FI MU

Faculty of Informatics
Masaryk University Brno

Evaluation of the Impact of Question Difficulty on Engagement and Learning

by

Jan Papoušek
Vít Stanislav
Radek Pelánek

FI MU Report Series **FIMU-RS-2016-02**

 March 2016

Publications in the FI MU Report Series are in general accessible via WWW:

`http://www.fi.muni.cz/reports/`

Further information can be obtained by contacting:

Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic

Evaluation of the Impact of Question Difficulty on Engagement and Learning

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

Vít Stanislav
Masaryk University
slaweet@mail.muni.cz

Radek Pelánek
Masaryk University
pelanek@fi.muni.cz

April 4, 2016

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 and highlight the issue of attrition bias. This paper is an extended version of the paper [13] presented at Intelligent Tutoring Systems conference.

1 Introduction

Open online educational systems like Khan Academy or Duolingo are on the rise. They often provide content to a huge number of learners and even a small decision like choosing the right value of a parameter can affect millions of learners. Although these systems should definitely be optimized with respect to learning, it is also necessary to optimize them with respect to engagement – they need to keep learners' attention and motivate them to return repeatedly. Learners are not forced to be active within these systems,

they often study in their free time on their own and they enter and quit the system when they want to.

Making practice suitably challenging is one of the key goals of adaptive educational systems and this issue has been previously addressed from several directions. The general idea that the best activity is neither too easy nor too difficult was formulated as Inverted-U Hypothesis [1]. More specifically, Lomas et al. [6] studied optimal difficulty in the case of a simple online educational game, they found higher engagement for easier questions and a conflict between learning and engagement. A similar research was done using Math Garden software [2]. The authors compared three conditions (target success rate 60%, 75%, 90%) and showed that the easiest condition led to the best learning (mediated by a number of solved tasks). Other authors have used more complex experimental techniques to find optimal parameter values (e.g., Bayesian optimization), but optimize only with respect to short term engagement [3] or short term transfer [4].

We report results of a randomized trial evaluating the 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 [11] (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 [10], but only with respect to engagement. Here we provide more detailed analysis including also learning. The used methodology is similar to a previous work [12], which compared an adaptive and a random version of the system. Here, we pay much more attention to issues related to high level of data aggregation, attrition bias, and a conflict between short and long term engagement.

Analyzing data from the experiment containing conditions targeting to 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 work [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.

2.1 The Used System

The system estimates learners' knowledge and based on this estimate it adaptively constructs multiple-choice (2–6 options) or open questions of suitable difficulty [11]. The adaptive behaviour of the system is based on models of learners' knowledge, which for each learner and item provide the current prediction of knowledge (probability of correct answer). These models have been described and evaluated in previous work [9, 11], here we use them as a 'black box'.

An important factor that influences the evaluation and interpretation of results are different contexts within the system. Learners can use the system with different maps and types of places (e.g., European states); 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 detailed analysis we use 10 contexts with most answers (listed in Figure 1). The system is available in Czech, English, German, and Spanish, but currently most users are from the Czech Republic.

2.2 Experimental Conditions

The system uses a target error rate and adaptively constructs questions in such a way that learners' achieved performance is close to this target [10]. In our experiment we evaluate four experimental conditions which differ only in one aspect – the target error rate: 5%, 20%, 35%, 50%. Learners were assigned to one of the conditions randomly when they entered the system for the first time. In the following text we denote the conditions as C5, C20, C35, and C50. The experiment was performed from November 2015 to January 2016, 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).

