

Properties and Applications of Wrong Answers in Online Educational Systems

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

Jiří Řihák
Masaryk University Brno
thran@mail.muni.cz

ABSTRACT

In online educational systems we can easily collect and analyze extensive data about student learning. Current practice, however, focuses only on some aspects of these data, particularly on correctness of students answers. When a student answers incorrectly, the submitted wrong answer can give us valuable information. We provide an overview of possible applications of wrong answers and analyze wrong answers from three different educational systems (geography, anatomy, basic arithmetic). Using this cross-system comparison we illustrate some common properties of wrong answers. We also propose techniques for processing of wrong answers and their visualization, particularly an approach to item clustering based on community detection in a confusion graph.

1. INTRODUCTION

A key advantage of computerized educational systems is their potential for personalization. By analyzing students' answers we can estimate their knowledge using student modeling techniques and adapt the behaviour of a system to the needs of individual students. Student models [6] typically utilize only information about correctness of answers. Online systems, however, collect (or can easily collect) much richer information, e.g., timing information [18] and specific details about answers and individual steps. In this work we focus on analysis of wrong answers.

Wrong or incomplete answers from online educational systems have been studied previously, but mostly just as a supplementary analysis to other research interests. For example, analysis of programming assignments in MOOCs [9, 14] shows that the distribution of wrong answers is highly skewed, containing few very common wrong answers. This research does not, however, focus on analysis of wrong answers, but rather on finding similar or equivalent solutions and their visualizations (as there are many ways how to write the same program) [7].

The observation that distribution of wrong answers is highly skewed holds not only for programming assignments, but also for other domains. For example, common wrong answers have been used for student modeling in mathematics [29], but this work uses only information about whether the wrong answer is common or not, it does not utilize actual values of wrong answers. Specific student answers were also modeled [8], but authors present only overall accuracy of the proposed model without discussion of specific mistakes.

Data analysis techniques has been used for analysis of mathematical errors with the goal of classification (explanation) of answers [13, 24]. The results show that it is possible to classify most wrong answers into one of few categories. Other data-driven techniques in educational data mining have focused mainly on programming assignments [10, 21]. Rather than “wrong answers” they utilize “incomplete solutions” and use them for automatic generation of hints (changes towards a correct solution).

In the wider context, wrong answers are related to misconceptions, which are intensively studied in pedagogical literature, e.g., misconceptions in mathematics [26] or chemistry [22]. This line of research focuses on understanding “buggy rules” used by students [4]. These rules are useful not just for educating teachers about student thinking, but also in development of intelligent tutoring systems. They can be also used as a basis of erroneous examples [1, 11]. Research in this direction is typically based on expert insight using only relatively small (and often qualitative) data and the focus is typically on complex skills.

In this work we focus on automatic techniques for analysis of large quantitative data, dealing with simple skills (learning of declarative knowledge and simple procedures). We describe analysis of wrong answers from three educational systems. Although the used systems share similar basic principles they cover widely different domains (geography, anatomy, basic arithmetic) and different learner populations (from kindergarten to university students). Thanks to the size of the used data set (millions of answers), results provide interesting insights into properties and potential of wrong answers. We describe specific examples of analysis and propose novel techniques for analysis and visualization of wrong answers. A key observation is that wrong answers in our three domain (geography, anatomy, basic arithmetic) share many properties and thus it should be feasible to carry insights and analysis techniques across domains.

2. POTENTIAL APPLICATIONS OF WRONG ANSWERS

In this section we outline potential applications of wrong answers. The presented applications are rather general and for a specific application they need to be more precisely quantified. In the next section we provide such specific analysis for three particular domains.

2.1 Student and Domain Modeling

Student and skill models [6] typically utilize only binary information about correctness of an answer (correct/incorrect). A more thorough analysis of wrong answers may improve student and skill modeling in several directions.

