Computers & Education 98 (2016) 169–179

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Applications of the Elo rating system in adaptive educational systems

Radek Pelánek

Masaryk University Brno, Czech Republic

ARTICLE INFO

Article history:
Received 22 May 2015
Received in revised form 24 March 2016
Accepted 29 March 2016
Available online 1 April 2016

Keywords:
Architectures for educational technology system
Interactive learning environments
Applications in subject areas
Student modeling

ABSTRACT

The Elo rating system was originally developed for rating chess players, nowadays it is widely used for ranking players of many other games. The system can be used in educational systems when we interpret student's answer to an item as a match between the student and the item. In this way we can easily dynamically estimate the skill of students and difficulty of items. We provide a systematic overview of different variants of the Elo rating system and their application in education. We compare the Elo rating system to alternative methods and describe a specific case study (an adaptive practice of geography facts) to illustrate the application of the Elo rating system in education. We argue that the Elo rating system is simple, robust, and effective and thus suitable for use in the development of adaptive educational systems. We provide specific guidelines for such applications.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Adaptive educational systems aim to provide students with individualized learning materials and thus to increase efficiency of the learning process. Specific examples of educational systems that incorporate adaptive behaviour and have been shown to increase efficiency of learning are Cognitive Tutor (Ritter, Anderson, Koedinger, & Corbett, 2007) and Math Spring (Arroyo et al., 2014).

In order to achieve adaptivity, systems need to be able to estimate the skill of students and difficulty of educational materials (items, questions, problems). One possible way to obtain skill estimates is to use models from item response theory (IRT). These models were, however, developed in the context of computerized adaptive testing and are typically based on the assumption of a constant skill. Another possibility is to use specialized models for modeling learning (Desmarais & Baker, 2012), e.g., Bayesian knowledge tracing (Corbett & Anderson, 1994) or Performance factor analysis (Pavlik, Cen, & Koedinger, 2009). All these models, however, are not easy to use: they may require calibration on large samples using nontrivial parameter estimation or are hard to use in an online educational system. These aspects make the development of adaptive systems more complicated, time consuming, and expensive.

An interesting alternative to specialized models is the Elo rating system (Elo, 1978). This system was originally proposed for rating chess players; for an overview of the history of chess rating systems see Glickman (1995). The basic principle of the system is simple: each player is assigned a rating, this rating is updated after each match, the update is proportional to the surprisingness of the match result. If a strong player beats a weak player, the results is not surprising and the update is small;

E-mail address: xpelanek@fi.muni.cz.

http://dx.doi.org/10.1016/j.compedu.2016.03.017
0360-1315/© 2016 Elsevier Ltd. All rights reserved.

whereas if the opposite happens, the update is large. The system has been extended in many ways, the most well-known extensions are Glicko (Glickman, 1999) and TrueSkill (Herbrich, Minka, & Graepel, 2006). Nevertheless, even the basic Elo rating system is still widely used for rating many contests, online games, and sports, e.g., football (Hvattum & Arntzen, 2010). The system provides a foundation for adaptive behaviour of gaming sites as it can automatically select opponents of similar skill. Moreover, the Elo rating system can be used not just for rating players. It has been used for rating patterns in the game of Go (Coulom, 2007), eliciting user preferences (Hacker & Von Ahn, 2009), assessing security and vulnerability risks (Pieters, van der Ven, & Probst, 2012), or for ranking posts in online forums (Das Sarma, Das Sarma, Gollapudi, & Panigrahy, 2010).

We can use the Elo rating system in educational systems if we interpret a solution attempt as a match between a student and an item. This approach provides an easy way to estimate the skill of students and difficulty of items. The use of the Elo rating system offers many advantages: it is a simple system, which is easy to implement in educational systems; it has only a small number of parameters that need to be set; it can be easily used in an online setting; and it also provides comparable performance to more complex models.

The Elo rating system has already been used in education. Previous studies, however, describe the use of the Elo rating system only briefly or study only a particular aspect of the system. Several studies discuss basic properties of the system: Brinkhuis and Maris (2009) provide analysis of statistical properties of the model; Antal (2013) used simulated data to compare parameter estimates by the Elo rating system with IRT estimates; Pelánek (2014) compares the model with other student models and discusses modification of the system for problem solving times.

Other studies focused on evaluations and applications. Wauters, Desmet, and Van Den Noortgate (2011) provide evaluation of an extension of the Elo rating system over real (but rather small) student data. Wauters, Desmet, and Van Den Noortgate (2012) focus on estimation of item difficulty (for items from computerized adaptive testing) and compares estimates by an IRT model, the Elo rating system, proportion correct, learner feedback, and user judgment. Klinkenberg, Straatemeier, and Van der Maas (2011) propose an extension of the Elo rating system and apply it in an online system for adaptive practice of mathematics. Papoušek, Pelánek, and Stanislav (2014) and Nižnan, Pelánek, and Řihák (2015) describe an application of the Elo rating system for adaptive practice of facts, particularly for prior knowledge estimation. Attali (2014) uses a ranking method similar to the Elo rating system for grading constructed responses.

In this paper we provide a systematic survey of those aspects and extensions of the Elo rating system that are relevant in educational settings. We also compare the Elo rating system to alternatives using an experiment with simulated data, and describe a specific case study of an application of the Elo rating system in a widely used educational system. Finally, we provide guidelines for the use of the Elo rating system in the practical development of adaptive educational systems.

2. The Elo rating system and its variants

In this section we describe in detail the Elo rating system and its variants relevant for student modeling.

