Impact of Data Collection on Interpretation and Evaluation of Student Models

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

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

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

ABSTRACT

Student modeling techniques are evaluated mostly using historical data. Researchers typically do not pay attention to details of the origin of the used data sets. However, the way data are collected can have important impact on evaluation and interpretation of student models. We discuss in detail two ways how data collection in educational systems can influence results: mastery attrition bias and adaptive choice of items. We systematically discuss previous work related to these biases and illustrate the main points using both simulated and real data. We summarize specific consequences for practice – not just for doing evaluation of student models, but also for data collection and publication of data sets.

Keywords

attition, bias, data sets, evaluation, parameter fitting, student modeling

1. INTRODUCTION

The way we collect data can have significant influence on results that we obtain by analysis of the collected data. A typical example is selection bias – if data are not representative of the studied phenomenon, results are not generalizable. In learning analytics research a typical example is self-selection in massive open online courses or voluntary questionnaires; techniques for reduction of such bias have been already studied [4]. The impact of data collection becomes particularly challenging issue when the data collection is done by an adaptive system. Student modeling techniques are developed with the aim of being applied in adaptive systems and are typically evaluated on data from such systems.

In student modeling research, however, the potential impact of data collection on results is typically not taken into account. That is unfortunate because uncritical use of historical data sets is prone to biases and misleading results. For example, intelligent tutoring systems often used mastery learning approach, which leads to attrition bias in logged data. System behaviour (e.g., choice of items or mastery detection) is typically done by a student model. The same model may be used for data collection and during model evaluation, which may bias the evaluation – it can happen that the used model does not collect data that would show its deficiencies. The presence of this feedback loop is an important difference compared to other forecasting domains. For example in weather forecasting models do not directly influence the system and cannot distort collected data. In student modeling they can.

Potential biases caused by data collection have been discussed in research in related domains. Research on evaluation of recommender systems [13, 14, 37] discussed potential biases in data collection, e.g., by filtering users, and train/test set issues when using offline data. Proposal for layered evaluation of adaptive systems [32] includes evaluation of data collection, but does not address specifically its impact on subsequent steps. The presence of a feedback loop between data collection and model evaluation has been previously discussed in the context of "exploration vs exploitation problem" (multiarmed bandits), with applications for news [20, 43] and advertisement [3, 19] selection.

In the context of student modeling, issues related to data collection have not been systematically studied before, although specific aspects have already been addressed by previous work, particularly in the context of learning curves [24] and mastery learning. When a tutoring system uses mastery learning, students with high skill drop out earlier from the system (and thus from the collected data), thus a straightforward interpretation of aggregated learning curves may be misleading [11, 16, 25, 27]. Previous work showed specific illustrations of confounded learning curves [27], discussed methods for disagreggation of learning curves [25], and proposed mastery-aligned models [16] to take this bias into account. Confounding effect of item ordering on learning and item difficulty has been mentioned in several works [12, 15, 17, 36], but only as a side note.

In this work we provide a systematic overview of potential biases caused by data collection. We provide discussion of previous works that mention specific biases over real data from tutoring systems and present some new illustrations on our data. We also present specific artificial scenarios, which are highly simplified (compared to real systems), but clearly demonstrate the core principles of discussed biases.

Our summary shows that the choice of data used for experiments can make important difference on fitted parameter values and results of evaluation. This has important consequences for research practice, since currently this issue is neglected and neither research papers nor descriptions of data set discuss in details the way in which the used data were collected. To contribute to the improvement of state of the art we conclude our overview with specific consequences for research practice.

2. BACKGROUND

Our aim is to illustrate the impact of data collection in many different contexts, and thus discussion of potential biases refers to many different student models and experimental settings. In this section we provide brief overview of used notions and pointers to more detailed explanations of used modeling techniques.

2.1 Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) [7, 40] is a model of learning which assumes a sudden change in knowledge. It is a hidden Markov model where skill is a binary latent variable (either learned or unlearned). The basic version of the model has 4 parameters: P_{init} is the probability that the skill is initially learned, P_{learn} is the probability of learning a skill in one step, P_{slip} is the probability of incorrect answer when the skill is learned, and P_{guess} is the probability of correct answer when the skill is unlearned. The estimated skill is updated using a Bayes rule based on the observed answers; the prediction of student response is then done based on the estimated skill. Estimation of model parameters (the tuple $P_{init}, P_{learn}, P_{slip}, P_{guess}$) can be done using several different techniques (the expectation-maximization algorithm, stochastic gradient descent, exhaustive search). For experiments in this work we use Yudelson's implementation [44]. The model has many extensions, but for our purposes (illustration of biases) the basic version is sufficient.