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

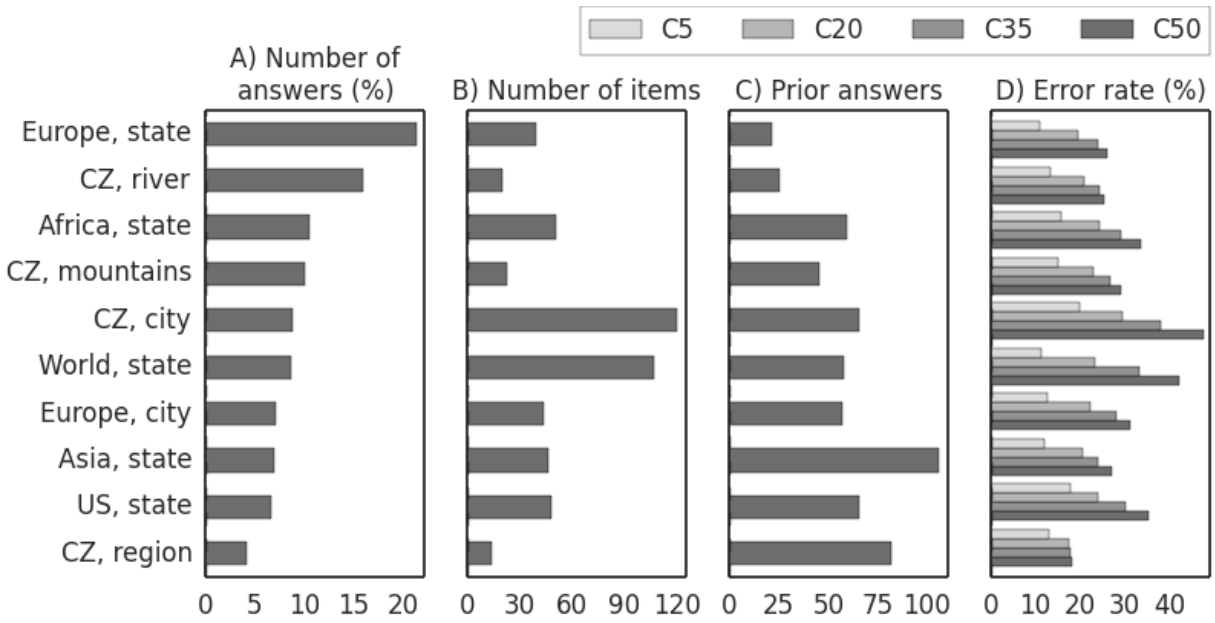

Figure 1: Top 10 mostly used contexts available for learners to practice. A) percentage of answers in the analyzed data set, B) number of items, C) average number of answers in other contexts prior to the first answer in given context, D) average error rate per experimental condition ignoring reference answers.

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 (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, 12].

2.3 Notes on System Behaviour

Although the system tries to achieve a specific error rate, the real error rate is not exactly the same. There are at least three causes – noisy user behaviour, imperfect predictive model, and insufficient number of appropriately difficult items. The achieved error rate depends on a specific context, see Figure 1. Figure 2 illustrates that largest differences among conditions can be observed at the beginning of the practice. These differences, however, decrease during the practice (all conditions except C5 are in most contexts really similar after 40 questions) as some learners quit their practice and others master

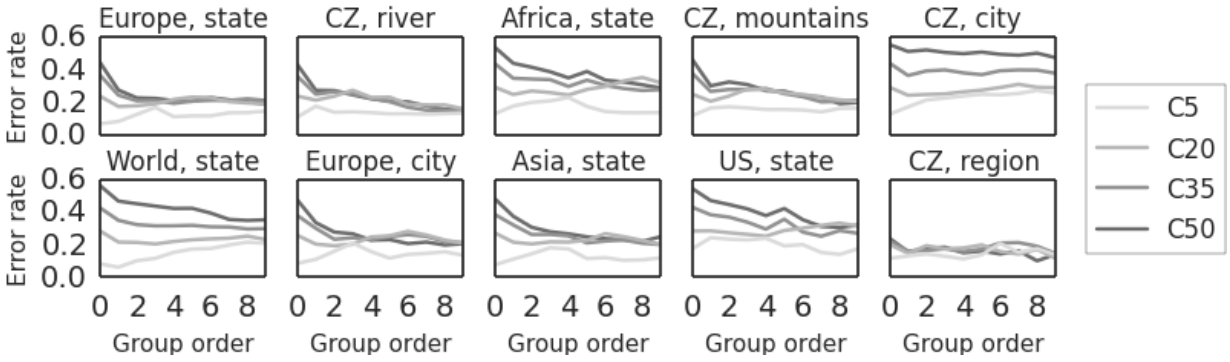

Figure 2: Error rate during the practice according to a number of learners' attempts (groups of 9, reference answers are excluded) for different conditions.

items from a particular context. Speed of this "convergence" differs for different contexts; it is lower for more difficult contexts with many items (e.g., Czech cities, world states).

