**User Modeling and User-Adapted Interaction manuscript No.**
(will be inserted by the editor)

# Elo-based Learner Modeling for the Adaptive Practice of Facts

**Radek Pelánek · Jan Papoušek · Jiří Řihák · Vít Stanislav · Juraj Nižnan**

Received: date / Accepted: date

**Abstract** We investigate applications of learner modeling in a computerized adaptive system for practicing factual knowledge. We focus on areas where learners have widely varying degrees of prior knowledge. We propose a modular approach to the development of such adaptive practice systems: dissecting the system design into an estimation of prior knowledge, an estimation of current knowledge, and the construction of questions. We provide a detailed discussion of learner models for both estimation steps, including a novel use of the Elo rating system for learner modeling. We implemented the proposed approach in a system for practising geography facts; the system is widely used and allows us to perform evaluation of all three modules. We compare the predictive accuracy of different learner models, discuss insights gained from learner modeling, as well as the impact different variants of the system have on learners' engagement and learning.

**Keywords** Learner modeling · Computerized adaptive practice · Elo rating system · Model evaluation · Factual knowledge

## 1 Introduction

Online educational systems like Khan Academy, Duolingo, or Coursera are used by millions of learners. Such systems offer great potential for exploiting the possibilities of adaptive behavior, i.e., to provide learners with materials and tasks that are most useful to them. This potential is currently only partially realized, because the development of adaptive learning systems is complex, lengthy, and expensive.

The general motivation of our work is to make the development of such systems as automated as possible, particularly to enable systems to learn relevant aspects of educational domains from data so that domain experts can

Faculty of Informatics, Masaryk University Brno; `pelanek@fi.muni.cz`, Phone +420-549 49 6991, Fax +420-549 491 820

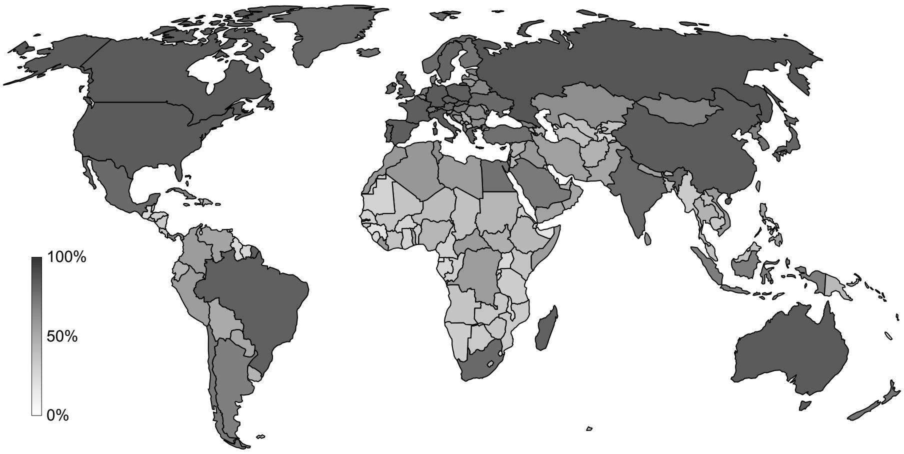

**Fig. 1** Map of the world colored by prior knowledge of countries. The shading corresponds to the estimated probability of a correct answer for an average user of `outlinemaps.org` (mostly Czech students).

focus on those parts of system development where their input is indispensable. This automation is especially important for developing systems for small target groups of learners, such as those that deal with specialized topics or languages spoken by relatively small numbers of people.

This work focuses on the development of adaptive systems for learning factual knowledge, i.e., for storing pieces of information in declarative memory. Using the terminology of the knowledge-learning-instruction framework (Koedinger et al, 2012), we focus on knowledge components that have a constant application condition and a constant response. We are particularly concerned with the learning of facts in areas where learners display great variation in their prior knowledge, e.g., geography, biology (flora and fauna), human anatomy, or foreign language vocabulary. To illustrate the usefulness of estimating prior knowledge, Fig. 1 depicts significant differences in prior knowledge of world countries.

The main contribution of this paper lies in how it integrates all of the steps necessary for the application of learner modeling in real adaptive educational systems, namely, methodical issues related to learner modeling, the evaluation of models, parameter fitting, and practical wide-scale applications. These issues have been studied before, but mostly in isolation. The integration of modeling in a real system forces us to consider practical aspects of learner modeling as well: we need to consider not just the predictive accuracy of models (the focus of most learner modeling papers), but also the computational efficiency and applicability of models in an online application.

Our approach is generic and can be applied to the adaptive practice of facts in any domain. We use a specific domain (geography) as our case study for which we developed a widely-used system (`outlinemaps.org`) and we used it to evaluate learner models and their impact.

We also provide several specific technical contributions:

- proposals for several novel learner modeling techniques or novel uses of models, particularly connected to the use of the Elo rating system (Elo, 1978) in the context of learner modeling,
- an evaluation of learner models over large scale historical data,
- interesting insights into the target domain and learner behavior,
- a proposal for and evaluation of techniques for adaptive question construction,
- an analysis of the relationship between question difficulty and learner motivation.

## 2 Related Work

To achieve effective adaptive learning in domain like geography it is necessary to address several interrelated issues, particularly the estimation of knowledge, modeling of learning, memory effects (spacing and forgetting), and question construction. These issues have been studied before, but separately and in different contexts.

Adaptation has been studied thoroughly in the context of *computerized adaptive testing* (CAT) with the use of item response theory (De Ayala, 2008). In CAT the primary goal is to determine the abilities of learners. Therefore, the focus is on precision and statistical guarantees. The research does not usually address learning since skills are not expected to change during a test, and motivation, which is typically extrinsic in the case of test taking. We focus on *computerized adaptive practice*. In this setting the primary goal is to improve learners' skills: the estimation of skills is a secondary goal, which helps to achieve the main one. Thus, we do not need to focus on statistical guarantees provided by the skill estimation as much as in CAT. On the other hand, the issues of learning, forgetting, and motivation are crucial for adaptive practice. An example of a typical computerized adaptive practice system is Math Garden (Klinkenberg et al, 2011), which focuses on practising basic arithmetical operations.

Adaptability in the context of learning is studied mainly in the area of *intelligent tutoring systems* (Vanlehn, 2006). These systems focus more on learning complex cognitive skills than on learning facts, e.g., mathematics (Koedinger and Corbett, 2006), physics (Schulze et al, 2000), or computational thinking (Basu et al, 2017). An important part of the research into intelligent tutoring systems includes issues like step-by-step solution monitoring, hints, scaffolding, and forms of feedback, which are issues not directly relevant to practising facts.

A fundamental part of all adaptive educational systems is *learner modeling* (Desmarais and Baker, 2012). A learner model provides an estimate of learners' knowledge based on their answers. The estimated knowledge is then used by other components of a system to adapt its behavior and provide feedback to learners. Two of the most popular approaches to learner modeling are

Bayesian knowledge tracing (Corbett and Anderson, 1994) and models based on a logistic function (which can be seen as extensions of the Rasch model from item response theory), e.g., Performance factor analysis (Pavlik et al, 2009). A lot of research focuses on the acquisition of skills while less attention is paid to prior knowledge and forgetting; exceptions include Pardos and Heffernan (2010); Qiu et al (2011). Learner modeling techniques most related to our approach are recent methods that integrate item response theory and knowledge tracing (González-Brenes et al, 2014; Khajah et al, 2014a,b). These methods can model both prior knowledge and learning and they do it in a principled way. However, they use algorithms that cannot be easily adapted for use in a realistic educational system (EM algorithm, Monte Carlo Markov Chain). We use methods based on the Elo rating system (Elo, 1978), which are more heuristic, but fast and easily applicable in an online setting. The Elo rating system was originally developed for rating chess players, and it has recently been adapted for use in educational systems (Klinkenberg et al, 2011; Pelánek, 2016; Wauters et al, 2011). We describe extensions of the Elo rating system related to learner models based on Bayesian networks (Conati et al, 2002; Käser et al, 2014; Millán et al, 2010).

We use the learner model to automatically construct suitable questions. Previous research has proposed many techniques for *automatic item generation* (Gierl and Haladyna, 2012), particularly using natural language processing techniques (Mitkov et al, 2006), ontologies, and domain models (Gierl et al, 2012). In contrast to this research we construct relatively simple multiple choice questions about factual knowledge, but we place greater focus on personalization (connecting the question construction to learner modeling).

The learning of facts is well studied in research on *memory*, e.g., in the study of spacing and forgetting effects (Pavlik and Anderson, 2005) and spaced repetition (Karpicke and Roediger, 2007). These studies are not, however, usually done in a realistic learning environment, but in a laboratory and in areas with little prior knowledge, e.g., learning arbitrary word lists, nonsense syllables, obscure facts, or Japanese vocabulary (Delaney et al, 2010; Pavlik and Anderson, 2005). Such an approach facilitates some interpretation of the experimental results, but the models developed so far are not easily applicable in educational settings where prior knowledge is an important factor. There are also many implementations of the spaced repetition principle using "flashcard software" (a well-known example is SuperMemo), but these implementations usually use scheduling algorithms with fixed ad-hoc parameters and do not try to learn from collected data (or only in a limited way). Spaced repetition was also studied specifically for geography (Zirkle and Ellis, 2010), but only in a simple setting.

Another important aspect of educational systems is *engagement*, which the adaptive system can influence for example by selecting suitably difficulty questions in order to aim at the flow state (Csikszentmihalyi, 1991). This is a typical general aim of adaptive systems, but the specification of adaptive behavior is usually based on the intuition of system developers without proper evaluation (Klinkenberg et al, 2011) or evaluated using only comparisons to a control group without any adjustments to the level of difficulty (Barla et al, 2010). The most relevant research is by Lomas et al (2013) who evaluated the "Inverted-U Hypothesis" by testing many variants of an educational game (number line estimation). However, they did not manage to find any U-shaped relationship between difficulty and engagement. For their study the relationship was a monotone function (simpler problems were more engaging). Explaining the results, they state that maybe they "never made the game easy enough" (Lomas et al, 2013). Our experiments are similar, the main difference being that we use a more realistic educational application. Another similar study was done using Math Garden software (Jansen et al, 2013). The authors compared three conditions (target success rate 60%, 75%, 90%) and showed that the easiest condition led to the best learning (mediated by a number of solved problems). Our results, in contrast, suggest that more difficult questions are better for learning facts.

