the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 72–87

www.wileyonlinelibrary.com

Dynamic estimation in the extended marginal Rasch model with an application to mathematical computer-adaptive practice

Matthieu J.S. Brinkhuis[1]* and Gunter Maris[2]
[1]Utrecht University, the Netherlands
[2]ACTNext, Iowa City, Iowa, USA

We introduce a general response model that allows for several simple restrictions, resulting in other models such as the extended Rasch model. For the extended Rasch model, a dynamic Bayesian estimation procedure is provided, which is able to deal with data sets that change over time, and possibly include many missing values. To ensure comparability over time, a data augmentation method is used, which provides an augmented person-by-item data matrix and reproduces the sufficient statistics of the complete data matrix. Hence, longitudinal comparisons can be easily made based on simple summaries, such as proportion correct, sum score, etc. As an illustration of the method, an example is provided using data from a computer-adaptive practice mathematical environment.

1. Introduction

As Savi, van der Maas, and Maris (2015) point out (online) learning requires, first, a detailed description of what a student can and cannot do and, second, what a student should do/learn next. We focus on their first point, describing a system that determines a student's current position on an educational map, a metaphor also used by Wainer (2000, p. xi).

In educational measurement, the rise of computer-adaptive learning (CAL) or computer-adaptive practice (CAP) applications (e.g., Brusilovsky, 2001; Eggen, 2012; Klinkenberg, Straatemeier, & van der Maas, 2011; Wauters, Desmet, & Van den Noortgate, 2010) underlines the importance of these points. In CAL and CAP, students respond frequently (e.g., daily) 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 statistical properties of both persons and items are expected to change during the course of measurement, as learning is the actual goal of such environments. The possibility of providing feedback directly to the learners, teachers or parents adds to the expectancy of changing parameters in applying models on such data. Hence, such applications require some sort of educational positioning to track developments of both items and students on an educational map.

*Correspondence should be addressed to Matthieu J. S. Brinkhuis, Princetonplein 5, 3584 CC Utrecht, The Netherlands (email: m.j.s.brinkhuis@uu.nl).

DOI:10.1111/bmsp.12157

An educational map can be regarded as a, possibly large, collection of items on which a position has to be determined. Such a collection can be regarded as a market-basket approach, where the set of items, the market basket, equals the subject domain (Mislevy, 1998, variation 3, p. 56).

Because students cannot respond frequently to all items in the construct, a possibly large part of the data will be missing. Zwitser, Glaser, and Maris (2016) develop a similar perspective for educational surveys, where development of cohorts of learners over administrations of the survey are evaluated on the basis of how simple summary statistics (over the market-basket) evolve over administrations of the survey.

We introduce a flexible and general response model that allows us to track (e.g., Brinkhuis, 2014; Brinkhuis & Maris, 2008) the summary statistics of a hypothetical complete data set over time, from incomplete data. We consider various special cases that smooth[1] the model parameters, and demonstrate how relevant properties of the hypothetical complete data (which correspond to the smoothed parameters) can be tracked over time. To deal with the missing data, data augmentation (DA) is used to complete the data matrix. Having obtained an augmented person-by-item data matrix, we can simply look at the distribution of, for example, sum scores to track the development of the student population, and to calculate the column means to track item difficulties. We illustrate our approach with data from an online practice environment.

2. Methods

The methodology introduced in this section is structured as follows. First, we introduce a flexible and general exponential family response model. Second, we discuss several possible restrictions and smoothers on this general model. Third, we demonstrate how Bayesian inference using a DA approach can be implemented for such models.

2.1. A general response model

When introducing a flexible response model, a couple of sufficient statistics are of interest. In a regular person-by-item matrix, we are interested in the distribution of row sums, the column totals and the item total regressions. The latter is of special interest as it is a natural means to evaluate the performance of the model.

The exponential family model, with these as sufficient statistics, is the general response model discussed by Andersen (1973, p. 127) in the context of goodness-of-fit tests for the Rasch model (RM):

$$P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{m} b_{ix_+}^{x_i}\lambda_{x_+}}{\sum_{s=0}^{m}\gamma_s(\boldsymbol{b}_s)\lambda_s}. \tag{1}$$

In equation (1), the probability to obtain response vector $\boldsymbol{x}$ depends on item difficulty parameters $\boldsymbol{b}_s$, the sum score parameters $\boldsymbol{\lambda}$ and the sum score x_+. The number of items and

[1] Smoothers in this context refer to restricting parameters, an approach clarified in Section 2 and applied in Section 3.2.4; note that this is distinct from smoothing parameters over time in a time-series context.

the maximum score are denoted by m, i is an item index and s is a score index. The denominator[2] makes the function a probability distribution, given all parameters are non-negative.

Obviously, the model in equation (1) is very general, with $(m+1)\cdot m$ parameters. It specifically allows for item parameters $\boldsymbol{b}_{x_+}$ to differ by an arbitrary amount between sum scores x_+. Clearly, contexts are available in which sufficient data are available to estimate this model but, for other applications, restrictions might be useful to smooth parameters. Such restrictions will be introduced hereafter.

