Chapter 5

Static, sequential and dynamic estimation in the extended marginal Rasch model

Summary

We present an extended marginal Rasch model (ERM) with a Bayesian estimation procedure that allows for estimation in static contexts using fully observed data matrices, in sequential adaptive contexts such as computer adaptive testing (CAT) where model parameters are estimated as data becomes sequentially available, and in dynamic contexts, such as computer adaptive learning (CAL) or computer adaptive practice (CAP) where model parameters can change over time. Such changes are expected due to learning taking place and the influences of feedback. The ERM is applied using data from a large scale fully adaptive on-line learning environment.
Chapter 5. Dynamic estimation in the extended Rasch model

5.1 Introduction

In psychometrics, the amount and the type of data that is available for analyses changed during the last decades, and is evolving still. The foundations of item response theory (IRT) were developed with the situation in mind where a group of students responds to a predetermined and fixed set of items (Rasch, 1960; Choppin, 1968; Lord & Novick, 1968). Later, methods were developed to deal with incomplete data matrices, e.g., due to a simple anchor test design (e.g., Hanson & Béguin, 2002), or by modeling non-ignorable missing data (e.g., Holman & Glas, 2005). The introduction of computer adaptive testing (CAT) created again a different structure of data, where every person could in principle make a unique item set adapted to his or her ability level (e.g., Wainer, 2000; van der Linden & Glas, 2002; Eggen, 2004). A more recent development is the rise of computer adaptive learning (CAL) or computer adaptive practice (CAP) (e.g., Wauters, Desmet, & Van den Noortgate, 2010; Klinkenberg, Straatemeier, & van der Maas, 2011), which are conceptually different in that students respond frequently to items over an extended period of time, e.g., years, with the same student potentially responding to the same question at different moments. The functioning of both persons and items are expected to change during the course of measurement, since learning is the actual goal of such environments. The possibility of providing feedback directly to the learners, teacher or parents adds to the expectancy of changing parameters in estimating models on such data.

In this writing, we present a simple psychometric model and an estimation method that allows to deal with all scenarios described above. It is able to deal with static data matrices, either fully observed or with structural missing data, sequential data as generated by CATs and dynamic data as generated by CAL or CAP, where both persons and item characteristics could be continually changing.

The approach we take is based on an extension of the extended marginal Rasch model (ERM) presented in (Maris, Bechger, & San Martin, submitted), where an efficient Markov chain Monte Carlo (MCMC) sampler for the ERM is used in the context of a single test with no missing data. We extend their approach in three directions. First, we want to estimate the parameters of the ERM for equivalent groups designs. Second, we extend the approach to random booklet designs so it can be used for sequential estimation, e.g., for sequentially updating the parameters in CAT as persons are added to the data. Third, a dynamic context is discussed where the model parameters are allowed to change over time to accommodate for learning.
5.1.1 Organization

This paper is organized as follows. The ERM is discussed, along with the implications of the equivalent groups design and the identification of the model, in section 5.2. Following in section 5.3, Bayesian estimation, sequential Bayesian estimation and dynamic Bayesian estimation are discussed here to estimate the model. Simulations for static and sequential adaptive estimation are presented in section 5.4, and dynamic estimation is performed using data from a fully adaptive on-line practice environment, in section 5.5. The paper ends with a discussion.

5.2 The extended Rasch model

The ERM is introduced by Tjur (1982) and coined as such and discussed by Cressie and Holland (1983). Relations to other generalized Rasch models are discussed by, amongst others, de Leeuw and Verhelst (1986) and Kelderman (1984). The probability distribution of the responses in the ERM can be denoted as follows (Maris et al., submitted):

\[
P(x|b, \lambda) = \prod_{i=1}^{m} b_i^{x_i} \lambda_{x_i} \sum_{s=0}^{m} \gamma_s(b) \lambda_s
\]

(5.1)

where \( x \) is a vector of \( m \) scored responses with \( x_+ \) as sum score and each item \( i \) has an item parameter \( b_i \). With every score, there is an associated score parameter \( \lambda_{x_+} \). Elementary symmetric functions (Verhelst, Glas, & van der Sluis, 1984; Baker & Harwell, 1996) of order \( s \) of the vector \( b \) are noted by \( \gamma_s(b) \) and are defined to be zero if \( s < 0 \) or if \( s > m \). Equation (5.1) concerns a complete design, hereafter we will discuss the ERM using incomplete designs.

5.2.1 The equivalent groups design

Incomplete designs commonly used in educational measurement. For example, one might construct different versions of a tests for security reasons, to prevent cheating or to suppress the exposure rate. The resulting data contains structural missing values, since respondents are offered only subsets of the complete set of test items. Though many methods exist to equate the test versions in incomplete designs, we only consider concurrent estimation (e.g., Hanson & Béguin, 2002).

A common incomplete design is the non-equivalent anchor test (NEAT), where groups are allowed to differ in their ability distribution and every test
version consists of a part that overlaps other versions, the anchor. Another common design, and the one used in this paper, is the equivalent groups design (e.g., von Davier, Holland, & Thayer, 2004). Here the groups of respondents should be equivalent in their ability distribution, but test versions are not required to contain an anchor. Clearly, a disadvantage of the equivalent groups design is that groups should be equal in their ability distribution. However, we discuss two specific advantages compared to the NEAT design.

First, the equivalent groups design in the ERM allows us to project any score on any version of the test to the complete collection of test items, i.e., any test score can be expressed as an item bank score. This is advantageous since these item bank scores are directly comparable regardless of the version they originate from and without having to assume a latent trait or specific score distribution. We note that this projection is not possible when using the NEAT design with the ERM, as discussed by Bolsinova and Maris (submitted). The projection from test version score to the whole collection of items is tightly connected to the identification of the model as discussed in section 5.A.1.

