Evaluation of an Adaptive Practice System for Learning Geography Facts

Jan Papoušek
Masaryk University Brno
jan.papousek@mail.muni.cz

Vít Stanislav
Masaryk University Brno
slaweet@mail.muni.cz

Radek Pelánek
Masaryk University Brno
xpelanek@mail.muni.cz

ABSTRACT

Computerized educational systems are increasingly provided as open online services which provide adaptive personalized learning experience. To fully exploit potential of such systems, it is necessary to thoroughly evaluate different design choices. However, both openness and adaptivity make proper evaluation difficult. We provide a detailed report on evaluation of an online system for adaptive practice of geography, and use this case study to highlight methodological issues with evaluation of open online learning systems, particularly attrition bias. To facilitate evaluation of learning, we propose to use randomized reference questions. We illustrate application of survival analysis and learning curves for declarative knowledge. The result provide an interesting insight into the impact of adaptivity on learner behaviour and learning.

Keywords

attrition bias, computerized adaptive practice, engagement, evaluation, learning curve, survival analysis

1. INTRODUCTION

Open online educational systems are becoming a key part of education – systems like Khan academy, Duolingo, or edX are today used by millions of learners and in the future the role of such systems is expected to grow. One advantage of computerized educational systems is adaptivity – their behaviour can be personalized for a particular learner. To assess the contribution of such educational systems and to tune their behaviour (e.g., choose a proper learner model for guiding the adaptive behaviour), we need to evaluate them. However, both openness and adaptivity significantly complicate the evaluation process.

The open nature of these systems means that they can be used by anybody, anywhere. This has several consequences for evaluation. Standard evaluation methods (like pre-test, post-test) are not applicable. The learner population is typically very heterogeneous, often comprising students using the system compulsory within classroom, students using the system voluntary as part of their preparation for an exam, adult learners who want to refresh their knowledge, and also people who just stumbled upon the system while browsing an internet or following a suggestion of a friend on a social network. The motivation of learners to use the system thus widely differs, the distribution of time in the system is typically highly skewed (most learners use the system for only a short time) and the departure from the system is not random. This creates attrition bias, which complicates evaluation of learning within the system. Adaptive behaviour of systems further complicates the evaluation – each learner proceeds through the system using different learning materials and questions and it is not easy to use these adaptively constructed questions for evaluation of learning gains. Moreover, feedback loop between a learner model and collection of data for evaluation [13] further complicates the evaluation.

Evaluation of adaptive systems has been studied before. Evaluation of recommender systems [5] faces many similar issues. Specifically for educational systems, previous research [10, 4] discussed wide coverage of different methods and evaluation aspect, but only on a high level without discussing specific details. Layered evaluation [20, 1] has been proposed as a basic framework for evaluation of adaptive systems. Current research, however, focuses mainly on evaluation of learner models (the first layer), which can be done using historical data and evaluation of predictive accuracy measured by metrics [21]. There has been attempts to use historical data to assess impact on learners [3], but reliable assessment of this impact needs a proper randomized control trial and such experiments are for open online adaptive systems currently rather rare.

Let us overview specific methods for evaluation of educational systems and discuss their properties and applicability in the context of open and adaptive systems. The "gold standard" for evaluation of educational interventions is a randomized control trial together with pre-test and post-test to evaluate learning gains. In our setting this is however not feasible due to the complete lack of control over learners using the system online. Another way to obtain high-quality evaluation would be to use results from external tests (e.g., university exams), but in most cases it is infeasible to obtain such data (learners using open systems typically do not take any external test). It may be possible to use voluntary tests within the system to asses learning, but the motivation to take the test influences results (self-selection bias).

A realistic approach is to use learning curves [11] which

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom

 ISBN 978-1-4503-4190-5/16/04…$15.00

DOI: http://dx.doi.org/10.1145/2883851.2883884

map learning during the usage of the system. Interpretation of learning curves is, however, complicated by issues with aggregation and attrition bias [12, 15]. Attrition bias is a key issue in evaluation of educational systems, methods from survival analysis may be useful [2]. We can also use learner modeling techniques and use model based detectors of learning. However, such approach is dependent on quality of models and their simplifying assumptions, which may influence results. Another relatively easily realizable (but imprecise) approach is to evaluate only aspects of the system which are more easily measurable than learning and use them as proxy metrics, e.g., number of answers or learner feedback [17]. However, previous research [9] suggests that such metrics may not be directly correlated with learning.

We propose to use periodic "reference questions" which are constructed fully randomly. Similar approach based on usage of random items have been used for evaluation previously in [7, 8]. We analyze reference questions using learning curves. Because of a random aspect of these questions, there is no influence from the adaptive algorithm, and thus we can fairly compare different conditions. However, there still remains attrition bias, which needs to be taken into account.