2.2. Smoothing

One of the interesting features of equation (1) is that several restrictions, or smoothers, on its parameters are possible, resulting in other well-known models. For example, it is interesting to restrict the difficulty parameters. The model in equation (1) allows these to differ in an arbitrary amount over sum scores s, although it would be unlikely that these are unrelated over s. A first restriction we consider is to have item difficulties independent of the sum score, $b_{is} = b_i$, such that the number of correct responses to item i is the sufficient statistic for b_i. In doing so, we obtain a marginal RM. Specifically, this model can be recognized as the extended marginal Rasch model (ERM) introduced by Tjur (1982), but also see Cressie and Holland (1983) and Maris, Bechger, and San Martín (2015):

$$\begin{aligned}\frac{\prod_{i=1}^{m} b_i^{x_i}\lambda_{x_+}}{\sum_{s=0}^{m}\gamma_s(\boldsymbol{b})\lambda_s} &= \frac{\prod_{i=1}^{m} b_i^{x_i}}{\gamma_{x_+}(\boldsymbol{b})}\frac{\gamma_{x_+}(\boldsymbol{b})\lambda_{x_+}}{\sum_{s=0}^{m}\gamma_s(\boldsymbol{b})\lambda_s}\\ &= \frac{\prod_{i=1}^{m} b_i^{x_i}}{\gamma_{x_+}(\boldsymbol{b})}\pi_{x_+}.\end{aligned} \tag{2}$$

However, it is easy to be more flexible than the ERM, such as allowing for some intercept and slope in $\log \boldsymbol{b}$ over s

$$\log b_{is} = \log b_i + \log c_i s. \tag{3}$$

Such a log-linear restriction restricts equation (1) to the flexible interaction model described by Haberman (2007, equation 13.5), which has the following form in our parametrization:

$$P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{m}\left(b_i c_i^{x_+-1}\right)^{x_i}\lambda_{x_+}}{\sum_{s=0}^{m}\gamma_s(\boldsymbol{b}\boldsymbol{c}^{s-1})\lambda_s}. \tag{4}$$

So far, we have restricted item difficulty parameters in the general model in equation (1) in two ways. In addition, restrictions can be applied on the score parameters $\boldsymbol{\lambda}$. For example, Cressie and Holland (1983) suggest imposing moment inequalities, which are a necessary prerequisite for the ERM to be a marginal RM. However, if we are

[2] The denominator sums over scores s and uses elementary symmetric functions $\gamma_s(\boldsymbol{b}_s)$ (Baker & Harwell, 1996; Verhelst, Glas, & van der Sluis, 1984) of order s of the vector $\boldsymbol{b}_s$, which are defined to be zero if $s < 0$ or if $s > m$.

interested in the mean and variance of the score distribution, by fitting the log score parameters with a quadratic function, we obtain exactly this

$$\log \lambda_s = \beta_0 + \beta_1 s + \beta_2 s^2, \tag{5}$$

with $\sum_p x_{p+}$ as sufficient statistic for β_1 and $\sum_p x_{p+}^2$ for β_2.

The model with the mean and variance of the score distribution along with item total scores as sufficient statistics is known as the Curie–Weiss network model in the field of statistical mechanics (e.g., Ellis & Newman, 1978). Connecting psychometric measurement models to network models from statistical mechanics has become a very fruitful area of active research (see, e.g., Epskamp, Maris, Waldorp, & Borsboom, 2016; Kruis & Maris, 2016; Marsman, Maris, Bechger, & Glas, 2015).[3] Following this literature, the relation between the Curie–Weiss network model and a marginal item response theory model derives from the following well-known identity:

$$\exp(\beta_2 x_+^2) = \int_{-\infty}^{\infty} \frac{\exp(2\sqrt{\beta_2} x_+ \theta - \theta^2)}{\sqrt{\pi}} d\theta. \tag{6}$$

This allows us to rewrite the Curie–Weiss model as:

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{b}, \beta_0, \beta_1, \beta_2) &= \int_{-\infty}^{\infty} \frac{\exp(\sum_{i=1}^m \log b_i x_i + \beta_0 + \beta_1 x_+ + (2\sqrt{\beta_2} x_+ \theta - \theta^2))}{Z\sqrt{\pi}} d\theta \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^m \frac{\exp(x_i(\log b_i + \beta_1 + 2\sqrt{\beta_2}\theta))}{1 + \exp(\log b_i + \beta_1 + 2\sqrt{\beta_2}\theta)} \\ &\quad \times \prod_{i=1}^m (1 + \exp(\log b_i + \beta_1 + 2\sqrt{\beta_2}\theta)) \frac{\exp(-\theta^2)}{Z\sqrt{\pi}\exp(-\beta_0}} d\theta, \end{aligned} \tag{7}$$

where Z is the normalization constant, β_0 cancels and β_1 can be absorbed in the item parameters. We recognize a regular marginal RM with an identified ability distribution. That is, an ERM with the mean and variance of the score distribution as sufficient statistics is a marginal RM.

2.3. Estimation

It is not simple to perform statistical inference for models, such as those discussed above, from data that are massively incomplete. We propose to make use of DA (e.g., Tanner & Wong, 1987) to overcome the missing data problem in combination with a Bayesian estimation procedure for the complete(d) data. For a discussion on ignorability in adaptive contexts, we refer to Mislevy (1998). To illustrate the approach, we focus on the ERM, for which Maris *et al.* (2015) have already developed a Bayesian estimation procedure using

[3] These authors all build on work by Emch and Knops (1970) and Kac (1968) in the statistical mechanics literature.

the Gibbs sampler. The ERM serves not only as a proof of concept but, as we will see later, it is a very robust yet simple model.

