Exploring the Utility of Response Times and Wrong Answers for Adaptive Learning

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

ABSTRACT

Personalized educational systems adapt their behavior based on student performance. Most student modeling techniques, which are used for guiding the adaptation, utilize only the correctness of student's answers. However, other data about performance are typically available. In this work we focus on response times and wrong answers as these aspects of performance are available in most systems. We analyze data from several types of exercises and domains (mathematics, spelling, grammar). The results suggest that wrong answers are more informative than response times. Based on our results we propose a classification of student performance into several categories.

INTRODUCTION

Personalization in adaptive learning systems is based on observing and modeling student performance. Using the observed performance, we estimate a student state and this estimate is then used to guide the behavior of the system. The observed student performance is useful also for other kinds of analysis, e.g., for analyzing item similarity and improving domain models.

Current research mostly utilizes only the primary aspect of student performance – the binary information about the correctness of answers. But are other aspects of student performance can also be easily collected and utilized. Some aspects can be quite specific to a particular system – for example, the usage of hints is a potentially valuable source of data about student knowledge, but systems significantly differ in the mode of presentation of hints, the richness of information they contain, and many other important details. In this work, we focus on two aspects of student performance that are available in wide range of systems in a relatively uniform way: response times and wrong answers.

Both response times and wrong answers have been studied before. The use of response times has been conceptually explored in the context of adaptive testing [7]. In adaptive learning, they have been used together with the correctness for student modeling [2] and for analysis of slip and guess behavior [1]. Wrong answers are known to have a highly skewed distribution [6], and their use for student modeling has also been explored [8]. However, the usage of both response times and wrong answers is still quite limited compared to the extensive research and application of models based only on the correctness of answers, e.g., most papers about the much studied Bayesian Knowledge Tracing model utilize only the correctness of answers [4].

One approach for utilizing response times, wrong answers, and other kinds of data about student performance is to incorporate them directly into particular applications, e.g., into a student model for estimating knowledge or into a technique for modeling similarity of items based on student performance. Another approach is to use the observed data to classify student performance into one of several classes (e.g., "great performance", "correct, but slow", "incorrect, but reasonable", "fast guessing") and then use these classes for specific applications. With this approach, we lose some nuances of the data, but with well-designed classification we may be able to lose only a little information and get significant simplification. Note that the currently dominant approach, where only the correctness of answers is used, is an extreme case of this "classification of performance".

We use data from an adaptive educational system to explore how much information do response times and wrong answers contain and whether they can be useful for adaptation. We use data from several types of exercises (multiple-choice, free answer, drag&drop) to explore how general are the results of the analysis. Based on the results of this analysis we propose a specific classification of student performance.

ANALYSIS

For our analysis, we use data from the Umíme educational system (`umimeto.org`), which is a system providing practice of mathematics, Czech grammar and spelling, English, and other domains. The system is targeted at Czech native speakers, particularly elementary school children. For our analysis, we use data from a variety of domains, exercise types and data sizes – see Table 1.

The system is adaptive as it uses mastery learning (the basic algorithm used in the system is described in [5]). The practice consists of repeatedly asking students questions from a particular knowledge component until mastery is reached – during the

L@S'18, London

ISBN .

DOI:

Table 1. Overview of used data.

domain	exercise type	answers (×1000)
math: expressions	free text question	658
math: expressions	choice from 2 options	904
math: word problems	free text question	73
Czech: spelling	choice from 2 options	10406
Czech: grammar	drag&drop	541
English: vocabulary	free text question	146
English: grammar	choice from 2 options	151

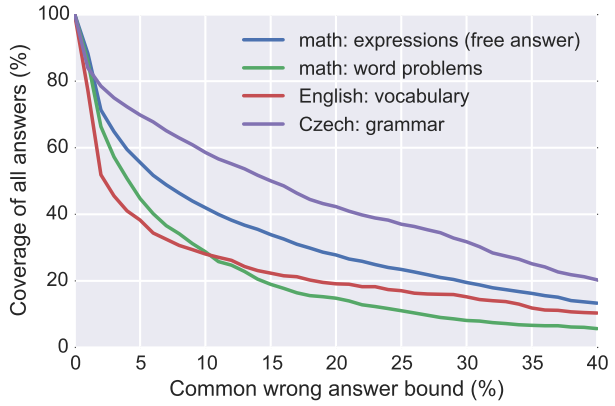

Figure 1. Coverage of wrong answers for different values of the bound that specifies which wrong answers are considered "common".

practice of a knowledge component items are chosen randomly from a set of (relatively homogeneous) items.