2.1. The basic Elo rating system

The core principle of the Elo rating system (Elo, 1978) can be summarized as follows. For each player i we have a rating estimate θ_i. Let $R_{ij} \in \{0,1\}$ be the results of a match between players i, j. The expected probability that the player i wins is given by the logistic function with respect to the difference of estimated ratings (Fig. 1 left):

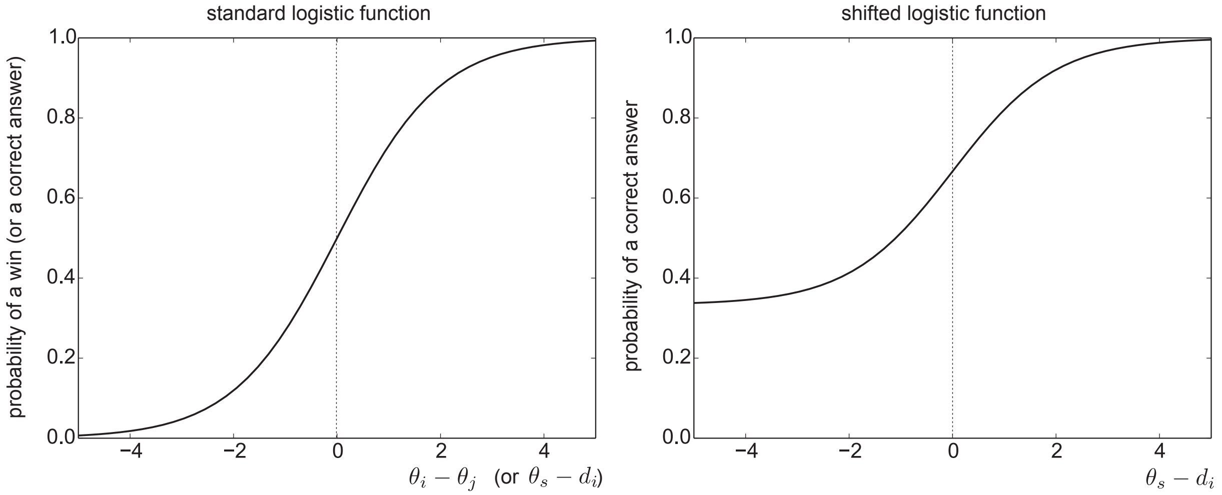

Fig. 1. The logistic function transforms the difference between opponents ratings (student skill and item difficulty) into a probability of a win (a correct answer). The shifted logistic function takes into account guessing (with the probability 1/3 in the example).

$$P(R_{ij} = 1) = 1\Big/\left(1 + e^{-(\theta_i - \theta_j)}\right)$$

Based on the result of a match the rating estimates are updated using the following update rule (K is a constant specifying sensitivity of the estimate to the last attempt):

$$\theta_i := \theta_i + K\left(R_{ij} - P\left(R_{ij} = 1\right)\right)$$

The used probability function can be seen as a reparametrization of the Bradley-Terry model for paired comparisons (Bradley & Terry, 1952). Under the Bradley-Terry model if two objects have true ratings (or preferences) π_1,π_2, then the first object is preferred (will rank higher in a comparison) with probability $\pi_1/(\pi_1 + \pi_2)$. Instead of the logistic function it is possible to use a normal cumulative distribution, which corresponds to the Thurstone-Mosteller model for paired comparisons (Glickman, 1995). Since the logistic function and normal cumulative distribution function have very similar shapes, the difference between these two variants is in practice not essential. Current realizations of the Elo rating system typically use the logistic function because it is easier to use in practical applications.

The original Elo rating system for chess uses a specific rescaling of the standard logistic function: $P(R_{ij} = 1) = 1/(1 + 10^{-(\theta_i - \theta_j)/400})$. This rescaling (i.e., the base 10 instead of e and the constant 400) is used for historical reasons (compatibility with the preceding system for rating chess players) and does not have any impact on the underlying principles. In this paper we consistently use the standard logistic function. The original Elo rating system also performs update of the skill level after a tournament (not after each match) and there are several extensions for modeling draws and the first-move advantage (Glickman, 1995). These aspects are not important for applications of the rating system in the educational context and thus we do not discuss them in more detail.

2.2. The Elo rating system in an educational setting

To apply the Elo rating system in the context of educational systems, we interpret a student's answer to an item as a match between the student and the item. In this context it is natural to slightly change the notation while keeping the basic principle the same. We denote skill of a student s as θ_s, difficulty of an item i as d_i, and the correctness of an answer of a student s on an item i as $\text{correct}_{si} \in \{0,1\}$. The probability of a correct answer is given by the logistic function with respect to the difference between skill and difficulty (Fig. 1 left):

$$P(\text{correct}_{si} = 1) = 1\Big/\left(1 + e^{-(\theta_s - d_i)}\right)$$

The skill and difficulty estimates are updated as follows:

$$\begin{aligned} \theta_s &:= \theta_s + K \cdot (\text{correct}_{si} - P(\text{correct}_{si} = 1)) \\ d_i &:= d_i + K \cdot (P(\text{correct}_{si} = 1) - \text{correct}_{si}) \end{aligned}$$

Initial values of θ_s and δ_i parameters are set to 0.

In this formulation the Elo rating system is closely related to the Rasch model (one parameter model) used in IRT. These two approaches share the same functional form of the model for predicting the probability of a correct answer, they differ only in the parameter estimation procedure. The standard approach in item response theory is to use maximum likelihood based methods (De Ayala, 2008), whereas the Elo rating system uses the above presented heuristic equations. They also differ in their basic assumptions. Basic IRT models assume that a student's skill is constant (such models are typically used in testing and it is assumed that there is no change of the skill during a test), whereas the Elo rating system is designed to track changing skill levels.