In extending the ERM to support incomplete data, we adapt our notation to support two non-overlapping sets of items, or booklets:

$$P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{b},\boldsymbol{c},\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{m} b_i^{x_i} \prod_{j=1}^{n} c_j^{y_j} \lambda_{x_+ + y_+}}{\sum_{u=0}^{m+n} \gamma_u(\boldsymbol{b},\boldsymbol{c})\lambda_u} \tag{8}$$

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

In a two-booklet example, we find two equations for the probabilities of observing responses in each of the two booklets: for booklet x,

$$P(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{c},\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{m} b_i^{x_i} \sum_{t=0}^{n} \left[\gamma_t(\boldsymbol{c})\lambda_{x_+ + t}\right]}{\sum_{u=0}^{m+n} \gamma_u(\boldsymbol{b},\boldsymbol{c})\lambda_u}; \tag{9}$$

and for booklet y,

$$P(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{c},\boldsymbol{\lambda}) = \frac{\prod_{j=1}^{n} c_j^{y_j} \sum_{s=0}^{m} \left[\gamma_s(\boldsymbol{b})\lambda_{y_+ + s}\right]}{\sum_{u=0}^{m+n} \gamma_u(\boldsymbol{b},\boldsymbol{c})\lambda_u}. \tag{10}$$

We see that, for both booklets, the implied model in equations (9) and (10) is again an ERM, with score parameter $\boldsymbol{\lambda}^1 = \sum_{t=0}^{n} \gamma_t(\boldsymbol{c})\lambda_{x_+ + t}$ and $\boldsymbol{\lambda}^2 = \sum_{s=0}^{m} \gamma_t(\boldsymbol{b})\lambda_{y_+ + s}$. Hence, the ERM is closed under marginalization.

Similarly, the ERM is closed under conditioning. For the conditional distribution of x given y, we find that

$$P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{m} b_i^{x_i} \lambda_{x_+ + y_+}}{\sum_{s=0}^{m} \gamma_s(\boldsymbol{b})\lambda_{s+y_+}}. \tag{11}$$

This conditional distribution is what we need to augment the observed data for learners that responded to booklet y with their missing responses on booklet x. Obviously, the conditional distribution of booklet y given booklet x has the same form. An algorithm for simulating from these conditional distributions is provided in the Appendix S1, where we also deal with the problem of parameter identifiability.

To summarize the approach, in every iteration of our DA–Gibbs sampler, we impute missing data according to the conditional distribution of missing responses conditionally on observed responses derived from an ERM for complete data, and we update the parameters of the ERM for complete data using the Gibbs sampler of Maris *et al.* (2015).

3. Results

3.1. Reconstructing sufficient statistics from incomplete data

To illustrate some of the properties of the DA procedure, we provide a simple example. We are specifically interested in demonstrating that from a data matrix involving a lot of missing data, we can reproduce the score distribution and the item proportions correctly, even when the statistical model does not fit.

The data we use for this illustration originate from a 2012 Dutch national test administration at the end of primary education, designed to advise on the suitability of secondary education tracks. The data matrix involves 200 items, which are constructed to measure a number of distinct abilities including mathematics, language and general study skills, and a total of 144,708 pupils. Because the whole data set is observed, we remove 80% of the data and compare the score distribution of the full data set with the augmented score distribution obtained by the algorithm, and we compare the item difficulties between the full and limited set. The algorithm was allowed to run a large amount of iterations (i.e., 500) to ensure convergence. This took about an hour on a mainstream laptop, using R-code (R Core Team, 2015) with some compiled functions.

As can be seen in Figures 1 and 2, with just 20% observed data, both the column means and the score distribution are conserved nearly perfectly. Correlations between parameter estimates on the augmented data and parameter estimates on the full data reach 1.000 for item difficulties and 0.999 for score parameters. The amount of autocorrelation is quite low for estimation on complete data, the average lag-1 autocorrelation on the difficulty parameters is about 0.096 in this example, which is expected from the work of Maris *et al.* (2015). As a result of the missing data, the average amount of autocorrelation increases to an average of 0.818, and is related to the fraction of missing information, a result described by Liu, Wong, and Kong (1995, lemma 3.2, p. 31). Repeating the procedure with a mere 10% of observed data introduces additional autocorrelation on the parameters, increasing it to 0.927. The correlation between parameters estimated on the full and reduced data matrix remain high: 0.999 for item difficulties and 0.998 for score parameters.

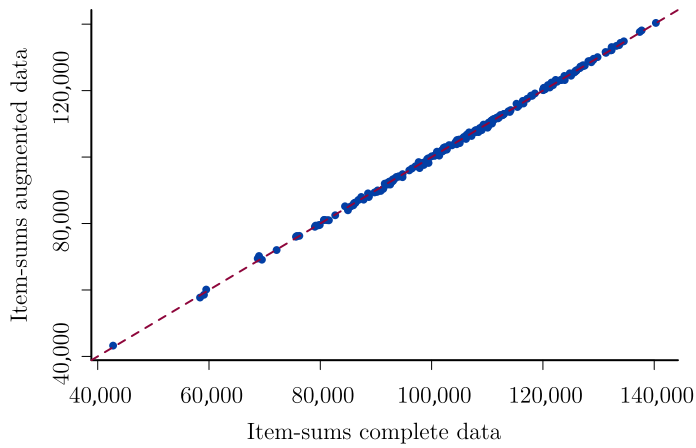

Figure 1. Scatter plot of the item difficulties based on the column sums of the complete data verses the column sums on the augmented data based on 20% of the observed data. Item sums are recovered very well, correlating to 0.9997. [Colour figure can be viewed at wileyonlinelibrary.com]