An interesting historical perspective is provided by the comparison of our system with a 45 year old computer-assisted instruction system called Scholar (Carbonell, 1970), whose principles were demonstrated in the domain of South American geography. On one hand, the Scholar system was more ambitious than the current system in that it was capable of a mixed-initiative dialog in a natural language and incorporated many geography facts (not just names and locations as in the system we are presenting). The system was, however, much more difficult to develop and required time consuming knowledge engineering. The main conceptual difference of our system is the "learning from data" approach, which makes the development of educational systems simpler and more scalable.

This paper is based on previously published conference papers (Nižnan et al, 2015; Papoušek and Pelánek, 2015; Papoušek et al, 2014, 2015; Pelánek, 2015; Papoušek et al, 2016b,c). It provides a systematic integration of previously published results with updated evaluations and several additional results.

## 3 System Description

The basic functionality of the proposed architecture is simple: the system provides a series of questions about items and learners answer them. Since we are dealing with learning factual knowledge, the structure of questions is also simple (e.g., multiple-choice questions) and the feedback consists only of information about correctness and a provision of the correct answer after a mistake. The core of the system lies in estimating learners' knowledge and selecting suitable questions.

### 3.1 General Structure

We break down the design of an adaptive practice system for facts into three steps and treat each of them separately.

1. *Estimating prior knowledge.* The system estimates the probability that a learner $l$ knows an item $i$ before the first question about this item. This is based on the learner's previous answers and on other learners' answers to questions about the item.
2. *Estimating current knowledge.* The system estimates the probability that a learner $l$ knows an item $i$ based on the estimation of prior knowledge and a sequence of previous answers of the learner $l$ on questions about the item $i$.
3. *Question construction.* Constructing a suitable question for a learner is based on the estimate of their current knowledge and a recent history of answers. The question construction phase also includes the choice of distractors for multiple choice questions.

Each of these issues is described and evaluated in a single section. While treating each of these steps independently is a useful simplification, inasmuch as it makes the development of systems and learner models more tractable, such a simplification has its limitations. For example, we are aware that estimating prior knowledge and current knowledge would be more accurate if they were more interconnected.

### 3.2 Modeling Approach

Although our focus is on modeling the learner's knowledge of facts, in the description of models we use the common general terminology used in learner modeling, particularly the notions of *items* and *skills*. In applying this to geography, items correspond to locations and names of places while skills correspond to the knowledge (memory activation) of these facts.

In all models we use the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ as a link between a skill and a probability that a learner answers correctly. In the case of multiple-choice questions the probability of a correct answer can be modeled naturally by a shifted logistic function $\sigma(x, n) = \frac{1}{n} + (1 - \frac{1}{n})\frac{1}{1+e^{-x}}$, where $n$ is the number of options. The same approach to modeling guessing is used for example, in the standard three-parameter logistic model of item response theory (De Ayala, 2008). We are only concerned with online models, i.e., those that are updated after each answer. Such models can adapt to user behavior quickly and are therefore very useful in adaptive practice systems.

### 3.3 Specific System – Geography

For experiments we use an adaptive educational system `outlinemaps.org` – an application for learning geography (Papoušek et al, 2014). Learners can choose a specific map (e.g., Africa, Germany) and a type of place (e.g., countries, regions, cities, rivers). The system uses just two simple types of questions: questions about the location of a selected place ("Where is France?") and questions about the name of a selected place ("What is the name of the highlighted country?"). The questions are either open (select any item from a given map) or multiple-choice with 2 to 6 options. The focus of the system is on adaptivity, thus the questions are selected according to the estimated knowledge of a particular learner.

Learners answer questions using an interactive 'outline map'. After a sequence of 10 questions, the system provides feedback on the learner's progress. Learners can also access a visualization of their knowledge using an open learner model.

The application is currently used by hundreds of learners per day, the majority of whom are from the Czech Republic ($> 85\%$) and Slovakia ($> 10\%$) since the interface was originally in Czech. English, Spanish, and German versions have since become available. The system is available to everyone, free of charge. We store no personal information about learners – we only log their IP address. We have no control over the number of answered questions, the time when learners practice, or whether they ever return to the system after one session of practice.

## 4 Estimation of Prior Knowledge

At first, we process the estimation of prior knowledge. Our aim in this step is to estimate the learners' knowledge before they start using the system. We specifically want to estimate the probability that a learner $l$ knows an item $i$ based on previous answers of the learner $l$ to questions about different items and previous answers of other learners to questions about the item $i$. For a simpler interpretation of the data, we use only the first answer about each item for each learner in this step and we assume that learner's knowledge of an item $i$ is not influenced by answering questions about other items – this is a simplification in the case of multiple-choice questions where the item $i$ can occur as a distractor in a question about other items.

### 4.1 Basic Model

The basic model assumes that both learners and studied facts are homogeneous. It assumes that learners' prior knowledge in the domain can be modeled by a one-dimensional parameter.

We model the prior knowledge using the Rasch model, which entails having a learner parameter $\theta_l$ corresponding to the global domain knowledge of a learner $l$ and an item parameter $d_i$ corresponding to the difficulty of an item $i$. The probability that the learner answers correctly is estimated using a logistic function of a difference between the global skill and the difficulty: $P(correct|\theta_l, d_i) = \sigma(\theta_l - d_i)$.

A common approach to parameter estimation for the Rasch model is joint maximum likelihood estimation. In its basic form this approach is an iterative procedure that is slow for large data, and is not suitable for an online application which needs to continuously adjust estimates of parameters.

Parameter estimation can be done efficiently using a variant of the Elo rating system (Elo, 1978). The Elo rating system was originally devised for chess rating, but we can use it in learner modeling by interpreting a learner's answer to a question about an item as a "match" between the learner and the item. The skill and difficulty estimates are updated as follows:

$$\begin{aligned} \theta_l &:= \theta_l + K \cdot (\mathit{correct} - P(\mathit{correct}|\theta_l, d_i)), \\ d_i &:= d_i + K \cdot (P(\mathit{correct}|\theta_l, d_i) - \mathit{correct}), \end{aligned}$$

where *correct* denotes whether the question was answered correctly and $K$ is a constant specifying the sensitivity of the estimate to the last attempt. An intuitive improvement, which is used in most Elo extensions, is to use an "uncertainty function" instead of the constant $K$ – the update should get smaller as we have more data about a learner or an item. We use an uncertainty function $U(n) = \alpha/(1 + \beta n)$, where $n$ is the number of previous updates to the estimated parameter and $\alpha, \beta$ are meta-parameters.

### 4.2 Bayesian Model

In our basic model, uncertainty is modeled as a simple function of the number of attempts. Such an approach is a simplification since some answers are more informative than others and thus the effect of answers on the reduction of uncertainty should be differentiated. This can be done using a Bayesian modeling approach. For this model we treat $\theta_l$, $d_i$ and *correct* as random variables. We can use Bayes' theorem for updating our beliefs about skills and difficulties:

$$P(\theta_l, d_i|\mathit{correct}) \propto P(\mathit{correct}|\theta_l, d_i) \cdot P(\theta_l, d_i).$$

We assume that the difficulty of an item is independent of a learner's skill and thus $P(\theta_l, d_i) = P(\theta_l) \cdot P(d_i)$. The updated beliefs can be expressed as marginals of the conditional distribution, for example:

$$P(\theta_l|\mathit{correct}) \propto P(\theta_l) \cdot \int_{-\infty}^{\infty} P(\mathit{correct}|\theta_l, d_i = y) \cdot P(d_i = y)\mathrm{d}y.$$

In the context of rating systems for games, the basic Elo rating system has been extended in this direction, particularly in the Glicko system (Glickman, 1999). It models prior skill by a normal distribution and uses a numerical approximation to represent the posterior by a normal distribution and to update the mean and the standard deviation of the skill distribution using closed form expressions. Another Bayesian extension is TrueSkill (Herbrich et al, 2006), which further extends the system to allow team competitions.

This approach is, however, difficult to modify for new situations, e.g., in our case we want to use the shifted logistic function for modeling answers to multiple-choice questions. Therefore, we use a more flexible particle based method to represent the skill distribution. The skill is represented by a skill vector $\boldsymbol{\theta_l}$, which gives the values of skill particles, and a probability vector

$\boldsymbol{p_l}$, which gives the probabilities of the skill particles (sums to 1). The item difficulty is represented analogically by a difficulty vector $\boldsymbol{d_i}$ and a probability vector $\boldsymbol{p_i}$. In the following text the notation $\boldsymbol{p_l}_k$ stands for the $k$-th element of the vector $\boldsymbol{p_l}$.

The skill and difficulty vectors are initialized to contain values that are spread evenly in a specific interval around zero. The probability vectors are initialized to proportionally reflect the probabilities of the particles in the selected prior distribution. During updates, only the probability vectors change, while the vectors that contain the values of the particles stay fixed. Particles are updated as follows:

$$\boldsymbol{p_l}_k := \boldsymbol{p_l}_k \cdot \sum_{j=1}^{n} P(correct|\theta_l = \boldsymbol{\theta_l}_k, d_i = \boldsymbol{d_i}_j) \cdot \boldsymbol{p_i}_j,$$
$$\boldsymbol{p_i}_j := \boldsymbol{p_i}_j \cdot \sum_{k=1}^{n} P(correct|\theta_l = \boldsymbol{\theta_l}_k, d_i = \boldsymbol{d_i}_j) \cdot \boldsymbol{p_l}_k.$$

After the update, we must normalize the probability vectors so that they sum to one. A reasonable simplification that avoids summing over the particle values is:

$$\boldsymbol{p_l}_k := \boldsymbol{p_l}_k \cdot P(correct|\theta_l = \boldsymbol{\theta_l}_k, d_i = E[\boldsymbol{d_i}]),$$
$$\boldsymbol{p_i}_j := \boldsymbol{p_i}_j \cdot P(correct|\theta_l = E[\boldsymbol{\theta_l}], d_i = \boldsymbol{d_i}_j),$$

where $E[\boldsymbol{d_i}]$ ($E[\boldsymbol{\theta_l}]$) is the expected difficulty (skill) particle value (i.e., $E[\boldsymbol{d_i}] = \boldsymbol{d_i}^T \cdot \boldsymbol{p_i}$). By setting the number of particles we can trade precision on one hand for speed and memory requirements on the other.

