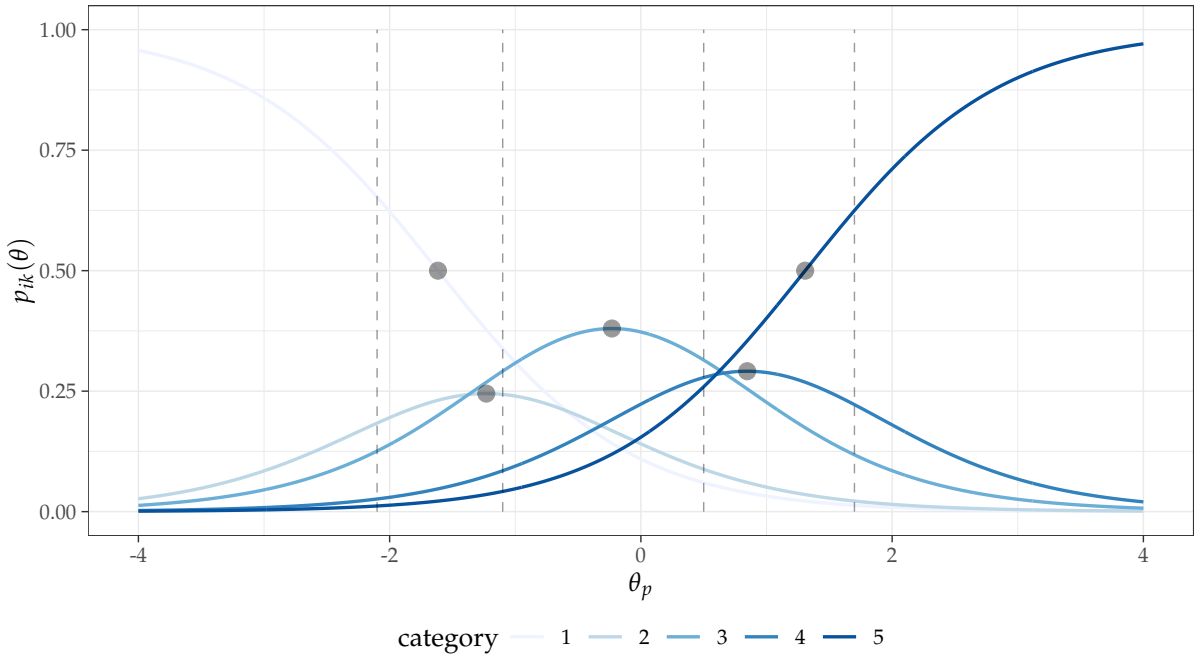

Figure 3: Category response curve (CRC) $p_{ik}(\theta_p)$ of an item i with threshold parameter vector $\gamma_i = (-2.1, -1.1, 0.5, 1.7)'$ and discrimination parameter $\alpha_i = 1.3$

Muraki et al. (1995) investigate multidimensional compensatory case with item discrimination parameter $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iD})' \in \mathbb{R}^D_{>0}$ for the linear predictor $\eta_i = \boldsymbol{\alpha}_i' \boldsymbol{\theta}$. This implies a cognitive setting in which a higher response requires at least one high value in the vector $\boldsymbol{\theta}$ and it does not necessarily have to be a specific one; they are able to compensate for each other in the weighted sum of the predictor. It might be useful to evaluate the Fisher information (e.g. for computerized adaptive testing, determine the optimal choice of item from a pool to present the examinee next) via Item Information and therefore the Item Category Information. The latter one is as usual defined as the expectation of the binary auxiliary variable

$$I_{ik}(\theta_p) = \frac{\left(\frac{\partial}{\partial \theta_p} p_{ik}(\theta_p)\right)^2}{p_{ik}(\theta_p)} - \frac{\partial^2}{\partial^2 \theta^2} p_{ik}(\theta_p), \tag{39}$$

and leads via summation of all category informations to the item information

$$I_i(\theta_p) = \sum_{k=1}^{K_i} I_{ik}(\theta_p). \tag{40}$$

Figure 4 illustrates the item information as a function of θ_p, which allows to make statements on how good an assessment differentiates between ability levels of different ranges on the support. The displayed item information leads to the conclusion, that this hypothetical item is especially informative about examinees with lower abilities. Only the category 5 is slightly informative regarding the ability parameters. In practical measurement situations, one would likely utilize more than one item to improve possibility to conduct valid inference for wider range of ability parameters.

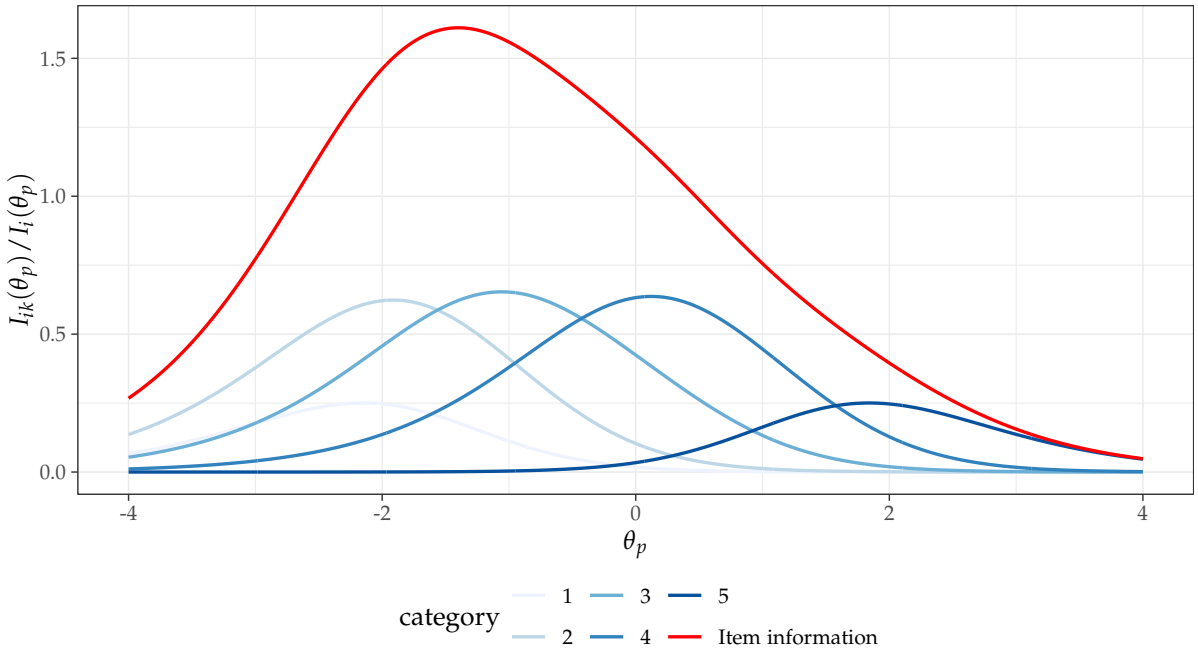

Figure 4: Item Category Information $I_{ik}(\theta)$ and Item Information $I_i(\theta)$ of an item i with threshold parameter vector $\gamma_i = (-2.1, -1.1, 0.5, 1.7)'$ and discrimination parameter $\alpha_i = 1.3$

2.2. B-Splines

Wherever data has to be fitted or curves of unknown shape have to be approximated, spline functions as a linear combination of a B-spline basis are a convenient choice, through their flexibility and simple computation. In this approach mentioned flexibility will be used to model the distributional characteristic of latent threshold variables $\gamma_{ik} \approx \gamma_i(k)$ as function of the category. This idea was already mentioned by Tutz (2021) in a frequentist context for a more flexible difficulty function. For the construction of a basis spline or B-splines, define a partition or knot sequence as non-decreasing values $\boldsymbol{t} = \{t_1, t_2, \ldots, t_m\}$ and an order $d \in \mathbb{N}$ on a real interval $[t_{\text{lo}}, t_{\text{up}}] \subset \mathbb{R}$. The B-spline of order $d > 1$ was defined by De Boor et al. (2001, p. 90) via the recurrence formula

$$B_{m,d,\boldsymbol{t}}(y) = \omega_{m,d,\boldsymbol{t}}(y) B_{m,d-1,\boldsymbol{t}}(y) + (1 - \omega_{m+1,d,\boldsymbol{t}}(y)) B_{m+1,d-1,\boldsymbol{t}}(y), \tag{41}$$

with

$$\omega_{m,d,\boldsymbol{t}}(y) = \frac{y - t_m}{t_{m+d-1} - t_m} \mathbb{1}_{\{t_m \neq t_{m+d-1}\}}, \tag{42}$$

although the classical way is by using divided differences or explicitly written piecewise polynomials. These B-splines have to conform to the condition, that they form a partition of the unity, meaning $\sum_{t \in \boldsymbol{t}} B_{m,d,\boldsymbol{t}}(t) = 1$. The first B-splines of order $d = 1$ are the indicator functions for the given knot sequence

$$B_{m,1,\boldsymbol{t}}(y) = \mathbb{1}_{[t_m, t_{m+1})}(y), \tag{43}$$

which are right-continuous and do satisfy the mentioned condition. From this point one can easily calculate the higher order B-spline functions with the help of $B_{m,1,\boldsymbol{t}}$. For the second order $d = 2$, the B-spline is therefore formulated as

$$B_{m,2,\boldsymbol{t}}(y) = \omega_{m,2,\boldsymbol{t}}(y) B_{m,1,\boldsymbol{t}}(y) + (1 - \omega_{m+1,2,\boldsymbol{t}}(y)) B_{m+1,1,\boldsymbol{t}}(y) \tag{44}$$

and analogously for $d = 3$ as

$$\begin{aligned} B_{m,3,\boldsymbol{t}}(y) = {} & \omega_{m,3,\boldsymbol{t}}(y)(\omega_{m,2,\boldsymbol{t}}(y) B_{m,1,\boldsymbol{t}}(y) + (1 - \omega_{m+1,2,\boldsymbol{t}}(y)) B_{m+1,1,\boldsymbol{t}}(y)) & (45) \\ & + (1 - \omega_{m+1,3,\boldsymbol{t}})(\omega_{m+1,2,\boldsymbol{t}} B_{m+1,1,\boldsymbol{t}}(y) + (1 - \omega_{m+2,2,\boldsymbol{t}}) B_{m+2,1,\boldsymbol{t}}(y)). & (46) \end{aligned}$$

In section A.2 of the Appendix, the functions explicitly calculated for d up to 3 can be found. This is because only splines using a B-spline basis of order 3 are being employed in this approach. This illustrates, that third order B-spline functions consist of 3 quadratic polynomial pieces that are smoothly joined at the knots. This polynomials are $d - 1$-times continuously differentiable at the knots, if there are no multiple knots. The interior knot sequence is usually extended by specifying boundary knots as the lower limit t_{lo} and upper limit t_{up} of the relevant interval. There are different approaches to ensure theoretical B-spline properties (replicating boundary knots infinitely many times for a bi-infinite view on theory of the sequence) that are described by De Boor et al. (2001) in detail, discussing the multiplicity of knots.

A spline for a fixed order of k and a fixed knot sequence $\boldsymbol{t}$ is defined as the linear combination of M B-splines $B_{1,d,\boldsymbol{t}}, \ldots, B_{M,d,\boldsymbol{t}}$

$$s(y) = \sum_{m=1}^{M} \lambda_m B_{m,d,\boldsymbol{t}}(y), \tag{47}$$

with B-spline coefficients $\lambda_1, \ldots, \lambda_M \in \mathbb{R}$. Every function s from the resulting functional space

$$\mathbb{S}_{d,\boldsymbol{t}} = \left\{ \sum_{m=1}^{M} \lambda_m B_{m,d,\boldsymbol{t}}(y) \;\middle|\; \lambda_m \in \mathbb{R} \right\}, \tag{48}$$

is a spline function. Figure 5 illustrates the local support of a B-spline basis and the smoothness of the corresponding spline function. In this special case, the B-spline knots are equally spaced and their are additional knots, which are equal to the boundary knots. The linear combination covers the function space equally.

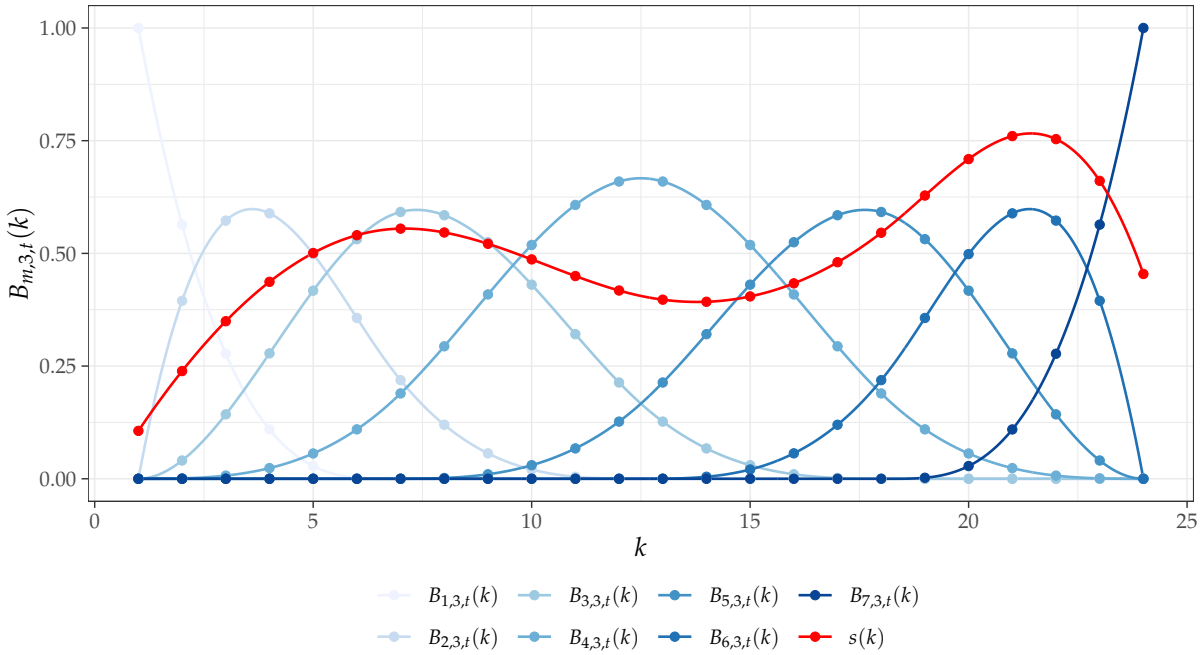

Figure 5: B-spline basis of degree $d = 3$ for $M = 7$ with randomly sampled $\lambda_m \sim U(0,1)$

In the case of modelling the distribution of the latent threshold variable in the GRM, it is necessary to impose a certain shape constraint on the spline. Threshold parameters and their functional approximation have to be monotonously increasing with increasing category index. One possible approach is to define monotonously non-decreasing basis functions, which was mentioned by Ramsay (1988) with the I-splines basically as integrals over B-splines. This would restrict the I-spline coefficient to be positive and would require an added constant to the spline function to model variables on the complete real functional space. An alternative possible which already was utilized by Leitenstorfer et al. (2007) and Brezger et al. (2008) exploits a property of splines and especially their derivatives. Abraham et al. (2015) constitutes with help of De Boor

et al. (2001, p. 116), that spline function in monotonously increasing if $\frac{\partial}{\partial k} s(k) \geq 0$. After the determination of the derivative

$$\frac{\partial}{\partial y} s(y) = \frac{\partial}{\partial y} \sum_{m=1}^{M} \lambda_m B_{m,d,\boldsymbol{t}}(y) \tag{49}$$

$$= \sum_{m=2}^{M} (d-1) \frac{\lambda_m - \lambda_{m-1}}{t_{m+d-1} - t_m} B_{m,d-1,\boldsymbol{t}}(y) \tag{50}$$

and the fact that $t_{m+d-1} > t_m$ holds, it can easily be seen after examining the term $\lambda_m - \lambda_{m-1}$, that

$$\frac{\partial}{\partial} s(y) \geq 0 \iff \lambda_1 \leq \cdots \leq \lambda_M. \tag{51}$$

Imposing the constraint on the B-spline coefficients leads to a modification

$$\widetilde{\mathbb{S}}_{d,\boldsymbol{t}} = \left\{ \sum_{m=1}^{M} \lambda_j B_{m,k,\boldsymbol{t}}(y) \;\middle|\; \lambda_m \leq \lambda_{m+1} \text{ for all } \lambda_m \in \mathbb{R} \right\}, \tag{52}$$

of the function space 48, which ensures monotonously increasing splines. Figure 6 illustrates the effect of monotonously increasing coefficients on the overall shape of the spline function. This realization enables to easily use well-known Bayesian modelling techniques only adding a constraint on the parameters. This can easily be integrated in the Bayesian workflow using `Stan`.

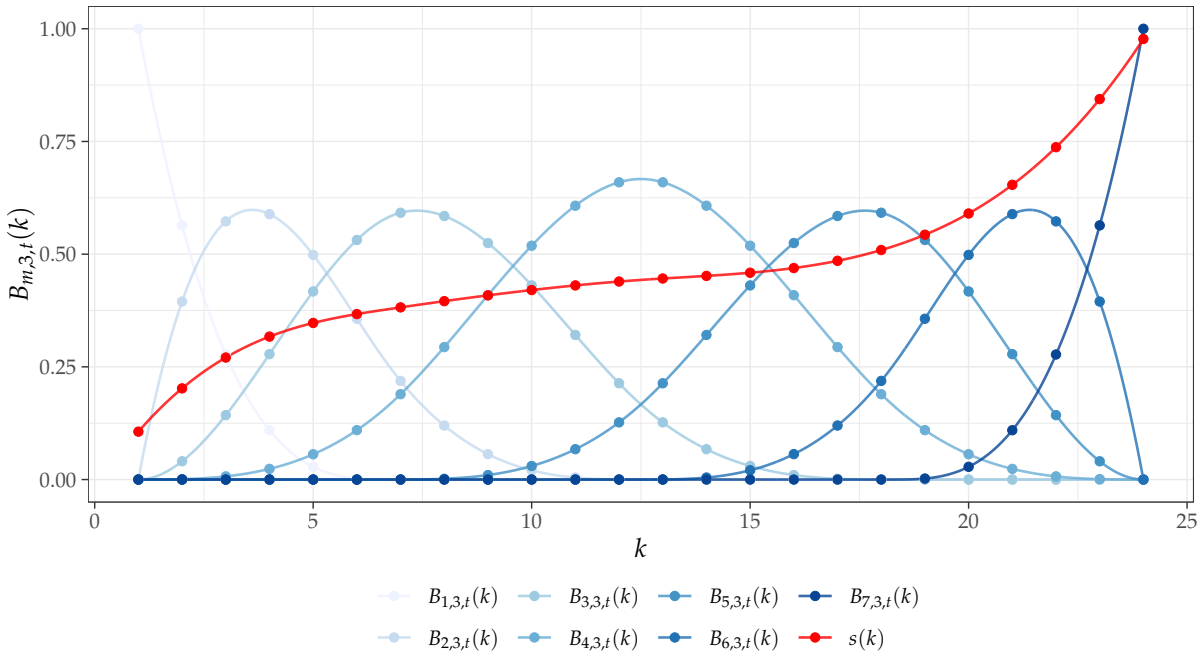

Figure 6: Monotonously increasing spline utilizing a B-spline basis of degree $d = 3$ for $M = 7$ with randomly sampled and increasingly sorted $\lambda_m \sim U(0,1)$

A simple approach to determining the knots would be to choose a simple equally spaced sequence which covers the relevant region of interest. This might be inefficient because areas with few responses and therefore less information would be covered

by a lot of knots. Therefore it's common to choose the knots for a B-spline in a data driven way via the sample quantiles of the observed responses. For very homogenous responses, caused by a heavily skewed or peaked response distribution, this approach would result in a knot sequence with knots of high multiplicity. Even if it is possible to deal with such high multiplicity, it is easier to deal with a strictly increasing sequence of knots without disadvantages. To achieve this given constraint, this approach attempts to use a smoothed version of quantiles by utilizing a kernel density estimate. The kernel density estimator

$$\widehat{f}_{\boldsymbol{y}_i}(t) = \frac{1}{Ph} \sum_{p=1}^{P} Kern\left(\frac{t - y_{ip}}{h}\right), \tag{53}$$

defined by Silverman (1986, p. 45), uses a weighted sum of kernel functions $Kern$, which are centred at the observed values $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iP})'$ of P examinees on one item i. This method is explicitly defined for continuous variable, but will be applied to categorical responses data because it is not used to directly model the distribution. Besides the Epanechnikov kernel, the Gaussian kernel function

$$Kern(x) = \frac{1}{\sqrt{2\pi}}\left(-\frac{x^2}{2}\right) \tag{54}$$

is the most commonly used. The scalar $h \in \mathbb{R}_{>0}$ is the bandwidth parameter, which controls the width of the kernel function and therefore regulates the overall smoothness. The optimal choice for the bandwidth can be very difficult and is based on different criteria. However, for the simple task of providing an alternative version of the sample quantiles, the common variation of the rule of thumb

$$\widehat{h} = 1.06\widehat{\sigma}_{\boldsymbol{y}_i} P^{-1/5}, \tag{55}$$

introduced by Scott (1992), will be sufficient. For this calculations the cumulative distribution function $F_{Y_i}(k)$ can be approximated via

$$\widehat{F}_{\boldsymbol{y}_i}(k) = \frac{1}{c} \sum_{t=1}^{k} \widehat{f}_{\boldsymbol{y}_i}(t) \tag{56}$$

with $c = \sum_{t \in A} \widehat{f}_{\boldsymbol{y}_i}(t)$ which ensures regular function values. To achieve a valid approximation one has to choose a fine partition over $A = [0, K_i] \in \mathbb{R}$ for the summation of density estimates. After estimation of this quantity one can exploit the definition

$$\widehat{Q}_{\boldsymbol{y}_i}(\tilde{p}) = \inf\{t \in \mathbb{R} | \widehat{F}_{\boldsymbol{y}_i}(t) \geq \tilde{p}\} \tag{57}$$

which provides a strictly increasing knot sequence. The actually chosen knot sequence decides about the quality of approximation and heavily depends on the number of knots. A too large number of knots could lead to an overfitting, capturing measurement errors rather than the actual functional relationship. A too large number of knots might not be able to capture important characteristics. In the threshold modelling

approach $\lfloor\sqrt{K_i}\rfloor + 2$ interior knots for the thresholds of item i seems a reasonable choice, because it is able to collect enough information about the polytomous response to map to the continuous parameter space. Including the boundary knot this leads to the calculation $M = \lfloor\sqrt{K_i}\rfloor + 4$. This is a rough rule of thumb, but proved superior to higher numbers in attempts to fit in a Bayesian framework, but might be evaluated mathematically and via more elaborate simulations. This B-spline modelling approach can be used on an item basis leading to I different version of the B-spline basis matrix $\boldsymbol{B}^i_{d,\boldsymbol{t}}$ and their matching coefficients $\boldsymbol{\lambda}_i$. For a fixed degree d and knot sequence $\boldsymbol{t}$ with length M including the boundary knots, the evaluated B-spline values for a specific set of finite arguments $\{1, \ldots, K_i - 1\}$ (corresponding to the threshold parameters in the GRM) can be presented in form of a matrix $\boldsymbol{B}_{d,\boldsymbol{t}} \in \mathbb{R}^{K_i \times M}$ of the appearance:

$$\boldsymbol{B}^i_{d,\boldsymbol{t}} = \begin{pmatrix} B_{1,d,\boldsymbol{t}}(1) & B_{2,d,\boldsymbol{t}}(1) & \cdots & B_{M,d,\boldsymbol{t}}(1) \\ B_{1,d,\boldsymbol{t}}(2) & B_{2,d,\boldsymbol{t}}(2) & \cdots & B_{M,d,\boldsymbol{t}}(2) \\ \vdots & \vdots & & \vdots \\ B_{1,d,\boldsymbol{t}}(K_i - 1) & B_{2,d,\boldsymbol{t}}(K_i - 1) & \cdots & B_{M,d,\boldsymbol{t}}(K_i - 1). \end{pmatrix} \tag{58}$$

This provides the easy calculation of the spline function via $\boldsymbol{B}_{d,\boldsymbol{t}}\boldsymbol{\lambda}$. A convenient implementation for B-splines in the `R`-package with the function `bs()` enables to easily specify a spline matrix providing only a sequence of function arguments, the knot sequence and the polynomial degree.