Second, the simplicity of the equivalent groups design proofs to be advantageous for sequential and dynamic estimation, as discussed in section 5.3.2 and 5.3.3.

Though anchoring or overlap is not required in the equivalent groups design, overlap can be considered beneficial for two reasons. It allows to test the assumptions that the groups are equivalent and a more connected design causes faster convergence of the estimation algorithms, cf. the discussion in 5.3.3 and 5.5 on the fraction of missing information and convergence.

In extending the ERM to support incomplete designs, we begin to generalize Formula (5.1) to support two test versions, or booklets. The number of booklets can easily be extended, but is limited to two for clarity in the following equations. We regard the booklets to be non-overlapping for now:

\[
P(x, y|b, c, \lambda) = \frac{\prod_{i=1}^{m} b_i^{x_i} \prod_{j=1}^{n} c_j^{y_j} \lambda_{x_i+y_j}}{\sum_{u=0}^{m+n} \gamma_u(b, c) \lambda_u}
\]  

(5.2)

Here, two booklets \(x\) and \(y\) are presented, with corresponding response vectors \(x\) and \(y\) with sum scores \(x_+\) and \(y_+\), and vectors of item parameters \(b\) and \(c\) with length \(m\) and \(n\). \(\lambda\) denotes the vector of score parameters. Elementary symmetric functions of order \(u\) of the vectors \(b\) and \(c\) are denoted by \(\gamma_u(b, c)\).

Under the equivalent group design, if only one booklet is made, the response probabilities depend on the other model parameters. In a two-booklet example, we find two formulas for the probabilities of observing responses in
5.3. Methods

In this section, we discuss three Bayesian estimation procedures for three different data structures, using the equivalent groups ERM. Each of these data structures includes a missing data problem.

The first missing data problem, in 5.3.1, concerns incomplete designs, where different test versions, or booklets, are administered to different students. In this scenario, the observed data matrix is fixed, or static, i.e., no data is added during estimation. We employ Bayesian estimation, specifically a Gibbs sampler (Casella & George, 1992; Maris et al., submitted), to estimate the ERM parameters in this situation.

The second missing data problem, in 5.3.2, concerns more complicated incomplete designs without a concise booklet structure, as those generated by CAT. Since every candidate receives a tailored set of items, the number of booklets can become large. The missing data problem is different compared to a more sparse booklet design, but in CATs we can regard the design to be ignorable (Eggen & Verhelst, 2011). Also, we can regard the data to be structurally different. Individual candidates can be added naturally to the data set, since tailored item sets are already considered for each other candidate as well. We employ sequential Bayesian estimation to sequentially estimate the parameters of the ERM as candidates are added to the data set.

The third missing data problem, in 5.3.3, is similar to the second in that
we also consider complicated incomplete designs, with item sets tailored to the individual candidates. However, we allow the data set to grow when new observations become available, and to shrink as older data gets purged. With the possibility to purge older data, we strike a balance in maintaining a sufficiently large data set for estimation, while minimizing the amount of bias, e.g., as would be introduced by ability growth, item drift, etcetera. In doing so, we allow for a dynamically changing data set, where data is added and purged as parameters are estimated. We estimate the equivalent groups ERM on this data using a tailored dynamic Bayesian estimation procedure. This dynamic approach would be especially suitable for situations where we expect a change in the model parameters, for example in pupil monitoring systems, were we expect item parameters to change, e.g., due to curriculum changes, and expect candidate parameters to change as well due to learning.

5.3.1 Static Bayes

In this first approach, a static data set is assumed, i.e., no changes occur to the data during estimation. We set up a tailored Gibbs sampler for a specific incomplete design. The approach is tailored in that a specific Gibbs sampler is required for a specific design. As a simple nontrivial example, two non-overlapping booklets are used. The approach can be extended easily to more booklets with possible overlap.

In the non-overlapping two booklet example, we sample from a specific proposal distributions for each of the two sets of item parameters, \( b \) and \( c \). In order to draw samples from the posterior distributions of (5.2), we choose a proper conjugate prior by adding two rows and two columns to the data set, a row and column with all zeros and a row and column with all ones. The posterior distribution follows:

\[
f(b, c, \lambda | x, y) \propto \frac{\prod_{i=1}^{m_x} b_i^{x+i} \prod_{s=0}^{m_y} \left[ \sum_{t=0}^{n_t} \gamma_t(c) \lambda_{s+t} \right]^{m_t^p}}{\left[ \sum_{u=0}^{m+n} \gamma_u(b, c) \lambda_u^{n_u-1} \right]^{m^p}} \\
\times \frac{\prod_{j=1}^{n_y} c_j^{y+j} \prod_{t=0}^{n_t} \left[ \sum_{s=0}^{m_s} \gamma_s(b) \lambda_{s+t} \right]^{n_t^p}}{\left[ \sum_{u=0}^{m+n} \gamma_u(b, c) \lambda_u \right]^{n_p}}
\]

(5.5)

where \( m^p \) and \( n^p \) are the number of respondents in booklets \( x \) and \( y \). Specifically, \( m_i^p \) is the number of respondents in booklet \( x \) with score \( s \), and \( n_t^p \) the number of people in booklet \( y \) with score \( t \). These numbers of respondents are all including the added rows from the prior, to simplify notation.
Similarly, we denote the sums with the added prior $x'_{+i} + 1$ and $y'_{+j} + 1$ simply as $x'_{+i}$ and $y'_{+j}$.