Using this particle model in a real-world application would require storing the probabilities for all the particles in a database. If we assume that our beliefs stay normal-like even after many observations, then we can approximate each of the posteriors by a normal distribution. This approach is called assumed-density filtering (Minka, 2001). Consequently, each posterior can be represented by just two numbers, the mean and the standard deviation. In this simplified model, each update requires the generation of new particles. We generate the particles in the interval $(\mu - 6\sigma, \mu + 6\sigma)$. Otherwise, the update stays the same as before. After the update is performed, the mean and the standard deviation are estimated in a standard way: $\mu_{\theta_l} := \boldsymbol{\theta_l}^T \cdot \boldsymbol{p_l}, \sigma_{\theta_l} := \|\boldsymbol{\theta_l} - \mu_{\theta_l}\|_2$.

The model can be extended to include multiplicative factors for items ($q_i$) and learners ($r_l$), similarly to the Q-matrix method (Tatsuoka, 1983; Barnes, 2005) or collaborative filtering (Koren and Bell, 2011). Let $k$ be the number of factors, then when $x$ is passed to the likelihood function $\sigma(x)$, it has the form: $x = \theta_l - d_i + \sum_{j=1}^{k} q_{i,j} \cdot r_{l,j}$. The updates are similar – we only need to track more variables.

## 4.3 Hierarchical Model

In the models discussed so far items were characterized only by their difficulty, otherwise the domain was assumed to be homogeneous. In the next model we try to capture the domain in more detail by relaxing this assumption. Items are divided into disjoint sets – usually called 'concepts' or 'knowledge components', e.g., the allocation of countries to continents. The model now uses a two-level hierarchy of skills: in addition to the global skill $\theta_l$, there are now concept skills $\theta_{lc}$. To estimate the model parameters we extend the Elo rating system. Predictions are done in the same way as in the basic Elo rating system, the global skill being corrected just by the concept skill: $P(\mathit{correct}|\theta_l, \theta_{lc}, d_i) = \sigma((\theta_l + \theta_{lc}) - d_i)$. The update of parameters is also analogical:

$$\begin{aligned}
\theta_l &:= \theta_l + U(n_l) \cdot (\mathit{correct} - P(\mathit{correct}|\theta_l, \theta_{lc}, d_i)),\\
\theta_{lc} &:= \theta_{lc} + \gamma \cdot U(n_{lc}) \cdot (\mathit{correct} - P(\mathit{correct}|\theta_l, \theta_{lc}, d_i)),\\
d_i &:= d_i + U(n_i) \cdot (P(\mathit{correct}|\theta_l, \theta_{lc}, d_i) - \mathit{correct}).
\end{aligned}$$

For the uncertainty function $U(n)$ we use the same function as before; $\gamma$ is a new meta-parameter specifying the sensitivity of the model to concepts.

The proposed model is related to several learner modeling approaches. It can be viewed as a simplified Bayesian network model (Conati et al, 2002; Käser et al, 2014; Millán et al, 2010). In a proper Bayesian network model we would model skills by a probability distribution and update the estimates using Bayes' theorem; equations in our model correspond to a simplification of this computation using only point skill estimates. The Bayesian network model can also model more complex relationships (e.g., prerequisites), which are not necessary in our case, i.e., learning factual knowledge. Other related modeling approaches are the Q-matrix method (Tatsuoka, 1983; Barnes, 2005), which focuses on modeling mapping between skills and items (mainly using $N : M$ relations), and models based on knowledge space theory (Doignon and Falmagne, 1999). Both these approaches are more complex than the proposed model. Our aim here is to evaluate whether even a simple concept-based model is practical for modeling factual knowledge.

The advantage of the hierarchical model is that learners' knowledge is represented in more detail and the model is thus less sensitive to the assumption of homogeneity among learners. However, to use the hierarchical model, we need to determine concepts, which involves dividing items into disjoint sets. This can be done in several ways. Concepts may be specified manually by a domain expert. In the case of the geography learning application some groupings are natural (continents, countries). In other cases the construction of concepts is more difficult, such as in the case of foreign language vocabulary where it is not clear how to determine coherent groups of words. It is also possible to create concepts automatically or to refine concepts provided by an expert with the use of machine learning techniques (Desmarais et al, 2012; Nižnan et al, 2014).

To determine concepts automatically, it is possible to use classical clustering methods. For our experiments we used the spectral clustering algorithm (Von Luxburg, 2007) with similarity of items $i, j$ defined as a Spearman's correlation coefficient $c_{ij}$ of correctness of answers (represented as 0 or 1) of shared learners – those who answered questions about both items $i$ and $j$. To take into account the use of multiple-choice questions, we decrease the binary representation of a response $r$ by the guess factor to $r - 1/k$ ($k$ being the number of options).

It is also possible to combine the manual and the automatic construction of concepts (Nižnan et al, 2014). With this approach the manually constructed concepts are used as item labels. Items with these labels are used as a training set of a supervised learning method for which we used logistic regression with regularization. For the item $i$, the vector of correlation with all items $c_{ij}$ is used as a vector of features. Errors of this classification method are interpreted as "corrected" labels; see Nižnan et al (2014); Nižnan et al (2014) for more details.

### 4.4 Networked Model

The hierarchical model enforces a strict division of items into groups. With the next model we bypass this division by directly modeling the relations between individual items, i.e., we treat items as a network, hence the name 'networked model'. For each item we have a local skill $\theta_{li}$. For each pair of items we compute $c_{ij}$ – the degree to which they are correlated, which is computed in the same way as in the concept detection. This is done from training data or – in the real system – once a certain number of answers has been collected. After the answer to the item $i$, all skill estimates for all other items $j$ are updated based on $c_{ij}$. The model still uses the global skill $\theta_l$ and makes the final prediction based on the weighted combination of the global skill $\theta_l$ and the local skill $\theta_{li}$: $P(\mathit{correct}|\theta_l, \theta_{li}) = \sigma(w_1\theta_l + w_2\theta_{li} - d_i)$. Parameters are updated as follows:

$$\begin{aligned}
\theta_l &:= \theta_l + U(n_l) \cdot (\mathit{correct} - P(\mathit{correct}|\theta_l, \theta_{li})), \\
\theta_{lj} &:= \theta_{lj} + c_{ij} \cdot U(n_l) \cdot (\mathit{correct} - P(\mathit{correct}|\theta_l, \theta_{li})) \quad \text{for all items } j, \\
d_i &:= d_i + U(n_i) \cdot (P(\mathit{correct}|\theta_l, \theta_{li}) - \mathit{correct}).
\end{aligned}$$

This model is closely related to the multivariate Elo rating system previously proposed in the context of adaptive psychometric experiments (Doebler et al, 2014).

For illustration of the model, Fig. 2 shows a selection of the most important correlations for European countries. Note that this automatically generated figure contains some natural clusters (from the perspective of a typical user of our system): Balkan countries (top center), Baltic countries (top left), Scandinavian countries (bottom right), and well-known countries (bottom left).

### 4.5 Evaluation

This section reports our experience with fitting model parameters and the comparison of different models with respect to the accuracy of their predictions. The experiments are based on a data set that is publicly available (Papoušek et al, 2016a). Our aim at this point is to model prior knowledge, so we selected

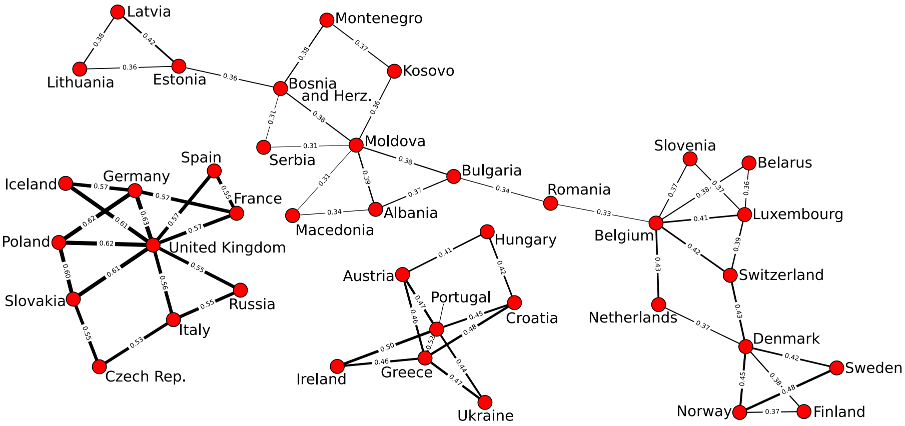

**Fig. 2** Illustration of the networked model on European countries. Only the most important edges for each country are shown.

for each learner and item the first answer only. The used data set contains approximately 3,900,000 answers of 91,000 learners. The data set was split into a training set (30%) and a test set (70%) in a learner-stratified manner. All the reported models work online. The training of models (parameters $\theta_l, d_i$) continues on the test set, but only predictions on this set are used to evaluate models.

### 4.5.1 Model Parameters

The training set was used for finding the values of the meta-parameters of individual models. The grid search was used to find the best parameters of the uncertainty function $U(n) = \alpha/(1 + \beta n)$. Optimal performance over the training set was achieved for values $\alpha = 1$ and $\beta = 0.06$; this exact choice of parameter values is not crucial as many choices of $\alpha, \beta$ provide very similar results. We also used these values for derived models that use the uncertainty function.

The basic Elo rating system with its uncertainty function provides both fast, rough estimates after a few answers and stability in the long run (see Fig. 3 left). It also provides nearly identical estimates as the joint maximum likelihood estimation (JMLE), which is the standard approach to estimating parameters of the Rasch model (Fig. 3 right, correlation 0.97). JMLE is an iterative procedure requiring several iterations over the whole data set, whereas the Elo rating system requires only a single pass of the data. More importantly, the Elo rating system can be easily used online (performing a simple update for each new observation). It is possible to modify the JMLE approach for online usage – learning item parameters offline and computing online only skill estimates, which are based on only a small subset of data. But such modification is still more complex than using the Elo rating system. Since the

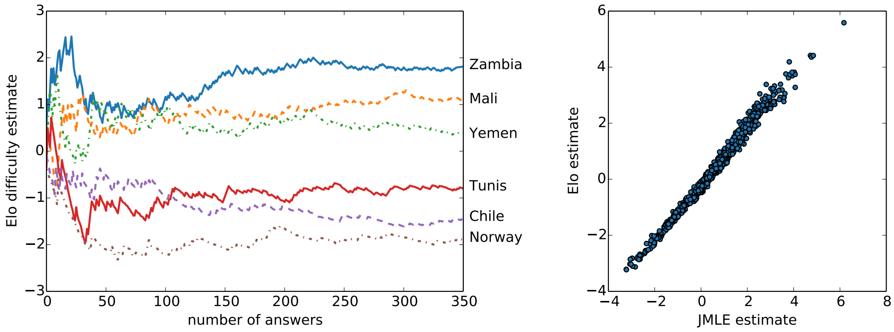

**Fig. 3** Estimation of prior knowledge: Development of estimates of difficulty of selected countries under the Elo rating system (left). Comparison of Elo and JMLE difficulty estimates (right).