2.3. Bayesian Estimation

In Bayesian estimation for item response models, as in the general approach of Bayesian statistic, the inference is made by interpreting the parameters $\boldsymbol{\kappa}$ of a given sampling model. This model tries to explain the random data generating process leading to a specific response pattern $\boldsymbol{Y}$ observed on I items and P examinees (M. S. Johnson et al., 2016). The main difference between a Bayesian approach to a Frequentist one is the assumption, that parameter vector $\boldsymbol{\kappa} \in \mathbb{R}^{N_{\text{par}}}$, for $N_{\text{par}} \in \mathbb{N}$ parameter overall, is seen as a random variable with a specific distribution and not as a fixed value. This specific distribution can quantify uncertainty or prior information about the parameter. Like in Frequentist (maximum-likelihood) methods, for Bayesian estimation the sampling model is expressed in a likelihood function for the observed response pattern $\boldsymbol{Y}$ given a set of model parameters $\boldsymbol{\kappa} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ in terms of a density function $f(y_{pi}|\boldsymbol{\kappa})$ of a single response of one participant

$$L(\boldsymbol{Y};\boldsymbol{\kappa}) = f(\boldsymbol{Y}|\boldsymbol{\kappa}) = \prod_{p=1}^{P} f(\boldsymbol{y}_p|\boldsymbol{\kappa}) = \prod_{p=1}^{P}\prod_{i=1}^{I} f(y_{pi}|\boldsymbol{\kappa}). \tag{59}$$

The conditional density of the model parameters $\boldsymbol{\kappa}$ given the observed data $\boldsymbol{Y}$ is called the posterior distribution and is the basis for all inference conducted in Bayesian analysis in Statistics, and is formulated in the eponymously Bayes' Theorem

$$f(\boldsymbol{\kappa}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y},\boldsymbol{\kappa})}{f(\boldsymbol{Y})} = \frac{f(\boldsymbol{Y}|\boldsymbol{\kappa})f(\boldsymbol{\kappa})}{\int f(\boldsymbol{Y},\boldsymbol{\kappa})d\boldsymbol{\kappa}}. \tag{60}$$

This equation reflects the joint distribution density

$$f(\boldsymbol{Y},\boldsymbol{\kappa}) = f(\boldsymbol{Y}|\boldsymbol{\kappa})f(\boldsymbol{\kappa}), \tag{61}$$

of the response pattern and all model parameters, which contains information on the sampling (Likelihood) and prior knowledge about the model parameter used in the specification. This prior knowledge is formulated in the prior distribution density $f(\boldsymbol{\kappa})$ and has to be chosen carefully to prevent affecting the statistical inference unintentionally in a significant way without using the observed data. The marginal distribution of the response pattern

$$f(\boldsymbol{Y}) = \int f(\boldsymbol{Y},\boldsymbol{\kappa})d\boldsymbol{\kappa} \tag{62}$$

can be obtained by integrating over the complete space of model parameters and therefore does not depend on these any more. In this regard, $f(\boldsymbol{Y})$ ensures that the posterior density is a proper density and integrates to one. For most inferential statements, the actual value of a density does not matter that much, as statistical moments of posterior distributions are often sufficient statistics and especially independent from a normalizing constant. Therefore the posterior is often simplified as proportional to Likelihood times prior:

$$f(\boldsymbol{\kappa}|\boldsymbol{Y}) \propto L(\boldsymbol{Y};\boldsymbol{\kappa})f(\boldsymbol{\kappa}). \tag{63}$$

In general the posterior density can not be easily obtained analytically and has to be approximated with Monte Carlo simulations. Only in specific cases of choosing a conjugate prior for the likelihood function, a convenient closed-form expression for the posterior density can be used without utilizing numerical integration. Modern implementations of sampling algorithms for Markov Chains provide a powerful tool for complex models and a robust statistical posterior inference. Once the posterior density is obtained the complete information about a parameters distribution, not only a single point estimate for a specific moment, can be used to provide meaningful inference. As the complete posterior distribution in its entirety is hard to interpret and compare it is useful to do inference based on distributional moments of parameters using simple point estimators. The most common Bayesian point estimate is the expected a posteriori (EAP) estimator which is the expected value

$$\mathrm{E}(\boldsymbol{\kappa}|\boldsymbol{Y}) = \int \boldsymbol{\kappa} f(\boldsymbol{\kappa}|\boldsymbol{Y}) d\boldsymbol{\kappa} \tag{64}$$

of the parameters $\boldsymbol{\kappa}$ given the actually observed response matrix $\boldsymbol{Y}$. Another point estimator could be the posterior median (0.5-quantile), which gives the smallest value for $\boldsymbol{\kappa}$ where its cumulative distribution exceeds the probability of 0.5. The posterior mode (MAP) represents another point estimate, which takes the value for $\boldsymbol{\kappa}$ where the posterior density $f(\boldsymbol{\kappa}|\boldsymbol{Y})$ reaches its maximum. In a lot of complex modelling situations, it is necessary to utilize MCMC methods to approximate the posterior density. In these situation, a simple approximation of the EAP estimator can be calculated by using the sampled values $\boldsymbol{\kappa}^{(m)}$ with $m \in \{1, \ldots, M\}$ of $M \in \mathbb{N}$ number of posterior draws (possibly after a warm-up phase of the MCMC algorithm, where the Markov Chain already converged to its stationary distribution). With these values only the mean

$$\widehat{\mathrm{E}}(\boldsymbol{\kappa}|\boldsymbol{Y}) = \overline{\boldsymbol{\kappa}}_{\text{draw}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\kappa}^{(m)} \tag{65}$$

has to be calculated. The more robust posterior median is as simply obtained by computation of the sample median of the non warm-up parameter draws. Another point estimate for the dispersion of the posterior distribution is the posterior covariance (for scalar $\boldsymbol{\kappa} = \kappa$ the variance)

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\kappa}|\boldsymbol{Y}) &= \mathrm{E}[(\boldsymbol{\kappa} - \mathrm{E}(\boldsymbol{\kappa}|\boldsymbol{Y}))(\boldsymbol{\kappa} - \mathrm{E}(\boldsymbol{\kappa}|\boldsymbol{Y}))'|\boldsymbol{Y}] && (66) \\ &= \int (\boldsymbol{\kappa} - \mathrm{E}(\boldsymbol{\kappa}|\boldsymbol{Y}))(\boldsymbol{\kappa} - \mathrm{E}(\boldsymbol{\kappa}|\boldsymbol{Y}))' f(\boldsymbol{\kappa}|\boldsymbol{Y}) d\boldsymbol{\kappa} && (67) \end{aligned}$$

of the posterior distribution. In the MCMC applications, it can be approximated analogously to the posterior mean via the empirical covariance by using the posterior draws in the equation

$$\widehat{\mathrm{Cov}}(\boldsymbol{\kappa}|\boldsymbol{Y}) = \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\kappa}^{(m)} - \overline{\boldsymbol{\kappa}}_{\text{draw}})(\boldsymbol{\kappa}^{(m)} - \overline{\boldsymbol{\kappa}}_{\text{draw}})'. \tag{68}$$

To provide more detailed information about the posterior distribution of model parameters the use of credible sets (in most applied cases intervals) as the Bayesian

equivalent of the confidence interval is mandatory. Reporting point estimates without their standard error and some sort of interval estimation to illustrate how confident one can be about the estimated values, lacks a foundation for a valid interpretation of the model or causal relationships. A $(1-\tau)\cdot 100\%$ credible set S_{cred} is a subset of the complete parameter space $\boldsymbol{S}$ of $\boldsymbol{\kappa}$, which satisfies the condition $P(\boldsymbol{\kappa} \in S_{\text{cred}}|\boldsymbol{Y}) = 1-\tau$. This condition can be met be different construction methods. A simple method for continuous parameters can be to use the posterior quantile function $F^{-1}_{\kappa_n|\boldsymbol{Y}}$ in

$$\boldsymbol{S}_{\text{cred},\tau}(\boldsymbol{\kappa}) = \left\{\kappa \in \boldsymbol{S} | F^{-1}_{\kappa|\boldsymbol{Y}}(\epsilon) < \kappa < F^{-1}_{\kappa|\boldsymbol{Y}}(1-\tau+\epsilon)\right\} \text{ for } \epsilon \in [0, 1-\tau], \tag{69}$$

for a two-sided credible interval. If the tail probabilities are chosen to be equal in form of $\epsilon = \tau/2$ the interval is called the equal-tailed credible interval. This is approach is especially attractive as the theoretical quantile function $F^{-1}_{\kappa|\boldsymbol{Y}}$ can be replaced by the approximative version $\widehat{F}^{-1}_{\kappa_{\text{draw}}|\boldsymbol{Y}}$ obtained via samples of κ_{draw} using Monte Carlo methods based of discrete Markov Chains.

2.3.1. Prior Distribution Choice

A central quantity in Bayesian statistics with which every specification starts are prior densities. These have to be chosen with great care to prevent affecting the posterior inference against the observed information of the data. In case of item response theory models the parameters $\boldsymbol{\kappa} = \{\boldsymbol{\theta}, \boldsymbol{\xi}\}, \theta_p \in \mathbb{R}$, can be further separated into ability/person/examinee parameters $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\xi}$. Fox (2010, p. 31-39) highlights the relevance of those parameters for the interpretation of estimated values. The within-individual heterogeniety in the responses provides especially information about the measurement instrument itself and can be reflected by examining posterior distributions of item parameters. The between-individual heterogeniety in responses on the other hand gives particular insight into the specific level of ability or latent trait; this can be mainly derived from ability parameters. As the interpretation of those topics in measurement contexts are highly relevant, the careful choice of prior information is a central aspect of statistical analysis. Levy et al. (2017, p. 273) approaches prior specification with the assumption, that the joint prior density for all model parameters can be factorized into ability and item parameters

$$f(\boldsymbol{\kappa}) = f(\boldsymbol{\theta}, \boldsymbol{\xi}) = f(\boldsymbol{\theta}) f(\boldsymbol{\xi}), \tag{70}$$

which implies independence of latent ability and item parameters of the measurement instrument. This enables convenient independent choice of priors but could potentially lead to potential misspecification of the true model. Generally, the strength of a prior distribution is controlled by its dispersion/variance (if existing), where a small value indicates a strong belief in a value and a large one, the lack of knowledge (Ames, 2018). The location/mean of the prior distribution on the other hand reflects the value for a specific parameter, which is assumed with the previously mentioned confidence quantified by the dispersion. Generally, it is popular to use weakly informative or even

subjective priors in specific applications, as the knowledge about some parameters is readily available based on previous surveys and their analyses. Also, the choice of priors does not have to be limited to conjugate priors (prior distributions, which lead in combination with the likelihood to a known family of posterior distribution), because modern sampling algorithm as implemented in `Stan` offer a performant framework for convenient Bayesian computation. The prior distribution choice for the person parameter $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_P\}$ in a univariate modelling approach are commonly chosen as normal in the form

$$\theta_p | \mu_\theta, \sigma_\theta^2 \overset{iid}{\sim} N(\mu_\theta, \sigma_\theta^2) \text{ for } p \in \{1, \ldots, P\}, \tag{71}$$

with unknown mean $\mu_\theta \in \mathbb{R}$ and unknown variance $\sigma_\theta^2 \in \mathbb{R}_{>0}$ as hyper parameters with their hyper prior distribution, which can be modelled on a higher level (Fox, 2010, p. 38). This implies stochastic independence conditional in the parameters and is not necessary for fully Bayesian inference. Assuming the individual examinees were randomly sampled from a larger population with the same probability, the aforementioned independence appears reasonable. To ensure identifiability and interpretability it is wide practice to select $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$, this identifies the scale of discrimination parameters and the location of threshold parameters in the GRM context, as described by Levy et al. (2017, p. 272-278). Additionally, this choice of prior enables better comparability to frequentist frameworks, which impose the above constraints for ability parameters estimation as well. This concept can be easily modified to a D-dimensional multivariate person parameter vector with

$$\boldsymbol{\theta}_p | \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta = (\theta_{p1}, \ldots, \theta_{pD})' | \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta \overset{iid}{\sim} N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \text{ for } p \in \{1, \ldots, P\}, \tag{72}$$

with mean $\boldsymbol{\mu}_\theta \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_\theta \in \mathbb{R}_{\geq 0}^{D \times D}$. This expresses correlations between the dimensions of the trait in more complex multivariate item response theory situations. In case of more elaborate multi-group or longitudinal study design modelling in the context of partial pooling, prior for hyper parameter and specifically fixed parameters are required and have to be specified for given application. These models are commonly approached in the framework of structural equation modelling (SEM), where complex relationships between observed and latent variables are modelled in systems of equations.

The choice of prior distributions for the item parameters $\boldsymbol{\xi} = \{\boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ depend heavily on the type and their specific role in the GRM. Similar to the independence of ability and measurement model parameters, the item parameters itself are commonly assumed independent simplifying the measurement joint prior density

$$f(\boldsymbol{\xi}) = f(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = f(\boldsymbol{\alpha}) f(\boldsymbol{\gamma}), \tag{73}$$

which imposes even stronger assumptions in the sense of assuming independence of discrimination and "difficulty"/"location" of the items. Theoretically, all parameters can be dependent on each other by sharing hyper priors or even assuming a common multivariate distribution, for illustration purposes of the chosen approach, this will

be neglected. For item discrimination parameters $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_I\}$ the strict positivity condition $\alpha_i \in \mathbb{R}_{>0}$ has to be reflected in the prior distribution. To achieve this positivity, log-normal, inverse-gamma or truncated normal are commonly used. More recently Y. Luo et al. (2018) utilized the Cauchy[4] distribution with its heavy tails. One approach analogous to the ability parameter is to assume independence of the items, which would imply the absence of a relation in sense of a common stimulus and assume for example half Cauchy

$$\alpha_i | \mu_\alpha, \sigma_\alpha \overset{iid}{\sim} Cauchy(\mu_\alpha, \sigma_\alpha) \mathbb{1}_{(0,\infty)} \text{ for } i \in \{1, \ldots, I\}. \tag{74}$$

The fixed choice of location $\mu_\alpha \in \mathbb{R}$ and scale $\sigma_\alpha \in \mathbb{R}_{>0}$ can be guided by expertise knowledge about typical values of discrimination parameters. Baker et al. (2017) describes typical values for α_i and what are extremely rare observed estimations in practical analyses. The prior distribution is therefore specified in a manner that all observed values and a wide margin are represented with a high probability. While Levy et al. (2017, p. 275) assumes $\alpha_i \overset{iid}{\sim} N(0, 2)\mathbb{1}_{(0,\infty)}$ and Fox (2010, p. 96) chooses $\alpha_i \overset{iid}{\sim} N(0, 1)\mathbb{1}_{(0,\infty)}$, Y. Luo et al. (2018) specified $\alpha_i \overset{iid}{\sim} Cauchy(0, 5)\mathbb{1}_{(0,\infty)}$, reflecting a less informative prior distribution. Another option would be to allow a relationship by choosing a common hyper prior for $\mu_\alpha \in \mathbb{R}$ and $\sigma_\alpha \in \mathbb{R}_{>0}$. In case of a common stimulus in sense of a testlet structure, it might be more appropriate to define a multivariate distribution, for example, a truncated multivariate normal distribution

$$\boldsymbol{\alpha} | \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha = (\alpha_1, \ldots, \alpha_I)' | \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha \overset{iid}{\sim} N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \text{ for } i \in \{1, \ldots, I\}, \tag{75}$$

with mean $\boldsymbol{\mu}_\alpha \in \mathbb{R}^I_{>0}$ and covariance matrix $\boldsymbol{\Sigma}_\alpha \in \mathbb{R}^{I \times I}_{\geq 0}$, which defines the exact relationship between items in greater detail. The threshold parameters

$$\gamma = \{\gamma_1, \ldots, \gamma_I\} = \{(\gamma_{i1}, \ldots, \gamma_{iK_i})', \ldots, (\gamma_{I1}, \ldots, \gamma_{IK_I})'\} \tag{76}$$

for $I \in \mathbb{N}$ items and $K_i \in \mathbb{N}$ categories for item i require a more elaborate prior specification due to their monotonicity constraint. One approach is to use truncated prior distribution for the k-th threshold parameter, whose support is bounded by the value of the $(k-1)$-th threshold parameter. Levy et al. (2017) and Y. Luo et al. (2018) utilize the half-normal distribution by defining

$$\gamma_{ik} | \mu_{\gamma_i}, \sigma^2_{\gamma_i} \sim N(\mu_{\gamma_i}, \sigma^2_{\gamma_i}) \mathbb{1}_{(\gamma_{ik-1}, \infty)} \text{ for } k \in \{1, \ldots, K_i - 1\} \text{ and } i \in \{1, \ldots, I\}, \tag{77}$$

with the mean threshold $\mu_{\gamma_i} \in \mathbb{R}$ and variance $\sigma^2_{\gamma_i} \in \mathbb{R}_{>0}$. Levy et al. (2017) uses fixed $\mu_{\gamma_i} \in \{2, 1, -1, -2\}$ for a 5-category response in an application example for every item in the same way. Y. Luo et al. (2018) in contrast defines a hyper prior for both threshold distribution parameters over all items. It might be helpful to utilize an

[4] A random variable Z is Cauchy distributed with location parameter $\mu \in \mathbb{R}$ and positive scale parameter $\sigma \in \mathbb{R}_{>0}$, if the probability mass function f is of the form $f(z) = (\pi\sigma)^{-1}(1 + (z - \mu)^2/\sigma^2)^{-1}$; The truncated or half Cauchy distribution is defined by the pmf $f(z) = 2(\pi\sigma)^{-1}(1 + (z - \mu)^2/\sigma^2)^{-1}\mathbb{1}_{[\mu,\infty)}(z)$

item-specific hyper prior, if the true distribution of threshold parameter are reasonably assumed to be very different between items. In `Stan` the order constraint $\gamma_{ik} < \gamma_{ik+1}$ for $k \in \{1, \ldots, K_i - 2\}$ of the vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iK_i}) \in \mathbb{R}^{K_i}$ is achieved directly by applying a transform g[5] to enable convenient sampling from the unconstrained space. By applying

$$\widetilde{\gamma}_{ik} = \begin{cases} \gamma_{i1} & \text{if } k = 1 \\ \log(\gamma_{ik} - \gamma_{ik-1}) & \text{if } k \in \{2, \ldots, K_i - 1\} \end{cases} \tag{78}$$

to the monotone sequence, $\boldsymbol{\gamma}_i$ is mapped to an unconstrained vector $\widetilde{\boldsymbol{\gamma}}_i \in \mathbb{R}^{K_i}$. The inverse is straightforwardly defined as

$$\gamma_{ik} = \begin{cases} \widetilde{\gamma}_{i1} & \text{if } k = 1 \\ \widetilde{\gamma}_{ik-1} + \exp(\widetilde{\gamma}_{ik}) & \text{if } k \in \{2, \ldots, K_i - 1\} \end{cases} \tag{79}$$

or alternatively as $\gamma_{ik} = \widetilde{\gamma}_{i1} + \sum_{j=2}^{k} \exp(\widetilde{\gamma}_{ij})$. To obtain the density function for the unconstrained version $\widetilde{\boldsymbol{\gamma}}_i$, applying the transformation for probability densities results in

$$f_{\widetilde{\boldsymbol{\gamma}}_i}(\widetilde{\boldsymbol{\gamma}}_i) = f_{\boldsymbol{\gamma}_i}(g^{-1}(\widetilde{\boldsymbol{\gamma}}_i)) \prod_{k=2}^{K_i - 1} \exp(\widetilde{\gamma}_{ik}), \tag{80}$$

as the Jacobian $\boldsymbol{J}^i$ for item i of the inverse g^{-1} is a lower triangular matrix with diagonal elements

$$J^i_{k,k} = \begin{cases} 1 & \text{if } k = 1 \\ \exp(\widetilde{\gamma}_{ik}) & \text{if } k \in \{2, \ldots, K_i - 1\} \end{cases}, \tag{81}$$

leading to a convenient determinant for triangular matrices of the form $|\det(\boldsymbol{J}^i)| = \prod_{k=1}^{K_i - 1} J^i_{k,k} = \prod_{k=2}^{K_i - 1} \exp(\widetilde{\gamma}_{ik})$. These are the transformations that are conducted in `Stan`, if a parameter vector is defined as a constraint variable class `ordered` and does not have to be addressed manually.

2.3.2. Model Identification

Although the posterior distribution does always exist, if the parameters are specified with proper priors (their density functions integrate to one over the complete parameter space; second Kolmogorov axiom), it is necessary to further specify the model, so the posterior inference is unambiguous and therefore interpretable meaningfully (San Martin, 2018). San Martin and Gonzalez (2010) defines a model as identified if the equation

$$f(\boldsymbol{Y}|\boldsymbol{\kappa}) = f(\boldsymbol{Y}|\widetilde{\boldsymbol{\kappa}}) \implies \boldsymbol{\kappa} = \widetilde{\boldsymbol{\kappa}} \tag{82}$$

of likelihood functions $p(\boldsymbol{Y}|\bullet)$ of a given response pattern $\boldsymbol{Y}$ holds true for parameters $\boldsymbol{\kappa}, \widetilde{\boldsymbol{\kappa}}$. In some common cases, specific constraints have to be made to achieve defined equality. In this approach, the identification of used models will be achieved by fixing the location and scale of the ability parameters.

[5] `https://mc-stan.org/docs/reference-manual/ordered-vector.html`

2.3.3. Posterior Predictive Checking

To evaluate if the specification of the fitted model is consistent with the observed data, Gelman et al. (2013, p. 143-153) describes posterior predictive checking. The basic assumption for the principle of posterior predictive checking is, that the data from the observed experiment or survey should look similar to the values of the posterior predictive distribution. Therefore thorough comparison should be conducted in a practical application, which could reveal a systematic difference caused by model misspecification. The posterior predictive distribution of a observed response pattern $\boldsymbol{Y} \in \{1, \ldots, K\}^{P \times I}$ and the replicated response pattern $\boldsymbol{Y}^{\text{rep}} \in \{1, \ldots, K\}^{P \times I}$, drawn in the MCMC method is defined as

$$f(\boldsymbol{Y}^{\text{rep}}|\boldsymbol{Y}) = \int f(\boldsymbol{Y}^{\text{rep}}|\boldsymbol{\kappa}) f(\boldsymbol{\kappa}|\boldsymbol{Y}) d\boldsymbol{\kappa}. \tag{83}$$

Based on this statistic quantity, a simple posterior predictive check is the visual comparison of observed and replicated data distribution. For the Graded Response Model case that could be the item-wise comparison of barplots reflecting the frequency of responses in each category. This is a rough but useful diagnostic to reveal serious problems, like replicated data outside the possible range of categories or severe differences in the location or scale of the posterior predictive distribution. Optimally there should be very little visually recognizable discrepancy of both barplots of ordered polytomous responses. Another possibility is to choose a suitable test statistic $T(\bullet, \boldsymbol{\kappa})$ for the response data to quantify the discrepancy by determining the probability

$$\begin{aligned} p_T &= P(T(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\kappa}) \geq T(\boldsymbol{Y}, \boldsymbol{\kappa})|\boldsymbol{Y}) && (84)\\ &= \int \int \mathbb{1}_{[T(\boldsymbol{Y},\boldsymbol{\kappa}),\infty)}(T(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\kappa})) f(\boldsymbol{Y}^{\text{rep}}|\boldsymbol{\kappa}) f(\boldsymbol{\kappa}|\boldsymbol{Y}) d\boldsymbol{\kappa} d\boldsymbol{Y}^{\text{rep}} && (85)\\ &= \mathrm{E}(\mathbb{1}_{[T(\boldsymbol{Y},\boldsymbol{\kappa}),\infty)}(T(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\kappa}))|\boldsymbol{Y}, \boldsymbol{\kappa})) && (86) \end{aligned}$$

This probability is called the posterior predictive p-value and it reflects the fit quality of the model by its deviation from the value 0.5. That means p-values near 0 or 1 imply that it is unlikely to recognize the observed pattern $\boldsymbol{Y}$ in the set of replicated patterns $\{\boldsymbol{Y}^{\text{rep}(1)}, \ldots, \boldsymbol{Y}^{\text{rep}(M)}\}$. Based on this observation the fitted model can be reasonably assumed as suitable or as not that different from the "true" model concerning the tested quantity. In practical applications there should be examined various test quantities, that might be reflecting interesting characteristics. In MCMC simulations the predictive p-value can be easily calculated by calculating the proportion

$$\begin{aligned} \widehat{p}_T &= \widehat{\mathrm{E}}(\mathbb{1}_{[T(\boldsymbol{Y},\boldsymbol{\kappa}),\infty)}(T(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\kappa})|\boldsymbol{Y}, \boldsymbol{\kappa}))) && (87)\\ &= \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{[T(\boldsymbol{Y},\boldsymbol{\kappa}),\infty)}(T(\boldsymbol{Y}^{\text{rep}(m)}, \boldsymbol{\kappa}^{(m)})), && (88) \end{aligned}$$

of simulated quantities which satisfies the condition $T(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\kappa}) \geq T(\boldsymbol{Y}, \boldsymbol{\kappa})$ for $M \in \mathbb{N}$ steps of the converged Markov chain. It might be useful in IRT models to further investigate quantities of item response vectors $\boldsymbol{Y}_i$ for the i-th item or even the single

responses Y_{ip} for examinee p, to evaluate suspicious items or specials examinees (outliers). This is also described by Sinharay (2006) and modified in a visual diagnostic, where the observed scores are plotted against the estimated p-value. This might be problematic for a lot of null categories in small samples with a wide response scale. The choice of test statistics is commonly basic statistics measures like the mean, standard deviance, quantiles or skewness and kurtosis to validate the overall shape and location of the response distribution. Posterior predictive checks should not be confused with model comparison and inference but only serves as a control step in checking model adequacy (Meng, 1994).