The advantage of the Elo rating system is that it can be easily modified for different situations. For example, in education we often use multiple choice questions (typically with 3–6 choices), where students have a significant chance of answering correctly just by guessing. We can easily incorporate this guessing behaviour into the Elo rating system. It is sufficient to change only the predicted probability of a correct answer by using the shifted logistic function, i.e., for question with k options the probability of a correct answer becomes (Fig. 1 right):

$$P(\text{correct}_{si} = 1) = 1/k + (1 - 1/k)\Big/\left(1 + e^{-(\theta_s - d_i)}\right)$$

Note that the main advantage of the Elo rating system is not expressive power, but its simplicity and versatility. Situations like multiple-choice questions or dynamically changing skill can be modeled by many methods, e.g., there exist suitable extensions of IRT models (Drasgow, Levine, Tsien, Williams, & Mead, 1995; Wang, Berger, & Burdick, 2013). The Elo rating system, however, offers a flexible student modeling approach, which can be easily modified.

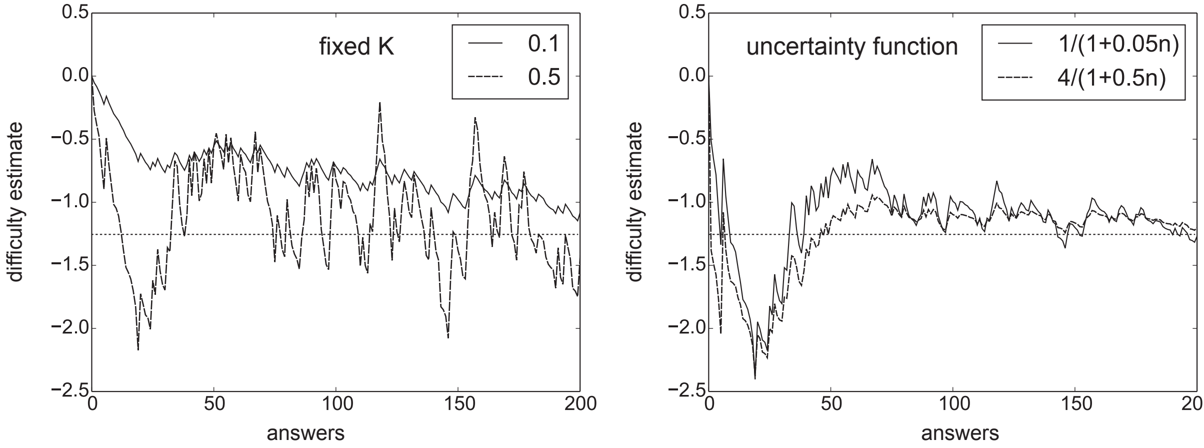

Fig. 2. Difficulty estimates for two items with different values of the *K* parameter (left) and different uncertainty functions (right). The dotted line represents the ground truth value of the difficulty parameter.

As noted above, the original Elo rating system is applied for a summary score of a tournament, i.e., the Elo rating system is not restricted to binary results. In education it can thus be used also with partial credit modeling of answers in the presence of hints (Wang & Heffernan, 2013); or with problem solving times for problems where time is the only measure of performance (Pelánek, 2014; Vaněk, 2014), e.g., time to find a correct solution of a Sudoku puzzle or to write a correct program which outputs a list of primes. In the case of problem solving times we have for each student a skill parameter θ_s and for each problem a difficulty parameter d_p. When a student s solves a problem p in the logarithm[1] of time t_{sp} we update parameter estimates as follows:

$$\theta_s := \theta_s + K\big(E(t|s,p) - t_{sp}\big)$$
$$d_p := \theta_p + K\big(t_{sp} - E(t|s,p)\big)$$

where $E(t|s,p)$ is the expected logarithm of a solving time for the student s and the problem p given as $E(t|s,p)=d_p-\theta_s$.

It is also possible to combine both the correctness of answers and response times into the update rule. For example, Klinkenberg et al. (2011) use a “high speed, high stakes” rule, where the answer is assigned a score based on the correctness correct_{si} and the response time t_{si} as follows (d_i is a time limit, a_i is a discrimination parameter):

$$S_{si} = (2\text{correct}_{si} - 1)(a_i d_i - a_i t_{si})$$

Under this rule stakes are high when a student answers quickly; with increasing response time the score goes towards zero (for both correct and incorrect answers). An appropriate way to include response times depends on the user interface of a particular system. The “high speed, high stakes” rule may be appropriate for systems showing the timing information and deadlines in the user interface — this is the case in the Math Garden software, where the rule has been applied (Klinkenberg et al., 2011). In other systems the response time may be measured without any explicit notification about its role to users. In such cases the response time may still contain interesting information (Papoušek, Pelánek, Řihák, & Stanislav, 2015), but a different scoring rule needs to be applied, see for example Řihák (2015) for a specific proposal.

2.3. Uncertainty

The value of the constant K in the update rule determines the behaviour of the system. If K is small, the estimation converges too slowly, if K is large, the estimation is unstable as it gives too large a weighting to the last few attempts, see Fig. 2 (left) for illustration. Previous work in student modeling used $K = 0.4$ (Antal, 2013; Wauters et al., 2012).

An intuitive improvement used in most Elo extensions[2] is to use an “uncertainty function” instead of a constant K. When we have a new player (student), the skill estimate is highly uncertain and thus the update should be large. As we get more data the size of the update should get smaller. A principled way to realize this idea is to use a Bayesian approach to parameter estimation. A well-known extension in this direction is the Glicko system (Glickman, 1999). It models prior skill by the normal distribution and uses numerical approximation to represent the posterior by the normal distribution and to perform the update of the mean and standard deviation of the skill distribution using closed form expressions. Another widely used extension based on a Bayesian approach is TrueSkill (Herbrich et al., 2006), which further extends the system to allow team competitions. There also exists Bayesian rating system for time-varying skill (Coulom, 2008).

