Evaluation of Learners' Adjustment of Question Difficulty in Adaptive Practice of Facts

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

Radek Pelánek
Masaryk University
Brno, Czech Republic
pelanek@mail.muni.cz

ABSTRACT

Personalized educational systems are able to provide learners questions of specified difficulty. Since learners differ, the appropriate level of difficulty may vary and it may be impossible to find an universal setting. We implemented a version of an adaptive educational system for geography practice that allows learners to adjust difficulty of questions. We evaluated this feature using a randomized control experiment. The overall results show only a small effect of the adjustment. A more detailed analysis, however, shows that for some groups of learners the effect can be important, although not necessarily advantageous. The collected data from the experiment provide insight into how to tune question difficulty automatically.

ACM Reference format:
Jan Papoušek and Radek Pelánek. 2017. Evaluation of Learners' Adjustment of Question Difficulty in Adaptive Practice of Facts. In *Proceedings of UMAP '17, Bratislava, Slovakia, July 09-12, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3079628.3079642

1 INTRODUCTION

User modeling allows us to develop personalized learning environments that make learning experience tailored towards individual learners. Using learner modeling techniques [2] we can estimate probability that a learner can answer a question or solve a problem. Based on these predictions we can automatically choose items of appropriate difficulty [9]. But what is an appropriate level of difficulty? This is typically a parameter that is specified externally by developers of a learning system. The choice of this parameter has been addressed in previous research, but without clear results [4, 5, 8].

A natural idea is to allow learners to manipulate the difficulty of questions. In addition to better tailored system behaviour, previous research suggests that a sense of control (or even perception of control, rather than the actual objective level of control) can increase engagement [6]. On the other hand, research on self-regulated study [1] shows that people are prone to mismanaging their own learning. A recent research explored self-adaptation of difficulty in math practice [3]. The authors did not find any impact of the availability of difficulty setting on learning, engagement, or students' self-belief. The study, however, has several limitations, e.g., the setting of an error rate interacted with gamification aspects of the user interface and the used sample size was small (48 students in each condition). We present a similar experiment, but for a different type of knowledge (learning geography facts) and using a large scale experiment with thousands of users and millions of answers.

2 EXPERIMENT

We use a system for an adaptive practice of geographical facts, e.g., names and locations of countries or cities. The system is available online at outlinemaps.org. Learners can use the system with different 'contexts' (combinations of a map and a type of place, e.g., European states) . The system collects data about the correctness of answers and based on the collected data it estimates the current knowledge of a particular learner and personalizes the provided practice [9]. A key parameter in the adaptive algorithm is "target difficulty", which sets the average error rate of learners that the system is aiming at.

In our experiment we let some learners modify the parameter based on their preferences. The practice within the system is presented in groups of 10 questions; after each series of 10 questions the systems shows a summary feedback to learners. At this moment we have inserted a new dialog box with a question "How difficult would you like the questions to be?" with 5 choices: "much harder", "harder", "same", "easier", "much easier". We call answers to these questions "ratings" (not "settings", because in a placebo condition they do not have any impact on the algorithm).

We have performed a standard randomized control trial with three experimental conditions: 1) *normal* – a control group, a standard version of the system without the dialog box; 2) *placebo* – the dialog is shown, but does not have any impact on the behaviour of the adaptive algorithm; 3) *adjustment* – the dialog is shown and the answer changes the target error rate (+20%, +10%, 0%, -10%, -20%).

In all cases the initial setting of the target error rate parameter is 35% (the value is based on results of the previous experiment [8]). The experiment was performed from October 2016 to January 2017 and we have collected roughly 8 200 000 answers from 85 000 learners. To make our research reproducible we make the analyzed data set available[1].

3 RESULTS

At first, we analyze ratings provided by users. Mostly, we have only one rating from a particular learner per context. Majority of learners do not provide any rating at all. Since the ratings are provided after finishing a practice set, we assume the main factor determining a learner's rating is an error rate achieved during the recent practice set, therefore we divide all ratings to buckets based

UMAP '17, Bratislava, Slovakia
 978-1-4503-4635-1/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3079628.3079642

[1] http://data.outlinemaps.org/2016-ab-user-difficulty-adjustment.zip

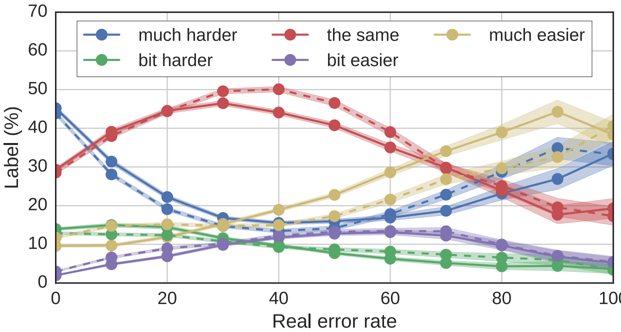

Figure 1: Learners' ratings with respect to the error rate achieved during the recent practice set. Solid lines stand for placebo condition, dashed lines stand for adjustment condition. The shaded areas shows 95% confidence intervals.

on the error rate. Figure 1 shows the relation between ratings and the recent error rate (based on the last 10 questions).