**Gibbs sampling**

The Gibbs sampler for the two booklet example is straightforward. First, item parameters $b$ and $c$ of the two booklets are sampled from the full conditional distribution for item parameters, see 5.A.2. The log-target distribution is found to be concave with linear tails. Sampling from such a distribution can be done in a multitude of ways. Two sampling approaches are provided in 5.A.2, using a Metropolis algorithm with a scaled beta prime distribution as a proposal, and a piecewise linear majorizing proposal distribution using adaptive rejection sampling.

Second, the score parameters are sampled. Since the full conditional distribution for score parameters in the equivalent groups ERM it is not an exponential family model, it is not guaranteed to be log-concave. Therefore, We choose for a data augmentation (DA) approach (Albert, 1992; Maris & Maris, 2002; van Dyk & Meng, 2001), where we impute the sum scores on unobserved items given the observed sum scores, i.e., we only do imputation of the marginal distribution. The use of DA allows us to sample $\lambda$ from the target distribution of the completed data.

Third, by adding the imputed and the observed marginals a completed marginal is obtained, from which $\lambda$ is sampled, see 5.A.2.

### 5.3.2 Sequential adaptive Bayes

When the design becomes complex and involves large numbers of different test forms, extending and evaluating the approach in section 5.3.1 becomes computationally prohibitive. Therefore, we discuss a marginal DA approach not only to facilitate the estimation of the score parameters $\lambda$, but also to estimate the item parameters $b$.

The procedure starts by imputing for each person the sum score on the missing data, as in the static data case in section 5.3.1. Second, the unique missing responses are imputed by generating plausible responses, using a method by Marsman (2014, chapter 3), details are provided in 5.A.3. Finally, new sufficient statistics are calculated using the imputed data, and the estimation procedure on complete data by Maris et al. (submitted) is used to estimate item and score parameters.

Though this approach introduces some additional autocorrelation on the parameter estimates, it is beneficial in that it needs no tailored approach for
different booklets designs. A simulation performed in section 5.4.2 demonstrates that the marginal DA approach can be useful in sequential parameter estimation, despite the increase in autocorrelation.

5.3.3 Dynamic Bayes

The method to perform dynamic Bayesian estimation in the ERM is very similar to performing sequential estimation in section 5.3.2, with some key differences. As mentioned in the introduction, in the context in which dynamic estimation is applicable changes in person and item parameters are expected. To facilitate these changes, we consider a data matrix of persons by items that not only grows as persons are added to the system, as with sequential estimation, but one that more generally only contains relevant responses. Responses are considered relevant if they are given recently according to some definition, e.g., a response given to an item a year ago is not considered informative for current parameter estimation.

In the context of practice it also occurs that questions are answered multiple times, in which the last scored response is considered most relevant for parameter estimation. The result is that the observed data in the data matrix is allowed to change continually, and therefore the augmented data as well. Since the ERM model parameters are estimated on the complete data matrix after the marginal DA has been applied, the model parameters can adapt to changes in the underlying observed data, with time lag depending on one’s definition of what data is still recent.

Though the data matrix is allowed to change, there are some limitations. First, the fraction of missing information is very relevant, since there is a lower bound to the fraction of observed information that is necessary for performing sequential DA and parameter estimation. Second, though it is possible to dynamically change the amount of persons, i.e., they can be added or removed between estimations, the amount of items is less flexible due to the score parameters $\lambda$. Clearly, extensions here are possible, for example by considering multiple booklets, as in section 5.3.2, but are not worked out in detail here.

A detailed application of dynamic estimation on a large data set is given in section 5.5. Due to this extensive application and the similarities between sequential and dynamic estimation, no additional simulations have been performed for dynamic estimation.
5.4 Simulations

In this section, simulations are used to demonstrate the workings of the static and sequential estimation using the ERM. First, estimation is performed on a static data matrix in section 5.4.1, where the influence of missing data on the estimated score distribution is plotted, and autocorrelation is discussed. Second, in section 5.4.2, simulation are performed for sequential estimation. In every iteration, a new person is added to the data matrix, and it is demonstrated how item parameters converge in this situation. Dynamic estimation is not simulated, but applied and discussed in section 5.5.

5.4.1 Static simulations

We consider the following simulation. We simulate 10,000 persons with abilities drawn from a normal distribution $\mathcal{N}(0.5, 1)$ and an item bank of 100 items with parameters from $\mathcal{U}(-2, 2)$. Every person generates a response pattern for 20 random items, the responses on the rest of the available items are considered missing.

To demonstrate the marginal DA described in 5.3.2, we perform a two step procedure. First, we impute the full marginal distributions, using uniform starting values $\mathcal{U}(0, 1)$ for both $\ln \lambda$ and $\ln b$. Second, we estimate the ERM on the augmented data matrix with 5 Gibbs iterations, which is a sufficient amount to ensure convergence of the parameter estimates, due to low autocorrelation (Maris et al., submitted).

Following this procedure, every iteration of the algorithm generates new marginals of the complete data set with new imputations for the missing data. On this augmented data set, the ERM parameters are estimated. We let this Gibbs data augmentation procedure run for 200 iterations, and consider the first 50 iterations as burn-in. In Figure 5.1, cumulative score distributions are plotted after 50, 75, 100, 125, 150, 175 and 200 iterations. Clearly, these distributions coincide with the reference distribution, which is the score distribution of the simulated complete data matrix without missing values. If the number of persons is reduced to 1,000 in this simulation, we obtain similar graphs with additional noise in the cumulative score distributions, as can be seen in Figure 5.2. Though the amount of autocorrelation is small for the parameter estimation on a complete data matrix (Maris et al., submitted), the use of DA on the sparse data matrix introduces additional autocorrelation. The amount of autocorrelation is clearly related to the fraction of missing information, and therefore to the proportion of missing responses in the data, which is simu-
lated to be .8. In Figure 5.3, the estimated item parameters $-\ln(b_t^i)$ of item $i$ have been plotted against their lag-1 estimates $-\ln(b_t^{i-1})$ for the 150 iterations after the burn-in period of 50 iterations, in the case where 10,000 persons are simulated. The correlation of this scatter plot is .71, which is quite high, for in the condition of 1,000 persons it is .81.