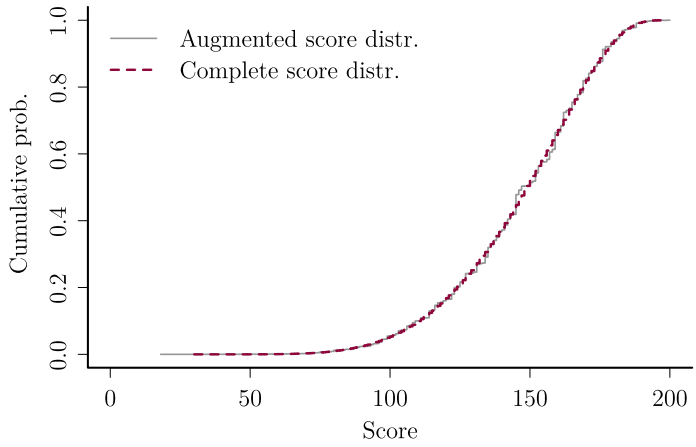

Figure 2. Cumulative distributions of the sum scores based on the complete data and of sum scores based on DA using 20% observed data. Both cumulative distributions nearly coincide, indicating a good recovery of the score distribution. [Colour figure can be viewed at wileyonlinelibrary.com]

Note that the high correlation between observed and augmented marginals is the case even though a RM is not expected to fit on this data set, which involves several constructs. The DA approach using the ERM successfully conserves its sufficient statistics, namely the column sums and the score distribution. Though not worked out in detail here, this follows the work of Jaynes (1982) concerning the relation between the sufficient statistics in exponential family distribution and Lagrange multipliers. The robustness of this approach in recapturing column means and row sums in a situation with a small percentage of observed data, and doubts on model fit, make this approach attractive for the use of dynamic estimation.

3.2. Using data from an online practice environment

3.2.1. The Math Garden

To illustrate the recapturing of column means and row sums in data involving changes in the underlying parameters, we use data from the Math Garden, an online CAP environment for arithmetic practice (Brinkhuis *et al.*, 2018). The Math Garden is adaptive in that it continually estimates item difficulties and pupil abilities, and adaptively selects new practice items based on someone's current ability estimate.

From the entire data set, we selected items from the tables of multiplication, a 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 response time of 20 s. A group of 1,000 users born between September 2003 and October 2004, who frequently use the system, was selected for this application. Together, they accounted for 552,248 responses between 3 September 2010 and 30 October 2013, a period of about 3 years. 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%, which is about the same as the aim of the adaptive item selection algorithm (Jansen *et al.*, 2013).

3.2.2. A stream of data slices

As we are interested in recapturing column means and row sums, the data are organized into slices of so-called wide data matrices. More specifically, the long-format data of a certain window (i.e., 50 days) are reshaped into a wide format, on which DA can be used to obtain the item scores and the distribution of sum scores. If such data slices are constructed daily, each on a window of the last 50 days, the development of the marginals can be tracked, as a kind of moving average over the (incomplete) slices.

The scheme in Table 1 illustrates this adding of new data and discarding of old data. The result is that the observed data set in the data matrix $\boldsymbol{X}_t$ is allowed to change continually, and hence the model parameters too.

The amount of data available to the estimation procedure can grow when new observations become available, and shrink as older data are discarded. With the possibility to discard 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, etc.). In doing so, we allow for a dynamically changing data set, where data are added and discarded as parameters are estimated. Because the estimates adapt to changes in the data set, one should interpret results regarding the relevant population, as this can change over time due to effects such as attrition, self-selection, etc. In our current implementation, sequential observations on a single person–item combination are not taken into account; the last response simply overwrites the previous response.

In Figure 3, the available proportion of individuals with recent responses out of the total group of 1,000 is plotted, and is shown to be limited in 2011, to increase in 2012 and to decrease again in 2013. Summer and winter vacations are displayed as vertical grey bars, where the former show a decrease in responses. 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 as a dashed black line, and hovers about 20–30% through time.

3.2.3. Data augmentation

With the help of DA to complete the matrix $\boldsymbol{X}_t$, we can easily obtain the distribution of person scores and the correct proportion of items, and we can track their development over time for each of the overlapping data slices. Questions regarding identification only deal with the identification of the model at a single time point t, as no model parameters are used to go from $\boldsymbol{X}_{t-1}$ to $\boldsymbol{X}_t$ – only the marginals of the completed data matrix are of interest. Details on model identification within a data slice are provided in the Appendix S1.

In what follows, we consider, first, the development of the recaptured score distribution and, second, the development of item difficulty, over time.