In this work we explore methodological evaluation issues using a specific case study – a widely used system for adaptive practice of geography facts [18, 17]. Using this system we performed a randomized control trial – a comparison of 4 different strategies for question construction ranging from fully adaptive to fully random. We explore factors influencing the length of stay within the system, where we employ techniques from survival analysis (particularly fit to Weibull distribution). We also explore an application of learning curves describing learners' progress of declarative knowledge (learning curves have previously been used mainly for evaluating procedural skills [11]). Using these techniques we illustrate when and how the adaptivity is important for the studied system. Our main point, however, is not limited for a particular system. Exploration of techniques and methodological issues of evalutation is generally relevant to any open and adaptive educational system, e.g., our results highlight the role of attrition bias and show how this bias can influence learning curves in different directions (in the context of a single educational system).

2. EXPERIMENTAL SETTING

2.1 The Used System

For our experiments we use an online adaptive system providing practice of geographical facts (names and location of countries, cities, ...), available at outlinemaps.org. The system estimates learners' knowledge and based on this estimate it adaptively constructs questions of suitable difficulty [18]. The system uses open questions ("Where is France?") and multiple-choice questions ("What is the name of the highlighted country?") with 2 to 6 options. Learners answer questions with the use of an interactive 'outline map'. These questions are asked in sequences of length 10, and although learners can quit a sequence anytime, they tend to finish it. After each sequence, a summary overview about practiced items is shown. Learners can also access a visualization of their knowledge using an open learner model. During a school year we collect roughly 1 000 000 answers from 10 000 users per month. Part of the data is publicly available[1] [16].

The adaptive behaviour of the system is based on models of learners' knowledge which provide for each learner and place current prediction of knowledge (probability of correct answer). These models have been described and evaluated in previous work [14, 22], here we use them as a 'black box'.

An important factor influencing the evaluation and interpretation of results are different contexts within the system. Learners can use the system with a lot of different contexts (maps, types of places) and these contexts differ widely in their difficulty (prior knowledge) and the number of places available to practice (from 10 to 120). Distribution of answers is highly uneven, most learners practice few popular maps (Figure 1).

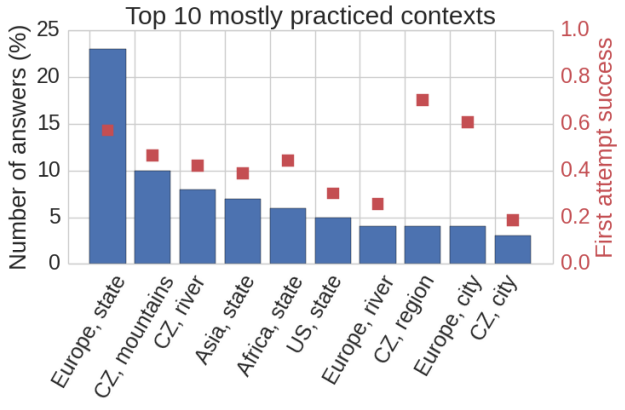

Figure 1: The most commonly answered contexts. For each of these context we also report the average error rate for the first answer of each learner.

2.2 Users of the System

This system is a typical open educational system available to anybody free of charge. 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. The majority of learners is from the Czech Republic (84%) and Slovakia (8%) since the interface was originally only in Czech. However, English, German, and Spanish are currently also available.

2.3 Experimental Conditions

The system uses a target error rate (20 %) and adaptively constructs questions in such a way that the learners' achieved performance is close to this target [17]. Firstly, the algorithm selects the stem of the question (which place to ask about). 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*). New learners were assigned to conditions randomly upon entering the system. Learners present in our system before this experiment are provided with the *A-A* condition and are not taken into account for further analysis.

[1] www.fi.muni.cz/adaptivelearning/data/slepemapy/

Adaptive version of item selection (*A-**) computes a score for each item taking into account its difficulty, number of a learner's answers about it and time elapsed since the last learner's answer about it. Random version of item selection (*R-**) picks the stem randomly. As for construction of options, adaptive version (**-A*) computes a number of options to make the question as close to the target difficulty as it is possible and uses the most competitive distractors. Random version (**-R*) chooses a number of options and options themselves fully randomly. Both version provide multiple-choice questions with from 2 to 6 options or completely open ones. It is worth noting that *A-A* condition prefers open questions, on the other side questions constructed by the *R-R* condition are mostly multiple-choice. Distractors for this condition are non-competitive, so it provides the easiest practice from all studied conditions (Figure 2, top).

To provide better intuition behind the used experimental conditions, we discuss specific example of question construction for a new learner who chooses to practice African countries. The first construction step in *A-** condition prefers Algeria (estimated error rate 25%) to Madagascar (6% – too easy) or Zimbabwe (55% – too difficult), whereas *R-** condition selects countries with uniform probability. In the second step, if *R-A* has Zimbabwe from the first step, it reduces its difficulty by selecting only 2 options (Zimbabwe and 1 competitive distractor - Zambia), whereas *A-A* has Algeria from the first step, Algeria has appropriate difficulty, and the algorithm thus selects either open question or a high number of options (6) with competitive distractors (Egypt, Libya, Dem. Rep. Congo, South Sudan, and Sudan). Regardless of whether the first step selected Algeria, Zimbabwe, or some other country, both **-R* conditions select random number of options and random distractors (e.g., 4-options question with distractors Morocco, Tanzania, and Ghana).