**Table 1** Comparison of models on the test set.

| Model | RMSE | LL | AUC |
|---|---|---|---|
| Mean of learner and item success rate | 0.4348 | $-1.5260 \times 10^6$ | 0.6797 |
| Elo ($\alpha = 1, \beta = 0.06$) | 0.4142 | $-1.4073 \times 10^6$ | 0.7431 |
| Bayesian model | 0.4147 | $-1.4106 \times 10^6$ | 0.7414 |
| Bayesian model (3 skills) | 0.4117 | $-1.3924 \times 10^6$ | 0.7503 |
| Hierarchical model | 0.4115 | $-1.3915 \times 10^6$ | 0.7514 |
| Networked model | 0.4118 | $-1.3918 \times 10^6$ | 0.7509 |

estimates of the two methods are nearly identical, we conclude that the Elo rating system is preferable in our context.

### *4.5.2 Accuracy of Predictions*

Table 1 shows the results of model comparison. As a baseline we report 'learner-item success rate': this prediction is given by averaging the success rate of previous learners on a given item and the success rate of a given learner on previous items. To compare models we use standard performance metrics. As a primary metric we consider the root mean square error (RMSE), since the application works with absolute values of predictions (see Pelánek (2015) for more details on choice of metric). In addition to RMSE we also report log-likelihood (LL) and the area under the ROC curve (AUC). The main results are not dependent on the choice of metric. In fact, predictions of models are highly correlated. For example, in the case of the basic Elo model and the hierarchical model, most predictions (95%) differ by less than 0.1.

The hierarchical model reported in Table 1 uses manually determined concepts based on both location (e.g., continent) and type of place (e.g., country). Both the hierarchical model and the networked model bring an improvement to the basic Elo model. The improvement is statistically significant (as determined by a t-test over results of repeated cross-validation), but it is rather small. Curiously, the Particle Bayes model is slightly worse than the simple Elo rating system, i.e., the more involved modeling of uncertainty does not improve predictions. The performance improves only when we use the multiple skill extension.

We hypothesize that the improvement of the hierarchical (respectively multiple skill) extensions model would be more significant for less homogeneous populations of learners. We probed this hypothesis by artificially creating heterogeneous data sets using location information from IP addresses. From the original data set we created two test sets. The first one consists of 6,000 Czech learners and represents a homogeneous population. The second one consists of 6,000 learners spread all over the world and represents a heterogeneous population. The results of the evaluation across these data sets show that the hierarchical and network model has the same performance on both data sets, whereas the basic model struggles with the heterogeneous data set and has significantly higher RMSE than for the homogeneous data set.

### 4.6 Using Models for Insight

In learner modeling we are interested not just in predictions, but also in getting insight into the characteristics of the domain and the learning process. The advantage of more complex models may lie in additional parameters that bring or improve such insights.

The extensions of the basic model (networked, hierarchical, Bayesian with multiple skills) bring insight into the domain thanks to the analysis of relations among items, e.g., by identifying the most useful clusters of items or by exploring relationships among items (see Fig. 2). Such results can be used for improving the behavior of an adaptive educational system. For example, the system can sequence the practice in such a way that items from one concept are practiced in a row (which is in many cases natural from the user experience perspective). Another possible use of concepts is for the automatic construction of multiple-choice questions with good distractors (falling under the same concept).

The hierarchical model can be used to evaluate the quality of different concepts. We used it to compare concepts obtained in three different ways: ‘manual’ (specified by authors using data about items type and location), ‘automatic’ (derived completely automatically and based on the available data), and ‘corrected’ (manually specified concepts refined using the data available). The methods used to realize the ‘automatic’ and ‘corrected’ approaches are described in Section 4.3. We used several approaches for specifying the concepts manually: based on type (e.g., countries, cities, rivers), location (e.g., Europe, Africa, Asia) and combination of the two approaches (e.g., European countries, European cities, African countries). Since we have the most answers for European countries, we also considered a data set containing only answers for European countries. For this data set we used two sets of concepts. The

**Table 2** Comparison of manual, automatically corrected manual, and automatic concepts ($C$ is the number of concepts). Quality of concepts is expressed as RMSE improvement ($\Delta$ RMSE) of the hierarchical model with these concepts over the basic model.

| All items | C | $\Delta$ RMSE |
|---|---|---|
| manual – type | 14 | 0.00144 |
| corrected – type | 14 | 0.00132 |
| manual – location | 22 | 0.00195 |
| corrected – location | 22 | 0.00183 |
| **manual – combination** | **56** | **0.00268** |
| corrected – combination | 56 | 0.00249 |
| automatic | 5 | −0.00004 |
| automatic | 20 | 0.00163 |
| automatic | 50 | 0.00156 |

| Europe | C | $\Delta$ RMSE |
|---|---|---|
| manual | 3 | −0.00009 |
| **corrected** | **3** | **0.00011** |
| manual | 6 | −0.00024 |
| corrected | 6 | 0.00004 |
| automatic | 2 | −0.00001 |
| automatic | 3 | 0.00009 |
| automatic | 5 | −0.00028 |

first is the partition into Eastern, Western, North-western, Southern, Central and South-eastern Europe, and the second concept set is obtained from the first one by the union of Central, Western and Southern Europe, since countries from these regions are mostly well-known by our Czech students, and then the union of South-eastern and Eastern Europe.

The quality of concepts was evaluated using the prediction accuracy of the hierarchical model using these concepts. Table 2 shows the results expressed as the RMSE improvement over the basic model. Note that the differences in RMSE are necessarily small, since the models used are very similar and differ only in the allocation of items to concepts. For the whole data set (1368 items), a larger number of concepts improves the performance. The best results are achieved by manually specified concepts (a combination of location and type of place), automatic correction does not lead to a significantly different performance. For the smaller data set of European countries (39 items), a larger number of both manual and automatically determined concepts causes an inferior performance – a model with too small concepts suffers from a loss of information. In this case the best result is achieved by a correction of manually specified concepts. The analysis shows that the corrections make intuitive sense, since most of them are shifts of well-known and easily recognizable countries such as Russia or Iceland to the block of well-known countries (the union of Central, Western and Southern Europe).

## 5 Estimation of Current Knowledge

We now turn to the estimation of a learner's current knowledge, i.e., knowledge influenced by repeatedly answering questions about an item. The input data for this estimation are an estimate of prior knowledge (provided by one of the models described above) and the history of previous attempts, i.e., the sequence of previous answers (correctness of answers, question types, timing information).

### 5.1 Basic Approach

Several models can be considered for estimating current knowledge. Bayesian knowledge tracing (Corbett and Anderson, 1994; van de Sande, 2013), a popular learner modeling technique, can be used in a straightforward way. In this context the probability of initial knowledge is given by the previous step. The probability of learning, guess, and slip are given either by a context (guess in the case of multiple choice questions) or can be easily estimated using an exhaustive search. However, in this context the assumptions of Bayesian knowledge tracing are not very plausible, as it assumes a discrete transition from the unknown to the known state. This may be a reasonable simplification for procedural skills, but for declarative facts the development of memory activation is more gradual.

Assumptions of Performance factor analysis (Pavlik et al, 2009) are more relevant for the learning of facts. Whereas Performance factor analysis (PFA) was originally formulated in the context of multiple knowledge components, we are using a simplified one-dimensional variant. In this model, the skill (memory activation) is given by a linear combination of an initial value and past successes and failures of a learner: $m = \beta + \gamma s + \delta f$, where $\beta$ is the initial activation, $s$ and $f$ are counts of previous successes and failures of the learner, $\gamma$ and $\delta$ are parameters that indicate the change of the skill associated with correct and incorrect answers. The basic disadvantage of this simple approach is that it does not consider the time between attempts; in fact, it even ignores the order of answers, as it uses only the summary number of correct and incorrect answers.

The ACT-R model (Pavlik and Anderson, 2005; Pavlik Jr et al, 2008) of spacing effects can be considered as an extension of this basic model. In this model the memory activation is estimated as $m = \beta + \log(\sum b_i t_i^{-d_i})$, where the sum is over all previous attempts, values $t_i$ are the ages of previous attempts, values $b_i$ capture the influence of correctness of answers, and $d_i$ is the decay rate computed by recursive equations (Pavlik and Anderson, 2005). The model also includes additional modifiers for treating time between sessions. The focus of the model is on modeling the decay rate to capture the spacing effect. Studies using this model (Pavlik and Anderson, 2005; Pavlik Jr et al, 2008) did not take into account the probability of guessing and variable initial knowledge of different items – initial activation was either a global constant or a learner parameter. Since detailed modeling of spacing effects has not been completely solved even in the case of simple 'laboratory' conditions, we currently omit modeling of spacing effects and focus on factors that are crucial in the context of our practical application, namely, guessing and variable initial knowledge.

A disadvantage of PFA is that it does not consider the order of answers and neither does it take into account the probability of guessing. Guessing is important particularly in our setting, where the system uses multiple choice questions with a variable number of options. To address these issues we propose combining PFA with some aspects of the Elo rating system, which in the following text we denote as PFAE – PFA Elo/Extended:

- $\theta_{li}$ is the estimated knowledge of a learner $l$ of an item $i$.
- The initial value of $\theta_{li}$ is provided by the estimation of prior knowledge, e.g., for the basic model it is $\theta_{li} = \theta_l - d_i$.
- The probability of a correct answer to a question with $n$ options is given by the shifted logistic function: $P(\mathit{correct}|\theta_{li}, n) = \frac{1}{n} + (1 - \frac{1}{n})\sigma(\theta_{li})$.
- After an answer to a question with $n$ options, the estimated knowledge is updated as follows:

$$\begin{aligned} \theta_{li} &:= \theta_{li} + \gamma \cdot (1 - P(\mathit{correct}|\theta_{li}, n)), \text{if the answer was correct,} \\ \theta_{lj} &:= \theta_{li} + \delta \cdot P(\mathit{correct}|\theta_{li}, n), \text{if the answer was incorrect.} \end{aligned}$$

### 5.2 Timing Information

To include timing information in this model, we increase the memory activation locally for the purpose of prediction, i.e., instead of $P(\theta_{li})$ we use $P(\theta_{li} + f(t))$, where $t$ is the time (in seconds) from the last attempt and $f$ is the *time effect function*.

