Impact of Question Difficulty on Engagement and Learning

Jan Papoušek, Vít Stanislav, and Radek Pelánek

Masaryk University Brno

Abstract. We study the impact of question difficulty on learners' engagement and learning using an experiment with an open online educational system for adaptive practice of geography. The experiment shows that easy questions are better for short term engagement, whereas difficult questions are better for long term engagement and learning. These results stress the necessity of careful formalization of goals and optimization criteria of open online education systems. We also present disaggregation of overall results into specific contexts of practice.

1 Introduction

Making practice suitably challenging is one of the key goals of adaptive educational systems. The general idea that the best activity is neither too easy nor too difficult was formulated as Inverted-U Hypothesis [1]. Lomas et al. [6] found that in the context of their simple educational game easier problems lead to higher engagement, but lower learning. A similar research was done using Math Garden software [2]. The authors compared three conditions and showed that the easiest condition led to the best learning (mediated by a number of solved tasks). Other authors have used more complex experimental techniques to find optimal parameter values (e.g., Bayesian optimization), but they have optimized only with respect to short term engagement [3] or short term transfer [4].

We report results of an online experiment evaluating impact of question difficulty on learning and engagement in the context of declarative knowledge and an open educational system. Specifically, we use a system for an adaptive practice of geographical facts [9] (e.g., names and location of countries or cities); the system is publicly available at `http://outlinemaps.org`. We have reported experiments with question difficulty in this system in previous work [8], but only with respect to engagement. Here we provide more detailed analysis including also learning. The used methodology is similar to a previous work [10] which compared an adaptive and a random construction of questions within the system. Here, we pay more attention to issues related to data aggregation and a conflict between short and long term engagement.

Analyzing data from the experiment containing conditions targeting 5%, 20%, 35%, and 50% error rate, we observe a conflict between learning and long term engagement on one side (more difficult is better), and short term engagement on the other (easier is better). These results demonstrate the risk hidden in optimizing only short term behaviour of the system (as done in [3,4]). Our results are also in contrast with previous studies [2,6], which concluded that easier questions are better (we are, however, using educational system from a completely different domain).

2 Experimental Setting

We have performed the evaluation using a randomized trial with four experimental conditions within a widely used adaptive system providing practice of geography. The system estimates learners' knowledge and based on this estimate it adaptively constructs multiple-choice (2–6 options) or open questions of suitable difficulty [9]. The adaptive behaviour of the system is based on models of learners' knowledge. These models provide a prediction of the current knowledge for each learner and item. This part has been described and evaluated in previous work [9], here we use these models as a 'black box'.

The system uses a target error rate and adaptively constructs questions in such a way that learners' achieved performance is close to this target [8]. In our experiment we evaluate four experimental conditions which differ only in one aspect – the target error rate: 5%, 20%, 35%, 50%. In the following text we denote the conditions as C5, C20, C35, and C50. Learners were assigned to one of the conditions randomly when they entered the system for the first time. The experiment was performed from November 2015 to January 2016 and we have collected almost 3 300 000 answers from roughly 37 000 learners. To make our research reproducible we make the analyzed data set available[1] (together with a brief description and terms of use).

To evaluate learning within the adaptive system we use "reference questions". The reference questions are open questions about a randomly chosen item from a particular context (independently of the experimental condition). The questions are used periodically (every 10th question is a reference question). The first reference question is the first question within a context, i.e. before the adaptive algorithm has any chance to influence the practice for the given context. A similar approach based on random items has been used for evaluation previously, for example in [4,10].

An important factor that influences the evaluation and interpretation of results are different contexts within the system. Learners can choose different maps and types of places to practice. These contexts differ widely in their difficulty (prior knowledge) and the number of items available to practice (from 10 to 170). Distribution of answers is highly uneven, most learners practice a few popular maps. For the analysis we use 10 contexts with most answers (listed in Fig. 1.). More detailed analysis of differences among contexts is available in the full version of the paper [11].

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

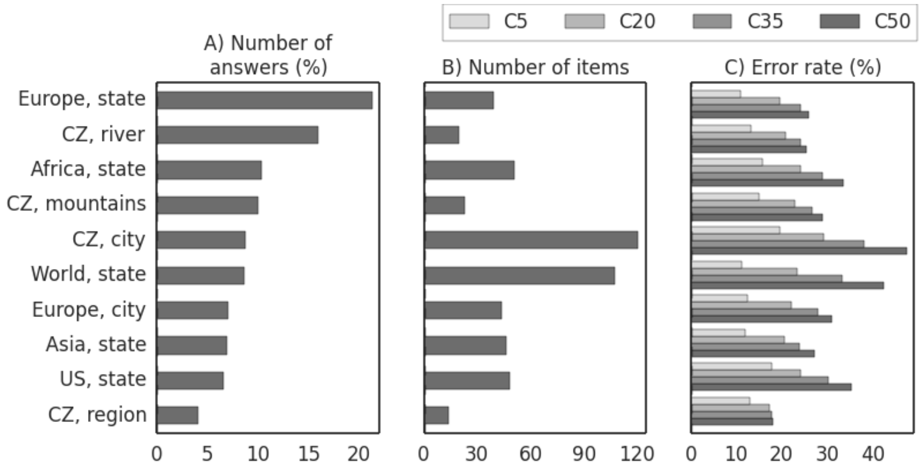

Fig. 1. Top 10 mostly used contexts available for learners to practice. A) percentage of answers in the analyzed data set, B) number of items, C) average error rate per experimental condition ignoring reference answers.