Figure 3 shows a median of the first presentation order according to the difficulty of items predicted by the currently used learner model. E.g., Serbia (the most difficult item from European states) is typically the second item the system ask about in case of C50; on the other side, Russia (the easiest item from European states) is typically the first item in case of C5. We see that in some cases conditions radically differ (e.g., for Asian states, C5 goes from the easiest item to the most difficult one, while others in the other direction), whereas in some cases the order is quite similar (Czech cities). We also note that C5 often asks questions with only 2 options which leads to a faster speed of answering, but we assume this feature is not fundamental for the presented analysis.

3 Engagement

To evaluate engagement we consider (*1*) survival rates (i.e., proportion of learners who answer at least k questions), (*2*) probability of returning to the system (after a delay of more than 10 hours; the specific duration of delay is not important for presented results), and (*3*) self-reported perception of practice difficulty (after 30 answers the system displays a dialog with question "What is the difficulty of asked questions?" and the following options: "Too Easy", "Appropriate", "Too Difficult"; in the present experiment learners provided more than 40 000 of these ratings). 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

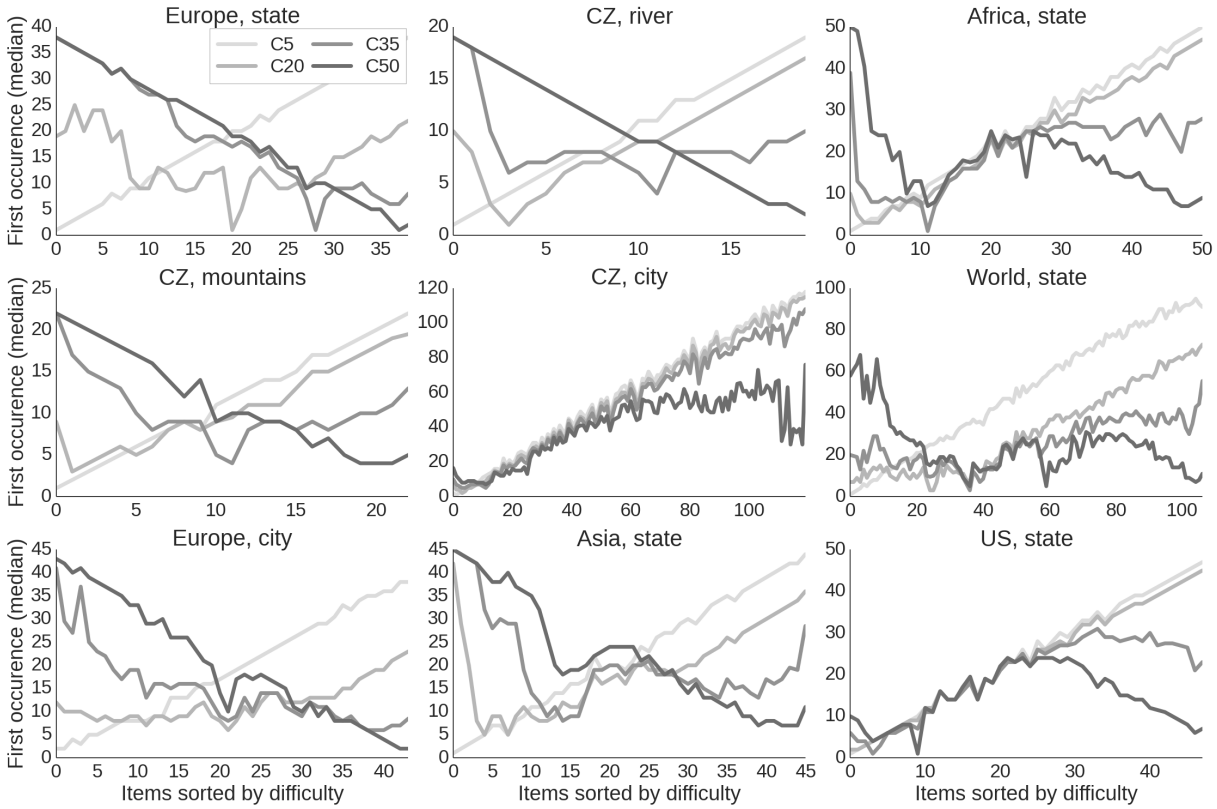

Figure 3: Median of the first presentation order according to the difficulty of items predicted by the currently used learner model for different contexts.

easier questions enhance engagement at the beginning, more difficult conditions engage more learners later on.