2.2 The Rasch Model and the Elo Rating System

The Rasch model is used typically in item response theory [9]. It assumes a constant student skill (no learning) and items with varying difficulty. Probability of correct answer for a student with skill θ and item with difficulty d is given by $\sigma(\theta - d)$, where σ is a logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$.

The Elo rating system [10] has been originally proposed for rating chess players, but recently it has been used also for student modeling [34]. It is closely related to the Rasch model since it also uses the same equation for predicting the probability of correct answers. The main difference is in the approach to parameter estimation. The estimation of parameters of the Rasch model is typically done using some iterative maximal likelihood procedure [9], whereas the Elo rating system uses simple update equations suitable for online updates. Previous research [34] showed that the obtained estimates are very similar.

2.3 Models of Learning based on Logistic Function

Models based on logistic function can be also extended to incorporate learning. For generating simulated data we consider a simple linear growth of the skill. More specifically, for the initial skill θ_0 we assume normally distributed skill $\theta_0 \sim \mathcal{N}(\mu, \sigma^2)$ and we model the change in skill by linear learning: $\theta_k = \theta_0 + k \cdot \Delta$, where Δ is either a global parameter or an individualized learning parameter (in that case we assume a normal distribution of its values). This model is a simplified version of the Additive Factors Model [5, 6, 16]; the original additive factor model uses multiple skills. A different variant of this model [38] uses "random walk learning": $\theta_{k+1} = \theta_k + \epsilon$, $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$.

For estimating student skills from data a commonly used technique based on logistic function is Performance Factor Analysis (PFA) [33]. The skill estimate is given by a linear combination of the initial skill[1] and past successes and failures of a student:

$$P(correct|k) = \sigma(\beta + \gamma \cdot s_k + \delta \cdot f_k)$$

where β is the initial skill, s_k and f_k are counts of previous successes and failures of a student during the first k attempts, γ and δ are parameters that determine the change of the skill associated with a correct and incorrect answer. Parameters β, γ, δ can be easily estimated using standard logistic regression. Note that originally the technique was formulated in terms of vectors, as it uses multiple knowledge components [33]. In our setting only the one-dimensional version is relevant.

3. TRACE LENGTH AND MASTERY ATTRITION BIAS

One commonly used approach to adaptation in educational systems is to let students solve varying number of items (problems, questions) of similar type and difficulty. Typical approach is mastery learning – students solve items until they master the topic. The "termination decision" (mastery detection) can be done by some simple rule (e.g., "3 correct in a row") or by a student model (e.g., "probability of knowing the skill is at least 95%").

A consequence for the collected data is that students have different trace length (solve different number of items) and this difference is not random. This creates a "mastery attrition bias"[2], which can influence several aspects of model evaluation and interpretation.

3.1 Learning Curves

One popular approach to evaluation of educational systems are learning curves [24]. Learning curve plots the error rate (or another student performance measure like response time) as a function of number of attempts. A decreasing learning curve is evidence of learning, curves can be used to compare models or find ill-specified knowledge components.

Learning curves are based on aggregating behaviour across students and this aggregation may complicate their interpretation [24]. Particularly, learning curves based on data from adaptive systems are prone to the mastery attrition bias. This issue has been discussed in recent research [16, 25, 27], but is not taken into account sufficiently in the current practice. For example a recent work on mixture modeling using learning curves [39] does assume constant length of trace and would require significant modification to work correctly in the presence of mastery attrition.

Figure 1 shows a specific example of the mastery attrition bias using simulated data. We use simulated students with probability of correct answer given by $\sigma(\theta + k \cdot 0.15)$, where $\theta \sim \mathcal{N}(-2, 2)$ is the initial skill of a student, k is the number

[1] This is usually denoted as item's difficulty, but in our setting the item difficulty and initial skill are interchangeable.

[2] Attrition bias is a type of selection bias, which is often present for example in medical experiments.