2.3.4. Model Evaluation and Comparison

Bayesian models can be evaluated and compared using their predictive accuracy. The model, which predicts new observations with the highest accuracy is likely the more appropriate one for further inferences. Gelman et al. (2013, p. 166-182) note that the evaluation of this quality of fit is optimally conducted by using alternative or future data $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_{ip})_{(i=1,\dots,I;p=1,\dots,\widetilde{P})}$ with $\widetilde{P} \in \mathbb{N}$ the number of new observations, which was in contrast to observed $\boldsymbol{Y} = (Y_{ip})_{(i=1,\dots,I;p=1,\dots,P)}$ not used to fit the model a sample from the posterior distribution. The accuracy of out-of-sample prediction can be quantified by using the log-likelihood values given the sampled parameters. This prediction is can be formulated as the expected logarithm of the pointwise predictive density

$$elpd = \sum_{i=1}^{I} \sum_{p=1}^{\widetilde{P}} \int f_{\widetilde{Y}_{ip}}(\widetilde{y}_{ip}) \log(f(\widetilde{y}_{ip}|\boldsymbol{Y})) d\widetilde{y}_{ip} \tag{89}$$

with true density $f_{\widetilde{Y}_{ip}}$ of future observations. The estimate of the LOO $elpd$ out-of-sample predictive fit in the Bayesian context is

$$elpd^{\text{loo}} = \sum_{i=1}^{I} \sum_{p=1}^{P} \log(f(y_{ip}|y_{-ip})) \tag{90}$$

$$= \sum_{p=1}^{P} \log\left(\int f(y_{ip}|\boldsymbol{\kappa}) f(\boldsymbol{\kappa}|y_{-ip}) d\boldsymbol{\kappa} \right) \tag{91}$$

in which y_{ip} stands for a specific response of examinee p to item i and y_{-ip} for all other response excluding this specific one. To find an appropriate estimate Vehtari, Gelman, et al. (2017) describes a Pareto smoothed version of importance sampling where the importance ratios defined by

$$r_{ip}^{(m)} = \frac{1}{f(y_{ip}|\boldsymbol{\kappa}^{(m)})} \propto \frac{f(\boldsymbol{\kappa}^{(m)}|y_{-ip})}{f(\boldsymbol{\kappa}^{(m)}|\boldsymbol{Y})}, \tag{92}$$

are improved in a 3 step smoothing procedure and resulting in weights $w_{ip}^{(m)}$ for each response y_{ip}. The approach is to fit a generalized Pareto distribution to the largest 20 percent of the importance ratios $r_{ip}^{(m)}$ for $m \in \{1, \dots, M\}$ (step 1). Replace the

$S = 0.2 \cdot M$ largest ratios by the expected values of the order statistics of the fitted generalized Pareto distribution[6] $F^{-1}((s-0.5)/S)$ for $s \in \{1, \ldots, S\}$ and labelling them as $\widetilde{w}_{ip}^{(m)}$ (step 2). Truncate each weight $w_{ip}^{(m)} = \min\{\widetilde{w}_{ip}^{(m)}, M^{-1/4}\sum_{m=1}^{M}\widetilde{w}_{ip}^{(m)}\}$ at the average of the smoothed weights (step 3). Finally, the estimated Pareto smoothed importance sampled LOO version of the $elpd$ is defined by

$$\widehat{elpd}^{\text{psis-loo}} = \sum_{p=1}^{P}\sum_{i=1}^{I}\log\left(\frac{\sum_{m=1}^{M} w_{ip}^{(m)} f(y_{ip}|\boldsymbol{\kappa}^{(m)})}{\sum_{m=1}^{M} w_{ip}^{(m)}}\right). \tag{93}$$

This procedure is readily available in the R-package `loo` by Vehtari, Gabry, et al. (2022), which utilizes results from sampled `Stan`-model. The estimated values for the shape parameter $\widehat{k}$ of the generalized Pareto distribution is a possible guide to evaluate how reliable the estimated accuracy measure is. Explicit values and how to interpret them meaningfully are described by Vehtari, Gelman, et al. (2017) and can be found in the vignettes of `loo`[7]. This is also an additional opportunity to assess model misspecification based on prediction accuracy and find particular influential responses compared to the rest. In the use of model comparisons of two different models A, B one would choose A over B, if $\widehat{elpd}_A^{\text{psis-loo}} > \widehat{elpd}_B^{\text{psis-loo}}$. For choosing a model the difference of $elpd$s

$$\widehat{elpd}_A^{\text{psis-loo}} - \widehat{elpd}_B^{\text{psis-loo}} =: \widehat{elpd}_{\text{diff}(A,B)}^{\text{psis-loo}}, \tag{94}$$

can be used, which is (for actual estimated differences) of a magnitude concerning the improvement of the chosen model A against B. As in most applications, simple model specifications are preferred to complex ones, if the $\widehat{elpd}^{\text{psis-loo}}$s are roughly the same. R-function `compare_loo()` of the `loo` package ranks differences for all models in an increasing order relative to the model with the largest $elpd$. The elpd formulation from above assumes a conditional log-likelihood, where the latent model parameters in the IRT context are kept, as conditional independence allows the simple summation. Merkle et al. (2019) addressed this approach for the WAIC in comparison to the marginal version, where the latent model parameters are integrated out. This method is more appropriate for psychometric applications, because the prediction of new observations is typically done for clusters (e.q. multiple items) of responses and should be modelled as such. The conditional version is implicitly implemented when the R-package `loo` is used and still is useful conducting model comparisons, especially in this crude and early development state of a Bayesian method. For elaborate Bayesian inference in applied research however, the marginal likelihood should be used for model comparison to achieve valid causal conclusions.

[6] A random variable $Z \in \mathbb{R}$ is generalized Pareto distributed with shape parameter $\xi \in \mathbb{R}$ if it has the probability density function $f_Z(z) = (1+\xi z)^{-(\xi+1)/\xi}\mathbb{1}_{\mathbb{R}\setminus\{0\}}(\xi) + \exp(-z)\mathbb{1}_{\{0\}}(\xi)$ and cumulative distribution function $F_Z(z) = (1-(1+\xi z)^{-1/\xi}\mathbb{1}_{\mathbb{R}\setminus\{0\}}(\xi) + (1-\exp(-z))\mathbb{1}_{\{0\}}(\xi)$, for values $Z \in \mathbb{R}_{\geq 0}$ for $\xi \geq 0$ and $Z \in [0, -1/\xi]$ for $\xi < 0$ (N. L. Johnson et al., 1995, p. 573-576).

[7] `https://mc-stan.org/loo/articles/online-only/faq.html`

2.3.5. Bayesian Computation

In Bayesian modelling the choice of conjugate prior for simple models, especially non-hierarchical ones, the posterior is obtainable analytically and the inference can be made using its exact distribution. In more complex applied statistical models, however it is necessary to use approximation methods to conduct Bayesian analyses. Aside from bias introducing deterministic algorithms, the field of Markov chain Monte Carlo (MCMC) methods enables the use of highly complex model specifications to achieve reliable inference. In MCMC methods the posterior $f(\boldsymbol{\kappa})$ is seen as the stationary distribution of a sequence of random variables $\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}, \boldsymbol{\kappa}^{(3)}, \ldots$ satisfying the Markov property

$$P(\boldsymbol{\kappa}^{(m+1)} = \boldsymbol{k} | \boldsymbol{\kappa}^{(1)} = \boldsymbol{k}_1, \boldsymbol{\kappa}^{(2)} = \boldsymbol{k}_1, \ldots, \boldsymbol{\kappa}^{(m)} = \boldsymbol{k}_m) = P(\boldsymbol{\kappa}^{(m+1)} = \boldsymbol{k} | \boldsymbol{\kappa}^{(m)} = \boldsymbol{k}_m), \quad (95)$$

which means, that the probability of the next state of $\boldsymbol{\kappa}^{(m+1)}$ depends only on the present state $\boldsymbol{\kappa}^{(m)}$ and not on the complete sequence/"history". There are a lot of sampling methods that simulate Markov chains meeting the conditions of irreducibility and aperiodicity to obtain the stationary distribution of the process, the target posterior distribution in a Bayesian context. Gibbs sampling and the basic Metropolis algorithm are prone to exhibit random walk behaviour of the Markov chain, which results in an inefficient exploration of the posterior density space (Gelman et al., 2013, p. 275-302). For the specific case of IRT models in fully Bayesian analyses, there are a lot of specialized sampling algorithms (Albert et al., 1993) proposed for the wide variety of possible models (dichotomous response: 1PL, 2PL, 3PL, etc.; polytomous response: nominal, ordinal). Hamiltonian Monte Carlo (HMC) on the other hand approaches this problem by introducing an auxiliary variable $\boldsymbol{r} \in \mathbb{R}^{N_{\text{par}}}$ with the same dimension as $\boldsymbol{\kappa}$. This augmented model can be seen as a Hamiltonian system in a physical sense describing the continuous movement of a particle in space in a discrete approximation. The variable $\boldsymbol{r}$ represents the momentum of the particle in the position $\boldsymbol{\theta}$ in the parameter space of dimension $\mathbb{R}^{N_{\text{par}}}$. The negative potential energy $\mathcal{L}(\boldsymbol{\kappa}) = \log(f(\boldsymbol{\kappa}|\boldsymbol{Y}))$ in the specific position $\boldsymbol{\kappa}$ and the kinetic energy $0.5\boldsymbol{r}'\boldsymbol{r}$ of the particle, can be used to define the negative energy $\log(f(\boldsymbol{\kappa}, \boldsymbol{r}|\boldsymbol{Y}))$, with $f(\boldsymbol{\kappa}, \boldsymbol{r}|\boldsymbol{Y}) \propto \exp(\mathcal{L}(\boldsymbol{\kappa}) - 0.5\boldsymbol{r}'\boldsymbol{r})$. The Markov chain can be simulated in an iterative algorithm starting with meaningful initial values for $\boldsymbol{\kappa}$ and sampling the first value for $\boldsymbol{r}$ in every iteration $m \in \{1, \ldots, M\}$. The Hamiltonian dynamic is reflected in the updating of $\boldsymbol{\kappa}$ and $\boldsymbol{r}$ after L leapfrog steps, in which $\boldsymbol{r}$ is updated based on gradient $\nabla_{\boldsymbol{\kappa}}\mathcal{L}(\boldsymbol{\kappa}^{(m)})$ for current $\boldsymbol{\kappa}^{(m)}$ and on the other hand $\boldsymbol{\theta}$ based on the current momentum. An iteration is completed by the Metropolis step of accepting the proposed value for $\boldsymbol{\kappa}$ with probability α_{acc} reflecting the change of energy $\log(f(\widetilde{\boldsymbol{\kappa}}, \widetilde{\boldsymbol{r}}|\boldsymbol{Y})) - \log(f(\boldsymbol{\kappa}, \boldsymbol{r}|\boldsymbol{Y}))$ in the simulated Hamiltonian system overall leapfrog steps from the last state and its current proposal.

Algorithm 1: Simple Hamiltonian Monte Carlo

input : initial parameter values $\boldsymbol{\kappa}^{(0)} \in \mathbb{R}$
step size $\epsilon \in \mathbb{R}$
number of steps $L \in \mathbb{N}$
logarithm of joint density of $\mathcal{L}(\boldsymbol{\kappa})$
number of samples drawn $M \in \mathbb{N}$

output: vector of sampled $\boldsymbol{\kappa} \in \mathbb{R}^M$

for $m \in \{1, \ldots, M\}$ **do**

  step 1: draw initial random sample $\boldsymbol{r}^{(0)} \sim N(\mathbf{0}, \boldsymbol{I})$

  step 2: set initial values

  set $\boldsymbol{\kappa}^{(m)}$ to $\boldsymbol{\kappa}^{(m-1)}$

  set $\widetilde{\boldsymbol{\kappa}}$ to $\boldsymbol{\kappa}^{(m-1)}$

  set $\widetilde{\boldsymbol{r}}$ to $\boldsymbol{r}^{(0)}$

  for $l \in \{1, \ldots, L\}$ **do**

    step 3: apply leapfrog integrator

    set $\widetilde{\boldsymbol{r}}$ to $\boldsymbol{r}^{(m)} + 0.5\epsilon \nabla_{\boldsymbol{\kappa}} \mathcal{L}(\boldsymbol{\kappa}^{(m)})$

    set $\boldsymbol{\kappa}^{(m)}$ to $\boldsymbol{\kappa}^{(m)} + \epsilon \widetilde{\boldsymbol{\kappa}}$

    set $\widetilde{\boldsymbol{r}}$ to $\widetilde{\boldsymbol{r}} + 0.5\epsilon \nabla_{\boldsymbol{\kappa}} \mathcal{L}(\boldsymbol{\kappa}^{(m)})$

  end

  step 4: metropolis acceptance with probability α_{acc}

  draw $u \sim U[0, 1]$

  set α_{acc} to $\min \left\{ 1, \frac{\exp(\mathcal{L}(\widetilde{\boldsymbol{\kappa}}^{(m)}) - 0.5\widetilde{\boldsymbol{r}}'\widetilde{\boldsymbol{r}})}{\exp(\mathcal{L}(\boldsymbol{\kappa}^{(m)}) - 0.5\boldsymbol{r}^{(0)\prime}\boldsymbol{r}^{(0)})} \right\}$

  if $u \leq \alpha_{acc}$ **then**

    set $\boldsymbol{\kappa}^{(m)}$ to $\boldsymbol{\kappa}^{(m)}$

    set $\boldsymbol{r}$ to $-\widetilde{\boldsymbol{r}}$

  else

    set $\boldsymbol{\kappa}^{(m)}$ to $\boldsymbol{\kappa}^{(m-1)}$

    set $\boldsymbol{r}$ to $-\widetilde{\boldsymbol{r}}$

  end

end

A simple HMC is illustrated in algorithm 1 in which the auxiliary momentum variable is sampled from a multivariate normal distribution with the identity matrix $\boldsymbol{I}$, which indicates independence of the single entries of $\boldsymbol{r}$ for each dimension. Another possibility is to define another matrix, for example a scaled to the posterior covariance version, specify the mass of the particle the physical point of view and might improve efficiency of the sampling procedure. Betancourt (2017) gives a thorough view of this dynamics from a physical point of view. The general issue is that the performance is strongly dependent on chosen step size ϵ and number of steps L, which have to be tuned for an efficient practical application. While too large ϵ values result in a low acceptance rate rendering the simulation inaccurate, too small values will produce little progress of exploring the complete posterior space and waste computation time. The sampled

Markov chain for too largely chosen values for L will exbibit looping trajectories with U-turns, which can even result in violation of the necessary ergodicity condition (Neal et al., 2011). Too few leapfrog integration results in very similar sampled values and will cause random walks without exploring the posterior space appropriately. The first problem of tuning the number of leapfrog steps L and the following slow progress in the difference of proposal $\widetilde{\boldsymbol{\theta}}$ and initial value for $\boldsymbol{\theta}$ is address by Hoffman et al. (2014). The approach is simulate the fictitious particle forward and backwards in time constructing a binary tree. This is done using the leapfrog integration until the distance between aforementioned values does not increase in a significant way and the generation of possible proposals stops. The sampler than selects a proposal randomly from all sampled values in the tree. The stopping condition prevents the particle to trace back the already travels trajectory. The second problem of adapting the step size ϵ, can be solved by using dual averaging, which is as modification of primal averaging described by Nesterov (2009), which utilizes stochastic optimization with vanishing adaption. Both modifications of the simple HMC are illustrated in detail by Hoffman et al. (2014) addressing mathematical properties, explicit implementation and the approaches for efficient choices of initial values in a comprehensible way. In this approach the NUTS algorithm by will be utilized as of its convenient implementation in the `Stan` package and its superior efficiency compared to other methods.

2.3.6. Sampling Diagnostic

The monitoring of convergence of above sampling algorithms can be done by assessing the variation of parameters between and within the multiple simulated Markov chains Gelman et al. (2013, p. 281-290). When the variation within the chain of samples is approximately equal to the variation between the chains, these Markov chains can be assumed to be converged to the target posterior distribution. The chains in themself converged to a stationary distribution and the chains have mixed well, meaning they converged to the same target distribution. If the Markov chains are stationary and well mixed, will be diagnosed for each theoretical parameter κ separately as a scalar. Initially all chains are split in half resulting in $C_{\text{no}} \in \mathbb{N}$ simulated half chains, each of length $L_{\text{chain}} \in \mathbb{N}$ (excluding the warm-up samples, commonly chosen to be the half of overall sampled values). The between chain variance B for $C_{\text{no}} \in \mathbb{N}$ can be calculated using single draws κ_{cl} for $c \in \{1, \ldots, C_{\text{no}}\}$ and $l \in \{1, \ldots, L_{\text{chain}}\}$ in the following manner:

$$B(\kappa) = \frac{L_{\text{chain}}}{C_{\text{no}} - 1} \sum_{c=1}^{C_{\text{no}}} (\overline{\kappa}_c - \overline{\kappa})^2. \tag{96}$$

To compute this measure of dispersion, the mean parameter samples within a chain $\overline{\kappa}_c = 1/L_{\text{chain}} \sum_{l=1}^{L_{\text{chain}}} \kappa_{cl}$ and the overall mean $\overline{\kappa} = 1/C_{\text{no}} \sum_{c=1}^{C_{\text{no}}} \overline{\kappa}_c$ are calculated. This variance measures the weighted sum of squared difference of the chain means and the

overall mean. Additionally the within chain variance can be easily computed via

$$W(\kappa) = \frac{1}{C_{\text{no}}} \sum_{c}^{C_{\text{no}}} \frac{1}{L_{\text{chain}}} \sum_{l}^{L_{\text{chain}}} (\kappa_{cl} - \overline{\kappa}_c)^2 \tag{97}$$

Finally the marginal posterior variance $\kappa|\boldsymbol{Y}$ can be estimated by averaging over both mentioned variances with weights resulting in the expression:

$$\widehat{\text{Var}}(\kappa|\boldsymbol{Y}) = \frac{L_{\text{chain}} - 1}{L_{\text{chain}}} W(\kappa) + \frac{1}{L_{\text{chain}}} B(\kappa). \tag{98}$$

In theory under stationarity condition for infinite number of samples $L_{\text{chain}} \to \infty$, the within chain variance W approaches the true marginal posterior variance $\text{Var}(\kappa|\boldsymbol{Y})$. Therefore a convergence behaviour of sampled values of κ can be diagnosed, if the expression

$$\widehat{R}(\kappa) = \sqrt{\frac{\widehat{\text{Var}}(\kappa|\boldsymbol{Y})}{W(\kappa)}} \tag{99}$$

is sufficiently close to 1. The condition of $\widehat{R}(\kappa) > 1.1$ as the potential scale reduction, indicates a severe divergence problem, in which it cannot be assumed, that the distribution of κ approaches the target distribution. Therefore a more thorough investigation of convergence behaviour with visual analyses (plot of the complete trace of the Markov chains) should be conducted. In practical applications it is recommended to consider a model reparametrization of inclusion of additional prior information to improve on convergence. $\widehat{R}(\kappa)$ in the definition in equation 99 is often called split-$\widehat{R}(\kappa)$, because it was calculated based on Markov chains split in half in contrast to an out-dated approach of using them in their entirety. The method of examining $\widehat{R}(\kappa)$ as monitoring tool of convergence simplifies the analysis in a substantial way. The alternative of plotting sampled parameter values against their index in the Markov chain (in the form of time series analysis) and visually determine mixing of chains and stationarity lacks objectivity and comparability. When the mixing and stationarity in context of the convergence is achieved, the effective number of independent simulated parameter values has to be diagnosed. To achieve stable estimation of scalar κ a necessary number of independent samples (effective sample size) has to be obtained within the converged Markov chain sampling procedure. To quantify the independence an estimate for the autocorrelations observed within and between simulated chains is needed. This can be calculated using the variogram estimator

$$V_t(\kappa) = \frac{1}{C_{\text{no}}(L_{\text{chain}} - t)} \sum_{c=1}^{C_{\text{no}}} \sum_{l=1}^{L_{\text{chain}}} (\kappa_{cl} - \kappa_{c-1l})^2, \tag{100}$$

for a lag $t \in \{0, \ldots, T\}$ with a stopping point $T \in \{1, \ldots, L_{\text{chain}}/4 - 1\}$. Based on this variogram the correlation can be computed with

$$\widehat{\rho}_t(\kappa) = 1 - \frac{V_t(\kappa)}{2\widehat{\text{Var}}(\kappa|\boldsymbol{Y})}, \tag{101}$$

which enables to calculate the estimated effective sample size

$$\widehat{L}_{\mathrm{ESS}}(\kappa) = \frac{C_{\mathrm{no}} L_{\mathrm{chain}}}{1 + \sum_{t=1}^{T} \widehat{\rho}_t(\kappa)} \tag{102}$$

In practice T is chosen as the largest $\widetilde{t}$, for which the sum of two successive of two autocorrelations $\widehat{\rho}_{\widetilde{t}} + \widehat{\rho}_{\widetilde{t}+1}$ is still positive. This gets rid of the difficulty of noise created by the sample correlation. The rule of thumb $\widehat{L}_{\mathrm{ESS}}(\kappa) > 5 \cdot 2C_{\mathrm{no}}$ ensures necessary stability for parameter estimation from the sampled values.

3. Simulation

The complete R-code and Stan-syntax for the simulation and the application can be found on GitHub[8]. R by **R** was used as fundamental programming environment for all simulations and analyses. For all handling of Bayesian models cmdstanr by Gabry et al. (2022) was used as an interface for the efficient sampling algorithm in Stan by Stan Development Team (2023) and Carpenter et al. (2017). All other R-packages can be seen in the code files.
Besides a lot of elaborate methods to asses the quality of estimation procedures (Luecht et al., 2018, for the cases of item response theory models), a parameter recovery study is the basic tool to evaluate the foundation of estimation techniques and determine a basis for further analyses. As a measure of quality in regard tof how good the point estimator, in this case the EAP $\widehat{\kappa}$, recovers the actually simulated parameters κ the theoretical quantity of the Mean Squared Error

$$\begin{aligned} \mathrm{MSE}(\kappa) &= \mathrm{E}((\widehat{\kappa} - \kappa)^2) && (103)\\ &= \mathrm{Var}(\widehat{\kappa}) + \mathrm{Bias}(\widehat{\kappa}, \kappa)^2, && (104) \end{aligned}$$

combines the bias $\mathrm{E}(\widehat{\kappa} - \kappa)$ of an estimator with its variance in a trade-off. In practical applications, the square root of this quantity $\mathrm{RMSE}(\kappa) = \sqrt{\mathrm{MSE}(\kappa)}$ is of interest, to achieve a better comparability. Harwell et al. (1996) recommends the estimated $\mathrm{RMSE}(\kappa)$ defined by

$$\widehat{\mathrm{RMSE}}(\kappa) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\widehat{\kappa}_n - \kappa)^2} \tag{105}$$

for $N \in \mathbb{N}$ the simulated data sets and estimated values $\widehat{\kappa}_n$. It might also be useful to diagnose the variance and the bias in a parameter recovery study on their own by estimating the empirical and the mean deviance from the true parameter. This can be done for every true parameters which the response data was simulated from, but in case of a excessive number of parameters like in the GRM with a large response scale and overwhelmingly many threshold parameters, it makes sense to summarize groups of parameters by averaging over parameter specific RMSEs. For a set of parameters $\boldsymbol{K} = \{\kappa_k\}$ for index $k \in \mathcal{I}_{\boldsymbol{K}}$ and the simulated estimation set $\{\widehat{\kappa}_{nk}\}$ the mean RMSE can be written as

$$\widehat{\mathrm{ARMSE}}(\boldsymbol{K}) = \frac{1}{|\mathcal{I}_{\boldsymbol{K}}|}\sum_{k\in\mathcal{I}_{\boldsymbol{K}}}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\widehat{\kappa}_{nk} - \kappa_k)^2}. \tag{106}$$

The ARMSE is a pure practical tool and is solely used to compare different models for the same subset of parameters. In the following simulation study the ARMSE among other diagnostics will be used for the parameter groups of ability, discrimination and threshold parameters to compare modifications of the GRM. To evaluate the quality

[8] https://github.com/robingrugel/Parsimonious-Modelling-of-Threshold-Parameters-for-Ordered-Polytomous-Response-Data

of model specification Gelman et al. (2013, p. 270) considers the use of the expected coverage of the credible interval S_τ

$$\mathrm{E}(\mathbb{1}_{S_\tau}(\kappa)) = \int \mathbb{1}_{S_\tau}(\kappa) f(\kappa|\boldsymbol{Y}) d\kappa, \tag{107}$$

for $\tau = 0.5$. In theory expected coverage should always be 0.5 or at least close to it. In practice this coverage can be estimated using the estimated credible intervals $S_{\tau,n}$ for each simulation run n by averaging over the indicator

$$\widehat{\mathrm{E}}(\mathbb{1}_{S_{0.5}}(\kappa)) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{S_{0.5,n}}(\kappa), \tag{108}$$

if the true parameter lies in-between the estimated bounds of $S_{0.5,n}$. The number of replications for a reliable estimation of these quantities has to be sufficiently high, which can be problematic in a Bayesian model fitting, where the computational costs are significantly larger than in the frequentists' counterpart. In this initial simulation study $N = 50$ will be seen as sufficient to compare defined quantities.

3.1. Simulation Setup

To achieve the maximized generalizability and replicability of the results, the initial simulation setup has to be specified in a way that covers a range of possible situations. This approach is especially addressed in practical applications where the number of examinees is relatively small compared to the number of categories given by the measurement instrument. To show the possible strength of a parsimonious threshold modelling approach the number of categories is $K_i \in \{25, 100\}$ following roughly the application in speeded C-Tests (response $k \in \{0, \ldots, 26\}$) and reading motivation example (response $k \in \{0, \ldots, 100\}$). To simplify the full factorial design of the simulation study and therefore to keep the computational cost to a reasonable extent, all items will have the same number of categories $K_1 = \cdots = K_I = K \in \mathbb{N}$. The number of simulated examinees will be $P \in \{25, 50\}$, which ensures the existence of null categories, especially for $K = 100$, and reflects the very low sample size of the reading motivation data in chapter 4.2. The ability parameter will be generated from a dense set of standard normal quantiles

$$\theta_p = \Phi^{-1}\left(\frac{p}{P+1}\right) \text{ for } p \in \{1, \ldots, P\} \tag{109}$$

to sufficiently characterize the whole scale of a continuous ability parameter distribution. Another location or scale could be chosen to imply misspecification in model fitting, but this is out of the scope of these initial evaluations and has to be conducted in a more advanced setting. The number of simulated items is $I = 6$, which is a realistic length for a short questionnaire or assessment and reflects the measurement tool of the speeded C-Test as well. In favour of computation time and power consumption, the simulation of more items will be omitted. Item discrimination parameters are

chosen to be equal among all items $\alpha_1 = \cdots = \alpha_6 = \alpha$ and reflect different levels $\alpha \in \{1, 3\}$. The simulated $\alpha = 1$ reflects an acceptable parameter value in practical applications and $\alpha = 3$ reflects an excellent value, which is not seen that often in real assessments and surveys. To evaluate the concept of parsimonious functional threshold parameter modelling, it important to cover very different shapes of threshold distributions in the simulation design. Generally, the simulated thresholds reflect the class of unimodal skewed, bimodal, and unimodal symmetric distributions, each with 2 items. The Beta-distribution[9] is capable of representing all these cases through its shape parameters $\beta = (\beta_1, \beta_2)' \in \mathbb{R}^2_{>0}$, if its values are centred around 0.5 and scaled by $p_{\text{scale}} = 20$ to match the Beta support $[0, 1]$ to quantiles of ability parameters. Analogous to the simulation of ability parameters the threshold parameters are chosen according to

$$\gamma_{ik} = 20 \cdot \left(I^{-1}\left(\frac{k}{K-1} | \beta_i \right) - 0.5 \right) \text{ for } k \in \{0, \ldots, K-2\} \text{ and } i \in \{1, \ldots, I\}, \quad (110)$$

with item-specific shape parameter vectors β_i. These shape parameter vectors are divided by their crude classification into:

$$\beta_1 = (4.5, 0.5)', \beta_2 = (2, 0.5)' \text{ for a skewed to the right distribution,} \quad (111)$$
$$\beta_3 = (0.25, 0.25)', \beta_4 = (0.5, 0.5)' \text{ for a unimodal distribution and} \quad (112)$$
$$\beta_5 = (7.5, 7.5)', \beta_6 = (2.5, 2.5)' \text{ for a bimodal distribution.} \quad (113)$$

This functional variety is displayed in Figure 7 for $K = 100$ categories and illustrates the difference between the very extreme cases of skewed distributions and the supposedly benign unimodal symmetry. In this simulation study, only the example of right-skewed distribution is selected instead of right- and left-skewed. It is assumed that only the skewness in itself impacts the modelling results, not the specific direction.