2.4 Collected Data

In case of this experiment running from the end of August to October 2015 we have collected more than 1 300 000 answers from roughly 20 000 learners. For each context separately every 1st, 11th, 21st, ... are reference questions, open questions about an item randomly chosen from the context. These questions result to 1st, 2nd, 3rd, ... reference answers which are used to track learning. It should be noted that 1st reference answer comes before the question construction algorithm has any chance to influence the practice for the given context.

We also ask learners to evaluate the difficulty of questions. After 30, 70, 120, and 200 answers the system shows the dialog "What is the difficulty of asked questions?", learners choose one of the following options: "Too Easy", "Appropriate", "Too Difficult". Within this experiment we analyze roughly 16 000 records.

To make our research reproducible we make the analyzed data set available[2], together with a brief description and terms of use.

3. EVALUATION: ENGAGEMENT

At first we evaluate impact of individual conditions on student engagement. Students engagement depends not only on system behaviour, but also on their motivation. As the system is open to anyone, its users vary in many aspects including their motivation for using the system.

Some learners use the system in school lessons. Learners in schools are mostly affected by external motivation factors as they are constrained by the time allocated by their teacher and might not be genuinely interested in practicing of geography. We can detect 'in-school' usage based on the IP address (a group of at least n learners who started using the system from the same IP address is identified as an 'in-school' group). The 'in-school' usage represents about 20% of the collected data.

There are also learners preparing for their school exams at home. These learners are probably more focused on mastering a particular map and not motivated to return to the system after the exam. Finally, some learners use the system just for fun. These learners do not have external motivation and thus are most likely to be affected by the system behaviour (e.g., leave the system if the practice is too difficult or not challenging enough). Although we have anecdotal evidence of these learner groups in the system, we do not have enough data to reliably distinguish between the latter two groups.

3.1 Statistics for Conditions

The experimental conditions differ in learners' error rate (Figure 2). Conditions *A-R* and *R-R* have overall lower error rate because they are more likely to use fewer options. *R-** conditions exhibit decline in the error rate throughout the use of the system (Figure 2, bottom), whereas *A-** conditions by definition keep the error rate more constant. Especially the *A-A* error rate is distributed closely around the target error rate (Figure 2, top). On the other hand, *R-R* error rate distribution is skewed towards 0%.

Error rate is influenced by average item difficulty which varies largely among different contexts. For the 10 most practiced contexts the error rate on the first reference question is between 30% and 80% (Figure 1). The relation of average item difficulty and error rate is different for different experiment conditions on different contexts. In *R-** conditions the error rate is highly influenced by average item difficulty of the context. *A-** conditions can decrease error rate by asking multiple-choice questions with fewer alternatives. However, when every item on a given contexts has prediction below the target error rate, then there is no way to increase the error rate.

3.2 Explicit Feedback

Figure 3 shows the results of learners explicit feedback about difficulty of questions. The most appropriately difficult questions among the experimental conditions are asked by the *A-A* condition. The other three conditions exhibit increased number of "Too Easy" evaluations. In particular, both **-R* conditions have increased number of "Too Easy" compared to their **-A* counterparts.

Explicit feedback also reflects error rate differences among contexts. *Random* conditions are more varied in this respect than *adaptive* conditions in both the first and the second question construction step. There is more space for adaptivity to make a difference in contexts with average item difficulty far from the target error rate. Especially in contexts with low average difficulty and thus non-trivial prior knowledge (e.g., Czech regions, European countries or cities), there are more "Too Easy" evaluations in *R-R* condi-

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

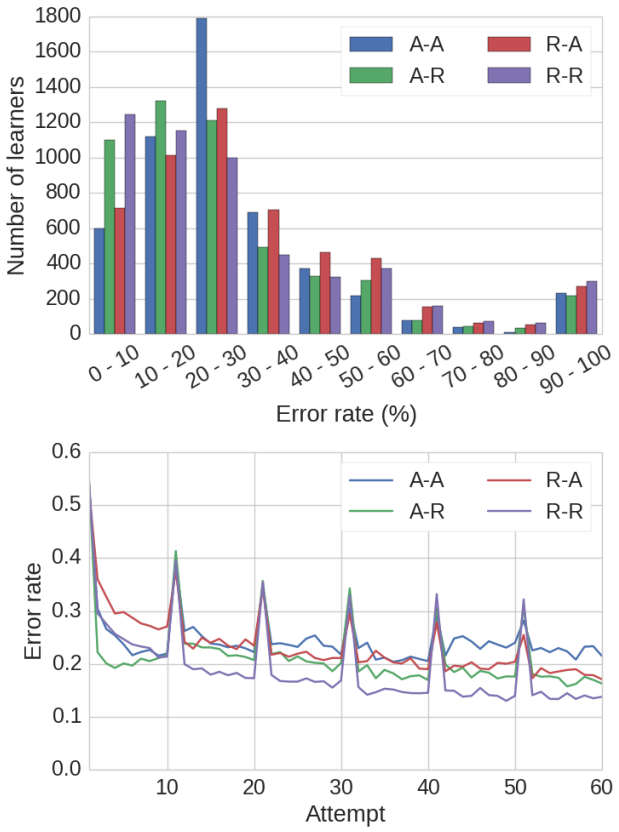

Figure 2: Comparison of error rates for the four conditions. Top: Histogram of the overall error rate. Bottom: Error rate as function of the number of attempts.