3 Engagement

To evaluate engagement we consider (*1*) survival rates (i.e., proportion of learners who answer at least k questions), and (*2*) probability of returning to the system (after a delay of more than 10 hours; the specific duration of the delay is not important for presented results). While analyzing differences among the conditions, we have identified opposite tendencies with respect to short term and long term engagement. The main trend is that while conditions with easier questions enhance engagement at the beginning, more difficult conditions engage more learners later on.

From the global viewpoint, short term engagement is better in case of easier questions. The survival rate after 10 answers is sorted according to question difficulty (C5: 89.2%, C20: 87.0%, C35: 84.0%, C50: 81.2%, confidence interval ±0.77%). The differences are decreasing with the number of answers, survival rates after 100 answers are very similar in all conditions (from 26.0% to 26.5%, confidence interval ±0.88%). Note that after 30 or more questions, the conditions C35 and C50 no longer achieve their target error rate in most contexts, since the items from these contexts are already mastered by learners. The return rate increases with the difficulty of questions, the largest difference being between C5 and other conditions (C5: 15.2%, C20: 16.0%, C35: 16.6%, C50: 16.8%, confidence interval ±0.75%).

There are quite large differences among individual contexts (see Fig. 2), most likely caused by learners' preferences and implementation details of the system, e.g., the system recommends 6 contexts (e.g., European states) as "quick start" options on the home page, which makes their survival rates lower than survival rates of "self-selected" contexts (e.g., Asian states). The magnitude of differences

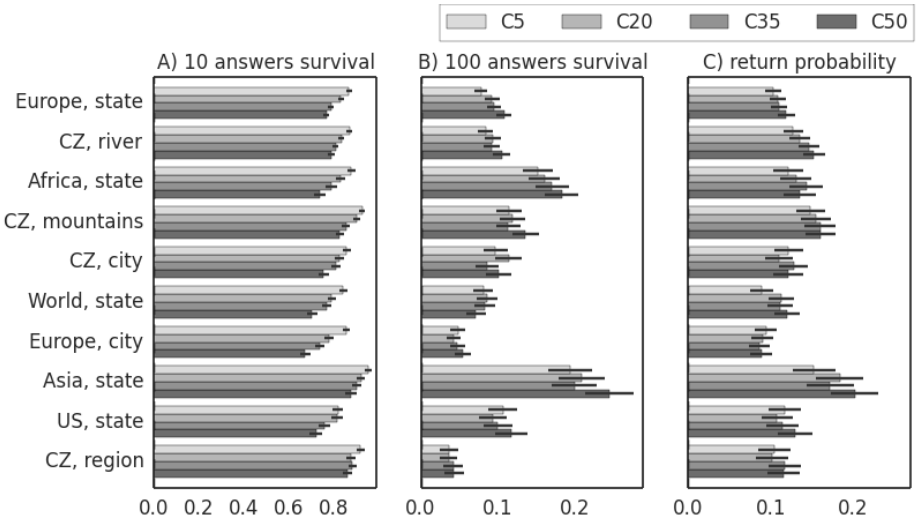

Fig. 2. Survival analysis (A, B) and probability of return after 10 hours (C) for 10 most practiced contexts and 4 experiment conditions. Error bars represent 95% confidence intervals.

between conditions is mostly aligned with differences in their behaviour in the particular context, e.g. its difficulty or number of items available to practice.

Short term survival (Fig. 2. A) differs in all contexts in favour of easier conditions. In case of long term survival (Fig. 2. B), the trend is quite opposite, although for individual contexts the differences are typically rather small. This contrast is best seen on European states (the context with most data), where we see a reliable difference between C50 and C5.