[9] A random variable $Z \in [0, 1]$ is Beta distributed with shape parameters $\beta_1, \beta_2 \in \mathbb{R}_{>0}$, if it has the pmf $f_Z(z) = z^{\beta_1 - 1}(1 - z)^{\beta_2 - 1} / B(\beta_1, \beta_2)$ with Beta-function $B(\beta_1, \beta_2) = \Gamma(\beta_1)\Gamma(\beta_2)/\Gamma(\beta_1 + \beta_2)$ and Gamma-function $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt$; the cumulative distribution function is the regularized incomplete beta function $I(z|\beta_1, \beta_2)$ and $I^{-1}(q|\beta_1, \beta_2)$ its quantile function

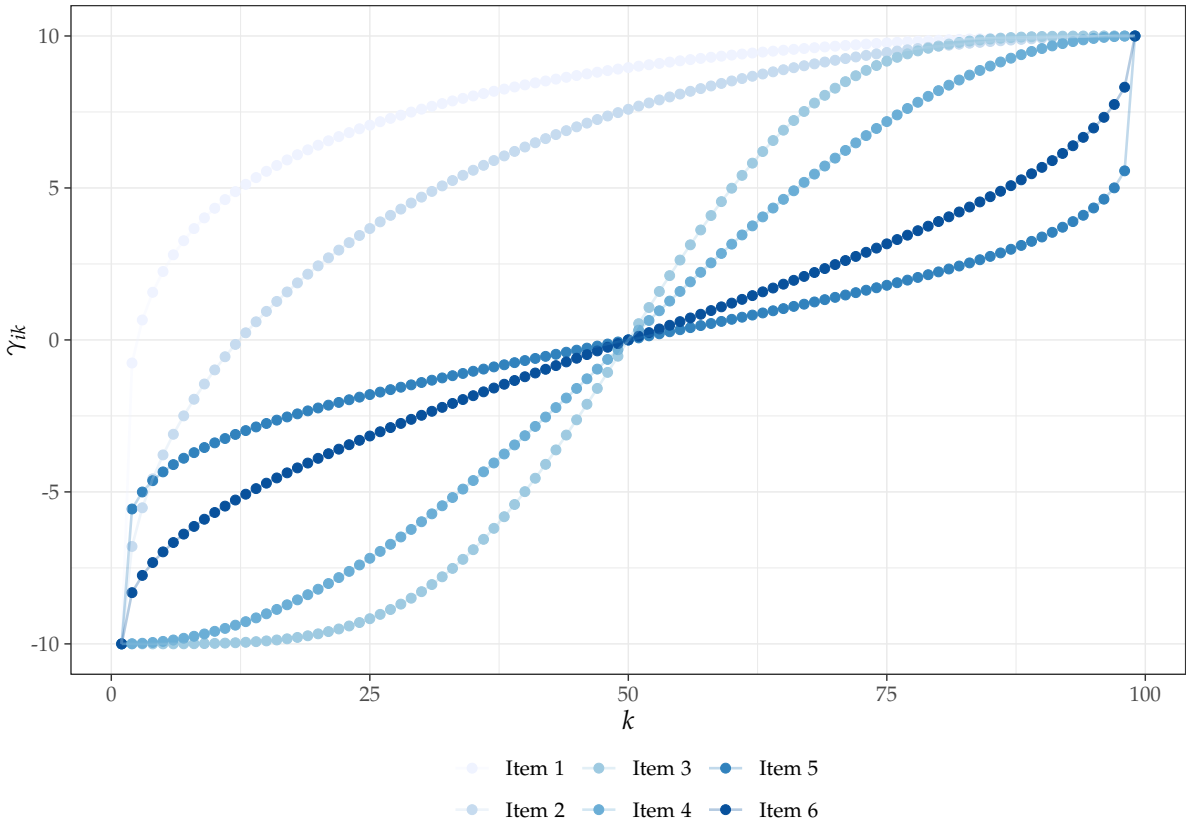

Figure 7: Threshold parameters γ_{ik} for items $i = 1, \ldots, 6$ and $k = 1, \ldots, 99$ for as an example

For the described factorial design $P \in \{25, 50\}$, $K \in \{25, 100\}$, $\alpha \in \{1, 3\}$ and defined β_i, the responses are simulated by calculating $P(Y_{ip} = k|\theta_p) = p_{ik}(\theta_p)$ for $i \in \{1, \ldots, 6\}$, $p \in \{1, \ldots, P\}$ and $k \in \{1, \ldots, K\}$ with help of the logistic distribution. This results in the vector of probabilities with which the actual sampling from the possible responses is conducted and therefore is the only source of randomness in this simulation, as the values for θ_p, α_i and γ_{ik} are fixed for every simulation run. This sampling is equivalent to the distribution statement 2 and is executed in `R` using the simple `sample()`-function using the aforementioned exact probabilities. In Figure 8 one can see the response frequency of $P = 10000$ simulated examinees by item and by category for both simulated discrimination parameters. These barplots highlight the extreme differences in the simulated responses to different items.

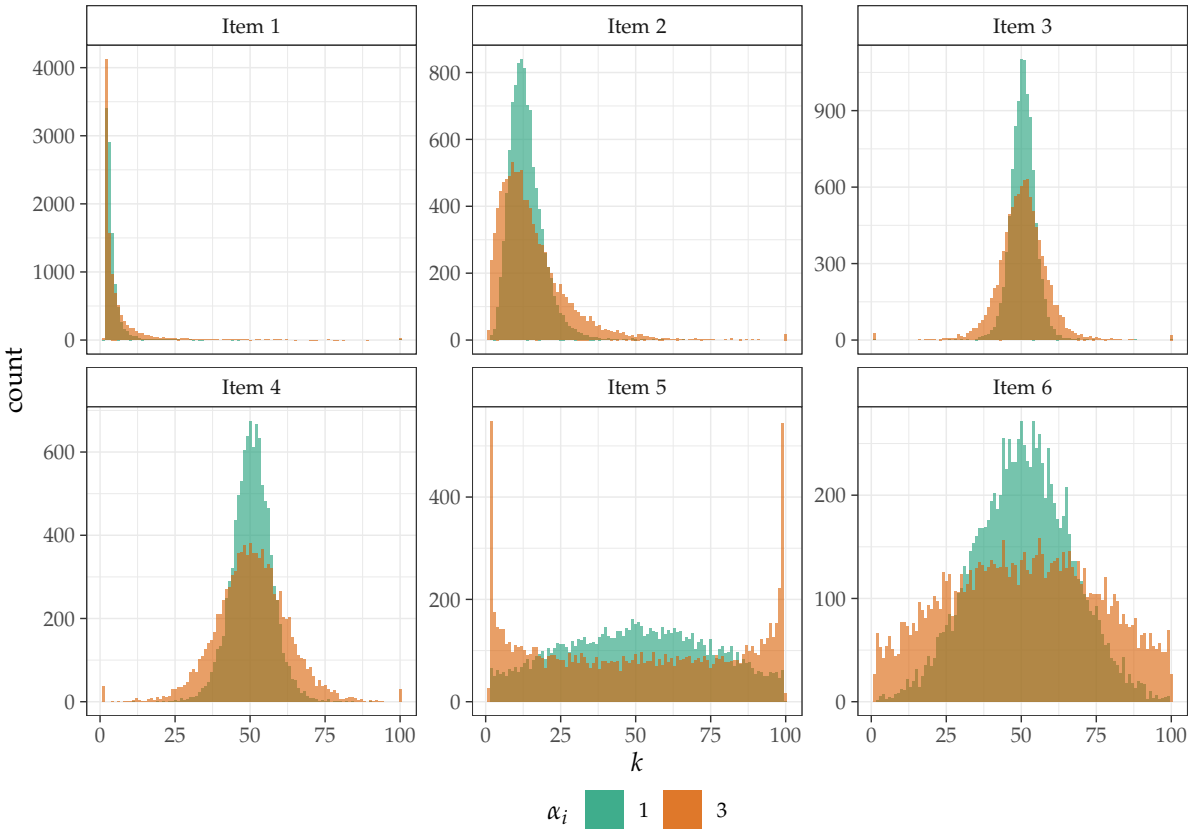

Figure 8: Barplot of $P = 10000$ simulated examinees responses to the aforementioned threshold distribution with discrimination $\alpha_i \in \{1, 3\}$ for items $i = 1, \ldots, 6$ and 100 categories

It is desired to keep simulation replications as high as possible to ensure reliable and credible results. Because Bayesian model fitting is computationally expensive, the number of simulation runs is set to $N = 50$. It is assumed that aggregating over so many replications as 50 yields stable and reliable results. The regular GRM and the GRM modified with monotonous increasing B-Splines are compared on this factorial simulation design. Both models' priors are specified identically regarding their ability and discrimination parameters in the following way:

$$\theta_p \overset{\text{iid}}{\sim} N(0, 1) \text{ for } p \in \{1, \ldots, P\} \tag{114}$$

$$\alpha_i \overset{\text{iid}}{\sim} Cauchy(0, 3)\mathbb{1}_{(0,\infty)} \text{ for } i \in \{1, \ldots, 6\}. \tag{115}$$

As mentioned before, the $N(0, 1)$ constraint distribution ensures identifiability and fixes the scale for the discrimination parameter through the multiplicative relationship $\alpha_i\theta_p$ in the GRM. The choice of truncated Cauchy distribution is inspired by Y. Luo et al. (2018) and represents a weakly informative prior for the discrimination, which usually exhibits values around 1. The threshold prior distribution will be set as

$$\gamma_{ik} \sim N(\mu_\gamma, \sigma_\gamma)\mathbb{1}_{(\gamma_{ik-1},\infty)} \text{ for } i \in \{1, \ldots, 6\} \text{ and } k \in \{1, \ldots, K-1\}, \tag{116}$$

for the regular GRM and analogously for the B-Spline coefficients of the parsimonious

functional modelling approach with $M = \lfloor \sqrt{K} \rfloor + 4$ basis splines:

$$\lambda_{im} | \mu_\lambda, \sigma_\lambda \sim N(\mu_\lambda, \sigma_\lambda) \mathbb{1}_{(\lambda_{im-1}, \infty)} \text{ for } i \in \{1, \ldots, 6\} \text{ and } m \in \{1, \ldots, M\} \tag{117}$$

$$\gamma_{ik} \approx \gamma_i(k) = \sum_{m=1}^{M} \lambda_{im} B_m(k). \tag{118}$$

The concepts of partial pooling and borrowing strength between threshold parameters are evaluated in this simulation. The first approach is to choose a common hyper prior distribution for mean and standard deviation of all normally distributed ordered thresholds. The second is an item-specific modelling of hyper prior information in the form of different prior distribution:

$$\text{simple hyper priors} \begin{cases} \mu_\gamma / \mu_\lambda & \sim N(0,5) \\ \sigma_\gamma / \sigma_\lambda & \sim Cauchy(0,5) \mathbb{1}_{(0,\infty)} \end{cases} \tag{119}$$

$$\text{item specific location and scale hyper priors} \begin{cases} \mu_{\gamma_i} / \mu_{\lambda_i} & \overset{\text{iid}}{\sim} N(0,5) \\ \sigma_{\gamma_i} / \sigma_{\lambda_i} & \overset{\text{iid}}{\sim} Cauchy(0,5) \mathbb{1}_{(0,\infty)} \end{cases}. \tag{120}$$

The idea behind this approach is to also enable to model the overall proficiency level more effectively. This can be loosely interpreted as overall difficulty of the test (in the simple hyper prior situation) or of the single item (in the item-specific location and scale hyper prior situation). This might complicate sampling from the posterior, but can be evaluated by convergence diagnostics afterward. To achieve efficient sampling, especially for the positively constraint Cauchy distributed parameters like the discrimination parameters or the variances of thresholds or B-spline coefficients respectively, the models are parametrized[10] by sampling from a uniform distribution and using the inverse Cauchy cumulative distribution function. In the case of normally distributed hyper parameters, the use of the non-centred parametrization[11] results in more efficient sampling for hyper prior in complex hierarchical modelling situations (Betancourt and Girolami, 2015; Betancourt, 2016).

3.2. Simulation Results

The quality of parameter recovery of ability parameters is based on the estimated RMSE, ARMSE and the expected coverage. The first two can be seen in Figure 9, which illustrates in which setting there is a clear difference. The estimation of ability parameters via the EAP for acceptable item discrimination parameters shows high values of RMSE for extremely low and extremely high values. This effect is significantly reduced by a very high value for the discrimination, which confirms the theory that it enables to discriminate between individual latent trait levels. In nearly every modelling situation, the B-spline approach outperforms the regular GRM. This is

[10] https://betanalpha.github.io/assets/case_studies/fitting_the_cauchy.html#7_the_half_cauchy_density_function

[11] https://mc-stan.org/docs/stan-users-guide/reparameterization.html

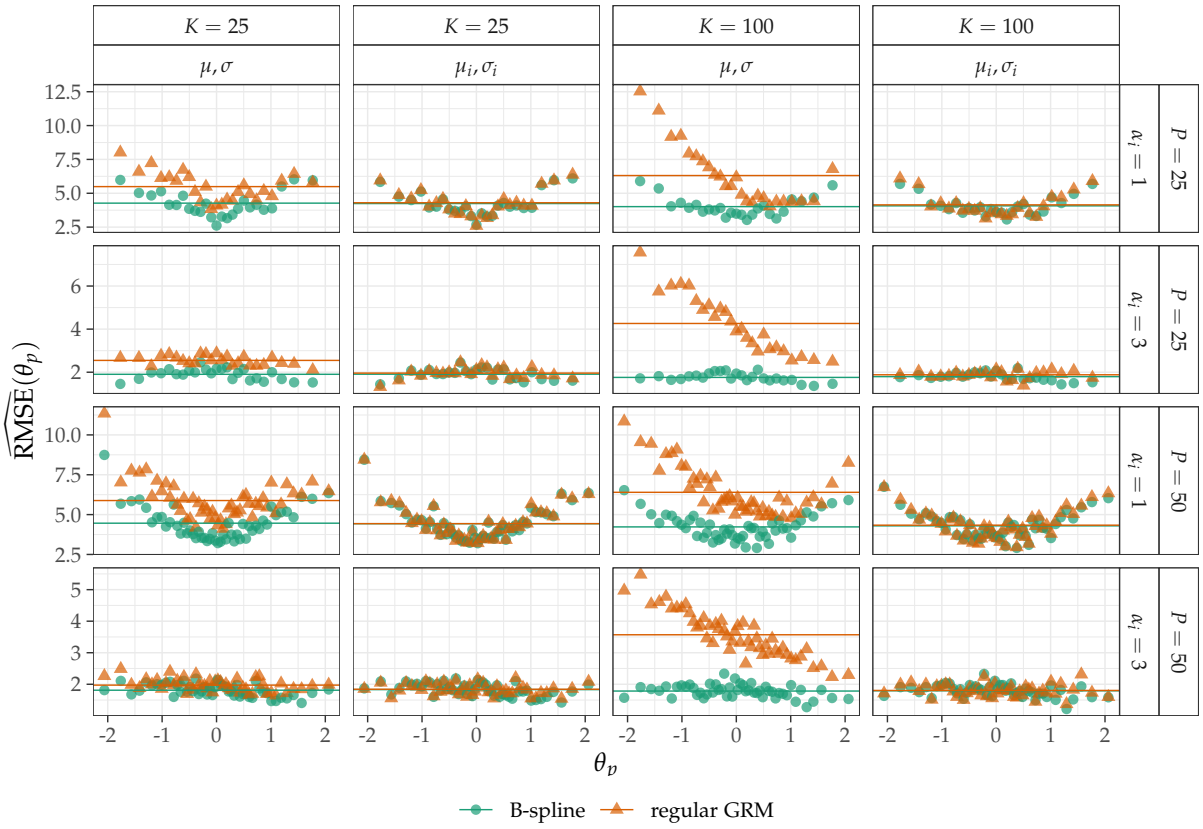

Figure 9: Estimated Root Mean Squared Error $\widehat{\text{RMSE}}(\theta_p)$ for the ability parameters θ_p with $p \in \{1, \ldots, P\}$ and Averaged Root Mean Squared Error $\widehat{\text{ARMSE}}(\theta)$ for $n \in \{1, \ldots, N\}$ and $N \in \mathbb{N}$ simulations as a horizontal line

more pronounced in smaller samples and a higher number of categories in the response scale. A huge improving effect can be seen through the use of item-specific hyper prior for the regular GRM. This improves the recovery almost to the level of the B-spline approach with global prior for its spline coefficients. This approach of item-specific hyper prior is not very commonly utilized in the Bayesian GRM literature, which could be due to the focus on bigger sample sizes, where it might be too computational expensive to sample from an additional hierarchical level. The curious values for RMSE in case of $K = 100$ and simple hyper prior for the regular GRM can be explained by a estimation bias, evident in Figure 10, possibly induced by the very skewed threshold distribution for 2 items. It can be seen that the absolute bias of the B-spline estimation is lower for every simulated data set. Especially for the well performing version of the regular GRM (with item-specific hyper prior), there is a more pronounced negative bias visible. This more noticeable for excellent discrimination parameters and large scales. Table 1 confirms this impression of a slightly more performant approach by using common hyper parameters for B-spline coefficients in comparison to all other approaches. This os also the case compared to the regular GRM with item-specific hyper priors, which performs surprisingly will for the simulated data sets.

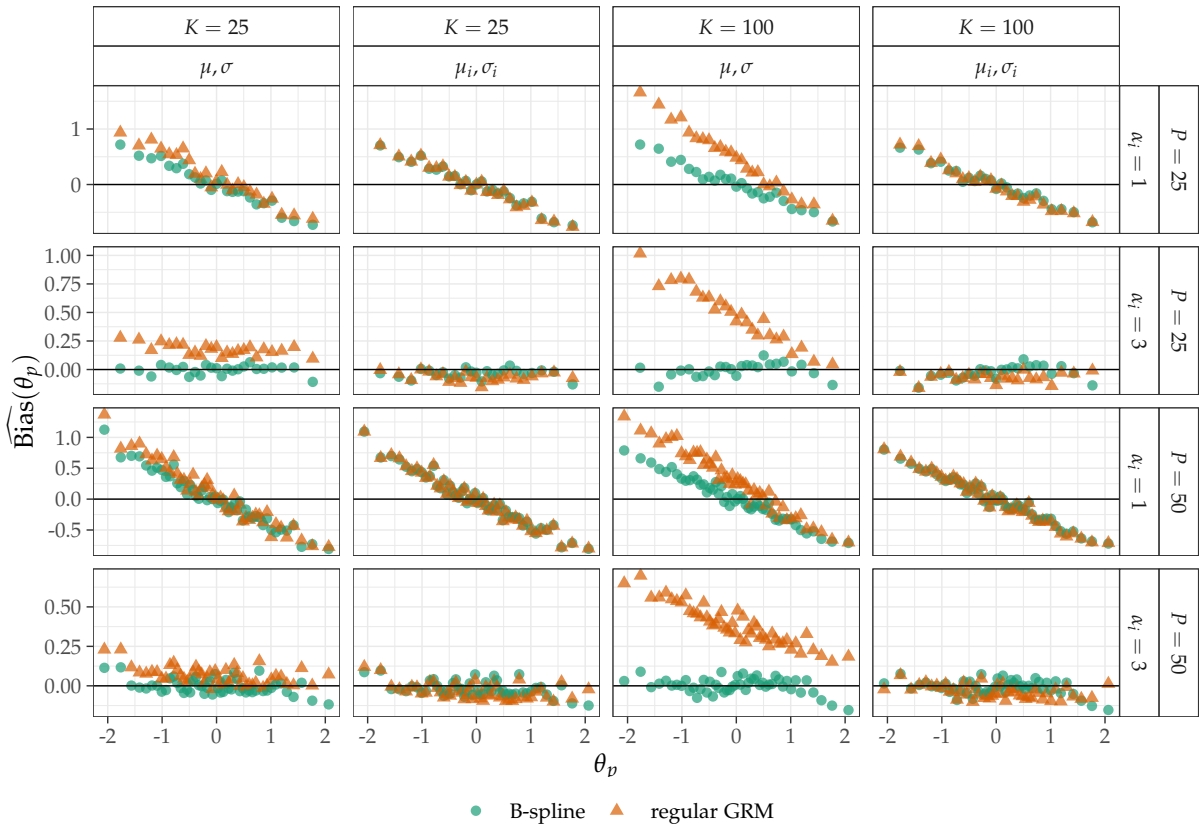

Figure 10: Estimated Root Mean Squared Error $\widehat{\text{Bias}}(\theta_p)$ for the ability parameters θ_p with $p \in \{1, \ldots, P\}$ and $N \in \mathbb{N}$ simulations

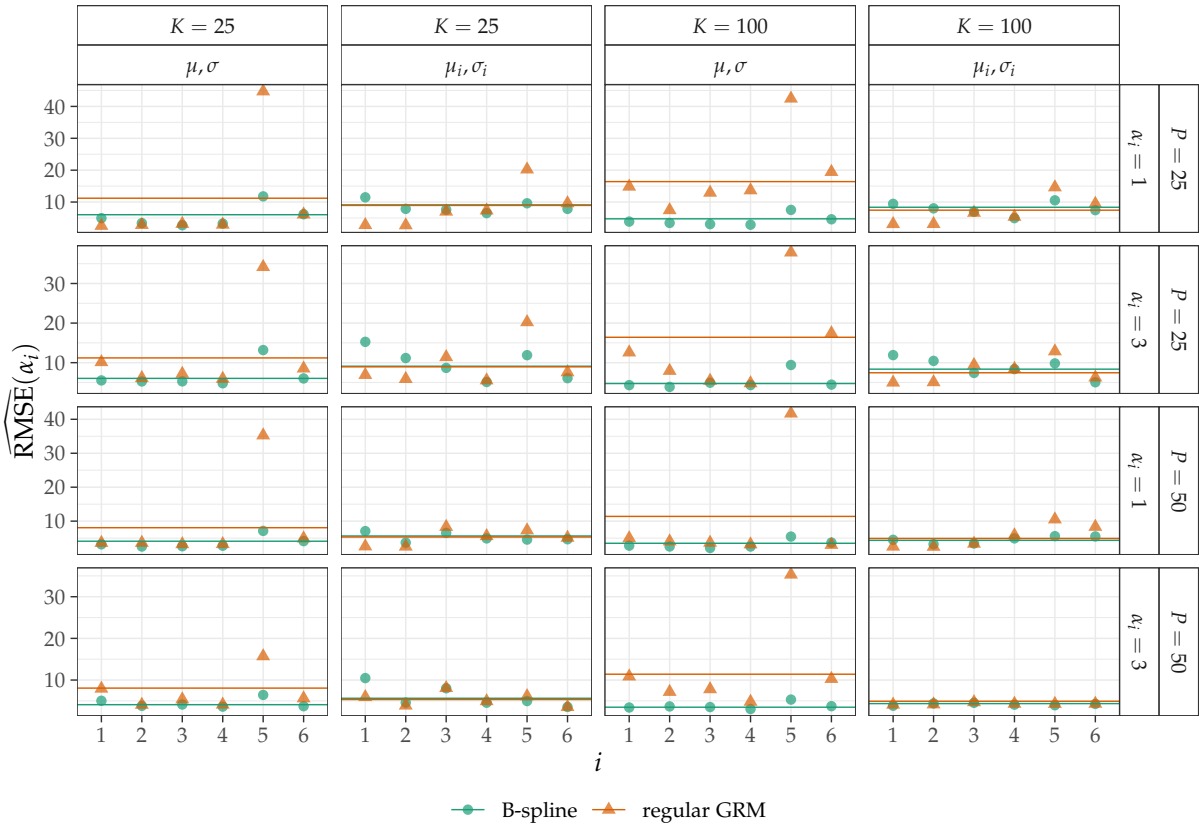

Figure 11: Estimated Root Mean Squared Error $\widehat{\text{RMSE}}(\theta_p)$ for the ability parameters θ_p with $p \in \{1, \ldots, P\}$ and Averaged Root Mean Squared Error $\widehat{\text{ARMSE}}(\theta)$ for $n \in \{1, \ldots, N\}$ and $N \in \mathbb{N}$ simulations as a horizontal line

			μ, σ		μ_i, σ_i	
P	K	α	B-spline	regular GRM	B-spline	regular GRM
$P = 25$	$K = 25$	$\alpha_i = 1$	4.271	5.49	4.241	4.298
		$\alpha_i = 3$	1.903	2.548	1.908	1.959
	$K = 100$	$\alpha_i = 1$	4.012	6.31	4.07	4.143
		$\alpha_i = 3$	1.759	4.262	1.794	1.882
$P = 50$	$K = 25$	$\alpha_i = 1$	4.462	5.882	4.431	4.432
		$\alpha_i = 3$	1.811	1.973	1.831	1.843
	$K = 100$	$\alpha_i = 1$	4.228	6.398	4.295	4.337
		$\alpha_i = 3$	1.785	3.569	1.793	1.804

Table 1: Averaged Estimated Root Mean Squared Error $\widehat{\text{ARMSE}}(\theta_p)$ for the ability parameters θ_p with $p \in \{1, \ldots, P\}$(minimal ARMSE highlighted in red)

Examining the mean coverage of an estimated 50% credible interval in Figure 24 in the Appendix draws a similar picture, that the B-spline approach leads to a more robust coverage of the true parameter values. The good improvement of the item-specific model approach for the regular GRM confirms a high priority in mindfully chosen prior structure in Bayesian statistics. Looking at recovery of the discrimination parameters reveals a very similar picture. The RMSE for the discrimination parameter illustrated in Figure 11 shows a superior behaviour of the B-spline approach using global threshold

hyper prior. It is interesting, that the recovery of the 5th item discrimination parameter was the most difficult one, but is easily explained by the distribution of matching threshold parameters. Item 5 has a very strong bimodal threshold distribution and exhibits an extreme bimodal response behaviour in the simulated population in Figure 8. This leads to a very high positive bias as can be seen in Figure 25 in the Appendix. Figure 26 in the Appendix validates the impressions of the ability parameter recovery and fortifies the idea of modelling item-specific threshold hyper prior in the GRM. Table 2 identifies the best model specification for the simulated data and reveals a consistent advantage of the B-spline approach. It would be interesting to simulate from a more diverse structure of discrimination parameters for each specific extreme threshold parameter distribution

			μ, σ		μ_i, σ_i	
P	K	α	B-spline	regular GRM	B-spline	regular GRM
$P = 25$	$K = 25$	$\alpha_i = 1$	5.381	10.382	8.491	8.296
		$\alpha_i = 3$	6.675	12.019	9.705	9.599
	$K = 100$	$\alpha_i = 1$	4.232	18.48	7.892	7.057
		$\alpha_i = 3$	5.229	14.347	8.836	7.847
$P = 50$	$K = 25$	$\alpha_i = 1$	3.72	8.987	5.23	5.235
		$\alpha_i = 3$	4.453	7.139	6.021	5.415
	$K = 100$	$\alpha_i = 1$	3.177	10.094	4.515	5.502
		$\alpha_i = 3$	3.779	12.704	4.178	4.315

Table 2: Averaged Estimated Root Mean Squared Error $\widehat{\text{ARMSE}}(\alpha_i)$ for the ability parameters α_i with $i \in \{1, \ldots, I\}$

Finally Figure 12 illustrates the aggregated RMSE values averaged over all thresholds of each item. It is obvious that the B-splines perform clearly better in every simulation setting although exhibiting a curious behaviour. One would assume, that the introduction of an item-specific hyper prior for the B-spline coefficients would improve the recovery for the thresholds of every single item, but the global hyper prior performs significantly better. This could be partly traced back to the data driven choice of knot sequence that was used, but was apparent as well for a knot sequence with equidistant spacing. Another curious characteristic is the bad parameter recovery for unimodal distributed thresholds in case of item-specific hyper priors. Especially for the regular GRM with thresholds explicitly normally distributed, it could be assumed, that the shapes might not be that far off.