Wrong Answers

In exercises where a student provides a free text answer, it may be useful in the case of a wrong answer to look at the particular answer of the student. This approach is also relevant for exercises where answers are not completely free, but the range of choices is large, e.g., drag&drop exercise where the answer is a permutation of words in a sentence.

Previous research has repeatedly shown that the distribution of wrong answers is highly skewed (e.g., [6]), i.e., for most items few typical wrong answers are covering most student mistakes. Our data confirm this pattern. Typically the most common wrong answer comprises 15-20% of all wrong answers for a particular item. In some cases, the ratio can be even over 70% (examples from mathematics: an item $12-6+4$ and an answer 2, an item 4^2 and an answer 8).

For utilizing wrong answers in student modeling, it would be optimal to have a mapping of wrong answers into several categories, e.g., "important misconception", "typing error", "numerical mistake in calculation". However, such mapping would be dependent on a particular knowledge component and would require extensive manual effort (or development of new automatic techniques). Here we explore whether even a simple division of wrong answers between "common" and "uncommon" can be useful.

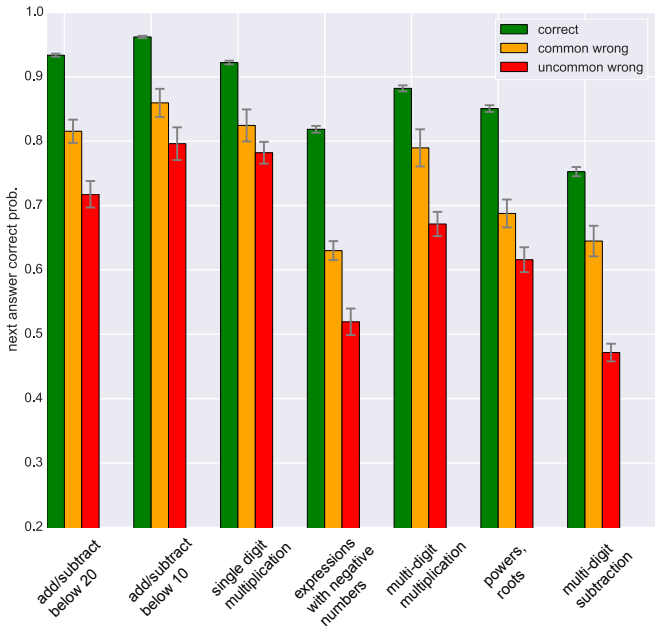

Figure 2. Probability that the next answer will be correct conditioned on the type of the current answer. Error bars show 95% confidence intervals. Results for math expressions with free answers.

The simplest way to classify wrong answers into common and uncommon is to consider as a common wrong answer any answer which comprises at least $B\%$ of all wrong answers. The question is how to choose the threshold B. To explore this question, we analyzed coverage of common wrong answers for different settings of this bound. Figure 1 shows the results for four types of exercises. For the analysis, we consider only items that have at least 50 wrong answers. Based on this analysis we suggest a bound 10% for the classification of an answer as a common wrong answer. With this bound common wrong answers comprise between one third and one half of all answers.

Can this classification of answers be useful for student modeling? To explore this question we use very simple analysis – we analyze the probability of the next answer being correct conditioned on the current answer. Figure 2 shows this analysis for several knowledge components in mathematics (an exercise with free text answers). We see that there is a consistent difference between common and uncommon wrong answers – students who give a common wrong answer are more likely to answer the next question correctly. We also see that the utility of this distinction depends on a particular knowledge component. For some knowledge components, particularly the easy ones like the addition of one digit numbers, the difference is negligible. For more complex knowledge components (e.g., multi-digit subtraction) the difference can be quite pronounced with common wrong answers being closer to correct answers than to uncommon wrong answers.