It is natural to use as a time effect function some simple analytic function, but the analysis of our data suggests that this approach does not work well. We experimented with two types of analytic functions: $f(t) = \frac{w}{t}$ and $f(t) = 1.6 - 0.1\log(t)$. The first function was used in the initial proposal of the system (Papoušek et al, 2014); the second function is based on previous research (Pavlik and Anderson, 2005), with parameters fitted to our data. Our analysis of these predictions shows that neither of these functions leads to well calibrated predictions (details are reported in Pelánek (2015)).

Since we were not able to find a simple time effect function that would provide a good fit, we derive the time effect function automatically from the data. To represent the function $f(t)$ we use a generic staircase function with fixed bounds $\mathbf{b}$ and values $\mathbf{v}$ which we learn from data:

$$f(t) = \begin{cases} v_i & \text{if } b_i \le t < b_{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Another type of timing information that could be potentially used to improve knowledge estimation is response time. The analysis of data from the system (Papoušek et al, 2015) shows that there is a relation between response time and correctness of the next answer for a question about the same item. Curiously, the effect of response time differs depending on whether the current answer is correct or incorrect. If the current answer is correct, then the probability of the next answer being correct is linearly dependent on the response time – it goes from 95% for very fast answers to nearly 80% for slow answers. If the current answer is incorrect, then the dependence on response time is weaker, but an approximately linear trend remains. Interestingly, in this case the trend is in the other direction (going from 60% to 65%). Response times have been studied extensively in psychology, for example in the context of perceptual learning. Specifically, previous work (Mettler et al, 2011) used response

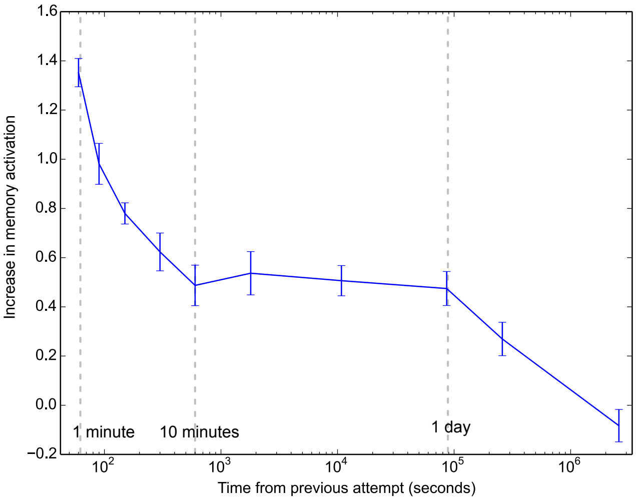

**Fig. 4** Time effect function – average from 10 independent data sets, error bars show standard deviations of parameter estimates.

time for adaptive scheduling of practice, but without considering prior knowledge. Incorporating response times into our modeling framework is beyond the scope of this paper, and offers an interesting direction for future work.

### 5.3 Evaluation

For this evaluation we consider only sequences where a learner answered at least 3 questions about an item. As an initial estimate of learner knowledge, we use outputs of the basic Elo model of prior skill. As the fixed bounds used in the staircase representation of the time effect function, we have chosen the following values (in seconds): 0, 60, 90, 150, 300, 600, 1800, 10800, 86400, 259200, 2592000. These values were chosen to be easily interpretable (e.g., 30 minutes, 1 day) and at the same time to have a reasonably even distribution of data into individual bins.

The model has the following parameters that have to be estimated from the data: update constants $\gamma, \delta$ and the vector $\mathbf{v}$ representing the time effect function. To estimate these parameters we use a greedy descent. To check the stability of the parameter estimation procedure we computed parameter values for 10 independent data sets. The results show that these parameters are very stable: $\gamma = 2.23 \pm 0.05$, $\delta = -0.89 \pm 0.04$; values $\mathbf{v}$ representing the time effect function are depicted in Fig. 4.

Since our data set is large and parameter estimates are stable, we can afford to do a more detailed analysis. Fig. 5 shows fitted time effect functions and $\gamma, \delta$ values when the parameters are fitted only for specific types of places. These parameters contain useful information about learners' learning in particular parts of the domain. Similar analyses show that there is quite a large difference

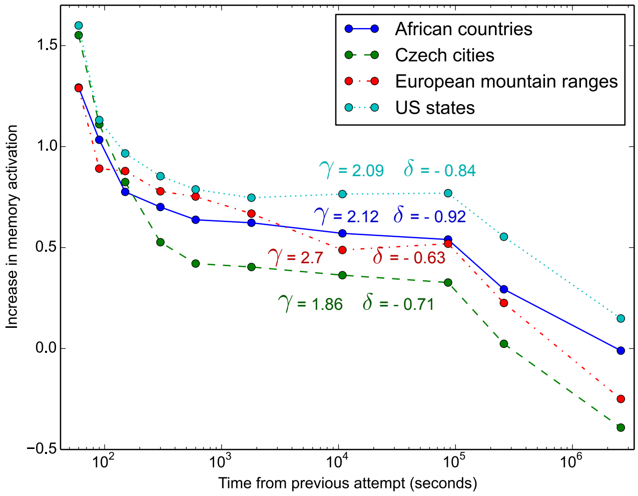

**Fig. 5** Time effect function and $\gamma, \delta$ parameters for different types of places.

between parameter values for cases with high and low prior knowledge. This suggests a possible improvement to the PFAE model – not just by including more parameters, but also by changing its functional form.

Our comparison of predictive accuracy of models (reported in detail in Papoušek et al (2014); Pelánek (2015)) shows that the PFAE model brings quite a large improvement over the basic Bayesian knowledge tracing and Performance factor analysis models. Differences between variants of the PFAE model show that the model with the fitted staircase function is better than models with prespecified analytic functions. These differences are statistically significant, but otherwise rather small. Individual predictions are actually highly correlated (correlation coefficient around 0.97).

## 6 Question Construction

Finally, based on the estimated knowledge of a learner we want to construct a suitable next question. In the context of our geography application the construction of a question consists of several partial decisions: what should be the target place (the correct answer); what question type to use (“Where is X?” versus “What is the name of this place?”); how many distractors to use; and what should these distractors be.

The question construction process should satisfy several criteria, which partly conflict with each other. The criteria and their weight may depend on a particular application, a target learner population, and the learners’ goals. It is therefore not feasible to formulate a universal algorithm for question construction, which led us to devise the following approach. The first step is to propose general criteria that the question construction should satisfy. We then discuss a flexible approach for achieving specified criteria. Finally, we present our evaluation of the final algorithm, illustrating how the parameters of the algorithm can be optimized.

### 6.1 Criteria

We propose the following main criteria. The selection of a question should depend on the estimated *difficulty* of a question (for a particular learner). From the testing perspective, it is optimal to use questions with expected probability of a correct answer close to 50%, because such questions provide the most information about a learners' knowledge. However, a 50% success rate is rather low and for many learners it could decrease their motivation. In our setting (adaptive practice), it therefore seemed better to aim for a higher success rate. In our experiments we evaluate different target success rates.

Another important issue is the *repetition* of questions. This aspect should ideally be governed by the research about spacing effects (Delaney et al, 2010; Pavlik and Anderson, 2005). It is rather complex to fully model the spacing effect, but a little consideration of spacing intervals is necessary; repeating the same question too soon is certainly not recommended.

What is recommended, however, is a *variety* of question types. Different question types are useful mainly as a tool for fine-tuning the difficulty of questions, but even if this is not necessary, the variability of question types may be meaningful criteria in itself, since it improves user experience, if used correctly.

### 6.2 Selecting a Target Item

We start by choosing a target item, which is the correct answer to a constructed question. As a general approach we have settled on a linear scoring approach. For each relevant attribute we consider a scoring function that expresses the desirability of a given item with respect to this attribute. These scoring functions are combined using a weighted sum; the item with the highest total score is selected as a target.

This approach is flexible and thanks to the choice of attributes and their weights it can be adjusted for a particular application. We take the following attributes into consideration:

1. the probability that the learner knows the item,
2. the time period since the last question about the same item,
3. the number of questions already answered by the learner about the item.

Fig. 6 illustrates the general shapes resulting from our choice of scoring functions for these attributes. Further we specify formulas that approximate these shapes using simple mathematical functions.

The first function takes into account the relation between the estimated probability of a correct answer ($P_{est}$) and the target success rate ($P_{target}$).

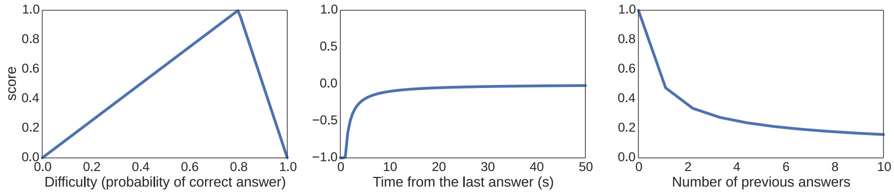

**Fig. 6** Desired contribution of different criteria to the selection of a target item.

Assume that our goal is to ask a question where the learner has 75% chance of a correct answer. The distance from the probability for the difficult items (nearly 0% chance of the correct answer) is higher than for easy ones (almost 100%), so it is necessary to normalize it:

$$S_{prob}(P_{est}, P_{target}) = \begin{cases} \frac{P_{est}}{P_{target}} & \text{if } P_{target} \geq P_{est}, \\ \frac{1-P_{est}}{1-P_{target}} & \text{if } P_{target} < P_{est}. \end{cases}$$

The second scoring function penalizes items according to the time elapsed since the last question about the same item – we do not want to repeat it when it is still in the short term memory. We use the function $S_{time}(t) = -1/t$, where $t$ is time in seconds. Using only the above mentioned attributes, the system would ask questions for only a limited pool of items. To induce the system to ask questions about new items we introduce the third scoring function that uses the total number $n$ of questions answered by the learner for the given item: $S_{count}(n) = 1/\sqrt{1+n}$. The total score is given as a weighted sum of individual scores, with the weights being set manually based on our experiences with the prototype version of the system: $W_{prob} = 10$, $W_{count} = 10$, $W_{time} = 120$. Ideally, values of these parameters should be optimized using experiments with the system, potentially using automatic experimentation techniques like multi-armed bandit algorithms (Lomas et al, 2016) or Bayesian optimization (Khajah et al, 2016). In Section 6.4 we report experiments analyzing the role of the target difficulty parameter.