### 5.4.2 Sequential adaptive simulations

In this section, a simulation similar to the one in section 5.4.1 is presented with one key difference, we consider sequential data. We consider again an item bank composed of 100 items, with items and persons drawn from the same distributions as in the previous simulation in section 5.4.1, and again every person is simulated to respond to 20 randomly selected items.

We start the procedure with 10 persons, and in every iteration we perform data augmentation, estimate the ERM parameters using 5 Gibbs iterations, and add a new person with 20 new responses. In the following iteration, data augmentation is performed on the increased data matrix. The number of persons is allowed to grow to 2,500 in total.

In Figure 5.4, an estimated item parameter $\delta_i$ is presented, as it develops over the adding of data. Every iteration the parameter changes since all model parameters are re-estimated. The dashed lines indicates the $\sqrt{n}$ convergence
5.5. Application of dynamic estimation

5.5.1 Using data from an on-line practice environment

To illustrate the use of dynamic estimation as introduced in section 5.3.3 on data involving changes in the underlying parameters, we used data from the Math Garden, an on-line CAP environment for arithmetics discussed by Klinkenberg et al. (2011). An adaptation of the Elo rating system (ERS) (e.g., Elo, 1978; Batchelder & Bershad, 1979), including both response times (Maris & van der Maas, 2012) and parameter variance estimation (e.g., Klinkenberg et al., 2011; Glickman, 2001), is used to provide parameter estimates.

From the entire data set, we selected items from the tables of multiplication, 100 items in total. Each of these were posed as open questions, for which an on-screen or real keyboard could be used to enter the response, with a maximum time of 20 seconds. A group of 1,000 users born between September 2003 and October 2004, which frequently use the system, were selected for this application. Together they accounted for 552,248 responses between September the 3rd, 2010 and October the 30th, 2013. The number of responses given by pupils is skewed, ranging between 327 and 2,227 items, with a median of 458 items.
and a rounded mean of 552 items. The number of responses per item ranges from 2,065 to 8,626, with a median of 5,800 observations. The mean percentage of correct responses over the entire data set is 70%.

In a dynamic practice environment such as the Math Garden, with data being collected over a period of 3 years, both item and score parameters are expected to change, making older observations less suitable for estimating their current value. We chose to purge observed responses in our 1,000 persons by 100 items matrix after 50 days. Clearly, the augmented data is based on the available observations and on the model parameters, and therefore the proportion of persons with recent observations and the proportion of responses is important. In Figure 5.5, the available proportion of persons with recent responses out of the total 1,000 is graphed, and shown to be limited in 2011, to increase in 2012 and and decrease again in 2013. Also, we observe sharp decreases during summer vacations, which are displayed as vertical gray bars. The proportion of relevant responses out of the total set of 100,000 responses from the 100 items by 1,000 persons matrix is also plotted. The fraction of observations on the reduced set of persons is plotted in a dashed black line, and hovers about 20-30% through time.
5.5. Application of dynamic estimation

5.5.2 Functional shape of model parameters

Under the assumption of a ERM, the expected a posteriori (EAP) estimates for ability can be directly obtained from the score parameters as follows:

$$\frac{\lambda_{s+1}}{\lambda_s} = \frac{\mathcal{E}(\exp((s+1)\theta)|x = 0)}{\mathcal{E}(\exp(s\theta)|x = 0)} = \mathcal{E}(\exp(\theta)|x_+ = s)$$  \hspace{1cm} (5.6)

which is a direct result from the Dutch identity (Holland, 1990; Hessen, 2012), and derived in (Maris et al., submitted). Though the model allows for the $\lambda$ parameters to take any functional form, we minimally expect that these EAP estimates are increasing in sum score, i.e., that all item-rest regressions are positive.

Both the ln $\lambda$ and the log EAP parameters for each sum score $x_+$ are plotted in Figure 5.6, where both the ln $\lambda$ parameters and the log EAP parameters are plotted. One can clearly observe that the log EAP estimates are generally increasing over sum scores. Notably, an approximately quadratic relation is observed for the ln $\lambda$ parameters, of which a fitted quadratic curve is plotted as a dashed line. Though this quadratic shape is only plotted for one specific time point in Figure 5.6, a quadratic curve fits well on most days, since on average $R^2 = .996$. The parameters of the quadratic approximations are given in Figure 5.7. The parameter $\beta_1$, represented by the solid line, is increasing over time, which indicates that the item pool is generally becoming easier. Also, one can
Figure 5.5: Development of the proportion of recent observations, persons and fraction of observed information.

Figure 5.6: Log expected a posteriori (EAP) estimates (with linear approximation) and log $\lambda$ parameters for each score (with quadratic approximation) at April 1st, 2013.
observe how there is a clear seasonal trend in $\beta_2$, increasing after the summer vacations, and slowly decreasing after the Christmas vacations. In the section hereafter, section 5.5.3, Formula (5.10) shows that $\beta_2$ is a general item pool discrimination parameter, where an increase in the parameter indicates more discrimination, therefore less noise in the data.