---

[1] The reason for the use of the logarithm of time is that problem solving times are log-normally distributed (Jarušek & Pelánek, 2012).

[2] In fact even the basic application of the Elo rating system in chess includes variable choice of K depending on the rating of a player. The specific choice of the K parameter has been subject of discussions (Sonas, 2002).

The approach used by Glicko and TrueSkill is, however, difficult to modify for new situations, e.g., when we want to use the shifted logistic function for modeling answers to multiple-choice questions. Such modifications significantly complicate derivation of equations for numerical approximation. An alternative approach to numerical approximation is to use particle filtering, i.e., to represent the skill distribution discretely by a vector of values (Nižnan et al., 2015). This approach has larger time and memory requirements than the closed form expression approach, but it can be easily used with a modified likelihood function.

From another point of view, the Elo rating system is closely related to stochastic gradient descent (SGD) Bottou (2010). The update rule in the Elo rating system corresponds to the update of parameters along the error gradient. The main difference is that SGD typically iterates over data several times, whereas the Elo rating system uses just one pass. The parameter K corresponds to the learning rate used in SGD. Similarly to the described use of uncertainty function, SGD often uses learning rate that decreases during time. Previous research identified principled ways to determine the change of learning rate, e.g., the AdaGrad algorithm (Duchi, Hazan, & Singer, 2011).

Experience suggests that in the case of student modeling, principled approaches to modeling uncertainty actually do not bring significant advantages – it may be sufficient to model uncertainty in a simpler, pragmatic way (Nižnan et al., 2015). Several works have explored this approach:

- Papoušek et al. (2014) and Nižnan et al. (2015) use an uncertainty function $U(n)=a/(1+bn)$, where n is the number of answered questions and a, b are meta-parameters fitted to data ($a=1$, $b=0.05$),
- Wauters et al. (2011) use an uncertainty function $U(n)=w_0/(1+ae^{bn})$, where n is the number of answered questions and w_0, a, b are meta-parameters fitted to data ($w_0=0.2$, $b=50$ and a between 0.01 and 0.15),
- Klinkenberg et al. (2011) use uncertainty initialized as $U=1$ and updated after each answer as follows: $U:=U-1/40+1/30D$ where D is a time from a previous attempt (in days).

Fig. 2 shows comparison of difficulty estimates with two different values of constant K and with two different uncertainty functions. The estimation was done for simulated data (described in more detail in Section 3.1), i.e., we can compare estimates to the ground truth item difficulty. The figure shows that using an uncertainty function we can quickly get coarse estimates, which are then fine-tuned.

In educational applications there is often an asymmetry in the number of available answers for items and students. In most systems we expect that (in the long run) each item is answered by many students, whereas for students the number of answer is typically smaller (often by orders of magnitude). It may thus be useful to use different uncertainty function for items and for students.

2.4. Multivariate extension

Another extension of the Elo rating system is to measure several related skills at once. Instead of a single skill for a student, we measure several correlated skills. This extension is not very relevant for application in games, but it was previously studied in the context of adaptive experiments in psychology (Doebler, Alavash, & Giessing, 2014) and for modeling prior knowledge of geography facts (Nižnan et al., 2015). Although the multivariate extension has not been used in mainstream applications of the Elo rating system, it is particularly relevant for educational applications where we often work with multiple knowledge components (Koedinger, Corbett, & Perfetti, 2012), e.g., in mathematics such knowledge components may be “fractions”, “polynomials”, and “trigonometric functions”.

Let us consider the following (simplified, but quite common) situation: each item belongs to a single knowledge component and skills for individual knowledge components are correlated, i.e., answers directly relevant for one knowledge component provide indirect information about other skills as well. The multivariate extension is used as follows. Based on available data we compute correlations c_{ij} between knowledge components. For each knowledge component i we have a student skill parameter θ_{si}. After a student s answers a question belonging to a knowledge component i we update estimates of all skills j as follows:

$$\theta_{sj} := \theta_{sj} + c_{ij}K(\text{correct} - P(\text{correct} = 1))$$

Previous applications of multivariate extensions (Doebler et al., 2014; Nižnan et al., 2015) combine this basic approach with modeling of the uncertainty (instead of K they use one of the above discussed uncertainty functions) and with the use of a global skill.

2.5. Extension with learning

As noted above, the basic difference between the Elo rating system and the basic IRT models is that IRT models typically assume a fixed skill, whereas the Elo rating system can track changing skill. An advantage of the Elo rating system is that by setting the parameter K (respectively the uncertainty function) we can "tune" the system for a particular situation. If the skill is expected to change quickly, K should stay large. If the skill is expected to be (nearly) constant, we can use an uncertainty function which approaches zero with an increasing number of attempts – in an educational setting this may be appropriate for example in the case of estimation of prior knowledge (Papoušek et al., 2014).

An important difference between applications of the Elo rating system in games and education is an asymmetry between students and items in education. For items we expect their difficulty to be approximately constant. In some cases it may be useful to track changes in difficulty, e.g., in geography general knowledge of some places may temporarily change due to their presence in media (Guinea during the Ebola epidemic). But such cases are exceptions and changes in difficulty are not expected to be large. On the other hand, changes in student skill are expected – after all, that is the aim of educational systems.