### 6.3 Choosing Options

Once the question's target item is selected, the question difficulty can be adjusted by using a multiple choice question with a suitable number of options. For a multiple choice question the probability of a correct answer is the combination of the probability of guessing the answer ($P_{guess}$) and knowing the target item ($P_{est}$): $P_{success} = P_{guess} + (1 - P_{guess}) \cdot P_{est}$. This is inevitably a simplification since a multiple choice question can also be answered by ruling out distractor options. But if the distractors are well chosen, this simplification is reasonable.

As our goal is to get $P_{success}$ close to $P_{target}$, we would like to make $P_{guess}$ close to

$$G = \frac{P_{target} - P_{est}}{1 - P_{est}}.$$

For $G \leq 0$, we use open question (no options), otherwise we use $n$ closest to $\frac{1}{G}$ as a number of options. For principled reasons the minimal possible value of $n$ is 2, and for practical reasons there is also an upper limit for $n$: presence of more than 6 options could make the user interface cluttered. The type of the question – "Where is X?" or "What is the name of this place?" is currently selected randomly. In the case of the second question type, open questions are transformed into questions with 6 options.

When using multiple choice questions, we also need to choose the distractor options. Unlike other systems for practice dealing with text (Mitkov et al, 2006; Mostow et al, 2002), we work with well-structured data, so the selection of distractors is easier. The choice of distractors can be based on domain information, e.g., geographically close countries or countries with similar names. However, the easiest way to choose good distractors is to simply base the choice on past answers. We can take items most commonly mistaken with the target item in open questions, and select from them randomly. The random choice is weighted by the frequency of mistakes with the given item – the distribution of wrong answers is typically highly skewed. For example, Kenya is most often confused with Tanzania (24%), Ethiopia (21%), South Sudan (9%), Uganda (5%), and Congo (3%).

### 6.4 Evaluation

Compared to the estimation of knowledge, question construction is much more difficult to evaluate since we do not have a single, clear, easily measurable goal. The overall goal of constructing questions is quite clear – it is the maximization of learning. But it is not easy to measure the fulfillment of this general goal, since it depends also on the context of the learning. An experiment with pre-test, post-test and fixed time in the system may provide a setting for an accurate evaluation of the different question construction strategies. Results of such experiments would, however, lack ecological validity, as many of the users of the system use it on their own without any time limits. The issue of engagement, for example, is much more important than in a controlled experiment.

To perform the evaluation we use randomized trials where learners are randomly assigned to one of several experimental conditions, which correspond to different variants of the question construction algorithm. We compare the experimental conditions by analyzing both learners' engagement and learning.

To measure engagement we consider both learners' objective behavior and subjective evaluation of the practice provided. To quantify the behavior we measure the total number of answered questions. The distribution of the number of answers across learners is highly skewed and is therefore not suitable to comparing conditions using averages, or even other measures of central tendency like the median. An analysis of the data (Papoušek et al, 2016b) shows that the length of stay within the system fits the Weibull distribution, which is a standard distribution in survival analysis: previous research has shown that this distribution also fits dwell time on web pages (Liu et al, 2010) well. Another approach to measure survival is to use survival rates, which express the proportion of learners that answer more than $k$ questions. These rates are both easier to interpret and provide similar insight as the parameters of the fitted Weibull distribution. The survival rates allow us to differentiate between short term and long term engagement. To measure long term engagement we also analyze the probability of returning to the system after more than 10 hours has elapsed, although the specific duration of the delay is not important for our results.

To measure the subjective perception of questions we ask learners to evaluate the difficulty of questions. After 30 answers the system shows the dialog "How difficult are the questions?" and learners choose one of the following ratings: "Too Easy", "Appropriate", "Too Difficult".

The evaluation of learning cannot be simply based on the success rate that the learners achieved, since the experimental conditions also influenced it. To measure learning we collect "reference questions" – every 10th question is an open questions about a randomly chosen item from the context being practiced, i.e., these questions are not influenced in any way by the experimental conditions. Based on these answers we construct learning curves which we use to compare learning in individual experimental conditions; see Papoušek et al (2016b) for more details.

### *6.4.1 Impact of the Question Construction Algorithm*

In the first experiment we compare the adaptive algorithm to a random construction of questions. The proposed adaptive algorithm for question construction consists of two main parts. Firstly, the algorithm selects the target item of the question (the correct answer). Secondly, it chooses the number of options for a multiple choice question and particular distractors. In our experiments we evaluate four versions of the question construction algorithm; for both construction steps we consider an adaptive condition and a random condition: *adaptive-adaptive* (*A-A*), *adaptive-random* (*A-R*), *random-adaptive* (*R-A*), *random-random* (*R-R*).

The experiment ran from August to October 2015, during which time we collected more than 1,300,000 answers from roughly 20,000 learners. The data set is available[1] (together with a brief description and terms of use).

Fig. 7 (top) gives an overview of different measures of engagement. The figure shows that adaptivity in the first question construction step is related to short term engagement (survival rates after 10 questions), whereas adaptivity in the second step is related to long term engagement (survival rates after 150

[1] `www.fi.muni.cz/adaptivelearning/data/slepemapy/2015-ab-random-parts.zip`

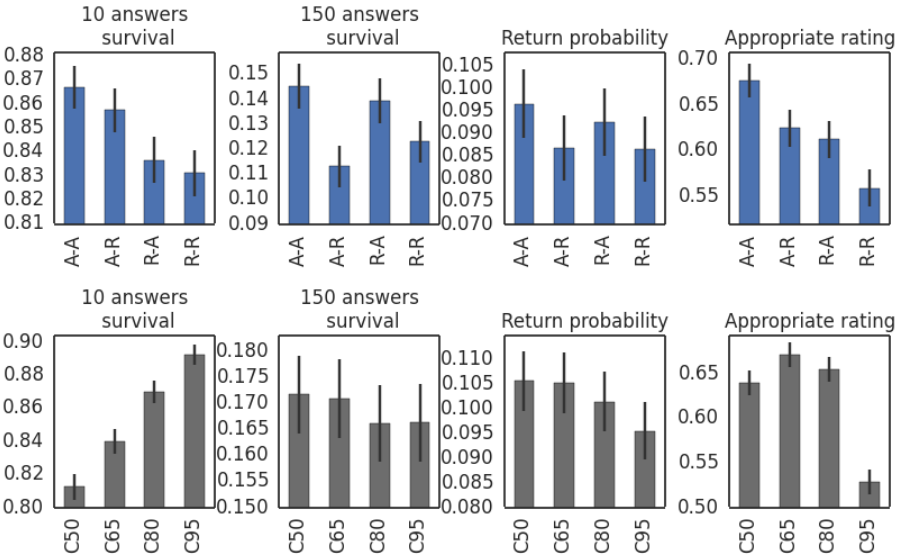

**Fig. 7** Summary of different engagement measures for the two performed experiments. Error bars show 95% confidence intervals.

questions, the probability of return). Note that with respect to the probability of return, the relative difference between *A-A* and *R-R* conditions is 15%, i.e., adaptability has a great impact on a learners' decision to use the system repeatedly.

Finally, Fig. 7 (top) also shows the results of learners' ratings of question difficulty – the most appropriately difficult questions among the experimental conditions are asked under the *A-A* condition. More detailed analysis shows that the other three conditions exhibit an increased number of "Too Easy" evaluations. In particular, both **-R* conditions have an increased number of "Too Easy" compared to their **-A* counterparts. The subjective evaluation reflects data on the success rate of learners in individual conditions. The random choice of options leads to a higher success rate with both *A-R* and *R-R* having an average success rate of 82% (excluding reference answers). In these cases learners can probably often guess the correct answer even when they are not sure. In the case of adaptive constructions of options, the success rate of most learners is close to the target success rate (75%) – both *A-A* and *R-A* have an average success rate of 78% (excluding reference answers).

The evaluation of learning using learning curves (illustrated in Fig. 8) is not straightforward due to attrition bias; see Papoušek et al (2016b) for more detailed discussion. The overall results, however, consistently show that the conditions with adaptive construction of options (*A-A*, *R-A*) surpass the conditions with random options (*A-R*, *R-R*). Item selection does not seem to have a great impact on learning. When we see differences between the *A-A* and *R-A*

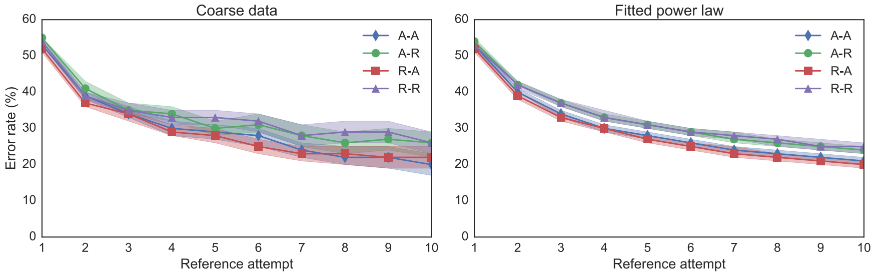

**Fig. 8** Learning curves. Left: Coarse data with 95% confidence intervals (indicated by filled areas). Right: Fitted power law curves.

conditions, the *R-A* condition is slightly better, i.e., it seems that with respect to learning, the adaptive choice of target item could be improved.

### *6.4.2 Impact of Difficulty*

In the second experiment we analyze the role of a key parameter in the question construction algorithm – the target success rate. The Inverted-U Hypothesis (Lomas et al, 2013) suggests that really easy and really hard questions should have negative impact on learners' engagement. In this experiment we compare several variants of the adaptive algorithm differing only in the target success rate: 50%, 65%, 80%, 95%. In the following text we denote the conditions as C50, C65, C80, C95. The experiment was performed between November 2015 and January 2016, during which time we collected almost 3,300,000 answers from roughly 37,000 learners. The data set is available[2] (together with a brief description and terms of use).