tion (Czech regions: 33%, European countries: 41%, European cities: 33%) than in *A-A* (Czech regions: 26%, European countries: 24%, European cities: 24%). For *A-R* and *R-A* conditions the percentages of "Too Easy" evaluations are somewhere in between *A-A* and *R-R*. On the other hand in contexts with lower prior knowledge (e.g., Czech mountains) the amount of "Too Easy" evaluations is less diverse among conditions (*A-A*: 23%; *R-R*: 25% for Czech mountains)

As we ask for the evaluations repeatedly, the proportion of "Appropriate" evaluations is slightly rising with the number of questions answered by the learner. It is most likely caused by dissatisfied learner leaving the system (attrition bias). This effect, however, does not occur in *A-R*, where proportion of "Appropriate" evaluations stays at the same level, but proportion of "Too Difficult" declines in favor of "Too Easy".

3.3 Survival Analysis

To compare 'attractiveness' of different conditions we analyze the number of answers within the system. The key observation is that the distribution of answers is very skewed and thus it is not suitable to compare conditions using averages (or even other measures of central tendency like the median). It is useful to employ techniques from survival analysis. Survival analysis deals with questions like "What

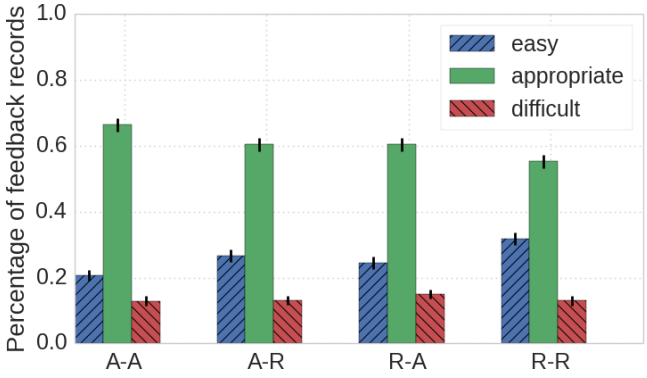

Figure 3: Explicit feedback by learners about question difficulty for the four conditions. Black lines show 95% confidence intervals.

proportion of population will survive past a given time?", typically in the context of medical data. Once we interpret "survival" as "active usage of a system", it is directly relevant to evaluation of educational systems.

Figure 4 (left) shows a survivor graph, i.e., proportion of learner population "surviving" the given number of questions. There are clear discrete steps after multiples of 10, these are due to the properties of the analyzed system described in Section 2.1 which presents a summary overview of a learner's progress after each sequence of 10 questions (and thus creates natural points to leave the system). Once the length of stay is analyzed for groups of 10 questions, the graph becomes smooth. Figure 4 (right) shows this preprocessed variant in the form of probability density function with fitted Weibull distribution. This is a standard distribution in survival analysis, previous research shows that it also fits well dwell time on web pages [6], and it has also been used to fit MOOC data [23]. Our results indicate that the Weibull distribution is useful also for fitting the number of answers within an open educational system.

Table 1: Fitted parameter of the Weibull distribution.

Condition	k	λ
A-A	0.762	6.673
A-R	0.793	5.849
R-A	0.746	6.349
R-R	0.751	5.882

Probability density function of the Weibull distribution is $f(x,k,\lambda) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(x/\lambda)^k}$ $(k,\lambda > 0, x \geq 0)$. The distribution has two parameters: k is the shape parameter and λ is the scale parameter. Values $k < 1$ correspond to negative aging ("infant mortality"), i.e., the probability of leaving decreases with the length of stay, for $k = 1$ we get exponential distribution (constant rate of leaving), values $k > 1$ correspond to positive aging. Fitted parameter values for our four conditions are in Table 1. In all cases we have $k < 1$, i.e., negative aging (which is typical for online systems [6]). The table shows that adaptivity in the first question construction step is related to the k parameter (adaptivity reduces "infant mortality"), whereas adaptivity in the second step is related to the λ parameter (the length of stay). This be-