In modeling of cognitive skills, wrong answers may help to distinguish between absence of understanding and slips (careless errors, typos). Highly uncommon wrong answer is more likely to be a careless error, whereas common wrong answer is more likely to be a genuine mistake (unless caused by poorly designed user interface). Wrong answers may also be indicative of the level of knowledge and strategies that students are using. Consider for example a multiplication 5×5 : a student A answers quickly 30, whereas a student B answers 24 after a long time. This may indicate that the student A retrieved the answer (incorrectly) from declarative memory, whereas the student B made an error in a procedural strategy. Wrong answers can thus be useful for modeling cognitive processes of learners [27]. Moreover, they may be useful also for modeling affect and motivation [29]. Irrelevant, highly uncommon wrong answers (particularly when repeated and quickly delivered) are probably indication of disengagement rather than lack of knowledge.

Wrong answer may be useful also for domain modeling. Common wrong answers may indicate relations between topics and thus may be used for automatic detection of knowledge components. Even through these may be misconceived relations, when they are common, they may be useful for student modeling. Relations between items based on wrong answers may also be taken into account in the design of the user interface or in the item selection algorithm. Wrong answers can also be used for student clustering – different groups of students make different types of mistakes and need different treatment from the educational system (e.g., students with dyslexia or dyscalculia).

2.2 Construction of Items and Hints

A basic observation about wrong answers, which seems to be valid in many different domains, is that the distribution of wrong answers is often highly skewed, i.e., some mistakes are much more common than others. This feature of wrong answers is potentially very useful for construction of questions and hints (both manual and automatic).

Common wrong answers may highlight student misconceptions and thus provide inspiration for new items (problems). In the case of items with simple structure, wrong answers may even be used automatically, e.g., as competitive distractors in multiple choice questions [16]. Previous work [1, 11] explored the possibility of using erroneous examples in education. Common wrong answers provide useful material for creation of such examples.

Wrong answers may also be useful for development of hints, feedback to students, and other scaffolding aids. If the hints are developed manually by experts, wrong answers provide good way to prioritize the expensive work of an expert. Due to the skewed distribution of wrong answers it may be possible to quickly provide answer-specific feedback to most answers even in open environments [9]. It is also possible to generate hints automatically based on actions of other students with the same wrong answer [23].

2.3 Feedback for Learners, Teachers, and Tool Developers

Analysis of wrong answers can also bring more pragmatic advantages. A useful feature of personalized educational systems is an overview of mistakes made by a learner or a class. Such an overview can serve for example as a base for a review session. Teachers may use such overview to quickly detect common problems of their students and thus focus on problematic parts in classroom time or in personal consultations.

For tool developers common wrong answers may be useful as an indicator of problems with a user interface. For example, in a prototype of one of the systems used in this study there was a common wrong answer “1” in cases where the answer should have been “10”. This turned out to be a user interface issue – the system was expecting a single click on a “10” button, whereas users were trying to click buttons “1” and “0”.

For these types of applications, basic analysis of wrong answers should be easily accessible in educational systems for both teachers and system developers. Since there can be a large number of mistakes, it is important to make the listing of mistakes easy to navigate. To achieve this goal, we need to understand common features of wrong answers.

3. ANALYSIS OF WRONG ANSWERS

After the general discussion of properties and possible applications of wrong answers, we turn to analysis of specific data.

3.1 Used Systems and Data

The used systems cover three different domains (geography, anatomy, basic arithmetic) and are used by very different learners, but they have been developed by the same research group and share the basic principles. All of them focus on adaptive practice of declarative knowledge or simple procedures. Systems estimate learners’ knowledge and based on these estimates they adaptively select questions of suitable difficulty. They use a target success rate (e.g., 75%) and adaptively selects questions in such a way that the learners’ achieved performance is close to this target.

The used questions are either multiple-choice questions or “open questions” – either a free text answer or selection of any item from a provided context (e.g., “select Rwanda on the map of all African states”). For the analysis we use only answers to open questions, since the used multiple-choice questions have adaptively chosen distractors and this feature makes analysis difficult (due to the presence of feedback loops [19]).