4 Learning

The evaluation of learning cannot be simply based on the achieved error rate of learners, since this error rate is by definition heavily influenced by the used experimental conditions. For this reason we collect previously described reference answers, which are not affected by any condition, and from these reference answers we construct learning curves. We construct a learning curve [7] in the same way as in [10]. We put together reference answers from all available contexts and compute an average error rate preserving their ordering within contexts (e.g., we put together all the first reference answers from all users and contexts to get the first point of the learning curve). We do not filter any data and users may quit their practice on their own, so for the first point of the learning curve we have more answers than for the second one and so on – the results thus may be influenced by attrition bias, this issue is discussed in the full version of the paper [11]. In accordance with previous research [7,10] we assume that the learning curve corresponds to the power law, i.e., the error rate can be expressed as

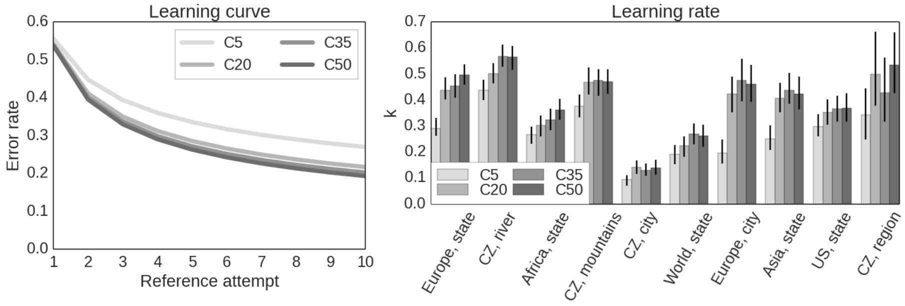

Fig. 3. Left: Global learning curve based on the power law ax^{-k}. Right: Learning rate k for different contexts. Error bars stand for 95% confidence intervals computed using bootstrapping.

ax^{-k}, where x is the number of attempts, a is the initial error rate, and k is the learning rate.

When we mix data from all contexts together and analyze learning only on the global level, more difficult practice seems to lead to better learning, see Fig. 3. (left). Fig. 3. (right) shows more detailed analysis for individual contexts. Instead of looking at the whole learning curves, we assume that the initial error rate a is the same for all conditions within the same context and we compare only their learning rate (the parameter k in the power law). The learning rate differs among some contexts (e.g., Czech cities vs. European states) due to differences in the number of items and other factors. Here, we are mainly interested in the comparison of our experimental conditions within individual contexts. The general trend is the same as in the case of the global learning curve with the largest differences being between C5 and other conditions. The size of differences is related to different behaviour of conditions within individual contexts – number of items available to practice and actually achieved error rate (e.g., Europen countries are much easier than Czech cities for most of our users).

5 Discussion

We performed an experiment with varied difficulty of items in a widely used open online educational system. The most interesting result is the difference between “short term engagement” (not leaving immediately) and “long term engagement” (prolonged usage of the system). Easy questions lead to better short term engagement, whereas difficult questions are better for the long term engagement. We also evaluated learning improvement, which is better for more difficult questions (the main difference being between very simple questions and others). These results are in contrast with previous research [2,6], which may be due to different learning domain (procedural knowledge in mathematics vs. declarative knowledge in geography). The issue of optimal difficulty thus warrants more attention in research.

These results have specific consequences for the studied system and for closely similar systems (e.g., vocabulary learning) – it seems that the system should start with easy questions "to hook learners up" and then switch to more difficult questions. But more importantly, the results have important methodological consequences for evaluation and optimization of educational systems. It is tempting to use "short term engagement" as a proxy for system quality, because this metric can be easily and quickly measured (as opposed to learning or long term engagement); this has been done for example in [3,8]. Our results show that this approach can be misleading and that it is important to use a "multi-criteria approach" (using techniques like [5]) since both engagement and learning are important in open online educational systems.

References

1. Sami Abuhamdeh and Mihaly Csikszentmihalyi. The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. *Personality and Social Psychology Bulletin*, 38(3):317–330, 2012.
2. Brenda RJ Jansen, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, 24:190–197, 2013.
3. Mohammad M Khajah, Brett D Roads, Robert V Lindsey, Yun-En Liu, and Michael C Mozer. Designing engaging games using bayesian optimization. In *Computer-Human Interaction*, 2016.
4. Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popović. Towards automatic experimentation of educational knowledge. In *Human Factors in Computing Systems*, pages 3349–3358. ACM, 2014.
5. Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *Educational Data Mining*, pages 161–168, 2014.
6. Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 89–98. ACM, 2013.
7. Brent Martin, Antonija Mitrovic, Kenneth R Koedinger, and Santosh Mathan. Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3):249–283, 2011.
8. Jan Papoušek and Radek Pelánek. Impact of adaptive educational system behaviour on student motivation. In *Artificial Intelligence in Education*, volume 9112 of *LNCS*, pages 348–357. Springer, 2015.
9. Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining*, pages 6–13, 2014.
10. Jan Papoušek, Vít Stanislav, and Radek Pelánek. Evaluation of an adaptive practice system for learning geography facts. In *Learning Analytics & Knowledge*, 2016. To appear.
11. Jan Papoušek, Vít Stanislav, and Radek Pelánek. Evaluation of the impact of question difficulty on engagement and learning. Technical Report FIMU-RS-2016-02, Masaryk University, 2016.