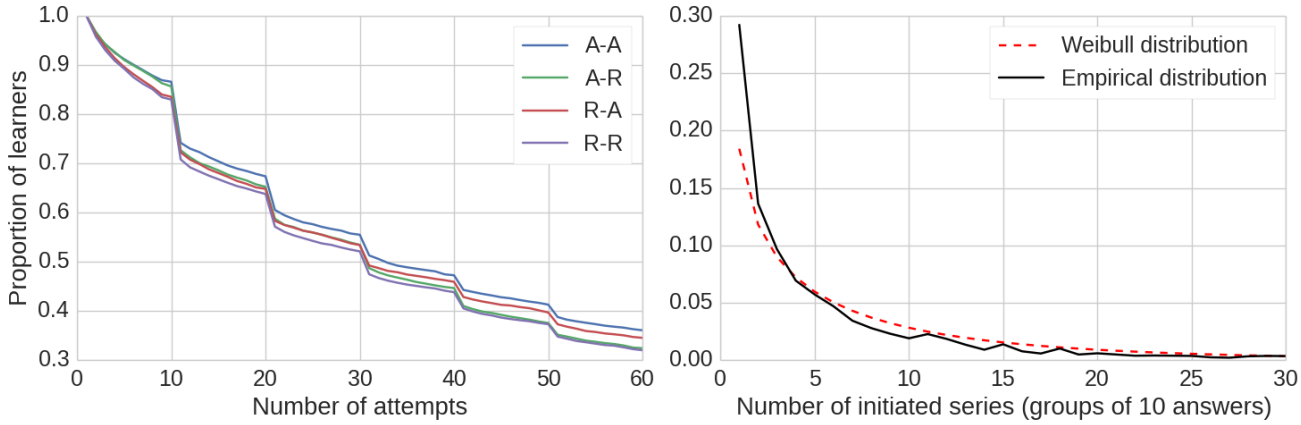

Figure 4: Left: Survivor graph (proportion of learners that answers at least k questions). Right: Probability density function for number of attempts (groups of 10) fitted by the Weibull distribution for the *A-A* condition (for other conditions the fit is very similar).

Table 2: Probability of return to the system. The 95% confidence interval is in all cases $\pm 0.9\%$ around the given value.

Condition	Probability
A-A	15.1%
A-R	13.9%
R-A	14.3%
R-R	13.1%

haviour can be seen also on the survivor graph (Figure 4 left), where at the beginning we have higher survivor rates for conditions *A-A* and *A-R*, whereas in the long run we have top performance for *A-A* and *R-A*.

We have also analyzed probability that a learner returns to the system (as a return to the system we consider occurrence of two attempts with pause between them of least 10 hours). Table 2 shows the comparison of conditions, we see that adaptivity increases chance of return. The relative difference between *A-A* and *R-R* condition is 15%, i.e., adaptability has large impact on learners' decision to return to the system.

4. EVALUATION: LEARNING

Our system provides practice of items which are independent on each other. We assume practice of an item A does not have any impact on knowledge of an item B, or the impact is negligible. Although the provided practice seems to be similar to testing, data collected using question construction algorithm can not be simply used to evaluate progression of learners' performance across different conditions, because these conditions differ in questions they ask, e.g., *R-R* condition provides easy, mainly multiple-choice questions, on the other side in case of *A-A* condition questions are more difficult and mostly open. For the same reason we can not use a model providing estimation of learners' performance based on data for this purpose, e.g., answers on too easy or too difficult questions do not contain the same amount of information as in case where the probability of correct answer is close to 50%, so it takes much longer time to estimate learners' knowledge. Estimation and improvement of learners' knowledge happen in the same time and we have to be sure one condition is not disadvantaged because of poor behaviour in estimation, since it can perform well in learning. In the following section we focus on analyzing answers to reference questions which represent objective data collected independently on studied conditions (they are constructed randomly).

4.1 Learning Curves

To measure learners' knowledge to analyze learning we look at learners' performance based on answers to reference questions. For each context we build series of learners' reference answers, we merge these series together and compute an average error rate for each attempt. By this technique we are able to analyze progression of performance for context(s) and a group of learners, but we can not simply do the same on a level of one item and one learner.

As the analysis in previous section shows, there is high attrition in the data (learners are leaving the system at different points of time). Due to this attrition it is not straightforward to construct and compare learning curves. We consider three approaches to construction of learning curves, each has different advantages and disadvantages:

1. *All learners*: In this case we include all learners. Since learners have varied number of answers, individual points of the learning curve are computed from different samples of learners. Previous research [12] has shown how this may lead to flat learning curves in case of mastery learning.

2. *Filtered learners*: We construct a curve for n answers and we include only learners having at least n answers. Now each point of the learning curve is computed from the same sample of learners, but this sample may be biased.

3. *Filtered learners, reverse*: Similarly to the previous case, but in case of learners who have more than n answers, we use the last n answers.

4.2 Attrition bias

The interpretation of learning curves is complicated by attrition bias. Attrition bias is a type of selection bias which is often present for example in medical experiments. In the context of educational systems and evaluation using learning curves, previous research identified mastery attrition bias [12, 15] – when learners, who master the studied topic stop practice (e.g., due to the use of mastery learning in the educational system), the learning curves become significantly flatter and can even mask learning.