Response Times

Medium response time for an item is related to labor-intensity of the item. It may be possibly orthogonal to the difficulty of the item, particularly for items with longer texts, where

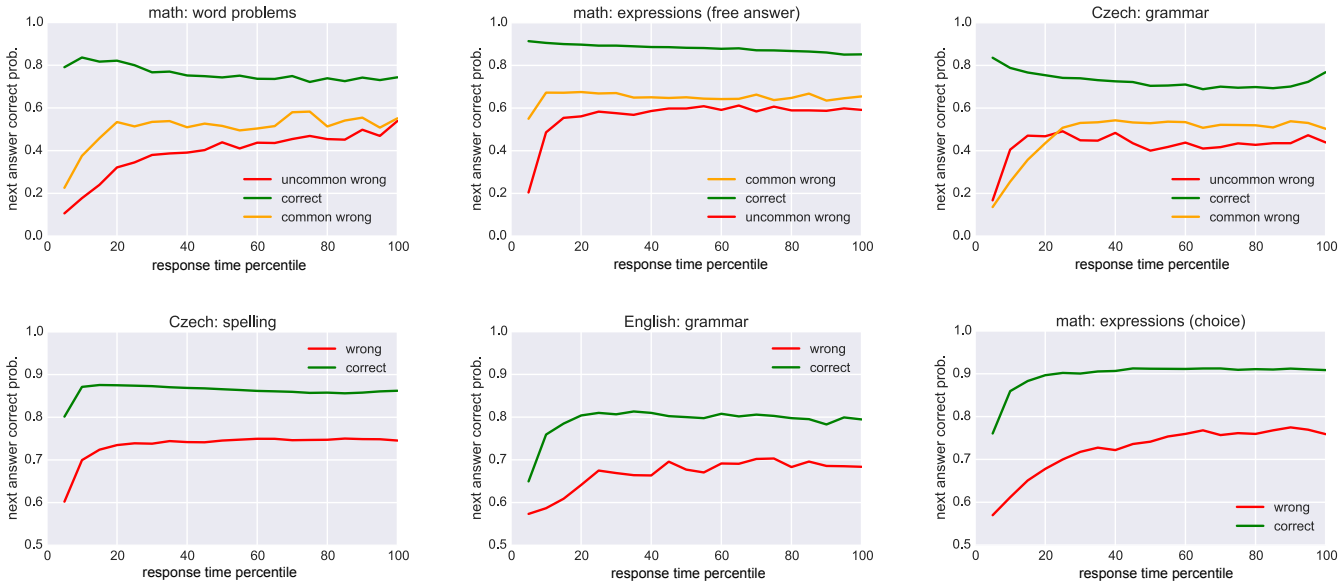

Figure 3. The relationship between the relative response time and the future performance.

the length of the text may significantly influence the response time. In our data for mathematics we typically see correlation around 0.7 between medium response time and success rate within one knowledge component, i.e., here the labor-intensity and difficulty of items are closely related.

To evaluate individual response times, it is useful to normalize them with respect to the particular item. To do so we transform raw response times into the "response time percentile", i.e., the percentage of other users that have faster response times on the item. To explore the usefulness of this information we analyze the probability that the next answer will be correct conditioned on the response time percentile for the current answer and for the classification of the current answer. Figure 3 shows results for several types of exercises. The top row shows exercises where students construct the answer, the bottom row shows exercises with a choice from 2 options (i.e., all wrong answers are the same). For exercises with constructed answers, we see that the differences between types of wrong answers are more important than response times.

For correct answers, the effect of response times is quite limited. In exercises with constructed answers we see a nearly linear relation between response times of correct answers and future performance. As can be expected, slower response means worse future performance. This effect is however minimal and probably not very useful for student modeling. For multiple-choice question we see a nonlinearity at the beginning: very fast answers are correlated with worse future performance compared to other correct answers – some of these fast correct answers are clearly obtained by guessing.

For wrong answers, the response times carries more information. Specifically, the distinction between very fast answers and other answers is now important also for exercises with constructed answers. Very fast answers are correlated with worse future performance – these are students who did not seriously tried to solve the problem. A boundary for these very fast answers is between the percentile 5% and 20%. Over the 20% percentile, there is a minimal effect for wrong answers. Interestingly, in cases where we can see a trend, it is in the opposite direction as in the case of correct answers – the longer response times are correlated with slightly better future performance (this trend has been already observed in a previous analysis of data from geography practice [3]).