The first system is Outline Maps (outlinemaps.org) for practice of geography facts (e.g., names and locations of countries, cities, mountains). Details of the behaviour of the system are described in [15, 16]. The used data set contains more than 10 million answers (with more than 1 million wrong answers) and is publicly available [17]. This data set is the largest of the three used data sets and it is at the core of the presented analysis. The application is currently used by hundreds of learners per day, majority of learners is from the Czech Republic since the interface was originally only in Czech. The geographical origin and language of students clearly influence interpretation of below presented results. However, our main point is not interpretation of particular results, but rather illustration of different insight that can be gained by the analysis of the data.

The second system is Practice Anatomy for practicing human anatomy (practiceanatomy.com). The main target audience of the system consists of junior medical students preparing for their anatomy exams. Currently, the system offers practice of more than 1800 items organized into 14 organ systems and 9 body parts. Learners can practice a selected organ system or a body part, or specify a more advanced practice filter as an intersection of a set of organ systems and a set of body parts. The system is available in Czech (with Latin terminology) and English. Most users are from the Czech republic. The analyzed data set contains over 380 000 answers.

The third system is MatMat (matmat.cz) for practice of basic arithmetic; its functionality is similar to for example Math Garden [24]. The system contains examples divided into 5 high level concepts (counting, addition, subtraction, multiplication, division), each of these concepts contains around 50-700 items, over 2 000 items in total. The system behaviour and the used student modeling approach are described in [28]. The analyzed data set contains over 180 000 answers.

Student knowledge and mistakes in the used domains have been analyzed before, e.g., recall and mistakes in knowledge of US states [20] or knowledge of Europe by Turkish students [25]. These works focused on difficulty of recall of individual countries and on factors which influence this difficulty (e.g., borders), they did not analyze wrong answers. Moreover, we use a data set that is orders of magnitudes larger than those used in previous research on geography knowledge. The domain of basic arithmetic has been studied intensively before, even with the focus on mistakes. A well-known example is the repair theory [4] with case study for subtraction problems. Particularly multiplication has been studied in detail, e.g., description of effects influencing difficulty (size effect, five effect, tie effect), connectionist model of retrieval [27], classification of errors [5, 24]. Our contribution in this domain is mainly in aligning the results with analysis from different domains (learning declarative knowledge in geography and anatomy).

3.2 Common Wrong Answers

Generally the distribution of wrong answers is highly skewed, most wrong answers are comprised from just few items. Analysis of commonly confused countries shows that the most important factors are whether the countries have com-

mon border, if they have similar size (important factor particularly if they have a common border) and whether their name starts with the same first letter (important factor particularly if they do not have a common border). There are differences between the skewness of the distribution of wrong answers for individual items. For some countries there are few very typical mistakes – for Bulgaria more than 40% of wrong answers are Romania, for Finland the two most common wrong answers (Sweden and Norway) comprise nearly three quarters of wrong answers. Some countries, however, have much more even distribution of wrong answers, e.g., for Switzerland or Croatia the most common mistake comprises only 10% of wrong answers.

The context of questions is also important. In the used system countries can be practiced either in the context of a single continent or of the whole world. In most cases the mistakes on the world map are within the same continent (i.e., the wrong answers on the world map are very similar to wrong answers within the continent map). There is, however, nontrivial number of exceptions, for example: countries with similar names, e.g., Guinea, Guyana, and Papua New Guinea, which have confusingly similar names and are on three different continents; countries close to continent borders, e.g., Turkey is confused with European countries and Arab countries in Africa and Asia confused; islands are confused together, e.g., Madagascar is not confused with other African countries, but with other islands. These examples illustrate the importance of proper practice context for some items, e.g., it is not very useful to practice Madagascar on the map of Africa, Madagascar should be practiced mainly on the map of the whole world. Such results can have direct consequences for the design of the behaviour of educational systems.