In many educational applications there is also an asymmetry between correct and incorrect answers, since students learn just by an exposure to an item. An answer to an item is not only evidence of student's knowledge, but also an opportunity for learning. In such situations we need to perform different updates for correct and incorrect answers. Papoušek et al. (2014) and Pelánek (2015) propose such modification of the Elo rating system under the name "Performance Factor Analysis Extended/Elo" (PFAE). This name reflects the fact that the modification can be viewed from different perspective as an extension of Performance factor analysis (PFA), which is a student modeling approach based on the logistic function (Pavlik et al., 2009). The basic principle of the PFAE system is that in the update rule instead of a single constant K we use different constants γ,δ for correct and incorrect answers:

$$\theta_{si} := \begin{cases} \theta_{si} + \gamma \cdot (1 - P(\text{correct}_{si} = 1)) & \text{if the answer was correct} \\ \theta_{si} + \delta \cdot P(\text{correct}_{si} = 1) & \text{if the answer was incorrect} \end{cases}$$

3. Difficulty estimates: comparison of Elo, IRT, and proportion correct

Wauters et al. (2012) provide a comparison of several methods for estimating difficulty of items: the proportion of correct answers, IRT models, the Elo rating system, and human judgment methods (e.g., student feedback, expert rating). Their conclusion is that with a sample size of 200 students the data-driven methods provide reliable and highly correlated estimates. One interesting aspect of this analysis is that even the simple proportion correct method provides good estimates. This result casts doubts on the practical usefulness of more complex methods like IRT or the Elo rating systems.

To explore this issue we analyze the relation among the proportion correct method, IRT models, and the Elo rating system using experiments with simulated data. We illustrate that the proportion correct method provides good estimates when the selection of items presented to students is random. In the case of adaptive selection of items, however, the proportion correct method fails. The Elo rating system is able to achieve in both cases very similar performance as the IRT approach, which is computationally much more demanding.

3.1. Experimental setting

We generated simulated data using the Rasch model with skills and difficulties generated from the standard normal distribution, i.e., $\theta_s \sim \mathcal{N}(0,1), d_i \sim \mathcal{N}(0,1)$; an answer of a student s on an item i is a Bernoulli random variable with $p = 1/(1 + e^{-(\theta_s - d_i)})$. We assume that students solve only a portion of available items; for selection of items we consider both random selection and adaptive selection. In the case of adaptive selection a student is presented with items having the closest difficulty to student's skill. We also consider transition between these two extremes (a "random factor" between 0 and 1 sets the weight of randomness in the selection).

From the generated data we want to recover the skill and difficulty parameters. To evaluate the quality of estimated parameter values we measure the correlation with the ground truth values (using the Spearman correlation coefficient). For estimation of parameters we compare three methods:

- The proportion of correct answers (it produces estimates on a different scale than the ground truth values, but this does not impact the Spearman correlation coefficient).
- Joint maximum likelihood estimation (JMLE) – the standard iterative procedure for estimating parameters of the Rasch model (De Ayala, 2008).
- The Elo rating system with the uncertainty function $U(n) = a/(1 + bx)$. We use parameters $a = 4$, $b = 0.5$; these values were optimized using grid search. The performance of the system is quite stable and the precise choice of parameter values is not fundamental to the presented results.

Note that all these methods estimate one parameter per item, i.e., they do not differ in model complexity (number of free model parameters).

3.2. Results

In the case that items are selected randomly all three methods provide very similar estimates and even the simple proportion correct method gives good prediction of item difficulty — this result is in line with conclusions by Wauters et al. (2012). However, when items are selected adaptively, results are quite different.

Fig. 3 shows results for scenarios with (partially) adaptive selection of items. For the reported experiments we used 150 items and the number of items solved by each student is chosen randomly (uniformly) between 10 and 100. The graph on the left shows the quality of estimated item difficulty parameters depending on the number of students (for fully adaptive selection of items). The graph on the right shows the quality of estimated item difficulty parameters for 200 students with different degree of randomness in the item selection process.

The results clearly demonstrate that in the case of adaptive selection of items the proportion correct method gives poor estimates. The quality of these estimates moreover does not improve with the increasing number of students. The Elo rating system, still a rather simple and efficient procedure, gives nearly the same estimates as joint maximum likelihood, which is a much more computationally demanding procedure (unsuitable for online updating of parameters). Also note that this scenario is optimistic for the JMLE method, since the simulated data adhere to the constancy of skill assumption, whereas real data typically contain at least some variability.

We have presented analysis of item difficulties, analogical results hold also for student skills — the Elo rating systems provides similar estimates as the JMLE estimation, the difference with respect to proportion correct is pronounced particularly for adaptive choice of items. Once the Elo rating system has good estimates of item difficulties, it can estimate the skill of new students rather quickly — under the simulation settings just 10 answers per student are sufficient to produce reasonable skill estimates (correlation 0.8 with the ground truth).

4. Case study: adaptive practice of geography

As an illustration of an application of the Elo rating system in education we describe an online adaptive system Outline Maps (outlinemaps.org) for practice of geography facts (e.g., names and location of countries, cities, mountains). Geography is a typical domain with widely varied prior knowledge of individual facts — Fig. 4 visualizes significant differences in the prior knowledge of African countries (of mostly Czech students using the system). The system contains over 1200 geographical items, has been already used by more than 100,000 students, who answered in total more than 15 million questions.

Similar architecture (including the use of the Elo rating system) has been used for adaptive practice of facts in other domains: anatomy for medical students (practiceanatomy.com), practice of driving licence test (autoskolachytre.cz), Czech grammar (umimecesky.cz), animal names (Drábek, 2016), vocabulary learning (Mikula, 2015), or early reading (Kühpastová, 2015). The Elo rating system has also been applied for practicing basic mathematical operations in the Math Garden system (Klinkenberg et al., 2011) and the MatMat system (Řihák, 2015). Applications for practice of mathematics use extensions of the Elo rating system that incorporate response time (it is an important indicator of knowledge in the case of basic arithmetic).