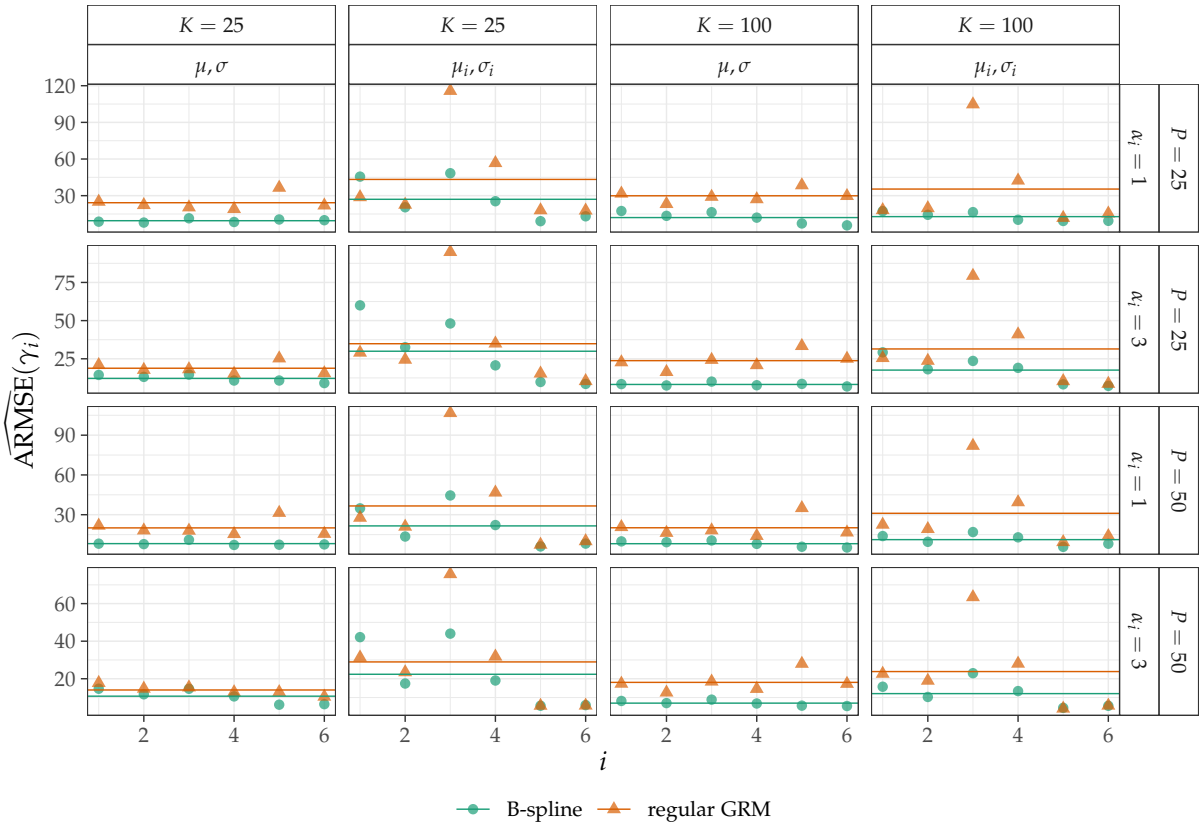

Figure 12: Averaged Estimated Root Mean Squared Error $\widehat{\text{ARMSE}}(\gamma_i)$ for the threshold parameters γ_i with $i \in \{1, \ldots, I\}$

			μ, σ		μ_i, σ_i	
P	K	α	B-spline	regular GRM	B-spline	regular GRM
$P = 25$	$K = 25$	$\alpha_i = 1$	9.52	24.24	27.03	43.31
		$\alpha_i = 3$	12.14	18.82	29.96	34.87
	$K = 100$	$\alpha_i = 1$	12.07	29.91	12.87	35.4
		$\alpha_i = 3$	8.2	23.79	17.6	31.44
$P = 50$	$K = 25$	$\alpha_i = 1$	8.2	20.07	21.56	36.61
		$\alpha_i = 3$	10.72	14.02	22.38	28.95
	$K = 100$	$\alpha_i = 1$	8.15	20.16	11.21	31.06
		$\alpha_i = 3$	7.01	18.07	12.09	23.83

Table 3: Averaged Estimated Root Mean Squared Error $\widehat{\text{ARMSE}}(\gamma_{ik})$ for the threshold parameters γ_{ik} with $i \in \{1, \ldots, I\}$ and $k \in \{1, \ldots, K_i - 1\}$

Table 3 reflects the optimal choice of model for the simulated settings clearly as the basic B-spline method. It should be stressed, that threshold parameters and their estimation are not the centre of a interpretable inference. Main focus should be the quality of ability and discrimination parameters which bear vital information for specific research questions. In general, the computation time was significantly shorter for the B-spline approach and very rare divergent transitions in the regular GRM sampling procedure were non-existent. In conclusion, quality of parameter recovery does not differ hugely between the newly introduced B-spline approach and the regularly utilized regular GRM. This might be due to a very benign Bayesian specification with weakly informative priors and very simple simulation settings. A more complex model with additional sources of noise and a severe misspecification like a multidimensional ability parameter would possibly show other pros and cons.

4. Application on Real Data

For the fitted models in this section, regular MCMC sampling behaviour was investigated using usual methods (divergent transitions, EBFMI, $\widehat{R}$, $\widehat{L}_{\text{ESS}}$ etc.) including visual ones. These checks were explicitly not included in the text to maintain readability and keep the text portion of this thesis to a reasonable length. When there were noteworthy diagnostics, it is reported and the results a treated with caution. To compare the parsimonious threshold modelling approach, the B-spline model with common hyper prior and the regular GRM with item-specific hyper prior were selected to analyse the real data sets. This decision was made due to the superior performance so the former and the surprising results of the latter one.

4.1. Speeded C-Test Data

When language abilities of examinees in various languages are evaluated in a research context, it is very likely to see the use of C-tests as a general measure (Heine, 2017). In C-tests examinees have to restore $K \in \mathbb{N}$ words previously deleted from text passages

leaving the gaps. For specific statistical approaches the correctly reconstructed gaps can be seen as separate items, which have a number of problems as consequence. A large number of missing values for low performing participants require a special and potentially difficult handling. An additional problem could be the for IRT models frequently used assumption of conditional independence, which is violated as the gaps and the words which have to be restored are elements of one and the same text (they share one stimulus). Therefore it is very common to use the number of correctly filled gaps in each passage as a ordered polytomous response with $K + 1$ possible categories. In C-tests with a high number of gaps this leads to lots of threshold parameters for the GRM. Forthmann et al. (2020) approached this problem by utilizing a count data model based on the Conway-Maxwell-Poisson distribution to get rid of the need for extensive parameter estimation. Another approach is the functional modelling of threshold parameters via spline functions with a B-spline basis. In this exemplifying application a comparison between a regular GRM and the proposed B-spline version will be conducted. The speeded C-test data utilized to illustrate this method was collected by Grotjahn (2010) and consists of responses from $271 = 80 + 50 + 89 + 52$ examinees, which can be divided into four different groups regarding the age they started to learn the German language. Goal of this research was to identify the impact of different ages at which the learning process of the German language started on the actual language proficiency. The test data consists of 6 scores calculated from 6 passages/items with 25 gaps which were to be filled in a time limit from 65 to 115 seconds. This leads to $K = 26$ correctly filled gaps and therefore possible responses $\{0, \ldots, 25\}$. The groups consist of monolingual German native speakers (1), early bilingual language learners who started learning when they were not older than 3 (2), late language learners who started after the age of 16 (3) and middle aged language learners who started within the age range of 4 to 15 years (4). Forthmann et al. (2020) decided to combine groups (1) and (2) into the population of early learners (a) and groups (3) and (4) into the late learners (b). In Figure 13 one can see the barplot of responses for the aggregated 2 groups for each item. It is apparent that the two groups exhibit a very different distribution of scores. While group (a)'s score distribution is skewed to the left for every passage, group (b)'s scores exhibit almost a symmetric pattern. The mean of scores in group (a) per item is higher and has a lower variance. Especially in item 2 and 4 a high number of maximal scores is noticeable interpretable as a ceiling effect for very proficient individuals. Group (b) shows very high variance ranging across the complete response scale, as well showing maximal scores.

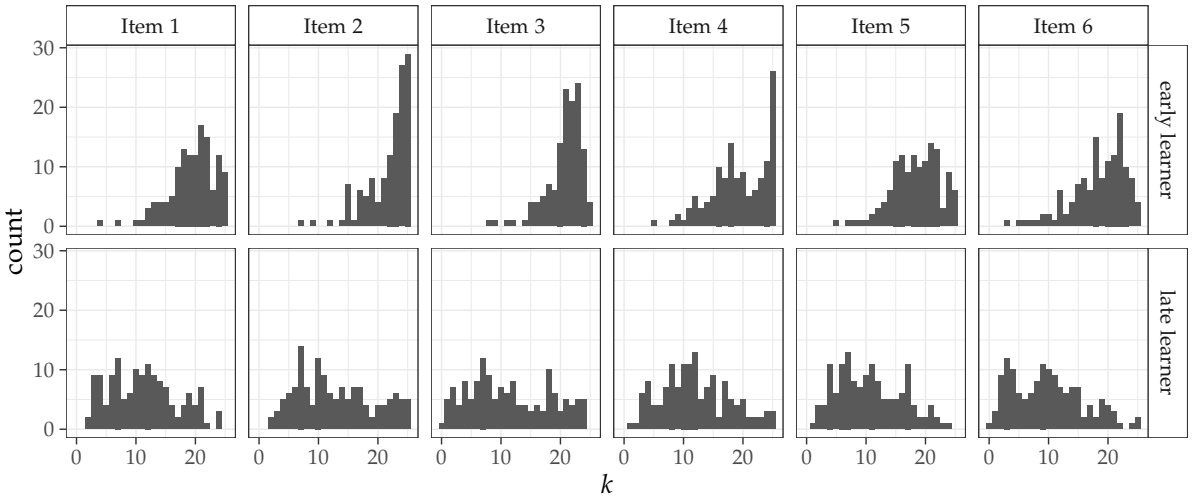

Figure 13: Barplot of the achieved scores per passage/item for both aggregated groups

Going back to barplots of the original partition of groups in Figure 14 it can be seen that the score distribution of the late language learners (4) shows a significant location shift to lower scores for every item compared to all other groups. This disaggregated presentation highlights the heterogeneity within the aggregated groups. While group (1) and (2) roughly represent the overall shape of its aggregated version, group (3) and (4) are very different to their aggregation.

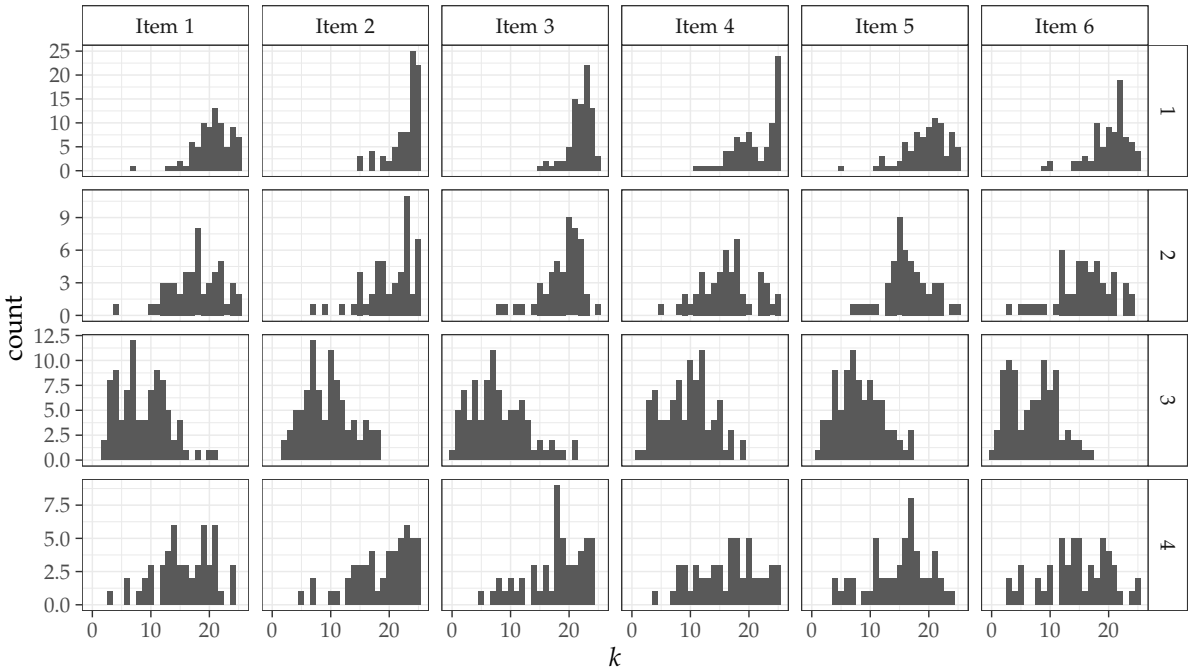

Figure 14: Barplot of the achieved scores per passage/item for the disaggregated original state of the data with initial grouping

This observation leads the urge to model the original groups and not their aggregated version. In this approach the comparison between models using both stratifications will be made. Two models per approach will be constructed by dividing all person indices into group-specific sets $\mathcal{I}_g$, so all examinees are included leading to $\bigcup_{g=1}^{n_g} \mathcal{I}_g =$

$\{1,\dots,P\}$ for a fixed $n_g \in \{2,4\}$. Every index for the examinee can only be in one set, providing the restriction $\bigcap_{g\in\{1,\dots,4\}} \mathcal{I}_g = \emptyset$. For both approaches, regular GRM and B-spline modelling, the prior specification for ability parameters and discrimination parameters will be the same:

$$\theta_{p_1} \sim N(0,1) \text{ for } p_1 \in \mathcal{I}_1 \tag{121}$$
$$\theta_{p_g}|\mu_g,\sigma_g \sim N(\mu_g,\sigma_g) \text{ with } p_g \in \mathcal{I}_g \text{ for group } g \in \{2,\dots,n_g\} \tag{122}$$
$$\mu_g \sim N(0,3) \text{ for group } g \in \{2,\dots,n_g\} \tag{123}$$
$$\sigma_g \sim Cauchy(0,3)\mathbb{1}_{(0,\infty)} \text{ for group } g \in \{2,\dots,n_g\} \tag{124}$$
$$\alpha_i \sim Cauchy(0,3)\mathbb{1}_{(0,\infty)} \text{ for item } i \in \{1,\dots,6\}. \tag{125}$$

To ensure identifiability of the scale and to improve interpretability, the first modelled group has a standard normal prior and will be interpreted as a reference group for detailed comparisons of group proficiency levels. The other groups will be specified with a normal distribution with unknown group-specific means and variances to consider performance level and heterogeneity within and between the groups. Threshold parameters for both approaches are modelled independently from the groups and only according to the performance in the simulation study. Therefore a item-specific hyper prior for location and variance will be utilized in the regular GRM:

$$\gamma_{ik} \sim N(\mu_\gamma,\sigma_\gamma)\mathbb{1}_{(\gamma_{ik-1},\infty)} \text{ for } i \in \{1,\dots,I\} \text{ and } k \in \{1,\dots,26\} \tag{126}$$
$$\mu_{\gamma_i} \overset{\text{iid}}{\sim} N(0,5) \text{ for } i \in \{1,\dots,6\} \tag{127}$$
$$\sigma_{\gamma_i} \overset{\text{iid}}{\sim} Cauchy(0,5)\mathbb{1}_{(0,\infty)} \text{ for } i \in \{1,\dots,6\}. \tag{128}$$

For the B-spline approach a overall hyper prior (which outperformed other methods in the simulation study) for all items will be used for the 9 spline coefficients

$$\lambda_{im}|\mu_\lambda,\sigma_\lambda \sim N(\mu_\lambda,\sigma_\lambda)\mathbb{1}_{(\lambda_{im-1},\infty)} \text{ for } i \in \{1,\dots,I\} \text{ and } m \in \{1,\dots,9\} \tag{129}$$
$$\gamma_{ik} \approx \gamma_i(k) = \sum_{m=1}^{9} \lambda_{im} B_m(k) \tag{130}$$
$$\mu_\lambda \sim N(0,5) \tag{131}$$
$$\sigma_\lambda \sim Cauchy(0,5)\mathbb{1}_{(0,\infty)}, \tag{132}$$

chosen by a simple rule of thumb $M = \lfloor\sqrt{26}\rfloor + 4 = 9$. After reparametrizations of normal and Cauchy hyper priors, the sampling from all four models was done using the `CmdStan`-interface `cmdstanr` by Gabry et al. (2022). The aforementioned diagnostic quantities of $\widehat{R}(\kappa)$ and $\widehat{L}_{\text{ESS}}(\kappa)$ does not indicate a sampling problem, based on the rule of thumbs in section 2.3.6. It should be noted, that more conservative values of $\widehat{R}(\kappa) > 1.05$ and $\widehat{L}_{\text{ESS}}(\kappa) < 10 \cdot 2C_{\text{no}}$ do indicate a problematic sampling behaviour especially for the regular GRM. To ensure that the model specification, using the GRM structure, is appropriate for the speeded C-test data, posterior predictive checks are conducted. A visual comparison of the observed data and replicated version can

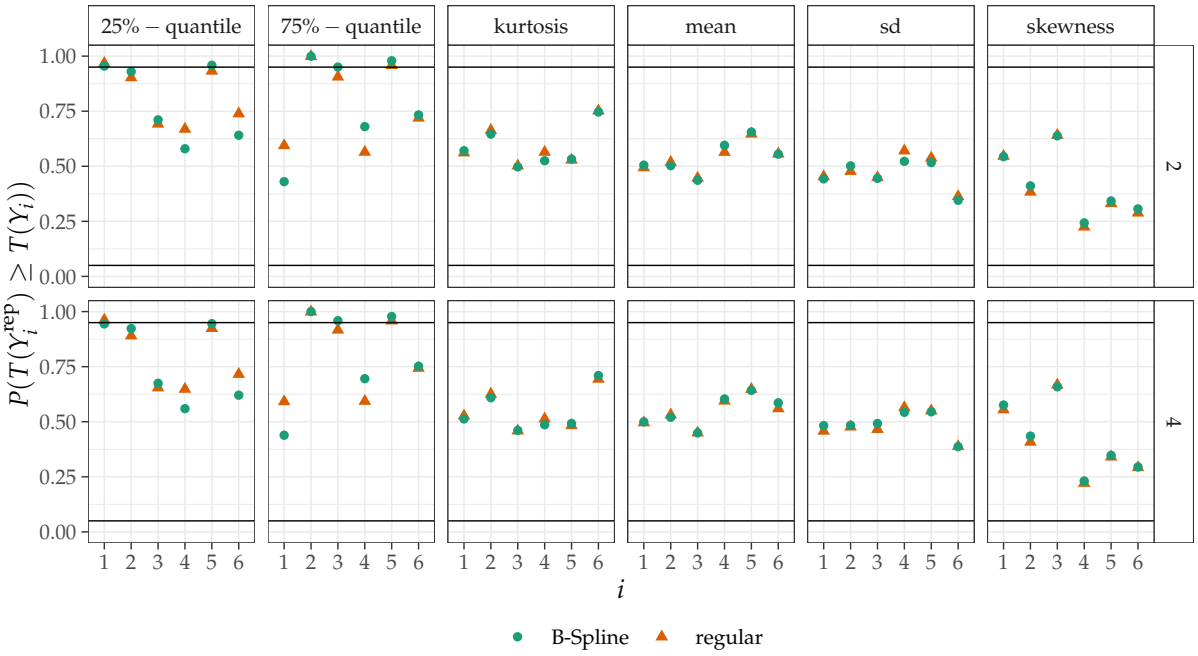

Figure 15: Posterior predictive p-value for various statistics

be seen in Figures 27, 28, 29 and 30 for all items separately. These plots all show no special differences of replicated response data. To consolidate this impression, posterior predictive p-values are calculated for the mean, standard deviation, 0.05- and 0.95-quantiles, skewness and kurtosis[12]. These values can be seen in Figure 15 and they indicate a problematic modelling approach in all four cases regarding the tail probabilities of the response quantified by the lower and upper quartile, because some values lie out of the reasonable range of $[0.05, 0.95]$, defined by Gelman et al. (2013, p. 151). All other simple statistics indicate a well behaved model specification reflected in the observed response matrix. Following a benign sampling diagnostic and a consistent model specification diagnostic in almost all cases, a further analysis and model comparison will be conducted. In Figure 16 one can see the EAP and the 90% equal tailed credible interval (using quantiles of the sampled values) of ability parameters for all 4 models.

[12] Joanes et al. (1998) reviewed estimates for the skewness $\widehat{\text{Skew}}(\boldsymbol{y}) = (1 - n^{-1})^{3/2} m_3 / m_2^{3/2}$ and kurtosis $\widehat{\text{Kurt}}(\boldsymbol{y}) = (1 - n^{-1})^2 m_4 / m_2^2$ with empirical moments $m_r = 1/n \sum_{p=1}^{P} (y_p - \overline{y})^r$ of degree r which will be utilized in this approach of posterior predictive checking, because they came to the conclusion that these estimates perform very well in case of mild skewness in the distribution.

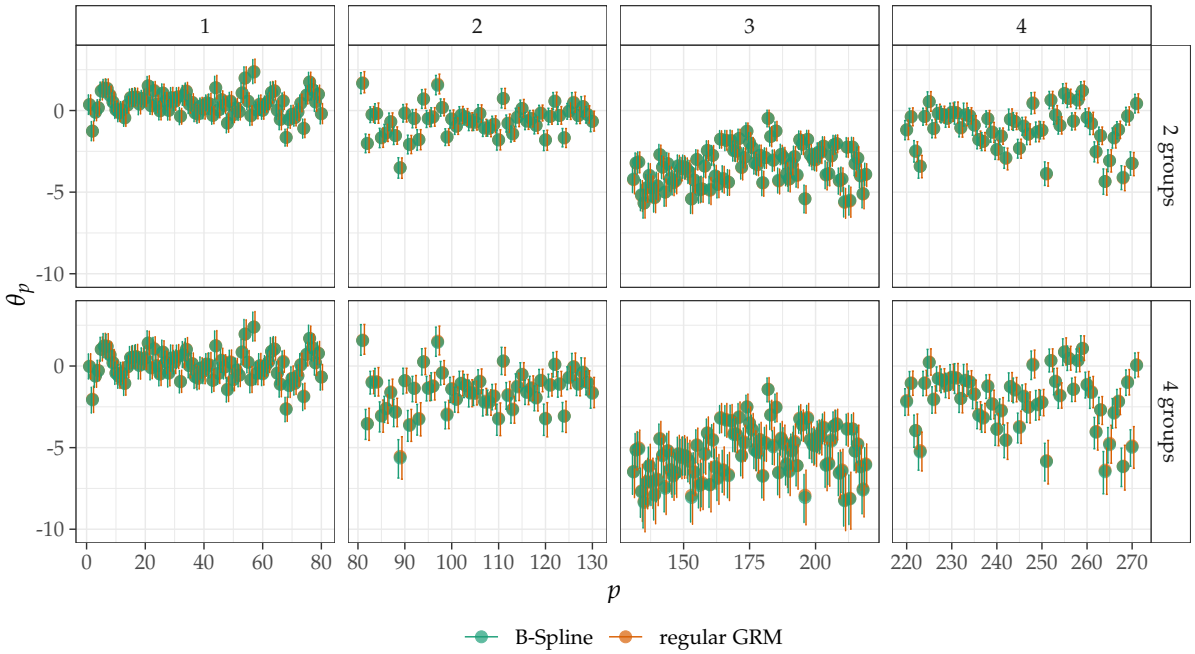

Figure 16: EAP estimation for ability parameters including a 90% credible interval for 2- and 4-groups modelling approach

It is apparent, that there is only small difference of ability parameter estimation and their credible intervals between the regular GRM and the B-spline approach. Therefore it does not come as a surprise, that the empirical reliability estimates[13] only differ very slightly, as can be seen in Table 4.

B-Spline		regular GRM	
2 groups	4 groups	2 groups	4 groups
0.9593	0.953	0.9598	0.9452

Table 4: Empirical reliability (Brown et al., 2014)

The difference in location between the 2-groups and the 4-groups models is very pronounced. Group (3) as the late language learner group starting after the age of 15 exhibit a clear shift towards lower values now being modelled as a separate group. While in the 2-groups modelling approach regular GRM and B-spline version do not show big differences in parameter point estimation or credible interval, the 4-groups method shows significant differences for group (2), (3) and (4). The regular GRM estimations are clearly lower compared to the B-spline counterpart. This could be an effect of the smaller sample size per group, so the prior influence increases while drawing Markov samples within the group ability distribution. Tables 5 and 6 illustrate the difference between the models confirming the mentioned location shift.