The data from the MatMat application contain similar patterns – the distribution of wrong answers is skewed, but the skewness of the distribution differs among items. Some items have very typical wrong answer (e.g., $1 \times 1 = 2$, $4 \times 9 = 32$), for other items wrong answers are more uniformly distributed (e.g., 6×8 with answers 42, 54, 56, 64, 78). Previous work [24] has analyzed classification of errors in basic arithmetic (particularly in multiplication), using categories like near miss (± 1), typo, operation error, or operand related error. In agreement with previous research [13, 24], large part of wrong answers fit into one of these categories, and the dominant categories are as expected – for counting and addition the dominant error type is “near miss”, whereas for multiplication a common error is operand related, e.g., $4 \times 9 = 32$ (which is 4×8). There are, however, interesting differences between items of the same type. For division the typical mistake is “near miss” (± 1). For division by 1 and 10, however, the typical mistakes are rather answers 1 and 10; for items of the type N/N common wrong answers are N or 0. For small operands (e.g., $4/2$) operation errors (multiplication instead of division) sometimes occur, whereas this does not happen for larger operands (e.g., $54/6$).

3.3 Categories of Wrong Answers

To provide a more quantitative analysis and comparison across educational systems, we define a coarse classification of wrong answers and analyze properties of individual categories. We propose the following classification of wrong an-

swers into four categories (note that the defined categories can be seen as “degrees of wrongness” of an answer with a natural ordering). *TopWA* is the most common wrong answer for a given item. *CWA* is a common wrong answer other than the most common one (as a definition of “common” we require that the number of occurrences is more than 5% of all wrong answers for the given item, it must also be larger than 1). *Other* is any nonempty answer that is not common. *Missing* is an empty answer. Previous research [29] used 10% bound for definition of common wrong answers, but they did not treat the top wrong answer separately.

Figure 1 (top) shows distribution of answers among these classes. Although there are some differences between the used systems (respectively specific maps in the geography system), overall the distribution is quite balanced, i.e., the used definitions of classes provide reasonable partition of wrong answers. The rest of Figure 1 shows characteristics of student behaviour related to answers from individual categories. Since in this work we are interested mainly in relative comparison among types of answers (and not among systems), the results are normalized with respect to correct answers (for each system). The reported characteristics are computed globally. We have also analyzed more detailed results (e.g., for specific practice contexts like European countries or one digit multiplication), the results show similar trends.

The results show clear trends that are very similar across the three used systems. The median response time is larger for more wrong answers, with the exception of missing answers. The probability of leaving the system directly after an answer is much higher for wrong answers than for correct answers. Also within the wrong answers there is a clear trend (the probability of leaving increasing with wrongness). Finally, the last two graphs analyze future success of a student; the probability of success on the next question about the same item, the probability of success on the next question within the system (global). In both cases there the probability of future success decreases with wrongness of the current answer.

We see that there are systematic differences between different types of wrong answers. The general nature of these differences is rather intuitive, the main interesting aspects of these results are the similarity of results across three different domains and the consistently linear nature of these relationships, i.e., we can say that the distance between *TopWA* and *CWA* is the same as the distance between *CWA* and *Other*. The bottom line is that the wrongness of answers can be treated as an interval variable and it may be useful to utilize it as such for student modeling (for modeling both knowledge and affect).

3.4 Confusion Graph and Item Clustering

So far we have analyzed wrong answers for each item separately. But mistakes for individual items are clearly interconnected. We can analyze these interconnections with a “confusion graph” (a similar analysis has been done previously for the domain of statistics [12], but for much smaller data). In a confusion graph nodes are individual items, and edges correspond to wrong answers – we consider a weighted graph where a weight of an edge (u, v) is given by a frequency

of a particular wrong answer v among all wrong answers on an item u . This definition leads to a directed graph, to obtain an undirected graph we compute the weight of an undirected edge by averaging the weights of the corresponding directed edges.

Figure 2 shows the confusion graph for European countries. The confusion graph contains distinct clusters of items, this observation holds also for confusion graphs of other practice contexts in the used systems. To automatically detect these clusters we use a community detection algorithm [3]. The resulting clusters are meaningful and can provide useful insight for teachers and developers of educational system (Figure 2 for an illustration). The presented clustering was obtained by off-the-shelf implementation of the community detection algorithm [2] without any tuning. For a specific application of such clustering it may be useful to experiment with different community detection algorithms and specific definitions of the confusion graph.

3.5 Other Properties of Wrong Answers

Wrong answer may help us to (quickly) differentiate between different groups of users. For example in the geography domain we can see some important differences in wrong answers of students of different geographical origin, e.g., confusions between Slovakia and Slovenia, which is much more common mistake for US students than for Czech students, or wrong answers for Belarus (Bulgaria for US students, Ukraine for Czech students).