Mastery is, however, not the only source of attrition. Particularly in open online systems, differences in motivation may also play a significant role. If a system, for example, offers rather difficult questions, this may disengage and deter weaker learners and we may obtain an opposite of mastery attrition (self-selection) which causes learning curves to be steeper.

Our results suggest that these two effects may be present at the same time within one system and it may be hard to disentangle them and interpret learning curves correctly. For the analysis presented in Figure 5 we consider different groups of learners based on the number of answers to reference questions. The figure shows for each group an average error rate on the first reference question (which corresponds to the prior knowledge for the group). The results show that there are differences between these groups, i.e., learners attrition is not random. Particularly interesting aspect of the figure is that the attrition differs between conditions. We see, for example, a big difference between learners having at least 3 reference answers in *R-A* and *R-R* conditions. Learners having at least 3 reference answers in *A-R* condition are those with above average prior knowledge, whereas in *R-A* condition those learners have below average prior knowledge. This phenomenon should be taken into account in case of learning curve based on learners having at least n answers (Figure 6, middle and right). It can easily happen we compare learning of different groups of learners, e.g., in this case low performers vs. high performers. The same phenomenon is present even though we do not use any filtering, but it is combined with mastery attrition bias.

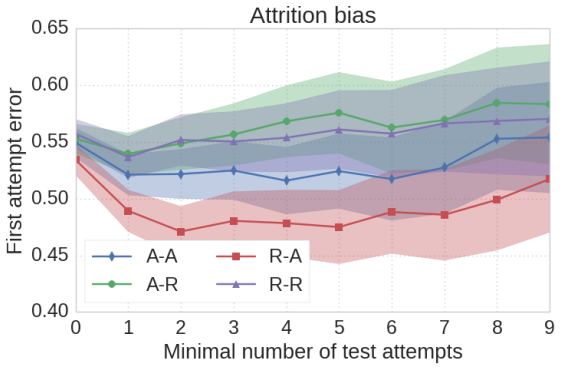

Figure 5: The first attempt error rate depending on how many reference questions the learner answered.

We do not come up with a solution to prevent the mentioned problems, but based on these observations we decided to analyze several types of learning curves and check how different methodologies affect the relative order of analyzed conditions with respect to learning. The advantage of the first approach without any filtering is that we do not arbitrarily omit any learner. In case of filtering out learners with less number of reference answers, the given number n we take the first n or the last n learners' answers (Figure 6, middle and right). In case of the last n answer we see how learners learn before they quit the practice. Since we assume many learners quit the practice once they master the topic, this curve has generally much lower error rate.

4.3 Results

Figure 6 (top) shows the resulting learning curves. Confidence intervals are computed for each data point independently. The confidence intervals for individual points overlap, but we get repeatedly similar results (ordering of experimental conditions). Based on previous research on learning curves [11], we fit the power law function to the data, i.e., $error\_rate(k) = ak^{-b}$, where k is the order of the attempt.

Another potential source of bias in learning curves is aggregation across different contexts (knowledge components) [11]. This issue is again more prominent in open educational systems, where learners can freely choose topics. Imagine that one of our experimental conditions is more interesting to learners for easier contexts, whereas other is more interesting for difficult contexts. This would mean a lower error rates in the aggregated learning curves for the first condition. However, this difference between learning curves would not be due to differences in learning, but due to different impact of conditions on engagement. It is thus useful to analyze also learning curves for individual contexts. A disadvantage of these disaggregated results is that they are constructed from smaller amount of data and thus the learning curves are more noisy. Figure 7 shows examples of such curves for a few popular contexts within our system.

The use of error rate as a measure of learning is a standard (and in our setting natural) choice. It is, however, not the only possible choice. We can take into account also other aspects of learners behaviour, e.g., the response time. Our previous research shows that the time learners spent by answering questions relates to their future success [19]. Even though the system does not motivate learners to have low response time (in fact it does not even indicate in any way that the response time is measured), we observe an improvement of response time and systematic differences between studied conditions (Figure 8). With respect to this measure we get the best results for the *R-A* condition.

Although none of the presented learning curves is ideal, the main results are consistent across different analysis methods. In all cases the conditions with adaptive construction of options (*A-A*, *R-A*) beat the conditions with random options (*A-R*, *R-R*). The item selection part does not seem to have large effect on learning. When we see differences between the *A-A* and *R-A* conditions, the *R-A* condition is slightly better, i.e., it seems that with respect to learning the adaptive choice of stems could be improved.