3.2.4. Score development

We have shown that the ERM reduces to a marginal RM with a particular distribution for ability, if the λ_s parameters are a log-quadratic function of the scores. From the Dutch identity (Hessen, 2012; Holland, 1990), and derived by Maris *et al.* (2015), it follows that:

Table 1. Scheme to illustrate the sequence of data matrices. In the center row, the current data matrix $\boldsymbol{X_t}$ is filled with data from the previous data matrix $\boldsymbol{X_{t-1}}$ and with new data $\boldsymbol{X_t^{\text{new}}}$. Observations deemed irrelevant for the current estimation are discarded, and hence they are overwritten by augmented data

$\boldsymbol{X}_0^{\text{new}}$		$\boldsymbol{X}_1^{\text{new}}$		$\boldsymbol{X}_2^{\text{new}}$		$\boldsymbol{X}_3^{\text{new}}$		
$\downarrow$		$\downarrow$		$\downarrow$		$\downarrow$		
$\boldsymbol{X}_0$	$\rightarrow$	$\boldsymbol{X}_1$	$\rightarrow$	$\boldsymbol{X}_2$	$\rightarrow$	$\boldsymbol{X}_3$	$\rightarrow$	$\ldots$
		$\downarrow$		$\downarrow$		$\downarrow$		
		$\boldsymbol{X}_1^{\text{discarded}}$		$\boldsymbol{X}_2^{\text{discarded}}$		$\boldsymbol{X}_3^{\text{discarded}}$		

$$\frac{\lambda_{s+1}}{\lambda_s} = \frac{\mathcal{E}(\exp((s+1)\theta)|\boldsymbol{x}=0)}{\mathcal{E}(\exp(s\theta)|\boldsymbol{x}=0)} = \mathcal{E}(\exp(\theta)|x_+ = s). \tag{12}$$

If we rewrite the EAP estimates in equation (12) using the quadratic expression of the ERM model in equation (5), we find the following log-linear expression for the log EAP estimates:

$$\mathcal{E}(\exp(\theta)|x_+ = s) = \exp[(\beta_1 - \beta_2) + 2\beta_2 s]. \tag{13}$$

The interesting result is that under this specific ERM, the log EAP is nothing but a linear transformation of the score.

Both the $\log \lambda$ and the log EAP parameters for each sum score x_+ are plotted in Figure 4. We can clearly observe that the log EAP estimates are generally increasing over sum scores, yet are very noisy. The linear approximation smooths this, and provides a logical increase of EAPs over sum scores.

In addition, an approximately quadratic relation is observed for the $\log \lambda$ 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 4, a quadratic curve fits well on most days, as the mean $R^2 = 0.996$ for the quadratic approximation. The parameters of the quadratic approximations are given in Figure 5. The parameter β_1, represented by the solid line, is increasing over time, which indicates that the item pool is generally becoming easier. Also, we can see that there is a clear seasonal trend in β_2, increasing after the summer vacations, and slowly decreasing after the Christmas vacations. Further research might be conducted by fitting models over time points, or by testing the fit of the quadratic model against the unconstrained model. We leave these options for future research.

Equation (7) shows that β_2 is a general item pool discrimination parameter, where an increase in the parameter indicates more discrimination, and therefore less noise in the data. We estimated a general ERM with a single parameter for each sum score, and we obtained an ERM with an identified ability distribution characterized by the item parameters $\boldsymbol{b}$ and two additional parameters β_1 and β_2. See San Martín and Rolin (2013) and San Martín, Rolin, and Castro (2013) for related work on the identifiability of such

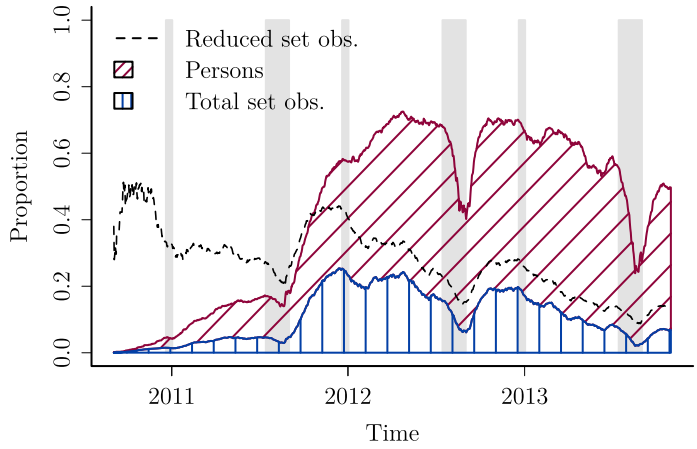

Figure 3. Development of the amount of available data. The development of the proportion of the 1,000 persons in the data is presented (diagonally striped area), together with the percentage of observations in the 1,000 persons by 100 items data matrix (vertically striped area). Disregarding persons without observations, the proportion of observations of a reduced n persons by 100 items data matrix is also presented (dashed line). [Colour figure can be viewed at wileyonlinelibrary.com]

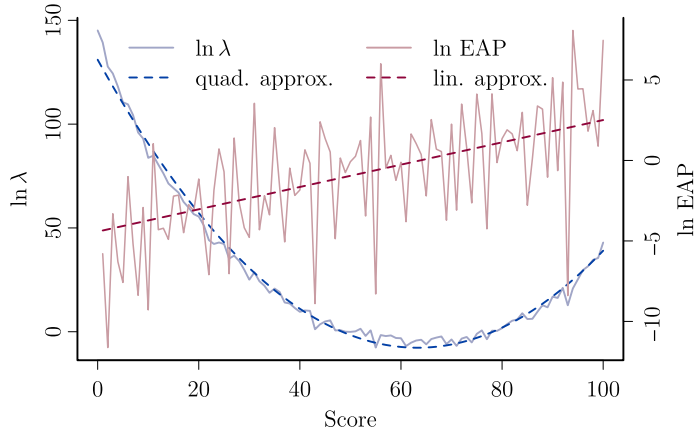

Figure 4. Log EAP estimates for each sum score with linear approximation and log λ parameters for each sum score with quadratic approximation on 1 April 2013. The quadratic approximation is quite close, and the linear approximation strongly smooths the log EAP estimates. [Colour figure can be viewed at wileyonlinelibrary.com]

models. It is possible to actually plot the population distribution characterized in equation (7), which is shown in Figure 6, using the last 50 days of responses.[4]

Clearly, the population distribution of this self-selected sample is bi-modal with two very identifiable groups, separated by a vertical dashed line. The estimated item

[4] The analyses were repeated using several different values for the number of days, ranging from 25 to 100. Results were found to be similar between these analyses.