Due to the quadratic shape $\ln \lambda x_+ = \beta_0 + \beta_1 x_+ + \beta_2 x_+^2$, the model in Formula 5.1 can be written as follows:

$$P(x|\mathbf{b}, \beta_0, \beta_1, \beta_2) = \frac{\exp\left(\sum_{i=1}^{m} \ln b_i x_i + \beta_0 + \beta_1 x_i + \beta_2 x_i^2\right)}{Z} \tag{5.7}$$

where $\beta_0, \beta_1$ and $\beta_2$ are the intercept, slope and quadratic term of the curve and $Z$ is the normalizing constant that makes the expression sum to one. Please note that the slope $\beta_1$ can be absorbed in the item parameters $\ln b$, and $\beta_0$ cancels.

If we rewrite the EAP estimates in Formula (5.6) using the quadratic expression of the ERM model in Formula (5.7), we get the following linear expression for the log EAP estimates:

$$\mathcal{E}(\exp(\theta)|x_+ = s) = \exp \left[ (\beta_1 - \beta_2) + 2\beta_2 s \right] \tag{5.8}$$

which is plotted in Figure 5.6 using a dashed line. Consequently, we have used
the quadratic relation of the score parameters to smooth the log EAP estimates. The interesting result is that under this specific ERM, the log EAP is nothing but a linear transformation of the score.

### 5.5.3 An identifiable ability distribution

If we further inspect the quadratic model in Formula (5.7), we can write the exponent of a quadratic term as an integral with an added parameter, e.g., $\theta$, using an idea of Emch and Knops (1970) and Kac (1968). Using the following well known identity:

$$\exp(\beta x^2) = \int_{-\infty}^{\infty} \frac{\exp(2\sqrt{\beta^2 x^2 + \theta^2})}{\sqrt{\pi}} d\theta$$  \hspace{1cm} (5.9)

which we can fill in in Formula (5.7), to obtain the following equation:

$$P(x|b, \beta) = \int_{-\infty}^{\infty} \frac{\exp\left(\sum_{i=1}^{m} \ln b_i x_i + \beta_0 + \beta_1 x_i + (2\sqrt{\beta^2 x_i + \theta^2})\right)}{Z\sqrt{\pi}} d\theta$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{m} \frac{\exp(x_i(\ln b_i + \beta_1 + 2\sqrt{\beta^2 \theta}))}{1 + \exp(\ln b_i + \beta_1 + 2\sqrt{\beta^2 \theta})}$$

$$\times \prod_{i=1}^{m} \frac{\exp(-\theta^2)}{Z\sqrt{\pi}} d\theta$$  \hspace{1cm} (5.10)

which is a regular marginal Rasch model (RM) with an identified ability distribution, a quite remarkable result. We estimated a general ERM with a single parameter for each sum score, and obtained an ERM with an identified ability distribution characterized by the item parameters $b$ and two additional parameters $\beta_1$ and $\beta_2$. It is possible to actually plot the population distribution characterized in Formula (5.10), which is shown in Figure 5.8, i.e. using the last 50 days of responses.

Clearly, the population distribution of this self-selected sample is bimodal with two well identifiable groups, separated by a vertical dashed line. The estimated item parameters $b$ at the same moment in time are displayed in the same metric as the population distribution, so one can observe the relative difficulty of these items for the low-performing group and the relative easiness of these items for the high-performing group. The set of items is split in two groups, where the *easy* item group are only the items involving multiplications with 1 and 10, and the *hard* item group consists of the rest of the items.

To evaluate that this bi-modal distribution is not an artifact of the dynamic Bayesian estimation procedure, we try to identify these two groups. There are
many ways to identify the two performance groups in Figure 5.8, we choose to not use model parameters directly, but to simply classify the persons according to two criteria, namely having responded to over 50% of the easy items, and having responded to over 50% of the hard items. Following the workings of the adaptive algorithm, one can loosely expect that persons who answer mostly easy items generally have a lower ability. The results are shown in Table 5.1. In this table, the number of persons with few responses on both easy and hard items is large, especially for the second moment in time. Given that there are some persons answering few question on this item set in general, this is expected. However, one can see that there are two large groups of persons at the first time point who make either many hard items or many easy items, a distinction that disappears at the second time point. This is consistent with the
idea that as pupils grow in ability, they are offered less items from the tables of multiplication, i.e., we expect the group answering few questions in general to increase and the more able group to disappear. The interpretation of Table 5.1 clearly has to be related to Figure 5.5, where the amount of pupils with recent observations is displayed.

5.5.4 Symmetric item development

The dynamic estimation technique using marginal data augmentation allows us to easily plot the development of proportion correct of persons or items over time. In Figure 5.9, the development of two symmetric items 4x9 and 9x4 is plotted. Clearly, one can see how the fraction of observed information in Figure 5.5 drives the amount of noise in Figure 5.9. Though we do not attempt to discuss any theory of learning the tables of multiplication, a few observations can be clearly made. First, the items become generally easier over time. Second, we see a clear change in item difficulty after the summer vacations. Third, the difference in difficulty between these items decreases over time, until the items are almost parallel in 2013. Similar observations can be made for other symmetric items in Figure 5.10 and 5.11. A difficulty in interpreting the development in these figures, is that the percentages are based on a changing sample of students. The sample changes because of the item selection algorithm, which prevents easier items begin administered to the more able students, causing
items to be made by specific subgroups of students. In addition, the sample is self-selected in that students themselves, their teacher, or parents, determine when practice takes place. If the best performing students practice the items in 2011, more regular students in 2012 and relatively weak students in 2013, than substantive differences in sub populations might cause parameters not to be comparable over time. The same can be happening during school vacations, where the occasional inversion in the item difficulties of symmetric pairs might be an indication of such a phenomena.