5. DISCUSSION

Our work focuses on evaluation of an open online educational system. In such systems it is important to evaluate impact of system behaviour on both learners motivation and learning. To analyze the length of the stay within the system we utilize techniques from survival analysis, particularly we fit the Weibull distribution and show that the fitted pa-

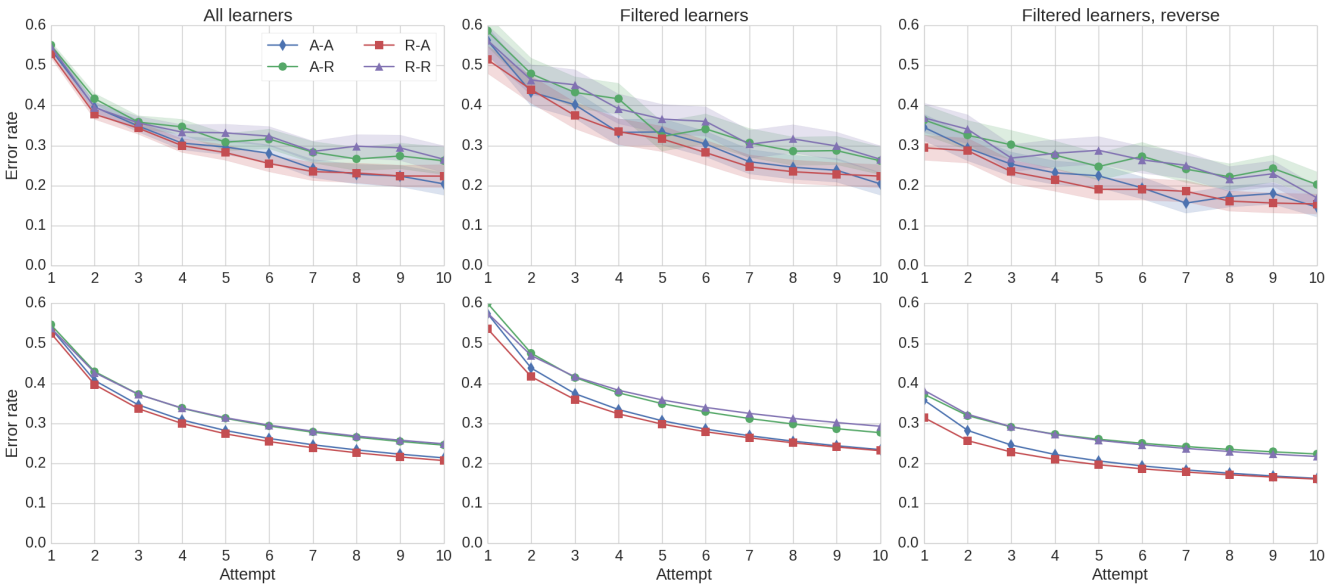

Figure 6: Learning curves based on reference questions. Top: Raw data with 95% confidence intervals. Bottom: Fitted power law functions. Columns correspond to different types of filtering.

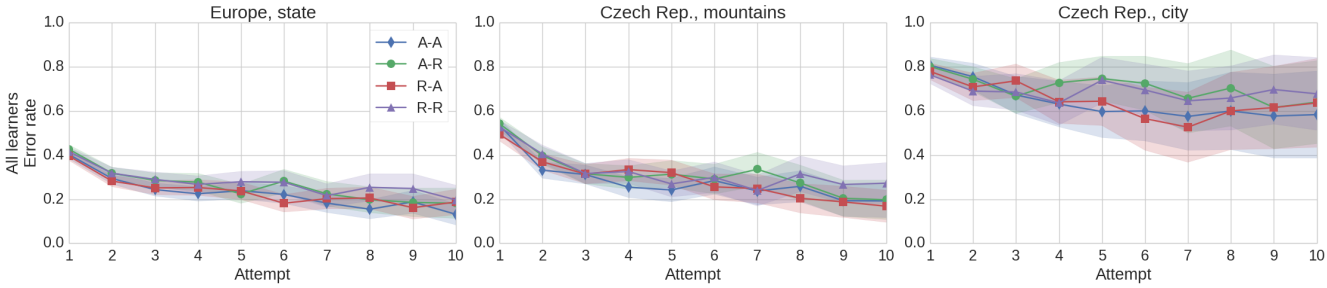

Figure 7: Learning curves (fitted power law functions) for specific contexts: European states, Czech mountains and Czech cities.

rameters provide insight into the impact of adaptive system behaviour on learners' use of the system. We also show that attrition differs for different conditions (versions of the used system), which creates attrition bias that complicates analysis of learning within the system. Previous work [12, 15] has considered mastery attrition bias, but our results show that attrition bias is not just due to mastery.

To evaluate learning we work with learning curves. To use learning curves in adaptive system we employ "reference" questions which are constructed randomly; this allows us to perform fair comparison of different question construction algorithms. However, the constructed learning curves still give a simplified view of learning. A particular disadvantage of learning curves is that they take into account only order of questions and not the time that passed between attempts. This aspect may be particularly important for declarative knowledge (as opposed to procedural skills for which learning curves have been used so far). In the current data set, practice is currently mostly massed (80% of answers are within the first session of a learner), so the used simplification should not have significant impact on presented results.

To incorporate the timing information into the analysis, it may be useful to study "learning surface", e.g., in the form of graph depicted in Figure 9, which visualizes data from our experiment. The figure shows error rate depending on both order of an attempt and the time from a previous attempt. Such analysis may help us to evaluate also long term effects of different learning situations. A proper way to construct and compare such learning surfaces is an interesting direction for future work.