We have also analyzed the relationship between the relative response time and the probability of leaving the exercise after the answer. This relation is mostly linear – higher response times correspond to a slightly higher probability of leaving. For some exercises the pattern is more complex – Figure 4 shows results for an exercise with a choice from 2 options (Czech spelling). Here for wrong answers we see a U-shaped pattern, i.e., the probability of leaving is higher for both short and long response times. These results suggest that response times may be more useful to modeling affect or behavior rather than the knowledge of students.

DISCUSSION

Our analysis suggests that wrong answers are more informative than response times, at least for modeling knowledge. Based on our analysis we propose to classify as a "common wrong answer" any answer that comprises more than 10% of all answers on an item. For response times the most important aspect seems to be the distinction between very fast answers and the remaining answers. Very fast answers are probably indicative of guessing and disengaged student behavior. A suitable boundary for "very fast" answer seems to be between 5% and 15% percentile of response times for the particular item.

Based on these results we suggest to classify student performance into one of the following categories:

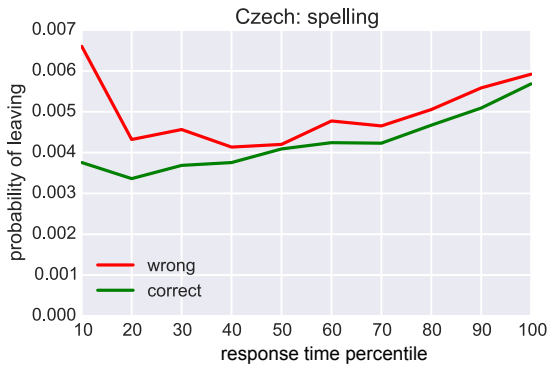

Figure 4. The probability of leaving the exercise after answering the current item.

- the correct answer,
- a very fast wrong answer,
- a common wrong answer and not very fast,
- other cases of wrong answers.

Potentially it may possible to further distinguish between the most common wrong answer and other common wrong answers (as suggested in [6]). Our data suggest that this distinction may be useful, but with the used data sets the differences are not yet stable.

Based on our results we hypothesize that using the proposed categorization in student modeling will lead to nontrivial improvement over models that utilize only the correctness of answers and to only a small disadvantage with respect to (necessarily more complex) models that utilize raw data about response time and wrong answers. This hypothesis needs to be further explored. We have performed our evaluation only in very simple setting (next answer correctness) and it is possible that the utility of response times or wrong answer may be different when used with more complex student modeling techniques.

Although we use data from several domains, all of them come from one type of system. Thus it is probable that the results are partially influenced by specific features of the used system, specifically the results may be influenced by the user interface of the system. In the used system there is no indication that response time is measured or that any attention is paid to specific values of incorrect answers. If such an indication would be available, the behavior of users may change. For example, the Math Garden software uses response times in the student modeling and indicates this in the user interface [2]. The effect of such user interface aspects on data used for student modeling needs to be explored.

ACKNOWLEDGMENTS

The author thanks Petr Jarušek for assistance with the data and fruitful discussions.

REFERENCES

1. Ryan SJ Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Proc. of Intelligent Tutoring Systems*. Springer, 406–415.
2. S Klinkenberg, M Straatemeier, and HLJ Van der Maas. 2011. Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education* 57, 2 (2011), 1813–1824.
3. Jan Papoušek, Radek Pelánek, Jiří Řihák, and Vít Stanislav. 2015. An Analysis of Response Times in Adaptive Practice of Geography Facts. In *Proc. of Educational Data Mining*. 562–563.
4. Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 313–350.
5. Radek Pelánek and Jiří Řihák. 2017. Experimental Analysis of Mastery Learning Criteria. In *Proc. of User Modelling, Adaptation and Personalization*. ACM, 156–163.
6. Radek Pelánek and Jiří Řihák. 2016. Properties and Applications of Wrong Answers in Online Educational Systems. In *Proc. of Educational Data Mining*. 466–471.
7. W.J. Van Der Linden. 2009. Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement* 46, 3 (2009), 247–272.
8. Yutao Wang, Neil T Heffernan, and Cristina Heffernan. 2015. Towards better affect detectors: effect of missing skills, class features and common wrong answers. In *Proc. of Learning Analytics And Knowledge*. ACM, 31–35.