The basic relation is intuitive – successful learners want more difficult questions, unsuccessful learners want easier questions. The "appropriate" ratings have the shape of inverted-U curve with the maximum at the error rate around 35%. This result is in agreement with our previous experiments that showed that target error rate 35% is suitable [8].

For high error rates the results are intriguing. As could be expected the ratio of "bit harder" ratings is very small. Unexpectedly, highly unsuccessful learners often provide "more difficult" rating. Although the number of highly unsuccessful learners is relatively small, this trend is statistically significant and consistent for both placebo and adjustment conditions. We interpret this trend as presence of a systematic "irony" in responses of a subgroup of users and we hypothesize that this behaviour is connected to disengagement with the system. This result should serve as a caution – preferences expressed by learners may reflect not just their true preferences with respect to the concerned question, but may also incorporate other aspects of their (affective) state.

The data also show a relation between ratings and context difficulty. The percentage of "much harder" ratings increases with decreasing context difficulty (e.g., European states are easier than African cities, at least for users of the used system). This observation indicates that our algorithm for adaptive practice is not adaptive enough and there is room for improvement – the algorithm could take into account the difficulty of a particular context.

The main point of the experiment is to find whether the dynamic adjustment leads to higher engagement and learning. As a measure of engagement we use a survival rate – the ratio of learners who answer at least 100 questions. To compare learning among conditions, we utilize "reference questions" – for each context separately every tenth question is selected randomly without any influence of the adaptive algorithm and we use data from these questions to construct learning curves. For more detailed discussion of the used evaluation approach see [7].

On the global level, we do not see any large differences among conditions (100 question survival rates are: normal 29.1%, placebo 27.5%, adjustment 28.3%). As expected, the engagement for the placebo condition is worse than for the control group. The dialog box asking for learners' ratings has negative impact on learners' engagement. For the adjustment condition the negative effect of the dialog is partially compensated by the benefits of the difficulty adjustment. However, the benefits of the difficulty adjustment are not sufficient to overweight the disadvantage of the additional dialog box. For learning we also do not have any significant differences in overall learning rates. Since many learners do not provide any ratings at all, it is not much surprising – most learners in placebo and adjustment conditions keep the original target error rate and thus their practice is the same as for the control group.

However, a more detailed analysis shows that for specific cases there are some trends. When we consider 30% easiest contexts (e.g., Europe states), the learning is actually worse for the adjustment condition. It is probably caused by learners' tendency to set lower difficulty even on easy contexts (e.g., by externally motivated learners from schools). Another more detailed analysis disaggregates the overall results with respect to learners – specifically based on their preference for easy or difficult questions. The results are interesting particularly for the group of learners who prefer easy questions – for this group the adjustment hampers the speed of learning, but increases engagement. The adjustment is clearly beneficial only for the group of learners who prefer difficult questions – for them it improves the speed of learning without any significant impact on engagement.

Overall, however, our experiment suggests that instead of giving learners control over the difficulty setting, we should rather develop better methods for automatic adjustment of target difficulty.

REFERENCES

[1] Robert A Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology* 64 (2013), 417–444.

[2] Michel C Desmarais and Ryan SJ Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 9–38.

[3] Brenda RJ Jansen, Abe D Hofman, Alexander Savi, Ingmar Visser, and Han LJ van der Maas. 2016. Self-adapting the success rate when practicing math. *Learning and Individual Differences* 51 (2016), 1–10.

[4] Brenda RJ Jansen, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas. 2013. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences* 24 (2013), 190–197.

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

[6] Thomas W Malone and Mark R Lepper. 1987. Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction* 3, 1987 (1987), 223–253.

[7] Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Evaluation of an Adaptive Practice System for Learning Geography Facts. In *Proc. of Learning Analytics & Knowledge*. ACM, 40–47.

[8] Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Impact of Question Difficulty on Engagement and Learning. In *Proc. of Intelligent Tutoring Systems (LNCS)*, Alessandro Micarelli, John C. Stamper, and Kitty Panourgia (Eds.), Vol. 9684. Springer.

[9] Radek Pelánek, Jan Papoušek, Jiří Řihák, Vít Stanislav, and Juraj Nižnan. 2017. Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction* 27, 1 (2017), 89–118.