3.1 Global Comparison

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, see Table 1. On the other side, 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 $\pm 0.88\%$). Note that after 30 or more questions, the conditions C35 and C50 no longer achieve their target error rate in a lot of contexts (see Figure 2). Because of these differences in behaviour among contexts, we further analyze survival in contexts separately. Return rate increases with the difficulty of questions, the largest difference being between C5 and other conditions, see Table 1.

Table 1: Global comparison of conditions with respect to engagement.

Condition	Survival rate*	Return rate†
C5	89.2%	15.2%
C20	87.0%	16.0%
C35	84.0%	16.6%
C50	81.2%	16.8%

* after 10 questions, confidence interval ±0.77%

† confidence interval ±0.75%

3.2 Comparison within Contexts

There are quite large differences among the contexts (see Figure 4), 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 between conditions is mostly aligned with differences in their behaviour in the particular context (confront Figure 4 with statistics in Figures 1, 2, and 3).

Short term survival (Figure 4 A) differs in all contexts in favour of easier conditions. Differences among conditions in case of some contexts are lower, probably because of attrition bias – number of other contexts practiced prior to a particular context. Extent of that effect depends on a number of items practiced in prior contexts, which varies among contexts (see Figure 1 C). In case of long term survival (Figure 4. B), the trend is quite opposite, although for individual contexts the differences are typically rather small. This contrast is best seen on European states (context with most data), where we see a reliable difference between C50 and C5.

Figure 4 C shows probability of return for different contexts. Differences are often not significant within particular context, but generally we have higher probability of return for more difficult questions (C35, C50).

3.3 Explicit Feedback

In all conditions most learners rate questions as "Appropriate" (note, however, that dissatisfied learners could have left before the first rating). Users are most satisfied