5.6 Discussion

In this paper, we have discussed the ERM under the equivalent groups design, as an extension to the work of Maris et al. (submitted). Three different data structures have been discussed, on which Bayesian estimation procedures were applied. First, a static data structure is discussed in section 5.3.1. Aspects of the equivalent groups ERM, such as the shape of the involved distributions, and how to sample from approximations, are discussed in the context of multiple test booklets. Second, a sequential data structure is discussed in section 5.3.2, were we consider data sets that increase, i.e., data is added during estimation. Typical applications that generate such data are found in the field of CAT, where every candidate generates a new and possibly unique set of
responses. It is illustrated how calibrations can be performed during data collection and simulations of this situation are provided in section 5.4.2, where the equivalent groups ERM is iteratively applied to an augmented data matrix. Third, a dynamic data extension is discussed in section 5.3.3. We again consider a data matrix involving observed and augmented data, where observed data is purged after a certain amount of time. Purged data is then augmented based on the other, more recent and therefore more relevant, observed data of the candidate. This method provides a data matrix which is always up-to-date, and on which iteratively the equivalent groups ERM can be estimated. Data involving possibly continually changing parameters, both of persons and items, such as CAL or CAP environments typically generate such dynamic data structures. An application using data from a large scale on-line arithmetic practice environment (Klinkenberg et al., 2011) is used to illustrate this method. A quadratic relation was found in the log score parameters of the equivalent groups ERM, which could be rewritten to a regular marginal RM involving an identified ability distribution parameterized by the item difficulties and two extra parameters.

We like to add a remark regarding the connection of this finding to network models. The specific ERM is a Curie-Weiss network model, common to the field of statistical mechanics (e.g., Ellis & Newman, 1978). The Curie-Weiss network model can be described as a network of nodes, in this case items, which are all interconnected. The relation is of interest because of at least two reasons.
First, in connection to the physical context in which the Curie-Weiss model is applied, one can estimate the temperature of the network. The temperature in this context is the degree of randomness in the item responses, or the general discrimination parameter, and one expects that the randomness decreases in education over time. Second, extensions to the Curie-Weiss model, such as the multi-dimensional Ising model, allow to simultaneous regard multiple dimensions in a single network. In CAP and CAL environments, and perhaps in education in general, we expect that at certain points in time different domains are represented by different dimensions, such as addition and multiplication. However, as a function of practice and learning, several dimensions are expected to collapse to form a more general arithmetic dimension, which is consistent with the findings by van der Maas et al. (2006). The relation of the marginal RM to network models from statistical mechanics might provide statistical tools to explore such hypotheses, which requires further research in this direction.

5.A Supplemental materials

5.A.1 Identification of the equivalent groups ERM

We discuss two issues in identification of the equivalent groups extended marginal Rasch model here. First, the relation of the score parameters $\lambda^1$ and $\lambda^2$ of booklet $x$ and $y$ to the complete set of score parameters $\lambda^*$. Second, the relation of the item parameters $b$ and $c$ in the just created booklets to the item parameters of the complete set $b^*$.

Identification of the score parameters

The relation between score parameters $\lambda^1$ and $\lambda^*$ will be described below, the relation between $\lambda^2$ and $\lambda^*$ follows easily. The fitted model to subset $x$ concerning only the score parameters of this subset $\lambda^1$ is:

$$P(x|b, \lambda^1, x_+) = \frac{\prod_{i=1}^{m} b_i^{x_i} \lambda^1_{x_+}}{\sum_{s=0}^{m} \gamma_s(b) \lambda^1_s}$$  \hspace{1cm} (5.11)

while the full model regarding the complete set of score parameters $\lambda^*$ and difficulty parameters $c$ of booklet $y$ is:

$$P(x|b, c, \lambda^*, x_+) = \frac{\prod_{i=1}^{m} b_i^{x_i} \sum_{t=0}^{n} \gamma_t(c) \lambda^*_{x_+ + t}}{\sum_{s=0}^{m} \gamma_s(b) \sum_{t=0}^{n} \gamma_t(c) \lambda^*_s + t}.$$  \hspace{1cm} (5.12)
We can substitute $\sum_{t=0}^{n} \left[ \gamma_{t}(c) \lambda_{s+t}^{*} \right]$ in the numerator and denominator of (5.12) with $\lambda^{1}$ from (5.11) if for all $s$:

$$\lambda^{1}_{s} = \sum_{t=0}^{n} [\gamma_{t}(c) \lambda_{s+t}^{*}] \quad (5.13)$$

and likewise concerning $\lambda^{2}$ for all $t$:

$$\lambda^{2}_{t} = \sum_{i=0}^{m} [\gamma_{i}(b) \lambda_{s+i}^{*}] \quad (5.14)$$

The vector of score parameters of the complete set $\lambda^{*}$ is identified if all its elements occur in one of the linear transformations $T^{1}$ and $T^{2}$:

$$\lambda^{1} = T^{1} \lambda^{*}$$
$$\lambda^{2} = T^{2} \lambda^{*} \quad (5.15)$$

which equals that the total matrix of transformations in (5.15) is of full column rank. The identification of the item parameters is discussed next.

**Identification of item difficulty parameters**

The identification of the difficulty parameters $b$ in booklet $x$ and $c$ in booklet $y$ in relation to the full set of item parameters $b^{*}$ is straightforward given the identification of the score parameters $\lambda^{1}$ and $\lambda^{2}$. The relation can be directly observed in (5.13) and (5.14), where $b$ and $c$ are related through the score parameters $\lambda^{*}$.