[13] Brown et al. (2014) defined an estimate for the empirical reliability $\widehat{\text{Rel}}_{\text{emp}} = 1 - \overline{(\widehat{\sigma}_\theta)^2} / \widehat{\text{Var}}(\theta)$.

	B-spline		regular GRM	
	mean	sd	mean	sd
μ_1	−2.540	0.2517	−2.546	0.2510
σ_1	1.686	0.1621	1.705	0.1653

Table 5: Hyper prior estimation for ability parameters in the 2-groups modelling

If one is willing to assume the sample size argument, it would be more appropriate to trust a more conservative approach, regarding the effect of starting age as a German language learner, of the B-splines. Examining the estimation of the discrimination parameters in Figure 17, a very similar picture regarding the differences can be seen.

	B-spline		regular GRM	
	mean	sd	mean	sd
μ_1	−1.561	0.2758	−1.522	0.2884
μ_2	−5.376	0.5375	−5.322	0.6356
μ_3	−2.040	0.3396	−2.001	0.3674
σ_1	1.409	0.2081	1.399	0.2241
σ_2	1.600	0.2100	1.592	0.2298
σ_3	1.907	0.2712	1.897	0.2973

Table 6: Hyper prior estimation for ability parameters in the 4-groups modelling

While both models perform roughly equal for the 2-groups model with slightly posterior means, the 4-groups models exhibit a reverse difference with much lower point estimates of the regular GRM. With the superior performance of the B-spline method from the simulation study in mind, the regular GRM results should be seen with caution and not be trusted unreflective.

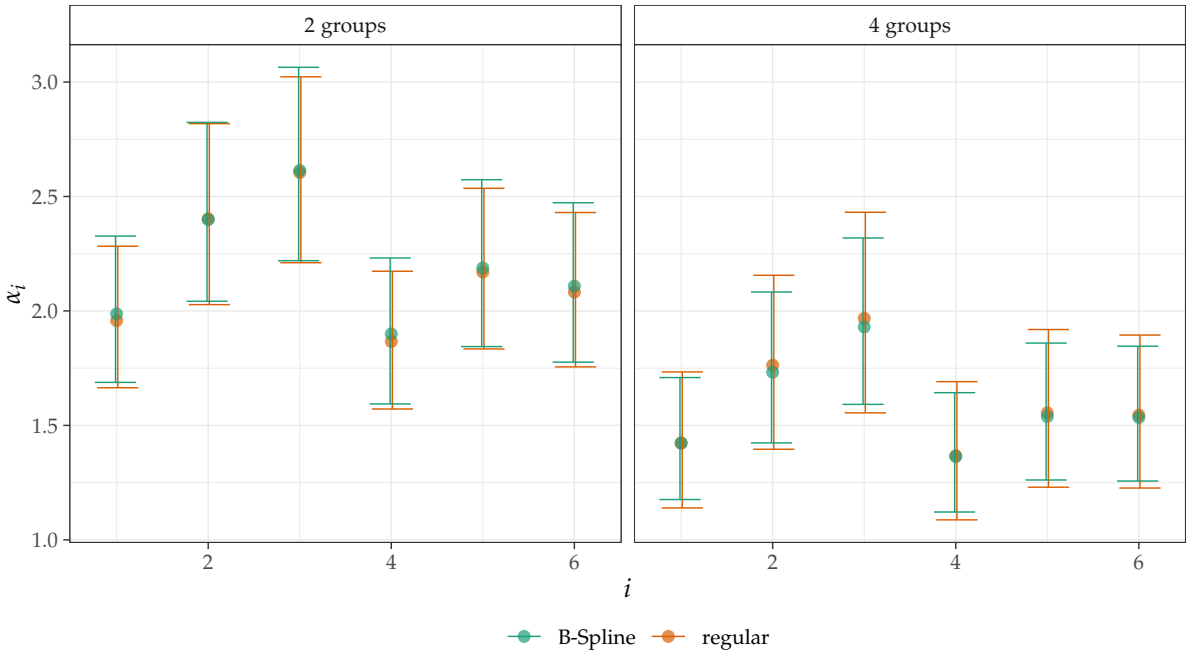

Figure 17: EAP estimation for discrimination parameters including a 90% credible interval for 2- and 4-groups modelling approach

	B-spline	regular GRM
2 groups	−7.543	−59.00
4 groups	0.000	−51.75

Table 7: Model comparison via differences of estimated elpd values

Finally the estimated threshold parameter means in Figure 18 do not show extraordinary differences except for threshold values in the lower end of the scale. In this area the regular GRM exhibits more extreme values and higher uncertainty expressed in the interval estimation. This can be explained by the very low frequency of low or minimal scores in the complete study population.

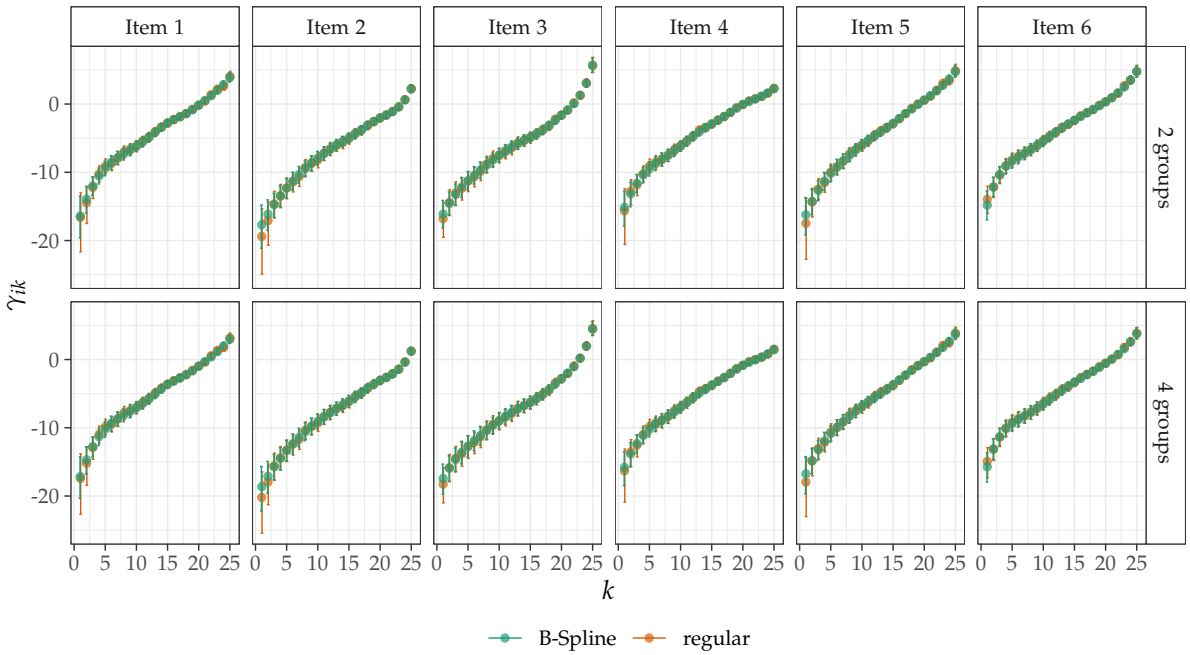

Figure 18: EAP estimation for threshold parameters including a 90% credible interval for 2- and 4-groups modelling approach

The estimated values for $elpd^{loo}$ and its difference to the best to all models can be seen in Table 7. One can see, that the B-spline version with four groups performs noticeably better compared to all other models. The difference between the two and four groups B-spline model however is not as pronounced, as compared to the regular GRM, but still existent. One could argue to choose the simpler model because the differences are small and it is desirable to keep modelling inference as simple as possible to draw general conclusions from it. In conclusion the model choice and every inference based on it should be treated carefully, due to the concerning self consistency check using the poster predictive p-values. However, a slight advantage of the B-spline can be seen through model comparison. A more conservative approach regarding ability and discrimination parameters should also motivate the use of a flexible threshold parameter modelling in general. Additionally it would make sense to introduce other covariates like the gender into a more complex model, which could improve model fit in general.

4.2. Reading Motivation Data

The second data set originates from an extremely small survey conducted as part of the teachers training in Göttingen by Katherina Warzecha in 2020. In her teaching practice she noticed, that the motivation of her class of 28 ninth grade students was extremely low when it came to reading English informative texts. The syllabus for ninth grade English class contains specific topics which have to be taught and are relevant for comprehensive schooling. Therefore it wasn't possible to exchange the text passages. To tackle this problem she decided to examine the effect of a pre-reading phase to improve the motivation to read informative non-fictional texts. Focus of this investigation was the motivation and not the reading proficiency, although the connection between motivation and comprehension is evident. A pre-reading phase has the purpose to present an alternative form to introduce a formal text with help of creative methods (audio or audio-visual media) to draw interest in this specific topic. This was done for the topic of the "Stolen Generations" as part of Australian history by using a short scene of the movie "Rabbit Proof Fence". A questionnaire with 10 different items which quantify the students motivation regarding the usefulness, joy and overall motivation to read the text, was used. As a scale the Visual Analogue Scale (VAS) scale already mentioned in the introduction, was used to capture granular differences in the responses in 101 categories. To measure the effect of the pre-reading phase, the students were randomly divided in two groups, one treatment group who watched the movie scene and a control group which was only given sparse information about the text they would read. The questionnaire was given out before and after the pre-reading intervention and the reading phase of the text. After a data clean-up, removing cases of missing values and a minor labelling issue, the responses of 20 students (10 in each group) remain to be modelled. This experimental design leads to the measurement at 2 different time points for two groups. A psychometric model for this situation can be visualized by the path diagram 19. Because of the response scale the GRM can be used and can be seen as a structural equation model (SEM) as explained by Rabe-Hesketh et al. (2004) or De Boeck et al. (2004). In this modelling approach a measurement invariance regarding equal item parameters is assumed. Therefore the discrimination and threshold parameters will stay the same over time and are independent from the specific group membership. The ability parameter however, will be assumed multivariate and incorporates the relevant effect of the pre-reading phase treatment.

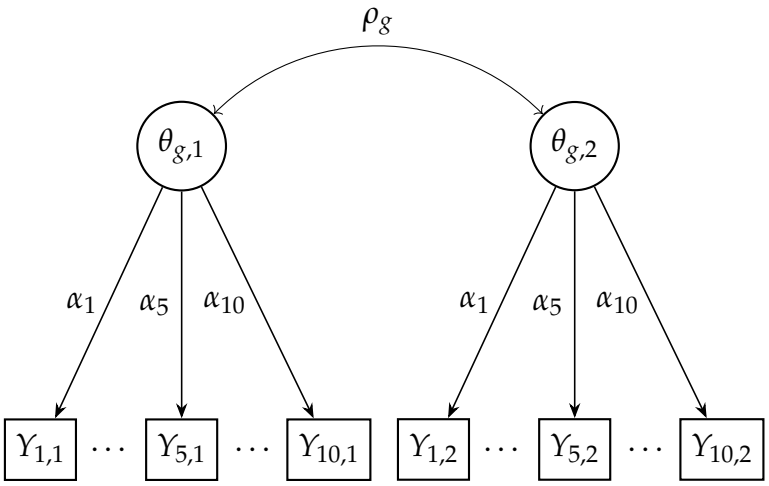

Figure 19: Path diagram for the repeated measure of reading motivation with group-dependent $\boldsymbol{\theta}_{g,p} = (\theta_{g,1}, \theta_{g,2})'$ and ρ_g

For the index sets $\mathcal{I}_t \cup \mathcal{I}_c = \{1, \ldots, P\}$ for the treatment and control group respectively, with $\mathcal{I}_t \cap \mathcal{I}_c = \emptyset$, the prior model specifications will be chosen as

$$\boldsymbol{\theta}_{t,p} = \begin{pmatrix} \theta_{t,1} \\ \theta_{t,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu_{all} + \mu_t \end{pmatrix}, \begin{pmatrix} 1 & \rho_t \sigma_t \\ \rho_t \sigma_t & \sigma_t^2 \end{pmatrix} \right) \text{ for } p \in \mathcal{I}_t \tag{133}$$

$$\boldsymbol{\theta}_{c,p} = \begin{pmatrix} \theta_{c,1} \\ \theta_{c,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu_{all} \end{pmatrix}, \begin{pmatrix} 1 & \rho_c \sigma_c \\ \rho_c \sigma_c & \sigma_c^2 \end{pmatrix} \right) \text{ for } p \in \mathcal{I}_c \tag{134}$$

$$\mu_{\text{all}} \sim N(0, 3) \tag{135}$$

$$\mu_t \sim N(0, 3) \tag{136}$$

$$\sigma_t \sim Cauchy(0, 3)\mathbb{1}_{(0,\infty)} \tag{137}$$

$$\sigma_c \sim Cauchy(0, 3)\mathbb{1}_{(0,\infty)} \tag{138}$$

$$\alpha_i \sim Cauchy(0, 3)\mathbb{1}_{(0,\infty)} \text{ for } i \in \{1, \ldots, 10\}. \tag{139}$$

The mean and variance of all ability parameters for the first measurement time point will be fixed to the values 0 and 1 respectively to ensure identifiability and interpretability of the treatment effect and its variance for both groups. The overall change of motivation of the distribution mean μ_{all} will be modelled to contrast the treatment effect μ_t of the pre-reading intervention. This enables to interpret μ_{all} as the time effect which arises solely due to the different measurement time points and could be of various behavioural or psychological causes. Variances for the second measurement time points are estimated freely for both groups, as well as their correlation. The

possible factorization of covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 \end{pmatrix} \tag{140}$$

$$= \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{141}$$

$$= \text{diag}(\sigma_1, \sigma_2) \cdot \Omega \cdot \text{diag}(\sigma_1, \sigma_2) \tag{142}$$

$$= \text{diag}(\sigma_1, \sigma_2) \cdot LL' \cdot \text{diag}(\sigma_1, \sigma_2), \tag{143}$$

first into the diagonal matrices of the standard deviations and the correlation matrix Ω and then into its Cholesky factors, enables an easy utilization of convenient priors for the correlation and fixing one variance to 1. In the following modelling approach, this parametrization enables to define L as the parameter class `cholesky_factor_corr` and utilize the Lewandowski-Kurowicka-Joe (LKJ) distribution (Lewandowski et al., 2009) with parameter $\eta = 2$ as a prior, which indicates a preference for higher values on the diagonal compared to the correlations. This additionally leads to a more efficient sampling from the multivariate normal distribution via specialized functions in `Stan`. The threshold parameters will be modelled analogously to the simulation study and the speeded C-test analyses for the regular GRM in the following specification:

$$\gamma_{ik} \sim N(\mu_\gamma, \sigma_\gamma)\mathbb{1}_{(\gamma_{ik-1},\infty)} \text{ for } i \in \{1, \ldots, 10\} \text{ and } k \in \{1, \ldots, 101\} \tag{144}$$

$$\mu_{\gamma_i} \overset{\text{iid}}{\sim} N(0, 5) \text{ for } i \in \{1, \ldots, 10\} \tag{145}$$

$$\sigma_{\gamma_i} \overset{\text{iid}}{\sim} Cauchy(0, 5)\mathbb{1}_{(0,\infty)} \text{ for } i \in \{1, \ldots, 10\}. \tag{146}$$

For the B-spline approach a overall hyper prior for all items will be used for the $\lfloor\sqrt{101}\rfloor + 4 = 14$ spline coefficients:

$$\lambda_{im}|\mu_\lambda, \sigma_\lambda \sim N(\mu_\lambda, \sigma_\lambda)\mathbb{1}_{(\lambda_{im-1},\infty)} \text{ for } i \in \{1, \ldots, I\} \text{ and } m \in \{1, \ldots, 14\} \tag{147}$$

$$\gamma_{ik} \approx \gamma_i(k) = \sum_{m=1}^{14} \lambda_{im} B_m(k) \tag{148}$$

$$\mu_{\boldsymbol{\lambda}} \sim N(0, 5) \tag{149}$$

$$\sigma_{\boldsymbol{\lambda}} \sim Cauchy(0, 5)\mathbb{1}_{(0,\infty)}. \tag{150}$$

The actual sampling via NUTS was very problematic for the regular GRM. There was no possibility to specify the model differently or parametrize model variables to achieve an unproblematic sampling procedure. No matter how wide or tight the priors or how the covariance structure was specified, the regular GRM exhibits divergent transitions, $\widehat{R}(\kappa) \gg 1.1$ and $\widehat{L}_{\text{ESS}}(\kappa) \ll 40$ for almost half of all modelled parameters. These could not be solved by changing the sampler tuning parameters. This might be due to the very high number of model parameters in comparison to the very sparse information provided by undersized sample. The B-spline method however did not show any problematic sampling behaviour and provides reliable estimates. Because

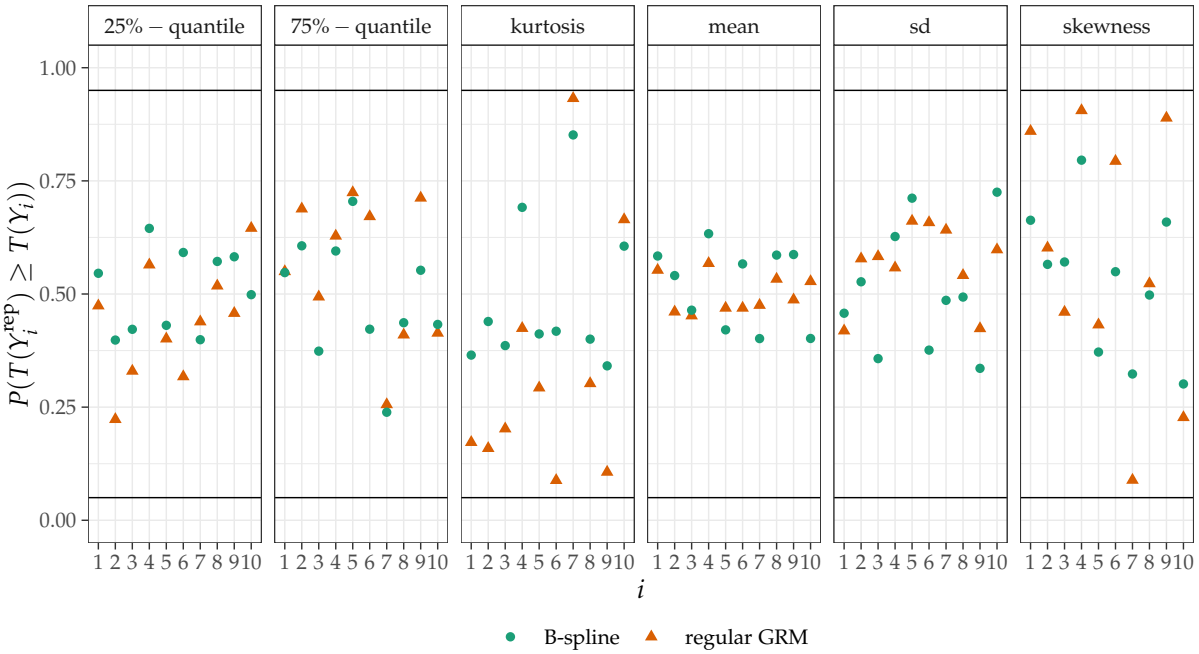

Figure 20: Posterior predictive p-value for various statistics

of failing posterior sampling for almost every parameter, in the following analyses of results, estimations from both models will be presented, but only the B-spline approach can be interpreted reliably. Examining the posterior predictive p-values in Figure 20 shows an appropriate model specification which is consistent with the actual observed responses. Examining the estimated ability parameter vectors for both groups in Figure 21, one cannot directly notice a specific connection of the overall levels between pre and post treatment measurements. It comes to no surprise that the ill-performing regular GRM exhibits quite different values. The basic assumption of an effect of the treatment cannot really be validated by this illustration.

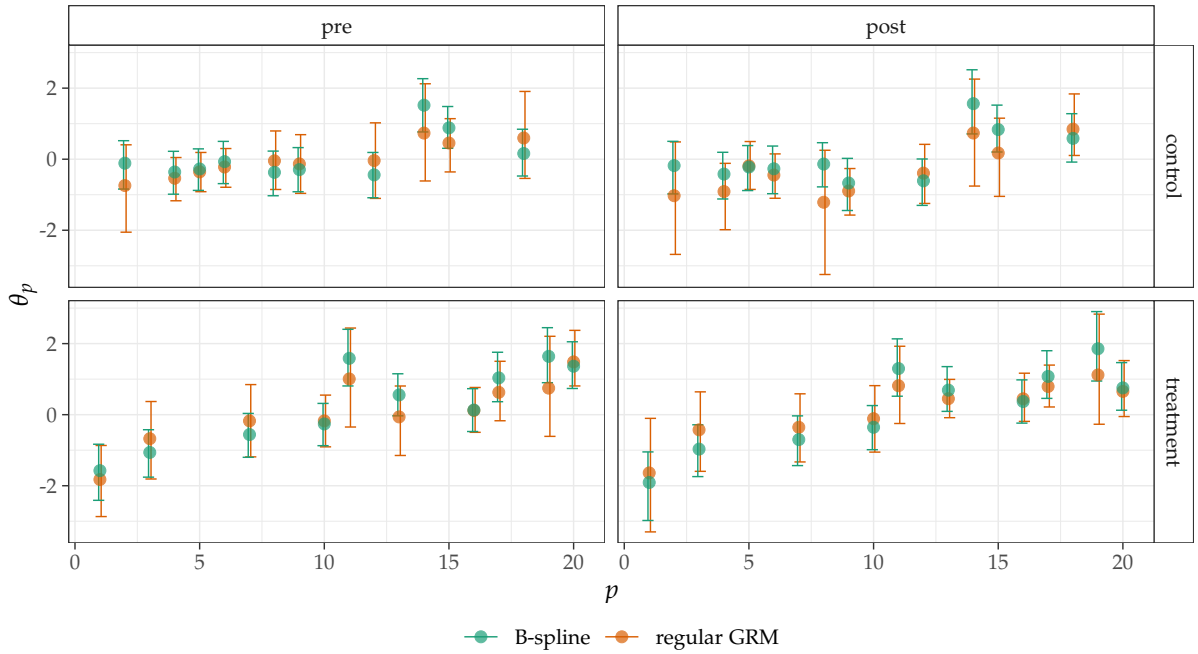

Figure 21: EAP estimation for ability parameters including a 90% credible interval

Table 8 illustrates estimates for the hyper parameter of the overall location effect between the measurements and the actual treatment effect μ_t. However it is still impossible to confirm neither a positive effect nor a negative effect. For B-spline modelling both estimates are nearly 0 with large standard errors. The regular GRM due to failed sampling might show a large bias and reveals a treatment effect with a high standard error.

	B-spline		regular GRM	
	mean	sd	mean	sd
μ_{all}	0.002945	0.2763	−0.2996	0.4551
μ_t	−0.033175	0.3852	0.3821	0.6633

Table 8: Location hyper parameter mean posteriors and standard error for ability parameters

The very small sample does not allow a positive inference regarding the treatment effect, even if very informative priors are chosen. Therefore it is more appropriate to report the absence of any revelations in favour of a robust and responsible modelling in form of the B-spline version. The covariance/correlation structure, which can be seen in Table 9 suggests smaller variance and therefore less heterogeneity in the treatment group. This interpretation is not reliable because of very big standard errors.

	regular GRM		B-spline	
	mean	sd	mean	sd
ρ_c	0.5784	0.3070	0.6686	0.2747
ρ_t	0.6316	0.2525	0.7710	0.1676
σ_c^2	1.4054	1.1135	1.0195	0.6860
σ_t^2	0.9703	0.8410	1.2795	0.6815

Table 9: Variance and correlation hyper parameter mean posteriors and standard error for ability parameters

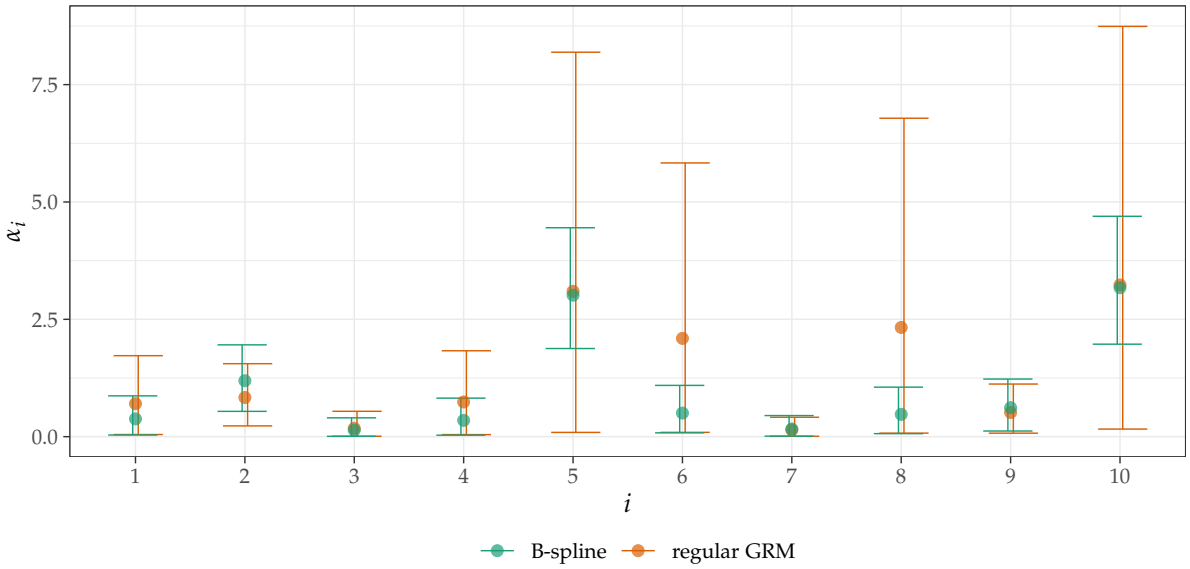

Figure 22: EAP estimation for discrimination parameters including a 90% credible interval

The discrimination parameters in Figure 22 reveal a possible problem in the questionnaire. The mean posterior values for item 5 (measures the interest in the topic) and 10 (measures the feeling to be readily prepared for the reading tasks) are significantly higher than for every other items. To inform more about the characteristics of these two items, a larger sample size would be necessary. The regular GRM again illustrates the poor MC sampling results in very high point estimates and huge standard errors. Finally the threshold parameters' posterior means in Figure 23 exhibit significantly different estimated distributions. It can also be seen, that the regular GRM estimates in case of Item 5, 6, 8 and 10 deviate significantly from the B-spline approximation. These items are the same 4 problematic ones as in the discrimination parameter estimation, which might be the cause for the bad performance in general. In conclusion it is not possible to state a nice treatment effect, but this analysis highlights, that the B-spline approach has special use for very small sample size to achieve robust and rather conservative inference.