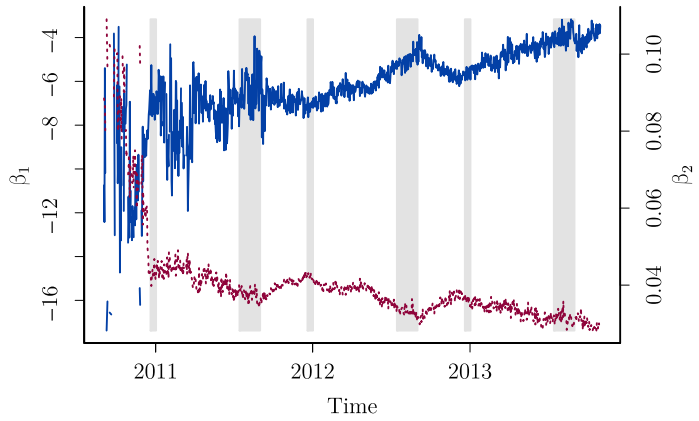

Figure 5. Development of the two quadratic approximation parameters to $\log \lambda$ over time, with parameters β_1 at the top of the graph (solid line) and β_2 at the bottom (dotted line). With the increase of β_1, the items in the item bank are generally becoming easier. [Colour figure can be viewed at wileyonlinelibrary.com]

parameters b at the same moment in time are displayed in the same metric as the population distribution, so we 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 into two groups, where the 'easy' item group consists only of 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 artefact 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 6; we choose not to use model parameters directly, but to simply classify the persons according to two criteria, that is, 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, we can loosely expect that persons who answer mostly easy items generally have a lower ability. The results are shown in Table 2.

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 questions on this item set in general, this is expected. However, we can see that there are two large groups of persons at the first time point who choose 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 fewer 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 2 clearly has to be related to Figure 3, where the amount of pupils with recent observations is displayed, and we have to consider that the composition of the population is also subject to change.

3.2.5. Symmetric item development

The dynamic estimation technique using marginal DA allows us to easily plot the development of proportion correct of persons or items over time, which we otherwise

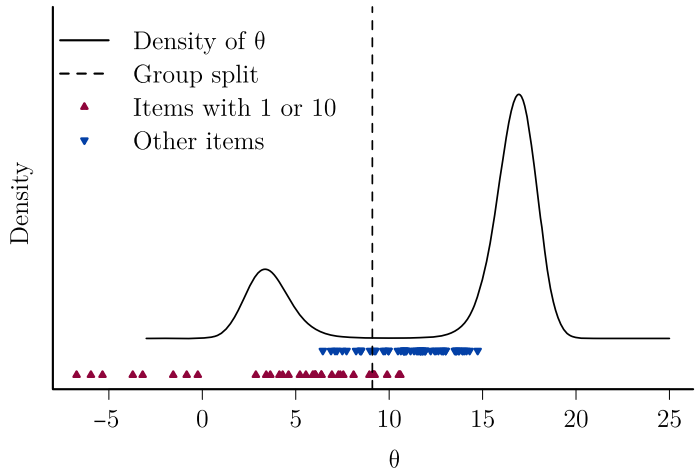

Figure 6. Bi-modal population distribution at 1 April 2013, separated by a vertical line (dashed). The 100 items are plotted at their parameter position, and are split between easy items with multiplication involving 1 and 10, and all others items. [Colour figure can be viewed at wileyonlinelibrary.com]

cannot easily observe directly. Figure 7 plots the development of two symmetric items 4×9 and 9×4.

Clearly, we can see how the fraction of observed information in Figure 3 drives the amount of noise in Figure 7. Though we do not attempt to discuss any theory of learning multiplication tables, 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 Figures 8 and 9.

A difficulty in interpreting the development in these figures is that the percentages are based on a changing sample of students including possible anomalies (Sosnovsky, Müter, Valkenier, Brinkhuis, & Hofman, 2018). The sample changes because of the item selection algorithm, which prevents easier items being administered to the more able students, causing items to be answered 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, then substantive differences in subpopulations might cause parameters to be incomparable over time.

Table 2. Identifying high and low performance groups according to the amount of responses (<50% and equal to or more than 50%) to easy and hard items at two time points

	1 April 2012		1 April 2013	
	Few hard	Many hard	Few hard	Many hard
Few easy	683	112	938	8
Many easy	156	49	46	8

The same might be happening during school vacations, where the occasional inversion in the item difficulties of symmetric pairs might be an indication of such a phenomenon. An alternative approach to detect differential development in symmetric item pairs can be found in Brinkhuis, Bakker, and Maris (2015).

4. Discussion

In this paper, we have introduced a general response model, which can be restricted to several well-known and useful models. As a proof of concept for educational positioning, the model is restricted to the ERM. Its use as a dynamic Bayesian estimation method to deal with CAP data streams is described.