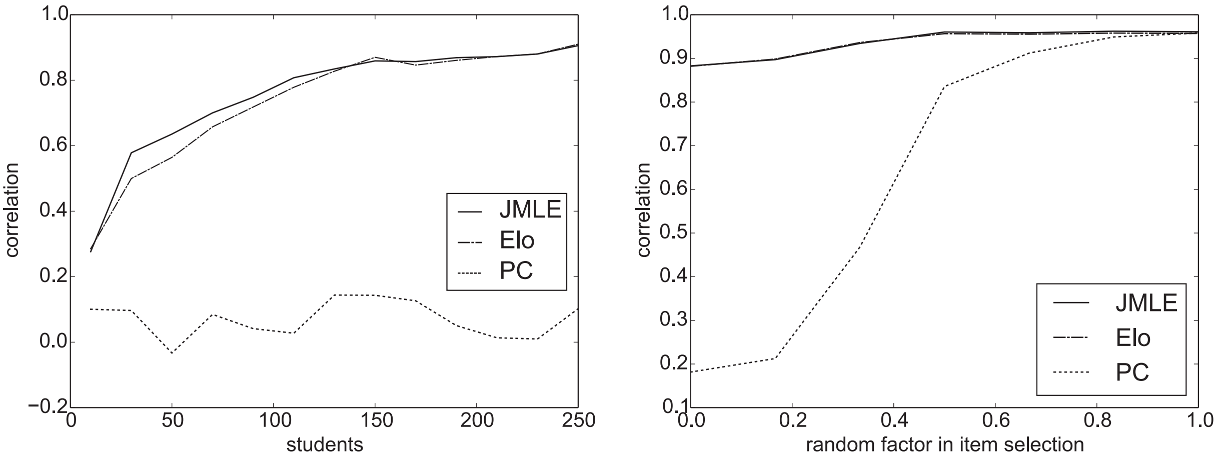

Fig. 3. Correlation between generated and estimated difficulty parameters (JMLE = joint maximum likelihood estimation, Elo = the Elo rating system, PC = the proportion of correct answers).

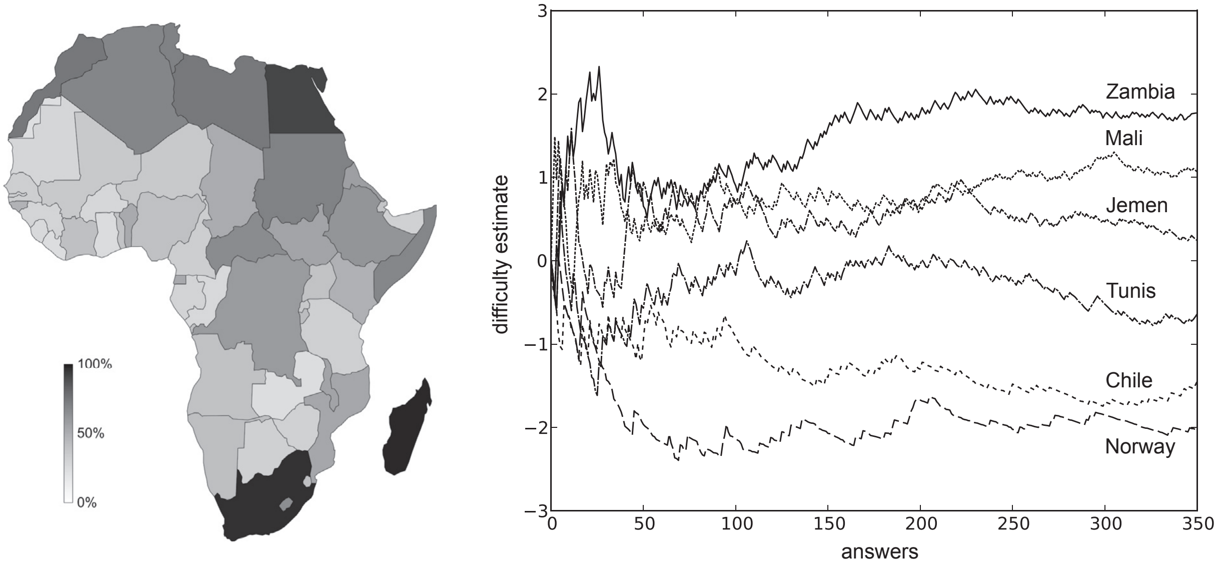

Fig. 4. Left: A map of Africa colored by prior knowledge of countries, the shade corresponds to the probability of a correct answer for an average user of outlinemaps.org. Right: Difficulty of countries – estimates by the Elo rating system.

4.1. System structure

The Outline Maps system estimates student knowledge and uses it to adaptively select questions of suitable difficulty (Papoušek et al., 2014). The system uses a target success rate (e.g., 75%) and adaptively selects questions in such a way that students' achieved performance is close to this target (Papoušek & Pelánek, 2015). The system uses open questions (e.g., “Where is Egypt?”) and multiple-choice questions with 2–6 options (e.g., “What is the name of the highlighted country?”). Students answer questions with the use of an interactive outline map. Students can also access a visualization of their knowledge using an open learner model.

The system has a modular structure consisting of three independent steps (Papoušek et al., 2014):

1. *Estimation of prior knowledge.* Estimation of the probability that a student s knows an item i before the first question about this item. The estimate is based on previous answers of the student s and on answers of other students about the item i.
2. *Estimation of current knowledge.* Estimation of the probability that a student s knows an item i based on the estimation of prior knowledge and a sequence of previous answers of the student s on questions about the item i.
3. *Question construction.* Construction of a suitable question for a student based on the estimation of knowledge and the recent history of answers.

The implementation of the first two steps uses some variant of the Elo rating system.