Wrong answers differ in their “persistence”, i.e., probability that the mistake will be repeated (by the same student) in future. For example, consider wrong answers for Ireland. United Kingdom is more probable mistake than Italy, but the second one is more likely to persists. Other similar examples are Moldova (answers Macedonia versus Kosovo) or Benin (answers Burundi versus Ghana). Some mistakes are very likely to be repeated, e.g., confusion between Zambia and Zimbabwe, Gambia and Senegal, or Guinea-Bissau and Burkina Faso.

4. CONCLUSIONS

Our analysis suggests that wrong answers are underused resource in online educational systems. They are easy to collect and can provide interesting insight applicable in many different ways (student modeling, automatic question and hint construction, feedback and inspiration for teachers and system developers). We provide a systematic overview of potential applications of wrong answers and many illustrative examples of interesting insights from educational applications.

We also propose specific novel approaches to analysis and utilization of wrong answers, particularly a classification of wrong answers into four categories (which can be treated as “degrees of wrongness”) and clustering of items using a confusion graph (based on wrong answers) and a community detection algorithm. The results of analysis from three different domains (geography, anatomy, basic arithmetic) show that properties of wrong answers are rather consistent and thus the developed approaches should be applicable also for other domains.

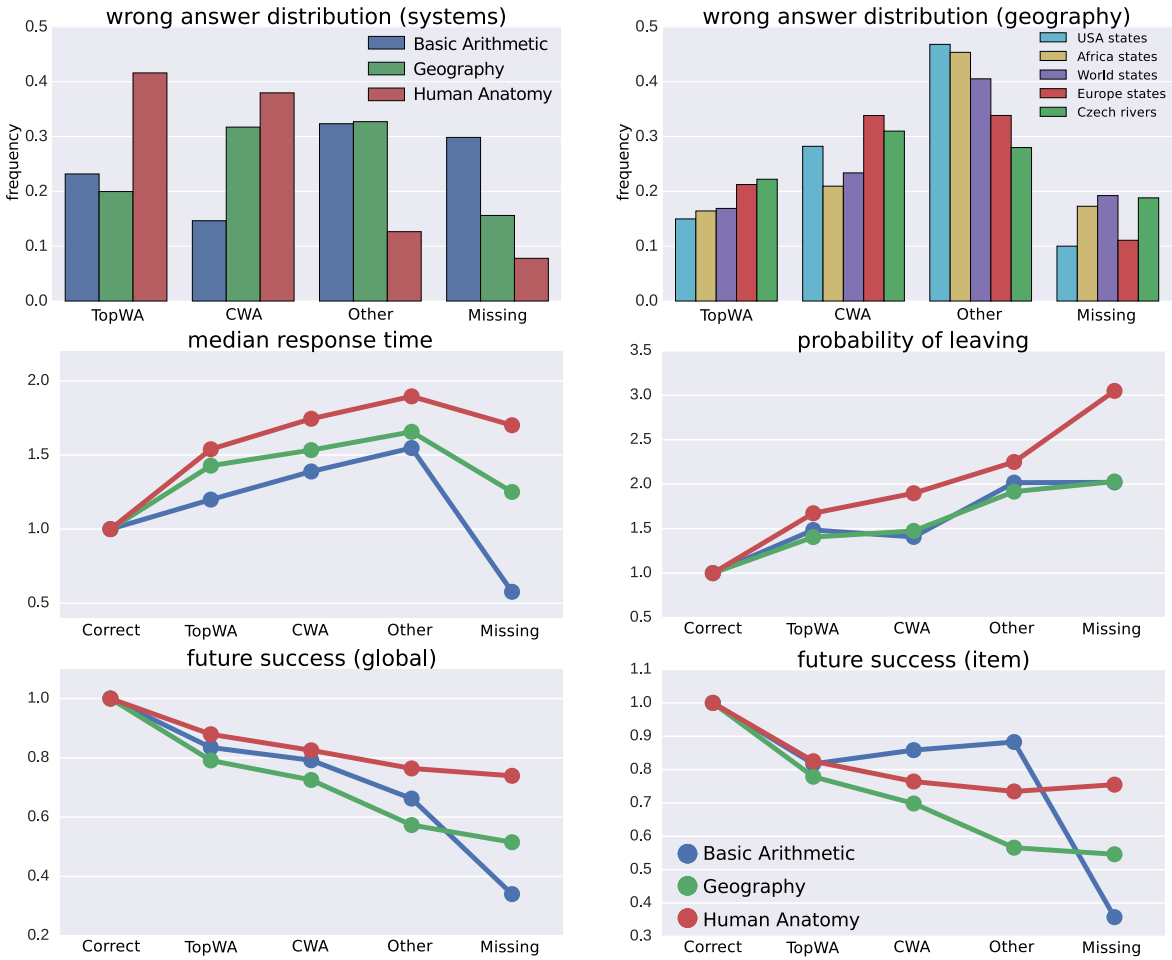


Figure 1: The first line shows frequency of different categories of wrong answers for different systems and for selected maps in geography system. The rest of the figure shows properties of different categories of answers normalized with respect to correct answers.

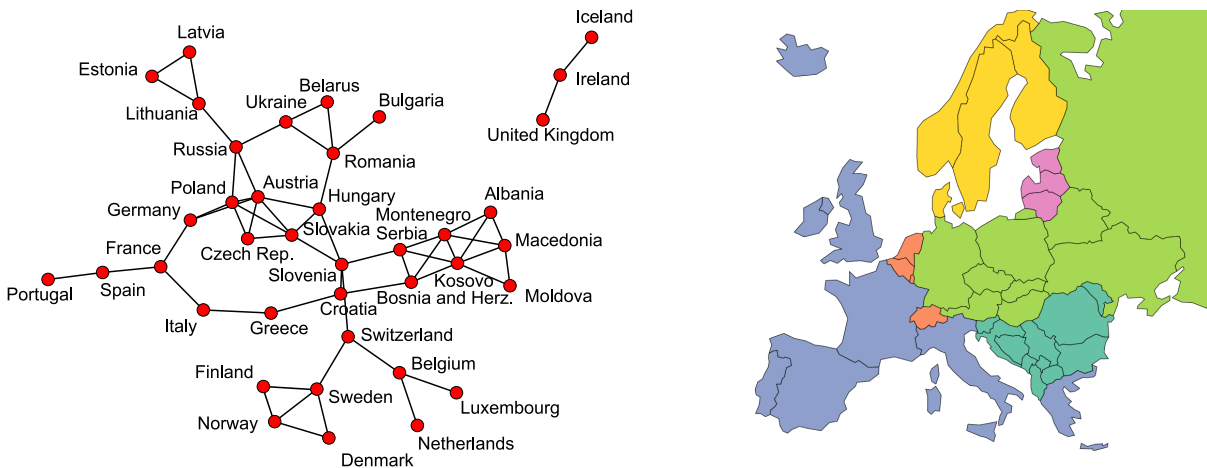


Figure 2: Left: A confusion graph for European countries (showing only the most significant edges). Right: Clustering of European countries based on community detection in the confusion graph.