CAP data streams are challenging because they are composed of large amounts of missing data, and because properties such as ability and item difficulty dynamically develop over time. In addition, data come in continually. We have provided a pragmatic solution by constructing wide data slices of observed data, to which new data are added continually and where data deemed too old to be relevant are removed. Discarded data are augmented based on the more recent observations. Thus, the resulting data matrix is composed of recent item responses and augmentations based on recent responses.

Despite dealing with data with large percentages of missing values, the proposed model and estimation procedure are able to reconstruct sufficient statistics such as row sums and column means, and to track their development over time. An important advantage of this approach is that our key outcomes (sum score distribution, proportion correct) are, at least in principle, directly observable, and hence can, at least in principle, be validated (as illustrated by our first example). Data involving possibly continually changing parameters, of both persons and items, such as CAL or CAP environments, typically generate such dynamic data structures. An application using data from a large-scale online arithmetic practice environment (Brinkhuis *et al.*, 2018) is used to illustrate this method.

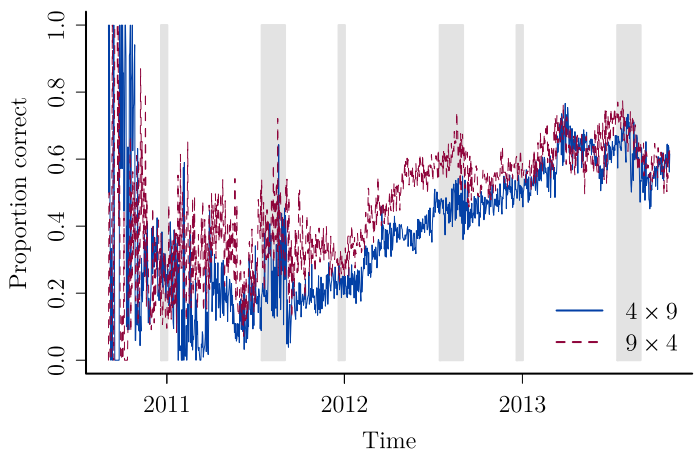

Figure 7. Development of item difficulty of items 4×9 and 9×4 in the correct proportion, with a positive overall trend. Note that 4×9 starts out more difficult and, after the summer break (grey horizontal bar) of 2012, they become parallel. [Colour figure can be viewed at wileyonlinelibrary.com]

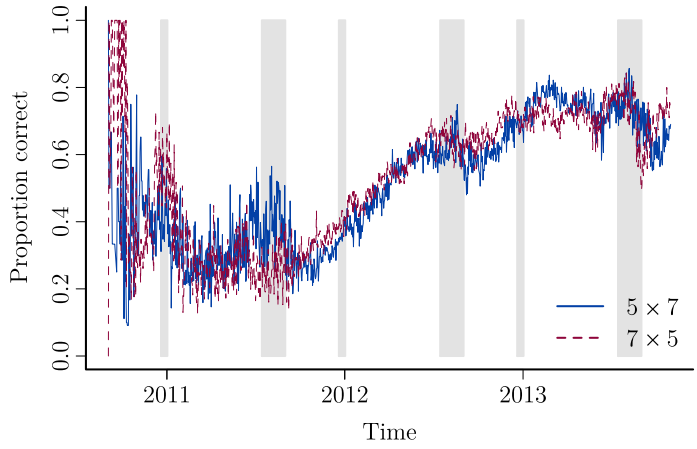

Figure 8. Development of item difficulty of items 5 × 7 and 7 × 5 in correct proportion. After 2012, the items are almost parallel in correct proportion, and seasonal trends are visible. [Colour figure can be viewed at wileyonlinelibrary.com]

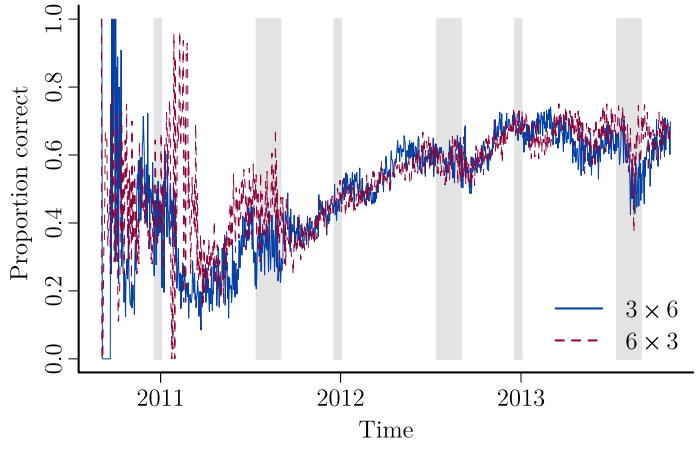

Figure 9. Development of item difficulty of item 3 × 6 and 6 × 3 in correct proportion. After 2012, the items are almost parallel in correct proportion, and seasonal trends are visible. [Colour figure can be viewed at wileyonlinelibrary.com]