4.2. Estimation of prior skill

For estimating prior skill we make a simplifying assumption that both students and items are homogeneous – we assume that we can model students' overall prior knowledge in the domain by a one-dimensional parameter. Currently the users of the system are mostly from the Czech Republic. For a system with students with different backgrounds it would be necessary to extend the used model or to make predictions for different groups of students independently. The homogeneity of items is addressed below.

We model students' prior knowledge of geography as a single global skill θ_s; difficulty of geography items is modeled by a difficulty parameter d_i. For estimation of these parameters we use the basic Elo rating system with the uncertainty function $U(n)=a/(1+bn)$, where n is the number of previous answers and a, b are meta-parameters. Using a grid search we have determined optimal values $a=1$, $b=0.05$. This exact choice of parameter values is not important, many different choices of a,b provide very similar results. Fig. 4 shows difficulty estimates for several countries. The figure illustrates how the Elo rating system with an uncertainty function provides both fast coarse estimates after few answers and stability in the long run. Estimates by the Elo rating system with the uncertainty function $U(n)=a/(1+bn)$ are also very similar to estimates obtained by a more principled Bayesian approach to modeling uncertainty (Nižnan et al., 2015).

Although the assumptions in this context are closer to the assumptions of the Rasch model (the global skill and the difficulty of items are rather constant), the Elo rating system is much more suitable for an online application. Similarly to the above reported experiments with simulated data, the used Elo rating system provides very similar estimates as the joint maximum likelihood estimation for the Rasch model (Papoušek et al., 2014). At the same time the Elo rating system is much faster and more suitable for online application than the iterative JMLE procedure.

As stated above, the model uses an assumption of a single global skill. We have tested this assumption by computing the skill for independent subsets of items (countries from different continents) and then checking the correlation between the obtained skill – resulting correlations are around 0.6. Given that there is some intrinsic noise in the data and that the skills are estimated from a limited amount of questions, this is quite high correlation. We have explored the use of extended models with a hierarchy of skills (e.g., separate skills for individual continents). These models are able to improve accuracy of predictions, the improvement is statistically significant, but small (Nižnan et al., 2015). In this context we can also use the multivariate extension of the Elo rating system. This model again leads to only small improvement in prediction accuracy, but the correlations used within the system can be of interest themselves (Nižnan et al., 2015).

4.3. Estimation of current skill

We now turn to the estimation of students' current knowledge, i.e., knowledge estimates based on repeated answers about a particular item. The input data for this estimation are an estimate of prior knowledge (provided by the above described model) and a history of previous attempts, i.e., a sequence of previous answers (the correctness of answers, information about question types, and timing information).

In this setting we use the extension of the Elo rating system with learning (i.e., asymmetric update in the case of correct and incorrect answer). It is also useful to incorporate forgetting, particularly when we are modeling declarative knowledge (in contrast to procedural skills, which are the typical focus of student modeling). We can do this by using time from previous attempt. We can model short term memory effect in the following way: the skill is “locally” increased by $f(t)$ where t is the time (in seconds) between attempts. For this function we used either a fixed function $f(t)=w/t$ with parameter $w = 80$ fitted to data (Papoušek et al., 2014) or a generic staircase function fitted to data (Pelánek, 2015). It should be possible to further improve the model by a more thorough treatment of forgetting and spacing effects, e.g., by incorporating some aspects of the ACT-R model (Pavlik & Anderson, 2005).

Papoušek et al. (2014) provide a comparison of the extended Elo rating system (PFAE) with standard student models: BKT (Corbett & Anderson, 1994) and PFA (Pavlik et al., 2009). The results show that the PFAE models achieves better predictive accuracy than both the standard PFA and BKT models. The results also show that the consideration of timing information further improves the performance of models.

5. Discussion and guidelines

To conclude, we provide a summary discussion of the Elo rating system and guidelines for its application in the practical development of adaptive educational systems.

5.1. When is the Elo rating system applicable?

In educational applications the Elo rating system is suitable mainly for adaptive practice or low stakes testing. The system provides reasonable estimates that are sufficient for guiding adaptive behaviour, but does not provide statistical guarantees on estimated skills (as opposed to well calibrated IRT models used in computerized adaptive testing).

The Elo rating system is particularly attractive when we want to build a reasonably behaving system quickly and cheaply. It allows us to get adaptive behaviour without expensive expert input, since the system can estimate difficulty of items (questions, problems) and skills of users just from data. Of course, to be able to learn from data, it needs to have enough data available. As our analysis and previous experience shows, the system needs at least 100 students to get good estimates of item difficulty.

Typical examples of potential applications of the Elo rating system are in domains with simple structure: learning of factual knowledge, foreign language vocabulary, or practice of basic skills (e.g., arithmetic). More complex domains, involving for example prerequisite relations among skills, require more involved student modeling approaches (Desmarais & Baker, 2012) or a novel extension of the Elo rating system.

The Elo rating system is attractive particularly for medium sized target groups, e.g., elementary and high school knowledge for speakers of smaller languages or company vocational training. In these cases it is usually not feasible to construct sophisticated educational systems by experts, particularly in the context of fast changes of the technological landscape. At the same time these target groups are sufficiently large to enable the ‘learning from data’ approach. The use of the Elo rating system provides a way to implement adaptive behaviour quickly and cheaply.