With respect to learning there is again an issue with attrition bias. Nevertheless the results suggest that more difficult practice leads to better learning, the difference being mainly between C95 and other conditions – see Papoušek et al (2016c) for a more detailed analysis of learning within individual contexts (maps).

For engagement the results are visualized in Fig. 7 (bottom) in the same way as for the previous experiment. The main observation is that there are opposing tendencies with respect to short term and long term engagement. Conditions with easier questions enhance engagement at the beginning, while more difficult conditions engage more learners later on. The survival rate after 10 answers is sorted according to question difficulty. The differences are decreasing with the number of answers, survival rates after 150 answers are similar in all conditions with slightly better results for more difficult questions. The return rate increases with the difficulty of questions, the largest difference being between C95 and other conditions. The subjective rating by learners is

[2] `http://www.fi.muni.cz/adaptivelearning/data/slepemapy/2016-ab-target-difficulty.zip`

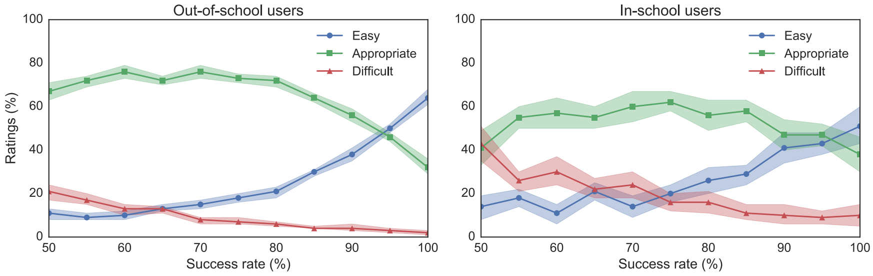

**Fig. 9** Ratings of question difficulty given by learners according to their achieved success rate, separately for out-of-school and in-school usage of the system.

best for the C65 conditions. The main difference is again between C95 and other conditions.

With respect to target success rates (as varied in experimental conditions), we do not have strong support for the Inverted U-hypothesis. One reason may be that we do not have sufficiently difficult questions. The success rate is influenced not only by the target rate, but also by other factors like learners choice of maps. Although the target probability is from the interval [50%, 95%], the average real success rate varies only from 65% to 90%. On several maps, such as countries in Europe (for which we have the most data), there are not enough difficult items to achieve a 50% success rate for most of our users. See Papoušek et al (2016c) for more detailed discussion.

However, the relation the between achieved success rate and the perceived difficulty of questions shows a clear U-shaped pattern (Fig. 9 left). For each learner we compute their success rate before they give us a rating, and we divide users into the buckets based on their success rate and then for each bucket we look at the percentage of “Too Easy”, “Appropriate” and “Too Difficult” records. The filled areas around curves represent confidence intervals. The curve does not have a sharp peak, but there are clear dynamics between the classes. With the decreasing difficulty the growth of the number of “Too Easy” votes is compensated for by the drop of “Too Difficult” votes. The peak of the “Appropriate” answers as well as the equal votes for “Too Easy” and “Too Difficult” occur between 60% and 70% success rates. This experiment therefore suggests that values around 65% may be a suitable target rate for this kind of application.

Previous research (Abuhamdeh and Csikszentmihalyi, 2012) suggests that the optimal difficulty may differ depending on the type of motivation (intrinsic, extrinsic), particularly in school-related activities as learners prefer lower levels of challenge. To examine this hypothesis we compared results for out-of-school usage of the system with in-school usage. To detect the ‘in-school usage’ we currently use only a coarse method based on IP address (as in-school usage we consider groups of at least 5 learners who started using the system from the same IP address). This in-school usage represents about 20% of the data. Fig. 9 shows that there is a substantial difference. The in-school group prefers easier questions with the optimal difficulty being around 75%, and they are also generally less satisfied with the practice in the system: the "Appropriate" ratings in Fig. 9 are generally lower for the in-school group than for the out-of-school group. Given that our approach to identifying in-school/out-of-school usage is quite simple, it is likely that the real difference is even higher.

## 7 Discussion

We present an integrated approach to building systems for the adaptive practice of facts, particularly for domains in which learners have varied prior knowledge. The proposed approach is based on "learning from data" and requires limited input from domain experts. This makes it a very cost-effective way to develop of adaptive educational systems. We illustrate and evaluate the approach on a specific case study from the field of geography. This approach can be directly applied to other domains, e.g., anatomy, biology, or vocabulary learning. Our group has already built several other systems that are based on the same principles as those described in the geography application, e.g., a system for adaptive practice of anatomy (`practiceanatomy.com`).

The adaptive behavior is fundamentally based on learner modeling. For learner modeling we use the Elo rating system. This model was originally developed for rating chess players (Elo, 1978), and only recently has it been used in educational systems. The Elo rating systems combines good predictive accuracy with simplicity and efficient of implementation. These aspects are often neglected in research papers, but are important for realistic applications of learner modeling.

We apply the Elo rating system for prior knowledge estimation and in combination with aspects of Performance factor analysis for current knowledge estimation. Our exploration of more complex models shows that they improve predictive accuracy, but only slightly. In an online educational system, the basic variants of the learner models we studied are preferable since they provide predictions of sufficient quality and are simple to implement and apply. More complex models are, however, still useful as they can provide additional insight into learner behavior and domain structure.

In this work we focus only on learning facts – the simplest type of knowledge component (Koedinger et al, 2012). For more complex knowledge components (e.g., rules) and domains with more involved structure (e.g, prerequisites among knowledge components), the basic Elo rating system is probably not sufficient. An interesting direction for future work is to explore possible extensions of the Elo rating system for more complex learning domains. On the other hand, it may be interesting to apply techniques developed in the context of complex learning domains to practice of facts. For example, recent methods for affective computing and open learner modeling (Grawemeyer et al, 2017; Long and Aleven, 2017) were evaluated in the context of learning mathematics (equations, fractions). It may be interesting to apply these methods for adaptive practice of factual knowledge.

We use predictions of a learner model to automatically construct questions of a suitable difficulty. For the evaluation of the whole question construction algorithm we performed two randomized trial experiments. The first experiment compares adaptive and random construction of questions. The results show that the adaptive behavior is beneficial (both for engagement and learning) and indicate which aspects of adaptivity are important – adaptive choice of the number of distractors rather than the choice of the target item. The second experiment studies the impact of the target difficulty of questions. The results of our experiments suggest that a suitable success rate is around 65 %. This is in contrast to previous similar research (Lomas et al, 2013; Jansen et al, 2013) that concluded that easier questions were preferable. This difference may have been due to different types of knowledge components. We have also detected differences between in-school and out-of-school usage: students using the system in schools prefer easier questions, which accords with previous literature (Abuhamdeh and Csikszentmihalyi, 2012). Nevertheless, this aspect is usually not studied or taken into account in the development of systems. Generally, the question of optimal difficulty requires further research.

Our evaluation also highlights several other issues deserving more attention. Our results show that learning, short term engagement, and long term engagement may not be aligned. Since all these aspects are important, evaluations should use a multi-criteria approach and study trade-offs between individual aspects of system performance. The evaluation of learning is complicated by attrition bias and the aggregation of results over different contexts of practice (Papoušek et al, 2016b,c). These issues should be studied in more detail not just for this system, but in the evaluation of educational systems in general.

## References

Abuhamdeh S, Csikszentmihalyi M (2012) The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. Personality and Social Psychology Bulletin 38(3):317–330

Barla M, Bieliková M, Ezzeddinne AB, Kramár T, Šimko M, Vozár O (2010) On the impact of adaptive test question selection for learning efficiency. Computers & Education 55(2):846–857

Barnes T (2005) The q-matrix method: Mining student response data for knowledge. In: American Association for Artificial Intelligence 2005 Educational Data Mining Workshop, pp 1–8

Basu S, Biswas G, Kinnebrew JS (2017) Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 17, this issue

Carbonell JR (1970) AI in CAI: An artificial-intelligence approach to computer-assisted instruction. Man-Machine Systems, IEEE Transactions on 11(4):190–202

Conati C, Gertner A, Vanlehn K (2002) Using bayesian networks to manage uncertainty in student modeling. User modeling and user-adapted interaction 12(4):371–417

Corbett A, Anderson J (1994) Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction 4(4):253–278

Csikszentmihalyi M (1991) Flow: The psychology of optimal experience. HarperPerennial New York

De Ayala R (2008) The theory and practice of item response theory. The Guilford Press

Delaney PF, Verkoeijen PP, Spirgel A (2010) Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of learning and motivation 53:63–147

Desmarais MC, Baker RS (2012) A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction 22(1-2):9–38

Desmarais MC, Beheshti B, Naceur R (2012) Item to skills mapping: deriving a conjunctive q-matrix from data. In: Trausan-Matu S, Boyer KE, Crosby ME, Panourgia K (eds) Proc. of Intelligent Tutoring Systems, Springer, LNCS, vol 8474, pp 454–463

Doebler P, Alavash M, Giessing C (2014) Adaptive experiments with a multivariate elo-type algorithm. Behavior Research Methods pp 1–11

Doignon JP, Falmagne JC (1999) Knowledge spaces. Springer

Elo AE (1978) The rating of chessplayers, past and present, vol 3. Batsford London

Gierl MJ, Haladyna TM (2012) Automatic item generation: Theory and practice. Routledge

Gierl MJ, Lai H, Turner SR (2012) Using automatic item generation to create multiple-choice test items. Medical education 46(8):757–765

Glickman ME (1999) Parameter estimation in large dynamic paired comparison experiments. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48(3):377–394

González-Brenes J, Huang Y, Brusilovsky P (2014) General features in knowledge tracing: applications to multiple subskills, temporal item response theory, and expert knowledge. In: Stamper J, Pardos Z, Mavrikis M, McLaren B (eds) Proc. of Educational Data Mining, pp 84–91

Grawemeyer B, Mavrikis M, Holmes W, Gutierrez-Santos S, Wiedmann M, Rummel N (2017) Affective learning: Improving engagement and enhancing learning with affect-aware feedback. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 17, this issue

Herbrich R, Minka T, Graepel T (2006) Trueskill: A bayesian skill rating system. In: Schölkopf B, Platt JC, Hofmann T (eds) Advances in Neural Information Processing Systems, MIT Press, pp 569–576