**5.A.2 Gibbs sampling in the static data case**

**Full conditional distributions for item parameters**

To set up a Gibbs sampler for the two booklet example, it is necessary to sample from the full conditional distributions of the item parameters. The full conditional distribution for an item parameter $b_{i}$ can be written as:

$$f(b_{i} | c, \lambda, x, y) \propto \frac{b_{i}^{x+i} \prod_{t=0}^{n} \left[ \sum_{s=0}^{m} \gamma_{s}(b) \lambda_{s+t}^{*} \right]^{r_{t}}}{\left[ \sum_{u=0}^{m+n} \gamma_{u}(b, c) \lambda_{u} \right]^{m+n+1}}. \quad (5.16)$$

Since it is not straightforward to sample from this distribution, we rewrite the elementary symmetric functions $\gamma$ in (5.16) as follows (e.g., Verhelst et al., 1984;
and we can rewrite $\gamma_u(b,c)$ likewise. Since these summations are linear in $b_i$, we can write (5.16) as:

$$f(b_i|\ldots) \propto\frac{b_i^{x+i} \prod_{t=0}^{n} (1 + k_t b_i)^{n_t^p}}{(1 + k b_i)^{m^p + n^p}}$$  (5.18)

with

$$k_t = \frac{\sum_{s=0}^{m} \gamma_s - 1(b(i)) \lambda_{s+t}}{\sum_{s=0}^{m} \gamma_s(b(i)) \lambda_{s+t}}$$

and

$$k = \frac{\sum_{u=0}^{m+n} \gamma_{u-1}(b(i),c) \lambda_u}{\sum_{u=0}^{m+n} \gamma_u(b(i),c) \lambda_u}.$$  (5.19)

With a change of variables $\delta_i = -\ln b_i$, we find that the target is log concave:

$$\ln f(\delta_i|\ldots) = -(x_{+i} + 1)\delta_i + \sum_{t=0}^{n} \left[n_t^p \ln \left(1 + k_t e^{-\delta_i}\right)\right]$$

$$- (m^p + n^p) \ln(1 + ke^{-\delta_i})$$  (5.20)

Furthermore, the tails of the log-target are readily found to be linear.

**Scaled beta prime proposal**  We choose as a log proposal distribution a scaled beta prime distribution:

$$\ln g(\delta_i|b^{(i)},c,\lambda,x,y;\alpha,\beta,\eta) = -A\delta_i - B \ln \left(1 + Ce^{-\delta_i}\right)$$  (5.21)

where

$$A = x_{+i} + \alpha_i$$

$$B = m^p$$

$$C = e^{\delta_i} \frac{D + A}{B - D - A}$$  (5.22)

$$D = \frac{d}{d\delta_i} \ln f(\delta_i|\ldots)$$

which guarantees that the tails are linear with an equal slope and the first derivatives of the target and the proposal distribution are equal at the support
Figure 5.12: The log target density and the log proposal density of item parameter $\delta_i$ with $\hat{\delta}_i = 2$.

One can see in Figure 5.12 how the proposal density for items (dashed line) approximates the target density (solid line) well. Since the proposal distribution is not exactly equal to the target distribution, an additional Metropolis step is needed to ensure reversibility (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Chib & Greenberg, 1995). However, since the approximation is close, a rejection rarely occurs and the algorithm introduces little autocorrelation.

**Piecewise linear majorizing proposal** As an alternative to using scaled beta prime proposal distribution, it is also possible to use a majorizing function (de Leeuw, 2006) of the log-concave target distribution, so simple acceptance-rejection (AR) sampling can be used. A majorizing function guarantees that:

$$
\frac{g}{f} \begin{cases} 
\leq 1 & \forall x \\
= 1 & x = s_1, s_2, \ldots 
\end{cases}
$$

which liberates us from using proportionality constants in AR sampling, because the acceptance probability is equal to unity at the support points $s_1, s_2, \ldots$, and below at all other $x$. The log-concave shape of the target density together with the linearity for extreme $x$ allow us to construct a simple piece-wise linear majorizing function as proposal log-density $g(x)$, which is illustrated in Fig-
Figure 5.13: Shape of the log density and a piece-wise log linear approximation with two initial support points, and a third support point added.

This linear approximation of the log-density uses two support points $s_1$ and $s_2$ (van Ruitenburg, 2005), where the slopes of these lines and the log-density are equal. More support points can be added, where the next support point can be iteratively added by considering the intersection of the two lines in Figure 5.13, resulting in $s_3$. This approach is equivalent to the adaptive rejection sampling by Gilks and Wild (1992), which is also implemented in the Bayesian inference using Gibbs sampling (BUGS) system (Gilks, Thomas, & Spiegelhalter, 1994; Lunn, Spiegelhalter, Thomas, & Best, 2009).

The piecewise linear proposal density has four parameters, the two intercepts $s_{\alpha 1}$ and $s_{\alpha 2}$ and two slopes $s_{\beta 1}$ and $s_{\beta 2}$ for the two lines:

$$g(x|\alpha, \beta, \nu) = e^{s_{\alpha 1}x + s_{\beta 1}^2}(x \leq s_{2-1}) + e^{s_{\alpha 2}x + s_{\beta 2}^2}(x > s_{2-1})$$  \hspace{1cm} (5.24)

which is a mixture of truncated negative exponential distributions. Sampling from this piece-wise linear distribution requires to choose the number and placement of support points, and to determine the sampling probabilities of the envelopes between the support points. A straightforward AR sampler can be used to draw weighted samples from these envelopes.