5.2. What are the advantages of the Elo rating system?

To appreciate advantages of the Elo rating system it is useful to compare it to alternative approaches to skill estimation. Basic item response theory models assume a constant skill. These models are thus applicable only for short tests (where we do not expect learning) or for modeling very “coarse-grained” skills (where the overall learning is slow). We can model finedgrained skill and learning over short term using models like Bayesian Knowledge Tracing or Performance Factor Analysis. These models, however, make specific assumptions about learning. The Elo rating system is flexible – it can model changing skill, but does not make fixed assumptions about the nature of learning. Since the Elo rating system does not make any specific assumptions, it should not be expected to bring an optimal performance for a particular situation. But it can be easily applied in a wide range of situations and provides reasonable accuracy. The system is also easy to modify for different situations, e.g., the use of multiple choice options or incorporation of timing information.

A practically important advantage of the Elo rating system is that it is very simple to implement. The basic Elo rating system needs to store just a single parameter for each student and each item; these parameters are updated after each answer using simple equations. The system requires us to set only the parameter K (respectively the uncertainty function) – other student models typically have more parameters and require calibration or complex parameter fitting.

The use of the Elo rating system allows easy updating of the content of the education system. When we change or add a new item into the system, we just (re)set its difficulty to 0 and the system learns the item difficulty from students' answers. Overall, the application of the Elo rating system is cheap as it requires expert input neither for domain knowledge nor for implementation.

5.3. Which variant of the Elo rating system is preferable?

The reported experience suggest that the basic version of the Elo rating system extended with a simple uncertainty function $U(n)=a/(1+bn)$ provides a good starting point (e.g., with values $a=1$, $b=0.05$; once enough data are collected these values can be easily fitted using a grid search). More complex variants can improve performance, but the improvement is often small and may not be worth the increase in implementation complexity. It is useful to focus only on extensions which are particularly suited for a specific educational application (e.g., incorporation of response times in the case of mathematics).

References

Antal, M. (2013). *On the use of elo rating for adaptive assessment* (vol. 58). Studia Universitatis Babes-Bolyai. Informatica.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education, 24*, 387–426.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement, 74*, 795–808.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 324–345.

Brinkhuis, M. J., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems*. Measurement and Research Department Reports (Rep. No. 2009-1). Arnhem: Cito.

Corbett, A., & Anderson, J. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Coulom, R. (2007). Computing elo ratings of move patterns in the game of go. *ICGA Journal, 30*, 198–208.

Coulom, R. (2008). Whole-history rating: a bayesian rating system for players of time-varying strength. In *Computers and games* (pp. 113–124). Springer.

Das Sarma, A., Das Sarma, A., Gollapudi, S., & Panigrahy, R. (2010). Ranking mechanisms in twitter-like forums. In *Proc. of the third ACM international conference on Web search and data mining, ACM* (pp. 21–30).

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

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*, 9–38.

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

Drábek, J. (2016). *Adaptabilní systém pro procvičování faktografických znalostí z biologie*. Master's thesis. Faculty of Informatics, Masaryk University Brno.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143–166.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research, 12*, 2121–2159.

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

Glickman, M. E. (1995). Chess rating systems. *American Chess Journal, 3*, 102.

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

Hacker, S., & Von Ahn, L. (2009). Matchin: eliciting user preferences with an online game. In *Proc. of the SIGCHI conference on human factors in computing systems, ACM* (pp. 1207–1216).

Herbrich, R., Minka, T., & Graepel, T. (2006). Trueskill: a bayesian skill rating system. In *Advances in neural information processing systems* (pp. 569–576).

Hvattum, L. M., & Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting, 26*, 460–470.

Jarušek, P., & Pelánek, R. (2012). Analysis of a simple model of problem solving times. In *Proc. of intelligent tutoring systems* (pp. 379–388). Springer.

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*, 1813–1824.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*, 757–798.

Kühpastová, M. (2015). *Adaptabilní systém pro procvičování čtení*. Master's thesis. Faculty of Informatics, Masaryk University Brno.

Mikula, D. (2015). *Adaptabilní systém pro procvičování anglické slovní zásoby*. Master's thesis. Faculty of Informatics, Masaryk University Brno.

Nižnan, J., Pelánek, R., & Řihák, J. (2015). Student models for prior knowledge estimation. In *Proc. of educational data mining* (pp. 109–116).

Papoušek, J., & Pelánek, R. (2015). Impact of adaptive educational system behaviour on student motivation. In *Proc. of artificial intelligence in education* (pp. 348–357).

Papoušek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *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 *Proc. of educational data mining* (pp. 562–563).

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cognitive Science, 29*, 559–586.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of artificial intelligence in education* (pp. 531–538). IOS Press.

Pelánek, R. (2014). Application of time decay functions and Elo system in student modeling. In *Proc. of educational data mining* (pp. 21–27).

Pelánek, R. (2015). Modeling students' memory for application in adaptive educational systems. In *Proc. of educational data mining* (pp. 480–483).

Pieters, W., van der Ven, S. H., & Probst, C. W. (2012). A move in the security measurement stalemate: Elo-style ratings to quantify vulnerability. In *Proc. of the 2012 workshop on new security paradigms, ACM* (pp. 1–14).

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: applied research in mathematics education. *Psychonomic bulletin & review, 14*, 249–255.

Řihák, J. (2015). Use of time information in models behind adaptive system for building fluency in mathematics. In *Educational data mining* (pp. 642–644). Doctoral Consortium.

Sonas, J. (2002). *The sonas rating formula – Better than elo?* Chessbase news.

Vaněk, L. (2014). *Elo systém a modelování času řešení*. Master's thesis. Masaryk University Brno.

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics, 7*, 126–153.

Wang, Y., & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: using continuous versus binary nodes. In *Prof. of artificial intelligence in education* (pp. 181–188). Springer.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2011). Monitoring learners' proficiency: weight adaptation in the elo rating system. In *Proc. of educational data mining* (pp. 247–252).

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: an auspicious collaboration between data and judgment. *Computers & Education, 58*, 1183–1193.