The results of our evaluation demonstrate the advantage of adaptive behaviour over a baseline, random selection of questions. More interestingly, the results show a part of the question construction for which the adaptivity is important. For the studied setting it turns out that adaptivity is important mainly for choosing number of options for multiple-choice questions and for choosing distractor (i.e., "fine-tuning" question difficulty), not in the choice of a question stem (the "target" factual knowledge). With respect to the length of the stay within the system the adaptive choice of a stem is related to initial mortality, whereas the adaptive choice of options is related to the overall length of the stay. With respect to learning the main factor is adaptivity

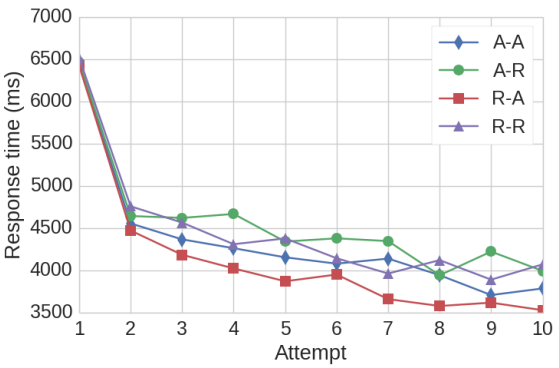

Figure 8: Learning curve for response times (the reported time for each attempt is the median of corresponding times).

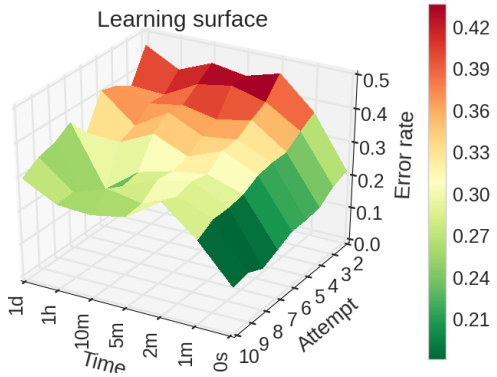

Figure 9: Learning surface showing error rate depending on both number of attempts and time from previous attempt.

in choice of options. For the choice of stem we get the same or even better results for random selection. This suggests that the algorithm for the choice of a question stem [18, 17] may need to be improved.

6. ACKNOWLEDGMENTS

This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

7. REFERENCES

[1] P. Brusilovsky, C. Karagiannidis, and D. Sampson. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Life Long Learning*, 14(4-5):402–421, 2004.

[2] M. Eagle and T. Barnes. Survival analysis on duration data in intelligent tutors. In *Intelligent Tutoring Systems*, pages 178–187. Springer, 2014.

[3] J. P. González-Brenes and Y. Huang. Your model is predictive – but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. *Educational Data Mining*, 2015.

[4] J. Greer and M. Mark. Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education*, pages 1–6, 2015.

[5] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[6] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 379–386. ACM, 2010.

[7] Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popović. Towards automatic experimentation of educational knowledge. In *Human Factors in Computing Systems*, pages 3349–3358. ACM, 2014.

[8] Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *Educational Data Mining 2014*, 2014.

[9] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 89–98. ACM, 2013.

[10] M. A. Mark, J. E. Greer, et al. Evaluation methodologies for intelligent tutoring systems. *Journal of Artificial Intelligence in Education*, 4:129–129, 1993.

[11] B. Martin, A. Mitrovic, K. R. Koedinger, and S. Mathan. Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3):249–283, 2011.

[12] R. C. Murray, S. Ritter, T. Nixon, R. Schwiebert, R. G. Hausmann, B. Towle, S. E. Fancsali, and A. Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

[13] J. Nižnan, R. Pelánek, and J. Papoušek. Exploring the role of small differences in predictive accuracy using simulated data. In *AIED Workshop on Simulated Learners*, 2015.

[14] J. Nižnan, R. Pelánek, and J. Řihák. Student models for prior knowledge estimation. In *Educational Data Mining*, 2015.

[15] T. Nixon, S. Fancsali, and S. Ritter. The complex dynamics of aggregate learning curves. In *Educational Data Mining*, 2013.

[16] J. Papoušek, R. Pelánek, and V. Stanislav. Adaptive geography practice data set. *Journal of Learning Analytics*, 2015. To appear.

[17] J. Papoušek and R. Pelánek. Impact of adaptive educational system behaviour on student motivation. In *Artificial Intelligence in Education*, volume 9112, pages 348–357, 2015.

[18] J. Papoušek, R. Pelánek, and V. Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining*, pages 6–13, 2014.

[19] J. Papoušek, R. Pelánek, J. Řihák, and V. Stanislav.

An analysis of response times in adaptive practice of geography facts. In *Educational Data Mining*, 2015.

[20] A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, 2010.

[21] R. Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2), 2015.

[22] R. Pelánek. Modeling students' memory for application in adaptive educational systems. In *Educational Data Mining*, 2015.

[23] D. Yang, T. Sinha, D. Adamson, and C. P. Rose. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, volume 11, page 14, 2013.