Without providing a detailed account on this approach, we want to elaborate on a few points. First, concerning the adding of support points. Support points can be added freely, but we consider a particularly efficient manner to
introduce them. One suggestion is to add them as proposed in Figure (5.13). The next best support point here is located at the intersection of the two two adjacent lines. The slope of the added line determines which newly introduced intersection is a new candidate for a support point. Following, the support points approximate the log-concave function best in the area where the posterior density is the highest. Adding support points in this case not only allows for more efficient sampling, but also provide a Bayesian modal selection algorithm for the ERM, squeezing the mode between sequential support points. The number of support points is required is solely a concern of efficiency, where the introduction of more support points comes at a computational cost, but reduced the number of rejections in AR sampling.

A second remark is made on the connection between the piece-wise linear approximation and quadratic Bézier curves. If we consider the support points for the linear approximation as support points for Bézier curves, we obtain a smooth approximating curve to the target density. Such an approach could be an alternative to the use of scaled beta prime distribution in 5.A.2, with much more general applications.

**Full conditional distribution for score parameters**

To obtain the full conditional distribution for score parameters, we might change the summations of (5.5) as follows:

\[
\lambda_{\nu}^{\eta_{\nu}^{-1}} \prod_{k=0}^{m} \left[ \sum_{k=0}^{m+n} \gamma_{k-s}(c) \lambda_{k} \right]^{m_{p}} \prod_{l=0}^{n} \left[ \sum_{k=0}^{m+n} \gamma_{k-t}(b) \lambda_{k} \right]^{n_{p}} \\
\left[ \sum_{u=0}^{m+n} \gamma_{u}(b,c) \lambda_{u} \right]^{m_{p}+n_{p}}
\]

(5.25)

where the elementary symmetric functions only exist if \(0 < k-s < m\) and \(0 < k-t < n\), elsewhere \(\gamma_{k-s}(c) = 0\) and \(\gamma_{k-t}(b) = 0\). Though the shape of (5.25) resembles the shape of (5.16), it is not an exponential family model, and therefore it is not guaranteed to be log-concave. Since sampling from this distribution is not straightforward, we choose for a different approach to estimate \(\lambda\).

We choose for a DA approach (Albert, 1992; Maris & Maris, 2002; van Dyk & Meng, 2001), where we impute the sum scores on an unobserved booklet given the observed sum scores, i.e., we only do imputation of the marginal distribution. This DA allows us to obtain the complete marginal sum score distribution, which allows to sample \(\lambda\) from the target distribution of the completed data.

We can write out the conditional probabilities for a specific response pattern
5.A. Supplemental materials

$y$ given $x$ as follows:

$$P(y|x) = \frac{P(x, y)}{P(x)} = \frac{\prod_{i=1}^{m} b_i^x_i \prod_{j=1}^{n} c_j^y_j \lambda_{x+y}}{\prod_{i=1}^{m} b_i^x_i \sum_{t=0}^{\gamma_t(c)} \lambda_{x+t}}$$

(5.26)

and use the sufficient statistics $x_+$ and $y_+$ to obtain:

$$P(y_+|x_+) = \frac{\gamma_{y_+}(c) \lambda_{x+y}}{\sum_{t=0}^{\gamma_t(c)} \lambda_{x+t}}$$

(5.27)

and likewise

$$P(x_+|y_+) = \frac{\gamma_{x_+}(b) \lambda_{x+y}}{\sum_{s=0}^{\gamma_s(b)} \lambda_{x+s}}$$

(5.28)

The imputed marginals provide us with $\phi_v^p$, the number of persons with sum score $v$, based on both the observed and imputed marginals. Using this complete marginal, we can simply use the complete data target distribution discussed in (Maris et al., submitted):

$$f(\lambda_v|\lambda(v), \ldots) \propto \frac{\lambda_v^{0p}}{(1 + k\lambda_v)^{mp + np}}$$

(5.29)

which is a scaled beta-prime distribution.

5.A.3 Simulating response patterns in the sequential case

To impute scores on the unobserved items, we make use of the sufficiency of the sum score $y_+$ in the RM:

$$P(y|\theta) = \prod_{j=1}^{n} \frac{e^{y_j(\theta - \ln(c_j))}}{1 + e^{\theta - \ln(c_j)}}$$

$$\Downarrow$$

$$P(y|y_+ = t) = \frac{\prod_{j=0}^{n} e^{-y_j \ln(c_j)}}{\gamma_t(e^{-\ln(c_j)})} \forall \theta$$

(5.31)
Marsman (2014, chapter 3) show a method to efficiently sample from conditional distributions. The main idea is, that if we keep sampling from the distribution in (5.30) until we by chance simulate the sum score we need, then the simulated realization is distributed according to (5.31), regardless of which value of $\theta$ was used for the simulation. Since the probability with which the needed sum score is generated depends on the choice of $\theta$, we choose it to be the maximum likelihood (ML) estimate corresponding to the imputed sum score.

The following block of pseudo-code generates plausible response patterns with $y_+ = t$:

\[
\hat{\theta} : \mathcal{E}(Y_+ | \hat{\theta}) = y_+
\]

\begin{verbatim}
repeat
  $y \leftarrow P(y | \hat{\theta})$
unti\[y_+ = t$
\end{verbatim}

We provide a simulation to demonstrate the performance of sampling response patterns. Response patterns with a score $t = 70$ are generated using an implementation based on the pseudo-code above. A sequence of 100 items, with item difficulties ranging between -2 and 2, is used, and $\hat{\theta}$ is estimated with ML estimation for $t = 70$. The distribution of waiting time in iterations is illustrated in Figure 5.14, using 10,000 replications. The graph illustrates that the
number of iterations required is limited due to the estimation of $\hat{\theta}$. In addition, iterations are computationally light, i.e., the average waiting time is about 15 ms on a mainstream laptop from 2010 using a script written in R (R Core Team, 2013).

**References**


Bolsinova, M., & Maris, G. (submitted). Can IRT solve the missing data problem of test equating?