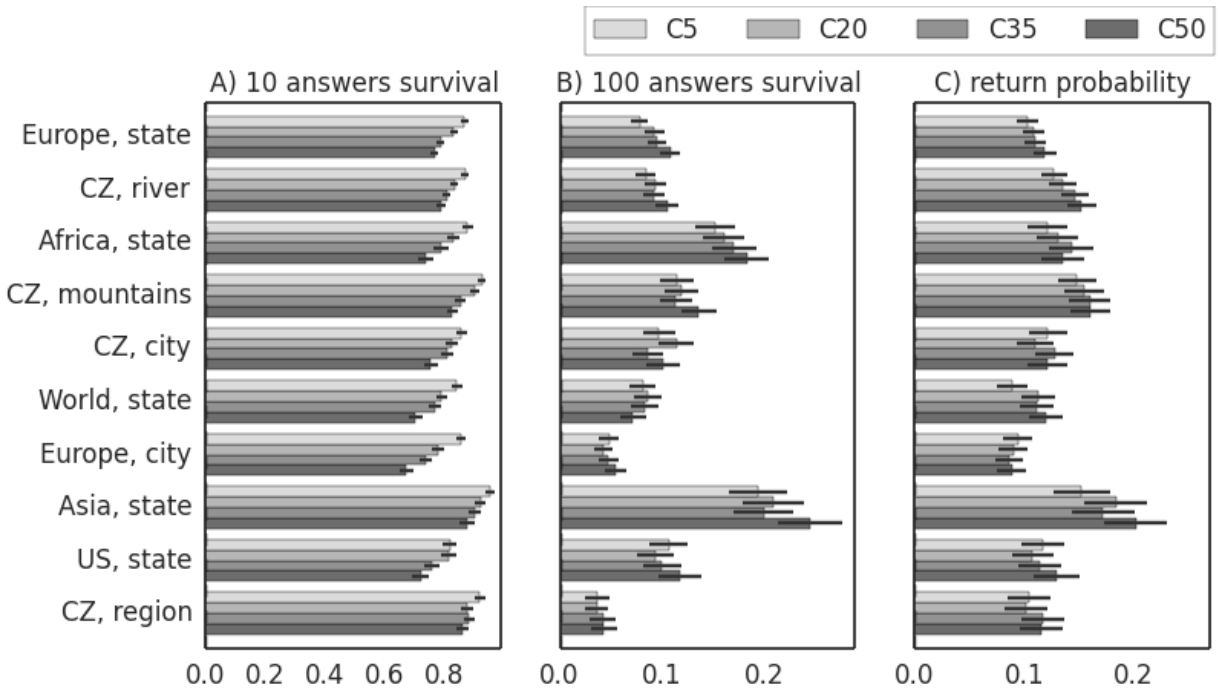

Figure 4: Survival analysis (A, B) and probability of return after 10 hours (C) for 10 most practiced contexts and 4 experiment conditions. Error bars represent 95% confidence intervals.

with C35 (66% of "Appropriate" ratings) and least satisfied with C5 (55%). Other two conditions are only slightly below C35. Unsurprisingly, C5 has more "Too easy" ratings (35%) than other three conditions (19% to 21%) and "Too difficult" ratings exhibit the opposite trend (with 10 to 17 % of ratings). So according to learners' feedback, the best condition is C35.

In addition to analyzing learners' feedback based on target error rate we have also considered learners' feedback based on their achieved error rate. Most learners report that question difficulty is appropriate when their achieved error rate is around 30% (without much regard to the condition). Even though there is an error rate when most learners are satisfied, at least a third of learners would still prefer easier or harder questions. An interesting direction for future work is to evaluate whether changing target error rate according to learners' rating enhances their engagement.

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. An alternative approach would be to use model based detectors of learning, i.e., to fit a learner model (e.g., Bayesian Knowledge Tracing or Performance Factor Analysis) to data and interpret the model parameters as an evidence of learning. Such results would be, however, influenced by violations of simplifying assumptions of learner models and by feedback loops between data collections and learner models [14].

4.1 Learning curves

We construct a learning curve [7] in the same way as in [12]. 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. In accordance with previous research [7, 12] we assume that the learning curve corresponds to the power law, i.e., the error rate can be expressed as ax^{-k}, where x is the number of attempts, a is the initial error rate, and k is the learning rate.

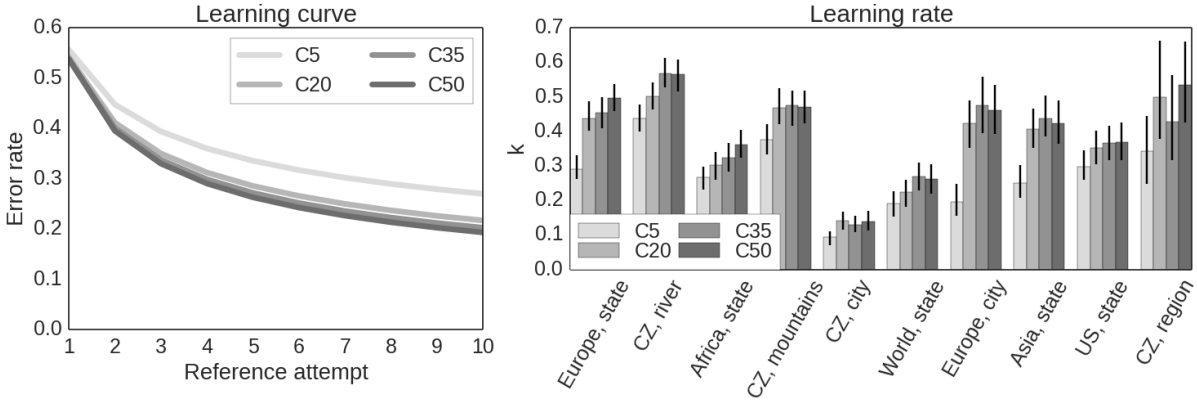

Figure 5: Left: Global learning curve based on the power law ax^{-k}. Right: Learning rate k for different contexts. Error bars stand for 95% confidence intervals computed using bootstrapping.

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 Figure 5 (left). Figure 5 (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 – which items are practiced in which order (see Figure 3) and what real error rate is actually achieved (see Figure 2).

We also performed other kinds of analysis, like filtering out learners having insufficient number of answers (to avoid attrition bias), looking only at answers after 10 hours delay (short term vs. long term learning), or constructing a learning curve for response time (time a learner spends by answering a question). Results are very similar in all cases – C5 is clearly the worst and differences among the others are small.