There exists a mathematical relation between, a special case of, the simplest of psychometric measurement models (RM/ERM) and the simplest network model from statistical mechanics (the Curie–Weiss model), which extends readily to less trivial cases (Emch & Knops, 1970; Epskamp *et al.*, 2016; Kruis & Maris, 2016; Marsman *et al.*, 2015). As we have demonstrated with our second illustrative application, exactly this special case of the ERM is consistent with real data. This particular special case is of interest, as it allows for expressing the manifest probabilities in closed form, whilst at the same time (in contrast to the general ERM) being a true marginal RM. As far as we know, this is the only known instance of the marginal RM for which the manifest probabilities have a closed form. Moreover, as one readily finds, the posterior distribution of ability is normal, with a variance that does not depend on the items or on the particular responses.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research, grant number CI1-12-S037. We sincerely thank the anonymous reviewers and the editor for their extensive comments that improved this manuscript.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140. https://doi.org/10.1007/BF02291180

Baker, F. B., & Harwell, M. R. (1996). Computing elementary symmetric functions and their derivatives: A didactic. *Applied Psychological Measurement*, *20*, 169–192. https://doi.org/10.1177/014662169602000206

Brinkhuis, M. J. S. (2014). *Tracking educational progress* (Doctoral dissertation, University of Amsterdam). Retrieved from http://hdl.handle.net/11245/1.433219

Brinkhuis, M. J. S., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*, 319–338. https://doi.org/10.1111/jedm.12078

Brinkhuis, M. J. S., & Maris, G. (2008). Student monitoring using chess ratings. In P. H. C. Eilers (Ed.), *Proceedings of the 23rd international workshop on statistical modelling* (Vol. *23*, pp. 137–142). Utrecht, the Netherlands.

Brinkhuis, M. J. S., Savi, A. O., Coomans, F., Hofman, A. D., van der Maas, H. L. J., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics*, *5*, 29–46. https://doi.org/10.18608/jla.2018.52.3

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, *11*, 87–110. https://doi.org/10.1023/A:1011143116306

Cressie, N., & Holland, P. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*, 129–141. https://doi.org/10.1007/BF02314681

Eggen, T. J. H. M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 14–25). Enschede, the Netherlands: RCEC. Retrieved from http://doc.utwente.nl/80211/

Ellis, R. S., & Newman, C. M. (1978). The statistics of Curie-Weiss models. *Journal of Statistical Physics*, *19*, 149–161. https://doi.org/10.1007/BF01012508

Emch, G. G., & Knops, H. J. F. (1970). Pure thermodynamical phases as extremal KMS states. *Journal of Mathematical Physics*, *11*, 3008–3018. https://doi.org/10.1063/1.1665087

Epskamp, S., Maris, G. K., Waldorp, L. J., & Borsboom, D. (2016). Network psychometrics. In P. Irwing, D. Hughes & T. Booth (Eds.), *Handbook of psychometric testing*. New York, NY: Wiley.

Haberman, S. J. (2007). The interaction model. In von Davier M. & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Statistics for social and behavioral sciences* (Chapter 13, pp. 201–216). New York, NY: Springer. https://doi.org/10.1007/978-0-387-49839-3

Hessen, D. J. (2012). Fitting and testing conditional multinormal partial credit models. *Psychometrika*, *77*, 693–709. https://doi.org/10.1007/s11336-012-9277-1

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18. https://doi.org/10.1007/BF02294739

Jansen, B. R. J., Louwerse, J., Straatemeier, M., van der Ven, S. H. G., Klinkenberg, S., & van der Maas, H. L. J. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197. https://doi.org/10.1016/j.lindif.2012.12.014

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*, 939–952. https://doi.org/10.1109/PROC.1982.12425

Kac, M. (1968). Mathematical mechanism of phase transitions. In M. Chrétien, E. P. Gross & S. Deser (Eds.), *Brandeis University Summer Institute in theoretical physics, 1966* (Vol. *1*, pp. 242–305). New York, NY: Gordon and Breach.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*, 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*, 34175. https://doi.org/10.1038/srep34175

Liu, J. S., Wong, W. H., & Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 157–169. https://doi.org/10.1111/j.2517-6161.1995.tb02021.x

Maris, G., Bechger, T., & San Martín, E. (2015). A Gibbs sampler for the (extended) marginal Rasch model. *Psychometrika*, *80*, 859–879. https://doi.org/10.1007/s11336-015-9479-4

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Nature Scientific Reports*, *5*, 1–7. https://doi.org/10.1038/srep09050

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education*, *11*, 49–63. https://doi.org/10.1207/s15324818ame1101_3

R Core Team (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Core Team. Retrieved from http://www.R-project.org/

San Martín, E., & Rolin, J.-M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*, 116–130. https://doi.org/10.1016/j.jspi.2012.06.014

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika*, *78*, 341–379. https://doi.org/10.1007/s11336-013-9322-8

Savi, A. O., van der Maas, H. L. J., & Maris, G. K. J. (2015). Navigating massive open online courses. *Science*, *347*, 958. https://doi.org/10.1126/science.347.6225.958

Sosnovsky, S., Müter, L., Valkenier, M., Brinkhuis, M., & Hofman, A. (2018). Detection of student modelling anomalies. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink & M. Scheffel (Eds.), *Lifelong technology-enhanced learning* (pp. 531–536). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-319-98572-5_41

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540. https://doi.org/10.1080/01621459.1987.10478458

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23–30. http://www.jstor.org/stable/4615850

Verhelst, N. D., Glas, C. A. W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly*, *1*, 245–262.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*, 549–562. https://doi.org/10.1111/j.1365-2729.2010.00368.x

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2016). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*, 210–232. https://doi.org/10.1007/s11336-016-9543-8

Received 30 January 2014; revised version received 3 December 2018

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Identification and simulating response patterns for data augmentation.