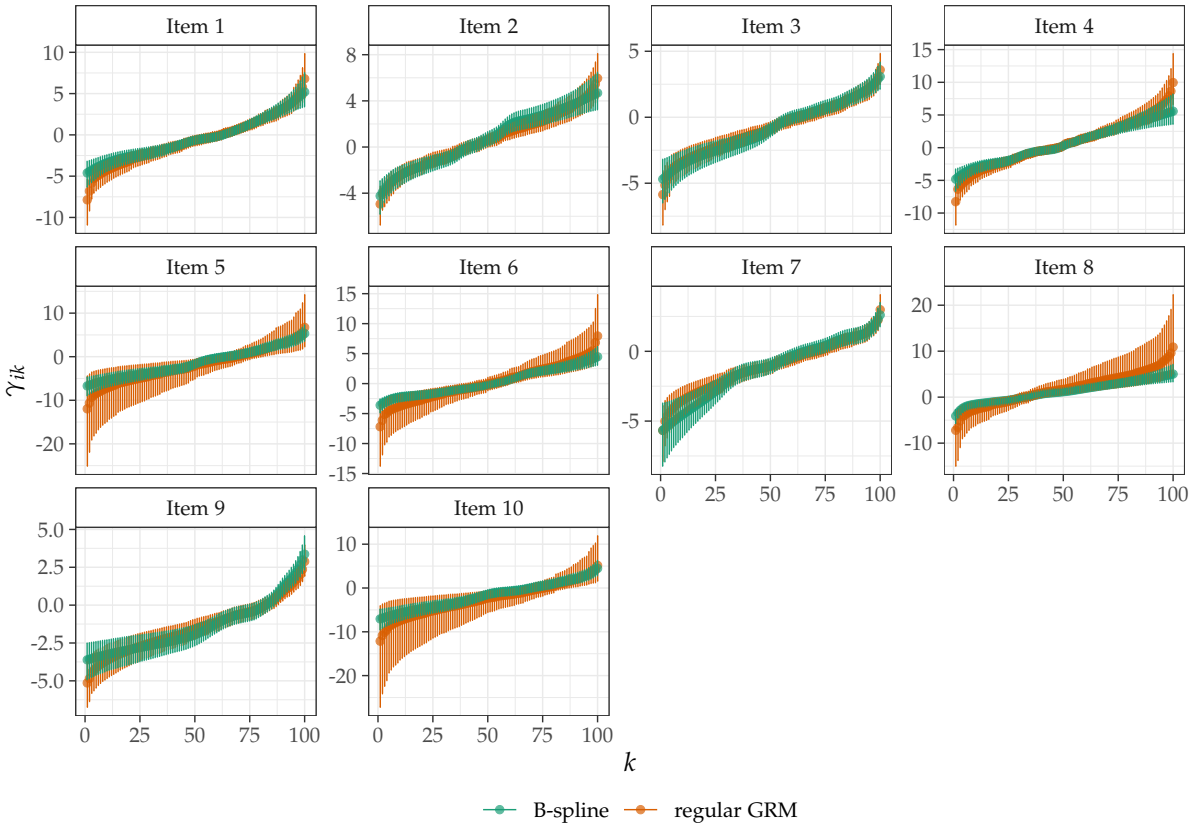

Figure 23: EAP estimation for threshold parameters including a 90% credible interval

5. Conclusion

Goal of this thesis was to propose a more parsimonious modelling approach for threshold parameters in ordered polytomous response models in a Bayesian Framework. The necessity for this arises in the context of large response scales (with many response categories) where a lot of parameters have to be estimated based on possibly very few responses resulting in null categories. The missing information about the threshold distribution is approximated by using a B-spline approach in context of the very commonly used GRM as a psychometric model. The special challenge is the shape constraint of a monotonously increasing assumption for the threshold parameters. The formulated approach was evaluated by a simulation study for small sample size and very extreme response patterns and applied to a well behaved real data situation of speeded C-test responses in comparison with its regular counterpart. Finally, an experimental attempt on a extremely small sample case for longitudinal data proved the only possible approach without extensive failing of MCMC sampling. In the simulation study, the B-spline methods performed clearly better in some factorial design settings regarding the averaged RMSE statistic. The proposed method is less biased for the EAP of ability and discrimination parameters and the interval estimation outperforms the regular GRM regarding the estimated coverage probability. Besides the faster computation it showed fewer problematic diagnostics warnings indicating a more efficient sampling properties. It would make sense to extend simulation settings by including severe misspecification of the general model, possibly revealing an even more robust behaviour. In the speeded C-test real data situation the results did not suggest any big improvements by using B-splines, which might be due to the smaller number of categories or the bigger sample size. However, a more conservative estimation of discrimination parameters and the superiority quantified by the $elpd^{loo}$ leads to the conclusion to favour the flexible model. In the extreme case of reading motivation data, only the B-spline model could be fitted without any problems with the simulation of the Markov chain and therefore could not be compared on a valid basis. In conclusion, one could state that the B-spline at least performs equally well, if not better. The easier implementation in a MCMC sampling algorithm with convergence and reliable non-biased estimates enables a more convenient out-of-the-box application on extreme measurement situation. Nonetheless, there are still many problems to tackle and alternative approaches to compare. In general, it would be a valid consideration to utilize a continuous response model as described by Mellenbergh (2018), which would be the most natural way to model a response scale like the VAS. This is especially valid, if a computerized more precise quantification is possible; in contrast to the mm-measure by hand. For ordered polytomous responses one could use the GPCM as basis for a flexible threshold parameter model approach. This would simplify the B-spline method, because there is no monotonicity constraint on the thresholds. Additionally, the parametrisation used by Andrich et al., 2003 would yield a interpretative meaning via the used "principle components". It would further make sense to also compare the method of collapsing null categories in these large scales

with the utilized methods regarding their item information. A possible alternative to B-splines could be a monotonously increasing polynomials revisited by Murray, Samuel Müller, et al. (2013) and modified (Murray, Müller, et al., 2016) and finally implemented in `Stan` for a fully Bayesian analysis by Manderson et al. (2017). For the B-spline modelling option alone there are innumerable possible topics, that could be addressed mathematically, with simulations or in practical applications. A thorough revision of a possible Bayesian penalized spline method via first and second order differences explained by Lang et al. (2004) could be one. Especially the sampling via random walk prior and the approximation in context of latent item parameters need a detailed examination. A further investigation of the knot distribution and the choice for the number of knots compared to the utilized approach in this thesis would also be an important topic for future research. Regarding the application to the real data sets, especially the longitudinal reading motivation data, further effort should be put into specifying a more robust model and evaluate when it fails precisely. As a concluding remark, it can be stated, that the parsimonious modelling via functional approximation is a performant approach and deserves recognition in applied analyses.

References

Abraham, Christophe and Khader Khadraoui (2015). "Bayesian regression with B-splines under combinations of shape constraints and smoothness properties". In: *Statistica Neerlandica* 69(2), pp. 150–170.

Agresti, Alan (2003). *Categorical data analysis*. John Wiley & Sons.

Albert, James H and Siddhartha Chib (1993). "Bayesian analysis of binary and polychotomous response data". In: *Journal of the American statistical Association* 88(422), pp. 669–679.

Ames, Allison J (2018). "Prior sensitivity of the posterior predictive checks method for item response theory models". In: *Measurement: Interdisciplinary Research and Perspectives* 16(4), pp. 239–255.

Andrich, David and Guanzhong Luo (2003). "Conditional pairwise estimation in the Rasch model for ordered response categories using principal components." In: *Journal of applied measurement* 4(3), pp. 205–221.

Baker, Frank B and Seock-Ho Kim (2004). *Item response theory: Parameter estimation techniques*. CRC press.

Baker, Frank B, Seock-Ho Kim, et al. (2017). *The basics of item response theory using R*. Springer.

Betancourt, Michael (2016). "Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1604.00695*.

Betancourt, Michael (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Betancourt, Michael and Mark Girolami (2015). "Hamiltonian Monte Carlo for hierarchical models". In: *Current trends in Bayesian methodology with applications* 79(30), pp. 2–4.

Brezger, Andreas and Winfried J Steiner (2008). "Monotonic regression based on Bayesian p–splines: An application to estimating price response functions from store-level scanner data". In: *Journal of business & economic statistics* 26(1), pp. 90–104.

Brown, Anna and Tim J Croudace (2014). "Scoring and estimating score precision using multidimensional IRT models". In: *Handbook of Item Response Theory Modeling*. Routledge, pp. 325–351.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell

(2017). “Stan: A probabilistic programming language”. In: *Journal of statistical software* 76(1).

De Boeck, Paul and Mark Wilson (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Vol. 10. Springer.

De Boor, Carl and Carl De Boor (2001). *A practical guide to splines*. Vol. 27. springer-verlag New York.

Forthmann, Boris, Rüdiger Grotjahn, Philipp Doebler, and Purya Baghaei (2020). “A comparison of different item response theory models for scaling speeded C-tests”. In: *Journal of Psychoeducational Assessment* 38(6), pp. 692–705.

Fox, Jean-Paul (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Gabry, Jonah and Rok Češnovar (2022). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin (2013). *Bayesian data analysis*. CRC press.

Grotjahn, Rüdiger (2010). “Gesamtdarbietung, Einzeltextdarbietung, Zeitbegrenzung und Zeitdruck: Auswirkungen auf Item-und Testkennwerte und C-Test-Konstrukt”. In: *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research*, pp. 265–296.

Harwell, Michael, Clement A Stone, Tse-Chi Hsu, and Levent Kirisci (1996). “Monte Carlo studies in item response theory”. In: *Applied psychological measurement* 20(2), pp. 101–125.

Heine, Simone (2017). *Fremd-und Zweitsprachenerfolg und seine Erklärung durch Erwerbsalter, kognitive, affektiv-motivationale und sozio-kulturelle Variablen: Eine empirische Studie*. kassel university press GmbH.

Hoffman, Matthew D, Andrew Gelman, et al. (2014). “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *J. Mach. Learn. Res.* 15(1), pp. 1593–1623.

Joanes, Derrick N and Christine A Gill (1998). “Comparing measures of sample skewness and kurtosis”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(1), pp. 183–189.

Johnson, Matthew S and Sandip Sinharay (2016). “Bayesian estimation”. In: *Handbook of item response theory* 2, pp. 237–257.

Johnson, Norman L, Samuel Kotz, and Narayanaswamy Balakrishnan (1995). *Continuous univariate distributions, volume 1*. Vol. 289. John wiley & sons.

Lang, Stefan and Andreas Brezger (2004). "Bayesian P-splines". In: *Journal of computational and graphical statistics* 13(1), pp. 183–212.

Leitenstorfer, Florian and Gerhard Tutz (2007). "Generalized monotonic regression based on B-splines with an application to air pollution data". In: *Biostatistics* 8(3), pp. 654–673.

Levy, Roy and Robert J Mislevy (2017). *Bayesian psychometric modeling*. Chapman and Hall/CRC.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of multivariate analysis* 100(9), pp. 1989–2001.

Lord, Frederic M and Melvin R Novick (2008). *Statistical theories of mental test scores*. IAP.

Luecht, Richard and Terry A Ackerman (2018). "A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals". In: *Educational Measurement: Issues and Practice* 37(3), pp. 65–76.

Luo, Guanzhong and David Andrich (2005). "Estimating parameters in the Rasch model in the presence of null categories". In: *J Appl Meas* 6(2), pp. 128–146.

Luo, Yong and Hong Jiao (2018). "Using the Stan program for Bayesian item response theory". In: *Educational and psychological measurement* 78(3), pp. 384–408.

Manderson, AA, E Cripps, K Murray, and BA Turlach (2017). "Monotone polynomials using BUGS and Stan". In: *Australian & New Zealand Journal of Statistics* 59(4), pp. 353–370.

Mellenbergh, Gideon J (2018). "Models for continuous responses". In: *Handbook of item response theory*. Chapman and Hall/CRC, pp. 153–163.

Meng, Xiao-Li (1994). "Posterior predictive p-values". In: *The annals of statistics* 22(3), pp. 1142–1160.

Merkle, Edgar C, Daniel Furr, and Sophia Rabe-Hesketh (2019). "Bayesian comparison of latent variable models: Conditional versus marginal likelihoods". In: *Psychometrika* 84, pp. 802–829.

Muraki, Eiji and James E Carlson (1995). "Full-information factor analysis for polytomous item responses". In: *Applied Psychological Measurement* 19(1), pp. 73–90.

Murray, Kevin, S Müller, and BA Turlach (2016). "Fast and flexible methods for monotone polynomial fitting". In: *Journal of Statistical Computation and Simulation* 86(15), pp. 2946–2966.

Murray, Kevin, Samuel Müller, and Berwin A Turlach (2013). "Revisiting fitting monotone polynomials to data". In: *Computational Statistics* 28(5), pp. 1989–2005.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2(11), p. 2.

Nesterov, Yurii (2009). "Primal-dual subgradient methods for convex problems". In: *Mathematical programming* 120(1), pp. 221–259.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles (2004). "Generalized multilevel structural equation modeling". In: *Psychometrika* 69(2), pp. 167–190.

Ramsay, James O (1988). "Monotone regression splines in action". In: *Statistical science*, pp. 425–441.

Samejima, Fumiko (1969). "Estimation of latent ability using a response pattern of graded scores." In: *Psychometrika monograph supplement*.

Samejima, Fumiko (1972). "A general model for free-response data." In: *Psychometrika Monograph Supplement*.

Samejima, Fumiko (1995). "Acceleration model in the heterogeneous case of the general graded response model". In: *Psychometrika* 60(4), pp. 549–572.

Samejima, Fumiko (1996). "Evaluation of mathematical models for ordered polychotomous responses". In: *Behaviormetrika* 23(1), pp. 17–35.

Samejima, Fumiko (2016). "Graded response models". In: *Handbook of item response theory, volume one*. Chapman and Hall/CRC, pp. 123–136.

San Martin, Ernesto (2018). "Identifiability of structural characteristics: How relevant is it for the Bayesian approach?" In.

San Martin, Ernesto and Jorge Gonzalez (2010). "Bayesian identifiability: Contributions to an inconclusive debate". In: *Chilean Journal of Statistics* 1(2), pp. 69–91.

Scott, David W (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, Bernard W (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Simonoff, Jeffrey S (2003). *Analyzing categorical data*. Vol. 496. Springer.

Sinharay, Sandip (2006). "Bayesian item fit analysis for unidimensional item response theory models". In: *British journal of mathematical and statistical psychology* 59(2), pp. 429–449.

Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Stan Development Team (2023). *The Stan Core Library*. Version 2.31.0. URL: http://mc-stan.org/.

Tutz, Gerhard (2011). *Regression for categorical data*. Vol. 34. Cambridge University Press.

Tutz, Gerhard (2021). "Item Response Thresholds Models". In: *arXiv preprint arXiv:2106.12784*.

Vehtari, Aki, Jonah Gabry, Mans Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.5.1. URL: https://mc-stan.org/loo/.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and computing* 27, pp. 1413–1432.

A. Additional Calculations, Figures and Tables

A.1. GRM Calculations

$$\frac{\partial}{\partial\theta_p}p_{ik}(\theta_p) = \begin{cases} \frac{\partial}{\partial\theta_p}\Psi(\gamma_{i1}-\alpha_i\theta_p) & \text{, for } k=0 \\ \frac{\partial}{\partial\theta_p}(\Psi(\gamma_{ik}-\alpha_i\theta_p)-\Psi(\gamma_{ik-1}-\alpha_i\theta_p)) & \text{, for } k\in\{1,\dots,K_i-1\} \\ \frac{\partial}{\partial\theta_p}(1-\Psi(\gamma_{iK_i-1}-\alpha_i\theta_p)) & \text{, for } k=K_i \end{cases} \tag{151}$$

$$= \begin{cases} -\alpha_i\psi(\gamma_{i1}-\alpha_i\theta_p) & \text{, for } k=0 \\ -\alpha_i(\psi(\gamma_{ik}-\alpha_i\theta_p)-\psi(\gamma_{ik-1}-\alpha_i\theta_p)) & \text{, for } k\in\{1,\dots,K_i-1\} \\ \alpha_i\psi(\gamma_{iK_i-1}-\alpha_i\theta_p) & \text{, for } k=K_i \end{cases} \tag{152}$$

$$\frac{\partial^2}{\partial^2\theta^2}p_{ik}(\theta_p) = \begin{cases} \frac{\partial}{\partial\theta_p}-\alpha_i(\psi(\gamma_{i1}-\alpha_i\theta_p) & \text{, for } k=0 \\ \frac{\partial}{\partial\theta_p}-\alpha_i(\psi(\gamma_{ik}-\alpha_i\theta_p)-\psi(\gamma_{ik-1}-\alpha_i\theta_p)) & \text{, for } k\in\{1,\dots,K_i-1\} \\ \frac{\partial}{\partial\theta_p}\alpha_i(\psi(\gamma_{iK_i-1}-\alpha_i\theta_p) & \text{, for } k=K_i \end{cases} \tag{153}$$

$$= \begin{cases} \frac{\partial}{\partial\theta_p}-\alpha_i\frac{\exp(\gamma_{i1}-\alpha_i\theta_p)}{(1+\exp(\gamma_{i1}-\alpha_i\theta_p))^2} & \text{, for } k=0 \\ \frac{\partial}{\partial\theta_p}-\alpha_i\left(\frac{\exp(\gamma_{ik}-\alpha_i\theta_p)}{(1+\exp(\gamma_{ik}-\alpha_i\theta_p))^2}-\frac{\exp(\gamma_{ik-1}-\alpha_i\theta_p)}{(1+\exp(\gamma_{ik-1}-\alpha_i\theta_p))^2}\right) & \text{, for } k\in\{1,\dots,K_i-1\} \\ \frac{\partial}{\partial\theta_p}\alpha_i\frac{\exp(\gamma_{iK_i-1}-\alpha_i\theta_p)}{(1+\exp(\gamma_{iK_i-1}-\alpha_i\theta_p))^2} & \text{, for } k=K_i \end{cases} \tag{154}$$

$$= \begin{cases} \alpha_i^2\frac{\exp(\alpha_i\theta_p+\gamma_{i1})(\exp(\alpha_i\theta_p)-\exp(\gamma_{i1}))}{(\exp(\alpha_i\theta_p)+\exp(\gamma_{i1}))^3} & \text{, for } k=0 \\ (\star\star\star) & \text{, for } k\in\{1,\dots,K_i-1\} \\ -\alpha_i^2\frac{\exp(\alpha_i\theta_p+\gamma_{iK_i-1})(\exp(\alpha_i\theta_p)-\exp(\gamma_{iK_i-1}))}{(\exp(\alpha_i\theta_p)+\exp(\gamma_{iK_i-1}))^3} & \text{, for } k=K_i \end{cases} \tag{155}$$

$$(\star\star\star) = \alpha_i^2\Big(\frac{\exp(\alpha_i\theta_p+\gamma_{ik})(\exp(\alpha_i\theta_p)-\exp(\gamma_{ik}))}{(\exp(\alpha_i\theta_p)+\exp(\gamma_{ik}))^3}- \tag{156}$$

$$\frac{\exp(\alpha_i\theta_p+\gamma_{ik-1})(\exp(\alpha_i\theta_p)-\exp(\gamma_{ik-1}))}{(\exp(\alpha_i\theta_p)+\exp(\gamma_{ik-1}))^3}\Big) \tag{157}$$

A.2. B-Spline Calculations

$$B_{m,1,\boldsymbol{t}}(y) = \mathbb{1}_{[t_m,t_{m+1})}(y) \tag{158}$$

$$= \begin{cases} 1 & ,y\in[t_m,t_{m+1}) \\ 0 & ,y\notin[t_m,t_{m+1}) \end{cases} \tag{159}$$

$$
\begin{aligned}
B_{m,2}(y) =& \omega_{m,2,\boldsymbol{t}}(y) B_{m,1,\boldsymbol{t}}(y) + (1 - \omega_{m,2,\boldsymbol{t}}(y)) B_{m+1,1,\boldsymbol{t}}(y) && (160)\\
=& \frac{y - t_m}{t_{m+1} - t_m} \mathbb{1}_{[t_m, t_{m+1})} + \left(1 - \frac{y - t_{m+1}}{t_{m+2} - t_{m+1}}\right) \mathbb{1}_{[t_{m+1}, t_{m+2})} && (161)\\
=& \frac{y - t_m}{t_{m+1} - t_m} \mathbb{1}_{[t_m, t_{m+1})} + \left(\frac{t_{m+2} - t_{m+1} - y + t_{m+1}}{t_{m+2} - t_{m+1}}\right) \mathbb{1}_{[t_{m+1}, t_{m+2})} && (162)\\
=& \begin{cases} \frac{y - t_m}{t_{m+k-1} - t_m} & , y \in [t_m, t_{m+1}) \\ \frac{t_{m+2} - y}{t_{m+2} - t_{m+1}} & , y \in [t_{m+1}, t_{m+2}) \\ 0 & , y \notin [t_m, t_{m+2}) \end{cases} && (163)
\end{aligned}
$$

$$
\begin{aligned}
B_{m,3}(y) =& \omega_{m,3,\boldsymbol{t}}(y)B_{m,2,\boldsymbol{t}}(y) + (1-\omega_{m+1,3,\boldsymbol{t}}(y))B_{m+1,2,\boldsymbol{t}}(y) && (164)\\
=& \omega_{m,3,\boldsymbol{t}}(y)(\omega_{m,2,\boldsymbol{t}}(y)B_{m,1,\boldsymbol{t}}(y) + (1-\omega_{m+1,2,\boldsymbol{t}}(y))B_{m+1,1,\boldsymbol{t}}(y))) && (165)\\
&+ (1-\omega_{m+1,3,\boldsymbol{t}})(\omega_{m+1,2,\boldsymbol{t}}B_{m+1,1,\boldsymbol{t}}(y) + (1-\omega_{m+2,2,\boldsymbol{t}})B_{m+2,1,\boldsymbol{t}}(y))) && (166)\\
=& \omega_{m,3,\boldsymbol{t}}(y)(\omega_{m,2,\boldsymbol{t}}(y)\mathbb{1}_{[t_m,t_{m+1})}(y) + (1-\omega_{m+1,2,\boldsymbol{t}}(y))\mathbb{1}_{[t_{m+1},t_{m+2})}(y)) && (167)\\
&+ (1-\omega_{m+1,3,\boldsymbol{t}}(y))(\omega_{m+1,2,\boldsymbol{t}}(y)\mathbb{1}_{[t_{m+1},t_{m+2})}(y) + (1-\omega_{m+2,2,\boldsymbol{t}}(y))\mathbb{1}_{[t_{m+2},t_{m+3})}(y)) && (168)\\
=& \omega_{m,3,\boldsymbol{t}}(y)\omega_{m,2,\boldsymbol{t}}(y)\mathbb{1}_{[t_m,t_{m+1})}(y) && (169)\\
&+ \omega_{m,3,\boldsymbol{t}}(y)(1-\omega_{m+1,2,\boldsymbol{t}}(y)\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (170)\\
&+ (1-\omega_{m+1,3,\boldsymbol{t}}(y))\omega_{m+1,2,\boldsymbol{t}}(y)\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (171)\\
&+ (1-\omega_{m+1,3,\boldsymbol{t}}(y))(1-\omega_{m+2,2,\boldsymbol{t}}(y))\mathbb{1}_{[t_{m+2},t_{m+3})}(y) && (172)\\
=& \frac{y-t_m}{t_{m+2}-t_m}\frac{y-t_m}{t_{m+1}-t_m}\mathbb{1}_{[t_m,t_{m+1})}(y) && (173)\\
&+ \frac{y-t_m}{t_{m+2}-t_m}\left(1-\frac{y-t_{m+1}}{t_{m+2}-t_{m+1}}\right)\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (174)\\
&+ \left(1-\frac{y-t_{m+1}}{t_{m+3}-t_{m+1}}\right)\frac{y-t_{m+1}}{t_{m+2}-t_{m+1}}\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (175)\\
&+ \left(1-\frac{y-t_{m+1}}{t_{m+3}-t_{m+1}}\right)\left(1-\frac{y-t_{m+2}}{t_{m+3}-t_{m+2}}\right)\mathbb{1}_{[t_{m+2},t_{m+3})}(y) && (176)\\
=& \frac{(y-t_m)^2}{(t_{m+2}-t_m)(t_{m+1}-t_m)}\mathbb{1}_{[t_m,t_{m+1})}(y) && (177)\\
&+ \frac{y-t_m}{t_{m+2}-t_m}\frac{t_{m+2}-t_{m+1}-y+t_{m+1}}{t_{m+2}-t_{m+1}}\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (178)\\
&+ \frac{t_{m+3}-t_{m+1}-y+t_{m+1}}{t_{m+3}-t_{m+1}}\frac{y-t_{m+1}}{t_{m+2}-t_{m+1}}\mathbb{1}_{[t_{m+1},t_{m+2})}(y) && (179)\\
&+ \frac{t_{m+3}-t_{m+1}-y+t_{m+1}}{t_{m+3}-t_{m+1}}\frac{t_{m+3}-t_{m+2}-y+t_{m+2}}{t_{m+3}-t_{m+2}}\mathbb{1}_{[t_{m+2},t_{m+3})}(y) && (180)\\
=& \begin{cases}
\frac{(y-t_m)^2}{(t_{m+1}-t_m)(t_{m+2}-t_m)} & ,y\in[t_m,t_{m+1})\\
\frac{(y-t_m)(t_{m+2}-y)}{(t_{m+2}-t_m)(t_{m+2}-t_{m+1})}+\frac{(y-t_{m+1})(t_{m+3}-y)}{(t_{m+3}-t_{m+1})(t_{m+2}-t_{m+1})} & ,y\in[t_{m+1},t_{m+2})\\
\frac{(t_{m+3}-y)^2}{(t_{m+3}-t_{m+1})(t_{m+3}-t_{t+2})} & ,y\in[t_{m+2},t_{m+3})\\
0 & ,y\notin[t_m,t_{m+3})
\end{cases} && (181)
\end{aligned}
$$