4.2 Attrition Bias

The interpretation of learning curves is complicated by attrition bias, which is a type of selection bias often present for example in medical experiments. Previous research identified attrition bias due to mastery learning [8] or differences in engagement [12]. Figure 6 shows that the selection of learners who stay for at least k reference attempts is different for individual conditions. In case of C5 and European states, learners who stay longer are associated with a lower initial knowledge. These learners have probably higher error rate later during the practice than learners with high initial knowledge (who left in the case of C5). The above described large difference in learning rate between C5 and the other conditions can be partially caused by this phenomenon.

Even more interesting is that the attrition differs among different contexts. Since users select a context for their practice themselves and some contexts are favored by the user interface, some contexts are more likely to be selected as the first one. Figure 1 B shows how many answer users have before they start to practice the given context (in average), e.g., Asian states are practiced by learners having already quite a lot of answers within the system. This filters out a certain subset of users, which leads to a different error rate in case of the first reference answer. Different error rate is present, even

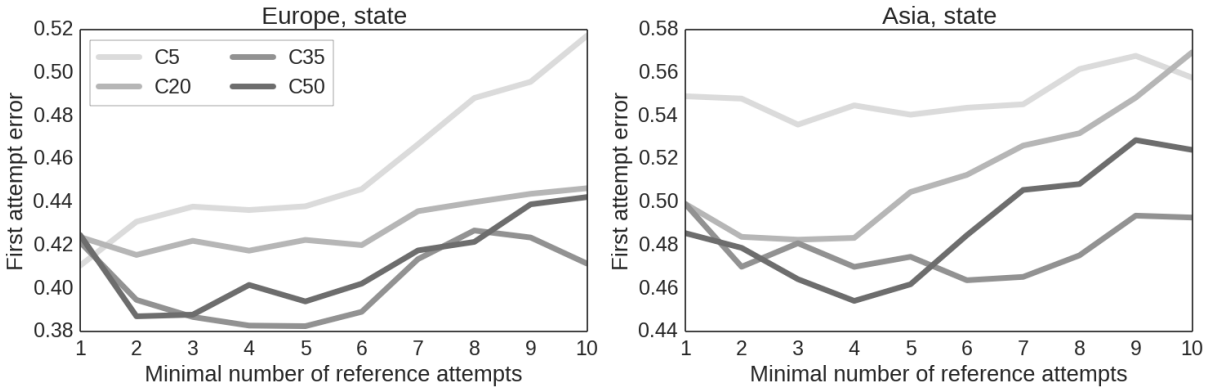

Figure 6: Attrition bias. The first attempt error rate depending on how many reference questions the learner answered.

though the studied conditions can not influence it, because the first reference question is random.

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); there is also a slight difference between different measures of the long term engagement (number of answers vs. probability of return). 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, 10]. 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.

Results of our experiment also highlight the issue of attrition bias. The population of users that stay within the system depends on the behaviour of the system and can be different for different variants of the system (e.g., with easier questions we have bias towards learners with lower prior knowledge). This issue is particularly pressing for open online education systems, which many learners use on their own (as opposed to usage in classes). As this kind of systems is currently on the rise, the evaluation of learning under the presence of attrition bias needs more research attention.

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] R Charles Murray, Steven Ritter, Tristan Nixon, Ryan Schwiebert, Robert GM Hausmann, Brendon Towle, Stephen E Fancsali, and Annalies Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

[9] Juraj Nižnan, Radek Pelánek, and Jiří Řihák. Student models for prior knowledge estimation. In *Educational Data Mining*, 2015.

[10] 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.

[11] 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.

[12] Jan Papoušek, Vít Stanislav, and Radek Pelánek. Evaluation of an adaptive practice system for learning geography facts. In *Learning Analytics & Knowledge*, 2016. To appear.

[13] Jan Papoušek, Vít Stanislav, and Radek Pelánek. Impact of question difficulty on engagement and learning. In *Intelligent Tutoring Systems*, 2016. To appear.

[14] Radek Pelánek, Jiří Řihák, and Jan Papoušek. Impact of data collection on interpretation and evaluation of student model. In *Learning Analytics & Knowledge*, 2016.