Jansen BR, Louwerse J, Straatemeier M, Van der Ven SH, Klinkenberg S, Van der Maas HL (2013) The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. Learning and Individual Differences 24:190–197

Karpicke JD, Roediger HL (2007) Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language 57(2):151–162

Käser T, Klingler S, Schwing AG, Gross M (2014) Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In: Micarelli A, Stamper JC, Panourgia K (eds) Proc. of Intelligent Tutoring Systems, Springer, LNCS, vol 9684, pp 188–198

Khajah M, Wing R, Lindsey R, Mozer M (2014a) Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In: Stamper J, Pardos Z, Mavrikis M, McLaren B (eds) Proc. of Educational Data Mining, pp 99–106

Khajah MM, Huang Y, González-Brenes JP, Mozer MC, Brusilovsky P (2014b) Integrating knowledge tracing and item response theory: A tale of two frameworks. In: Proc. of Personalization Approaches in Learning Environments Workshop

Khajah MM, Roads BD, Lindsey RV, Liu YE, Mozer MC (2016) Designing engaging games using bayesian optimization. In: Kaye J, Druin A, Lampe C, Morris D, Hourcade JP (eds) Proc. of CHI Conference on Human Factors in Computing Systems, ACM, pp 5571–5582

Klinkenberg S, Straatemeier M, Van der Maas H (2011) Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. Computers & Education 57(2):1813–1824

Koedinger KR, Corbett A (2006) Cognitive tutors: Technology bringing learning sciences to the classroom. In: Sawyer K (ed) The Cambridge Handbook of the Learning Sciences, Cambridge University Press

Koedinger KR, Corbett AT, Perfetti C (2012) The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science 36(5):757–798

Koren Y, Bell R (2011) Advances in collaborative filtering. Recommender Systems Handbook pp 145–186

Liu C, White RW, Dumais S (2010) Understanding web browsing behaviors through weibull analysis of dwell time. In: Crestani F, Marchand-Maillet S, Chen H, Efthimiadis EN, Savoy J (eds) Proc. of Research and development in information retrieval, ACM, pp 379–386

Lomas D, Patel K, Forlizzi JL, Koedinger KR (2013) Optimizing challenge in an educational game using large-scale design experiments. In: Wendy E Mackay SB Stephen A Brewster (ed) Proc. of Human Factors in Computing Systems, ACM, New York, USA, pp 89–98

Lomas D, Forlizzi J, Poonwala N, Patel N, Shodhan S, Patel K, Koedinger K, Brunskill E (2016) Interface design optimization as a multi-armed bandit problem. In: Kaye J, Druin A, Lampe C, Morris D, Hourcade JP (eds) Proc. of CHI Conference on Human Factors in Computing Systems, ACM, pp 4142–4153

Long Y, Aleven V (2017) Enhancing learning outcomes through self-regulated learning support with an open learner model. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 17, this issue

Mettler E, Massey CM, Kellman PJ (2011) Improving adaptive learning technology through the use of response times. In: Carlson L, Holscher C, Shipley T (eds) Proc. of Conference of the Cognitive Science Society, Cognitive Science Society, pp 2532–2537

Millán E, Loboda T, Pérez-de-la Cruz JL (2010) Bayesian networks for student model engineering. Computers & Education 55(4):1663–1683

Minka TP (2001) A family of algorithms for approximate bayesian inference. PhD thesis, Massachusetts Institute of Technology

Mitkov R, Ha LA, Karamanis N (2006) A computer-aided environment for generating multiple-choice test items. Natural Language Engineering 12(2):177–194

Mostow J, Tobin B, Cuneo A (2002) Automated comprehension assessment in a reading tutor. In: Proc. of ITS 2002 Workshop on Creating Valid Diagnostic Assessments, pp 52–63

Nižnan J, Pelánek R, Řihák J (2014) Using problem solving times and expert opinion to detect skills. In: Stamper J, Pardos Z, Mavrikis M, McLaren B (eds) Proc. of Educational Data Mining, pp 434–434

Nižnan J, Pelánek R, Řihák J (2015) Student models for prior knowledge estimation. In: Santos OC, Boticario JG, Romero C, Pechenizkiy M, Merceron A, Mitros P, Luna JM, Mihaescu C, Moreno P, Ventura AHS, Desmarais M (eds) Proc. of Educational Data Mining, pp 109–116

Nižnan J, Pelánek R, Řihák J (2014) Mapping problems to skills combining expert opinion and student data. In: Hlinený P, Dvorak Z, Jaros J, Kofron J, Korenek J, Matula P, Pala K (eds) Proc. of Mathematical and Engineering Methods in Computer Science, Springer, LNCS, vol 8934, pp 113–124

Papoušek J, Pelánek R (2015) Impact of adaptive educational system behaviour on student motivation. In: Conati C, Heffernan NT, Mitrovic A, Verdejo MF (eds) Proc. of Artificial Intelligence in Education, Springer, LNCS, vol 9112, pp 348–357

Papoušek J, Pelánek R, Stanislav V (2014) Adaptive practice of facts in domains with varied prior knowledge. In: Stamper J, Pardos Z, Mavrikis M, McLaren B (eds) Proc. of Educational Data Mining, pp 6–13

Papoušek J, Pelánek R, Řihák J, Stanislav V (2015) An analysis of response times in adaptive practice of geography facts. In: Santos OC, Boticario JG, Romero C, Pechenizkiy M, Merceron A, Mitros P, Luna JM, Mihaescu C, Moreno P, Ventura AHS, Desmarais M (eds) Proc. of Educational Data Mining, pp 562–563

Papoušek J, Pelánek R, Stanislav V (2016a) Adaptive geography practice data set. Journal of Learning Analytics Submitted, data available at http://www.fi.muni.cz/adaptivelearning/

Papoušek J, Stanislav V, Pelánek R (2016b) Evaluation of an adaptive practice system for learning geography facts. In: Gasevic D, Lynch G, Dawson S, Drachsler H, Rosé CP (eds) Proc. of Learning Analytics & Knowledge, ACM, pp 40–47

Papoušek J, Stanislav V, Pelánek R (2016c) Impact of question difficulty on engagement and learning. In: Micarelli A, Stamper JC, Panourgia K (eds)

Proc. of Intelligent Tutoring Systems, Springer, LNCS, vol 9684

Pardos ZA, Heffernan NT (2010) Modeling individualization in a bayesian networks implementation of knowledge tracing. In: User Modeling, Adaptation, and Personalization, Springer, pp 255–266

Pavlik PI, Anderson JR (2005) Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science 29(4):559–586

Pavlik PI, Cen H, Koedinger KR (2009) Performance factors analysis-a new alternative to knowledge tracing. In: Dimitrova V, Mizoguchi R, du Boulay B, Graesser AC (eds) Proc. of Artificial Intelligence in Education, IOS Press, Frontiers in Artificial Intelligence and Applications, vol 200, pp 531–538

Pavlik Jr P, Bolster T, Wu SM, Koedinger K, Macwhinney B (2008) Using optimally selected drill practice to train basic facts. In: Woolf BP, Aïmeur E, Nkambou R, Lajoie SP (eds) Proc. of Intelligent Tutoring Systems, Springer, LNCS, vol 5091, pp 593–602

Pelánek R (2015) Metrics for evaluation of student models. Journal of Educational Data Mining 7(2)

Pelánek R (2015) Modeling students' memory for application in adaptive educational systems. In: Santos OC, Boticario JG, Romero C, Pechenizkiy M, Merceron A, Mitros P, Luna JM, Mihaescu C, Moreno P, Ventura AHS, Desmarais M (eds) Proc. of Educational Data Mining, pp 480–483

Pelánek R (2016) Applications of the elo rating system in adaptive educational systems. Computers & Education 98:169–179

Qiu Y, Qi Y, Lu H, Pardos ZA, Heffernan NT (2011) Does time matter? modeling the effect of time with bayesian knowledge tracing. In: Pechenizkiy M, Calders T, Conati C, Ventura S, Romero, C, Stamper J (eds) Proc. of Educational Data Mining

van de Sande B (2013) Properties of the bayesian knowledge tracing model. Journal of Educational Data Mining 5(2):1

Schulze KG, Shelby RN, Treacy DJ, Wintersgill MC, Vanlehn K, Gertner A (2000) Andes: An intelligent tutor for classical physics. Journal of Electronic Publishing 6(1)

Tatsuoka KK (1983) Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement 20(4):345–354

Vanlehn K (2006) The behavior of tutoring systems. International Journal of Artificial Intelligence in Education 16(3):227–265

Von Luxburg U (2007) A tutorial on spectral clustering. Statistics and computing 17(4):395–416

Wauters K, Desmet P, Van Den Noortgate W (2011) Monitoring learners' proficiency: Weight adaptation in the elo rating system. In: Pechenizkiy M, Calders T, Conati C, Ventura S, Romero, C, Stamper J (eds) Proc. of Educational Data Mining, pp 247–252

Zirkle DM, Ellis AK (2010) Effects of spaced repetition on long-term map knowledge recall. Journal of Geography 109(5):201–206

**Radek Pelánek** received his Ph.D. degree in Computer Science from Masaryk University for his work on formal verification. Since 2010 his research interests focus on areas of educational data mining and learning analytics. Currently he is the leader of the Adaptive Learning group at Masaryk University and is interested in both theoretical research in user modeling and practical development of adaptive educational systems.

**Jan Papoušek** is a Ph.D. candidate in Computer Science at Masaryk University where he received also his masters degree. He worked in the industry as a developer of a high level analytical language for business intelligence. Currently, he is a member of the Adaptive Learning group at Masaryk University where he focuses on systems providing adaptive practice of factual knowledge. His research interests include educational data mining and evaluation methods of intelligent tutoring systems.

**Jiří Řihák** received his masters degree Mathematics from Masaryk University. Currently, he is a Ph.D. candidate in Computer Science at Masaryk University and he is a member of the Adaptive Learning group where he focuses on adaptive educational systems. His primary interests lie in the areas of machine learning, educational data mining, and image recognition.

**Vít Stanislav** received his master's degree in Computer Science from Masaryk University. As a member of the Adaptive Learning group he participates in the development of several educational systems.

**Juraj Nižnan** received his master's degree in Computer Science from Masaryk University, where he participated in the research of the Adaptive Learning group.