5. REFERENCES

- [1] Deanne M Adams, Bruce M McLaren, Kelley Durkin, Richard E Mayer, Bethany Rittle-Johnson, Seiji Isotani, and Martin van Velsen. Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior*, 36:401–411, 2014.
- [2] Thomas Aynaud. Community detection for networkx, 2009.
- [3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [4] John Seely Brown and Kurt VanLehn. Repair theory: A generative theory of bugs in procedural skills. *Cognitive science*, 4(4):379–426, 1980.
- [5] Brian Butterworth, Noemi Marchesini, Luisa Girelli, and AJ Baroody. Basic multiplication combinations: Passive storage or dynamic reorganization? *The Development of Arithmetic Concepts and Skills: Constructive Adaptive Expertise*, pages 187–201, 2003.
- [6] Michel C Desmarais and Ryan SJ d Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.
- [7] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. Overcode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction*, 22(2):7, 2015.
- [8] George Gogvadze, Sergey Sosnovsky, Seiji Isotani, and Bruce McLaren. Evaluating a bayesian student model of decimal misconceptions. In *Educational Data Mining 2011*, 2010.
- [9] Jonathan Huang, Chris Piech, Andy Nguyen, and Leonidas Guibas. Syntactic and functional variability of a million code submissions in a machine learning mooc. In *AIED 2013 Workshops Proceedings Volume*, page 25, 2013.
- [10] Barry Peddycord Iii, Andrew Hicks, and Tiffany Barnes. Generating hints for programming problems using intermediate output. In *Educational Data Mining*, 2014.
- [11] Seiji Isotani, Deanne Adams, Richard E Mayer, Kelley Durkin, Bethany Rittle-Johnson, and Bruce M McLaren. Can erroneous examples help middle-school students learn decimals? In *Towards Ubiquitous Learning*, pages 181–195. Springer, 2011.
- [12] Jaclyn K Maass and Philip I Pavlik Jr. How spacing and variable retrieval practice affect the learning of statistics concepts. In *Artificial Intelligence in Education*, volume 9112 of *LNCS*, pages 247–256. Springer, 2015.
- [13] Thomas S McTavish and Johann Ari Larusson. Labeling mathematical errors to reveal cognitive states. In *Open Learning and Teaching in Educational Communities*, pages 446–451. Springer, 2014.
- [14] Andy Nguyen, Christopher Piech, Jonathan Huang, and Leonidas Guibas. Codewebs: scalable homework search for massive open online programming courses. In *Inter. conf. on World Wide Web*, pages 491–502. ACM, 2014.
- [15] 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, 2015.
- [16] Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining (EDM)*, pages 6–13, 2014.
- [17] Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive geography practice data set, 2015. <http://www.fi.muni.cz/adaptivelearning/>.
- [18] Radek Pelánek and Petr Jarušek. Student modeling based on problem solving times. *International Journal of Artificial Intelligence in Education*, 25(4):493–519, 2015.
- [19] Radek Pelánek, Jiří Řihák, and Jan Papoušek. Impact of data collection on interpretation and evaluation of student model. In *Learning Analytics & Knowledge*, pages 40–47. ACM, 2016.
- [20] James A Reffel. Cued vs. free recall in long-term memory of the fifty united states. *Current Psychology*, 16(3-4):308–315, 1997.
- [21] Kelly Rivers and Kenneth R Koedinger. Automatic generation of programming feedback: A data-driven approach. In *Workshop on AI-supported Education for Computer Science*, page 50, 2013.
- [22] Hans-Jürgen Schmidt. Students’ misconceptions—looking for a pattern. *Science education*, 81(2):123–135, 1997.
- [23] John Stamper, Tiffany Barnes, Lorrie Lehmann, and Marvin Croy. The hint factory: Automatic generation of contextualized help for existing computer aided instruction. In *Intelligent Tutoring Systems Young Researchers Track*, pages 71–78, 2008.
- [24] Marthe Straatemeier. *Math Garden: A new educational and scientific instrument*. PhD thesis, Universiteit van Amsterdam, Faculty of Social and Behavioural Sciences, 2014.
- [25] Ilkay Sudas and Cemil Gokten. Cognitive maps of europe: geographical knowledge of turkish geography students. *European Journal of Geography*, 3(1):41–56, 2012.
- [26] Dina Tirosh. Enhancing prospective teachers’ knowledge of children’s conceptions: The case of division of fractions. *Journal for Research in Mathematics Education*, pages 5–25, 2000.
- [27] Tom Verguts and Wim Fias. Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & cognition*, 33(1):1–16, 2005.
- [28] Jiří Řihák. Use of time information in models behind adaptive system for building fluency in mathematics. In *Educational Data Mining, Doctoral Consortium*, 2015.
- [29] Yutao Wang, Neil T Heffernan, and Cristina Heffernan. Towards better affect detectors: effect of missing skills, class features and common wrong answers. In *Learning Analytics And Knowledge*, pages 31–35. ACM, 2015.