A.3. Simulation Results

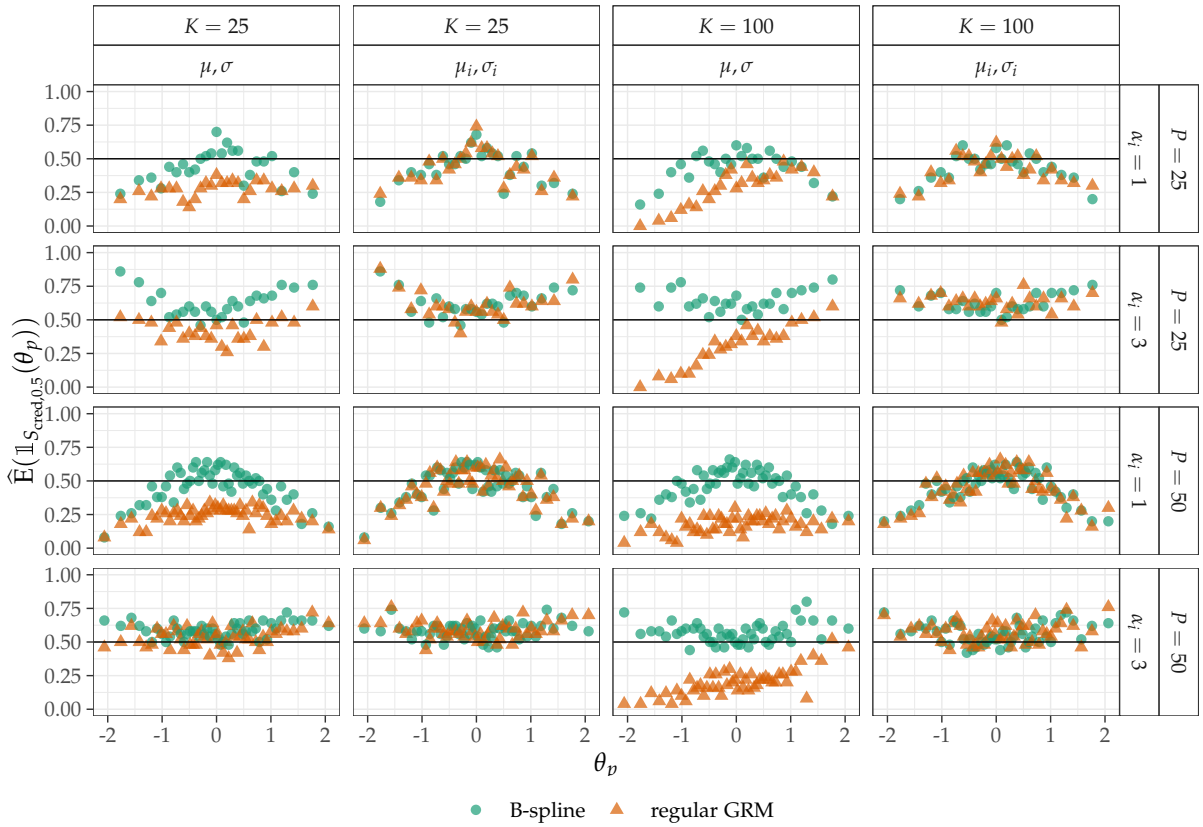

Figure 24: Estimated Mean Coverage $\widehat{\mathrm{E}}(\mathbb{1}_{S_{\mathrm{cred},0.5}}(\theta_p))$ of 50% credible interval

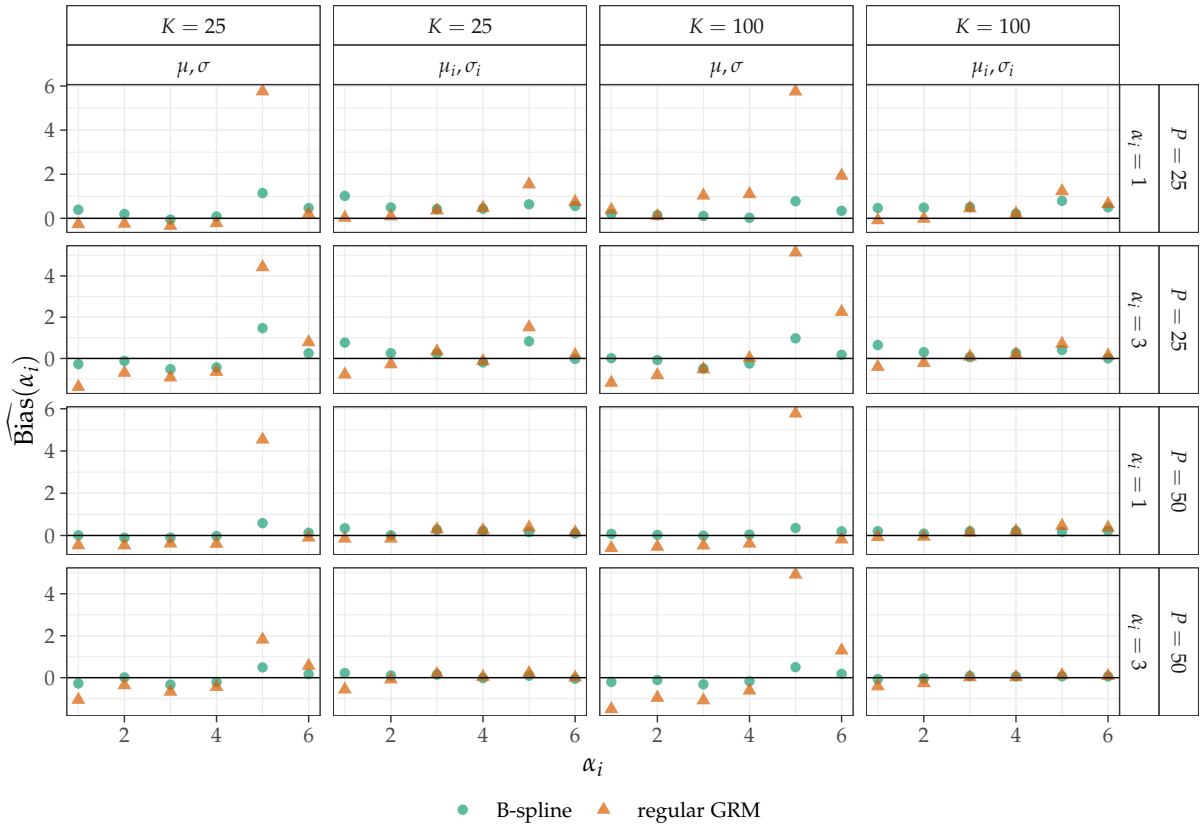

Figure 25: Estimated Root Mean Squared Error $\widehat{\text{Bias}}(\alpha_i)$ for the discrimination parameters α_i with $i \in \{1, \ldots, 6\}$ for $n \in \{1, \ldots, N\}$ and $N \in \mathbb{N}$ simulations

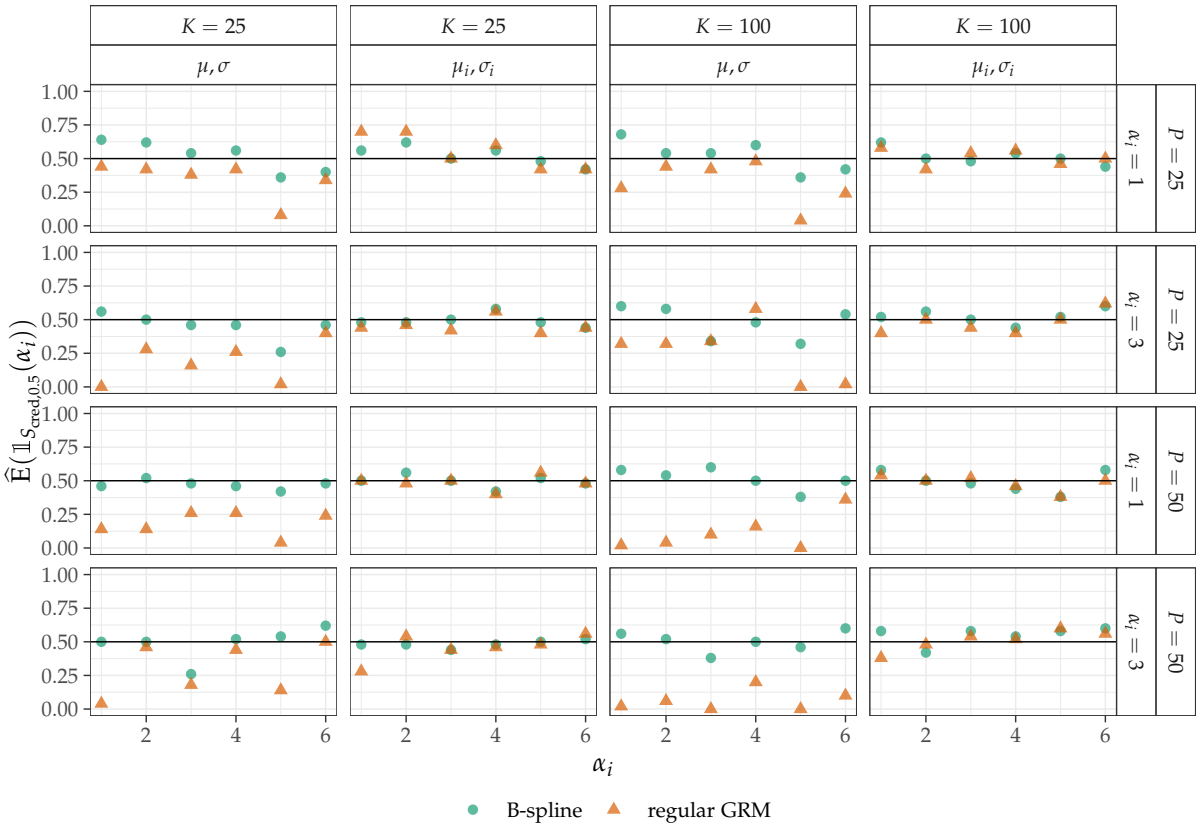

Figure 26: Estimated Mean Coverage $\widehat{\mathrm{E}}(\mathbb{1}_{S_{\mathrm{cred},0.5}}(\alpha_i))$ of 50% credible interval

A.4. C-Test Data

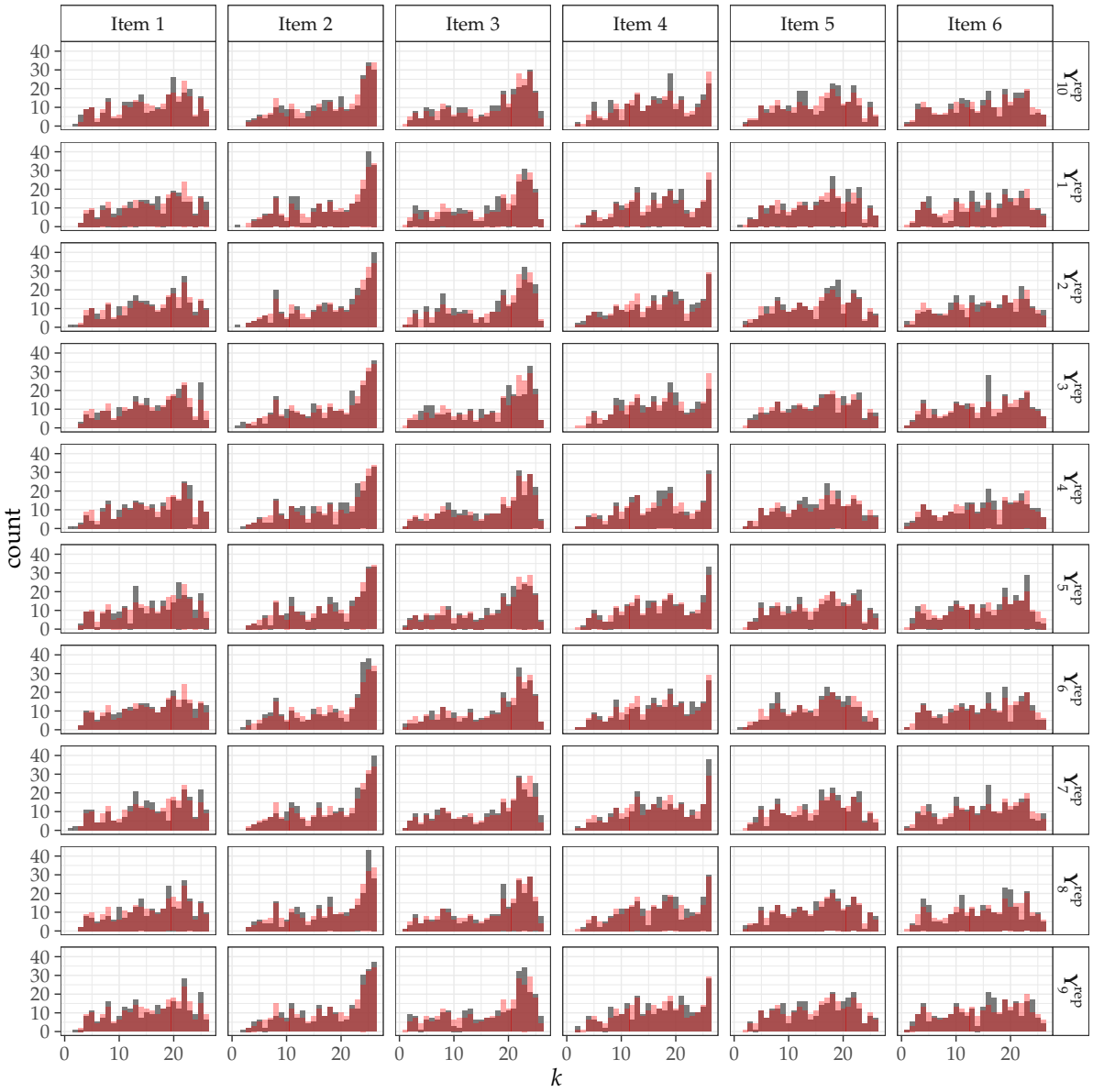

Figure 27: Barplots of 10 replicated response pattern $\boldsymbol{Y}^{\text{rep}}$ of the regular GRM per item in gray and the observed response pattern $\boldsymbol{Y}$ as red overlay

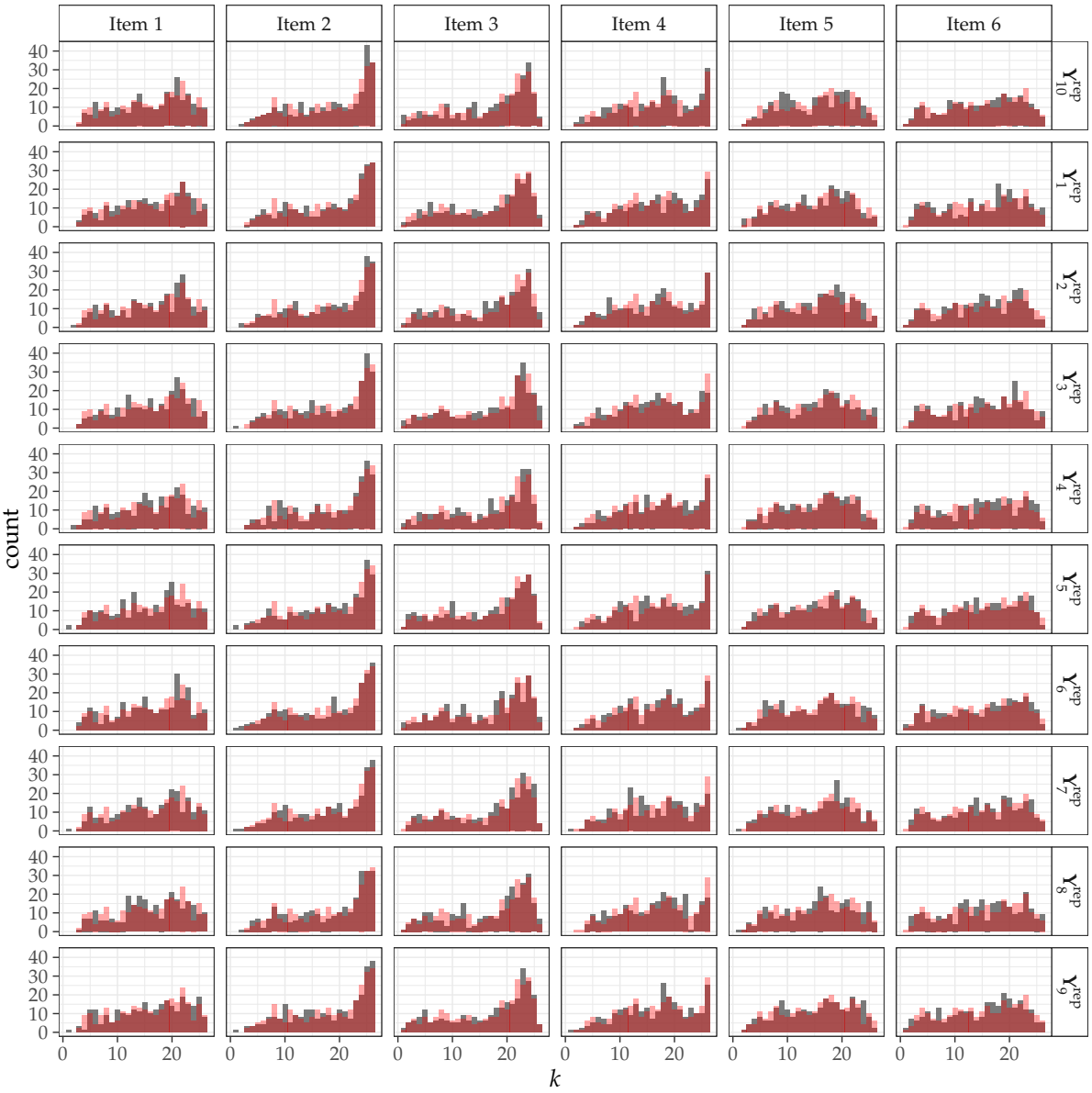

Figure 28: Barplots of 10 replicated response pattern $\boldsymbol{Y}^{\text{rep}}$ of the B-Spline version of the GRM per item in gray and the observed response pattern $\boldsymbol{Y}$ as red overlay

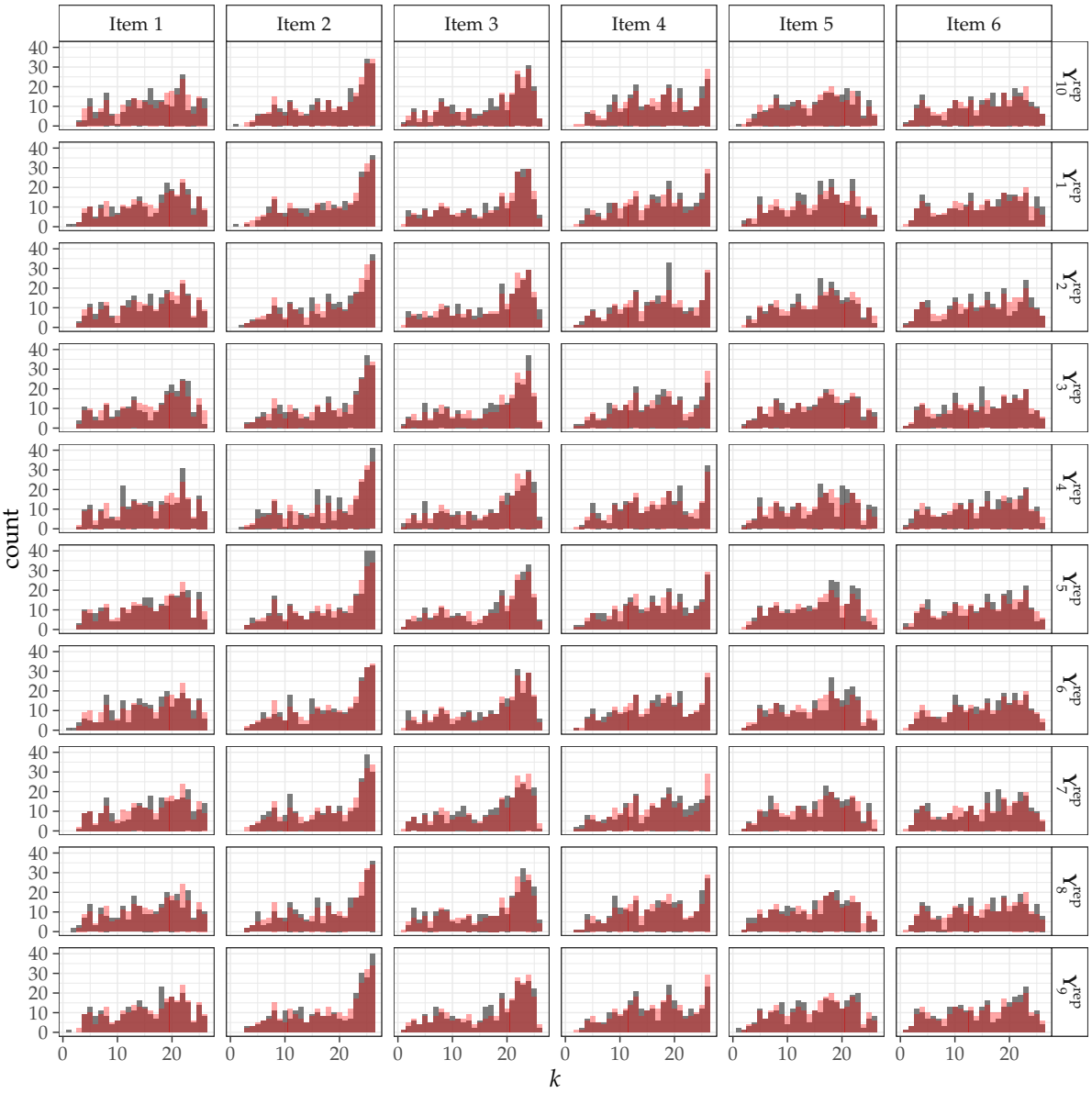

Figure 29: Barplots of 10 replicated response pattern $\boldsymbol{Y}^{\text{rep}}$ of the regular GRM per item in gray and the observed response pattern $\boldsymbol{Y}$ as red overlay

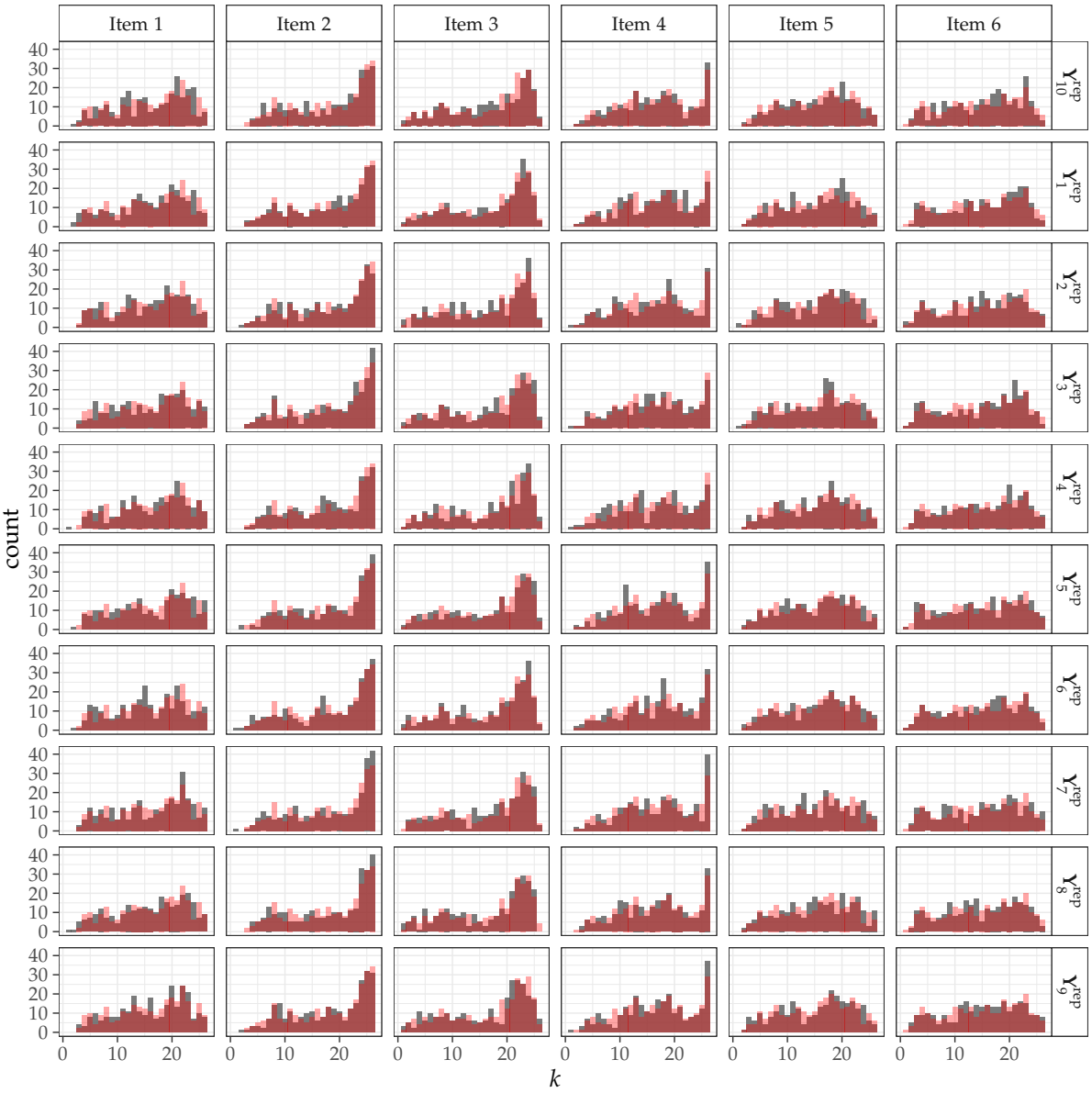

Figure 30: Barplots of 10 replicated response pattern $\boldsymbol{Y}^{\text{rep}}$ of the B-Spline version of the GRM per item in gray and the observed response pattern $\boldsymbol{Y}$ as red overlay