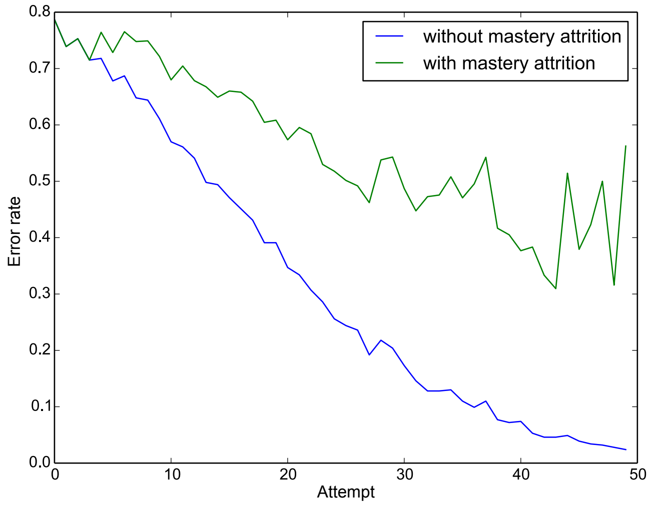

Figure 1: Learning curves – impact of mastery attrition.

of attempts, and σ is the logistic function. When all students solve all problems, we get a nice learning curve. When we use mastery attrition (in this case realized by a simple "3 correct in a row" rule), we get much flatter and noisier curve, which underestimates student learning.

In educational systems, particularly those that are used by students voluntary, we have have self-selection bias, which can work in opposite direction to mastery attrition. In many systems the length of trace is decided by students (rather than mastery learning or other system rule). In such cases the length of trace depends on student motivation, which can be influenced for example by success. As a simple model scenario consider the following case of heterogeneous student population consisting of two subpopulations. Students in the first subpopulation have constant success rate 50 %, students in the second subpopulation have constant success rate 80 %, i.e., students do not learn, they just differ in their prior knowledge. Now if the length of trace is correlated with success rate (as for example in [22, 29]) and we plot the aggregated learning curve, the curve will show improvement in student performance – but this is just an illusory effect caused by student attrition. In some systems it may even happen that both the mastery attrition bias and self-selection occur [31], which can make the interpretation of data highly challenging.

3.2 Fitted Parameters

The length of trace and mastery attrition bias also influence values of parameters of student models. For illustration we use two commonly used student modeling techniques described in Section 2: Bayesian Knowledge Tracing (BKT) and Performance Factor Analysis (PFA).

The trace length can have large impact on fitted parameters particularly when there is a mismatch between model assumptions and characteristics of data. If we generate simulated data by the BKT model and then fit the data by the PFA model, there is an interesting impact of the trace length. Fitted PFA parameters (and thus also predictions of the model) may differ significantly; Figure 2 shows calibration graphs for different trace lengths (data were generated by the BKT model with $P_{init} = 0.15, P_{learn} = 0.35, P_{slip} = 0.18, P_{guess} = 0.25$). PFA is able to fit either the beginning of the trace or the end of the trace, and parameter values depends on the trace length used for parameter fitting. We obtain similar result in the opposite direction. We generate data using logistic function $\sigma(\theta + k \cdot 0.4)$, where $\theta \sim \mathcal{N}(-1, 1)$. When we fit the data using BKT, we get significantly different values for some parameters for trace of length 5 ($P_{init} = 0.25, P_{learn} = 0.25, P_{slip} = 0.29, P_{guess} = 0.19$,) and length 25 ($P_{init} = 0.08, P_{learn} = 0.24, P_{slip} = 0.03, P_{guess} = 0.21$).

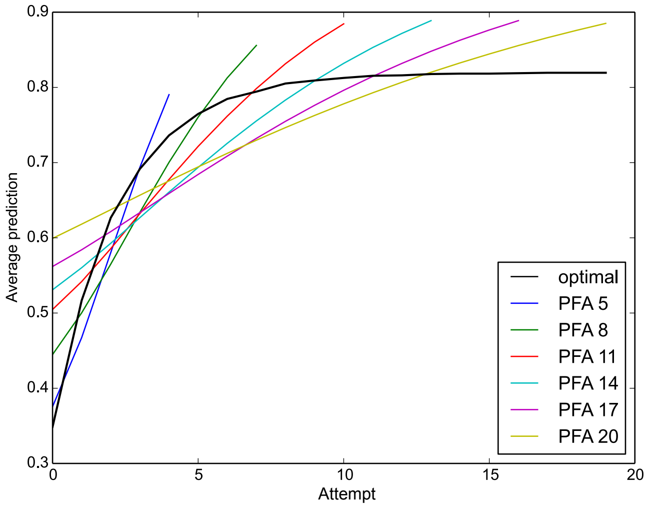

Figure 2: Calibration of different variants of the PFA model for data generated by BKT. 'PFA X' denotes a model with parameters fitted on traces of length X.

When we consider mastery attrition, data collection impacts fitted parameters even in the case when the fitted model exactly corresponds to the way data were created. Consider data generated by BKT with parameters $P_{init} = 0.25, P_{learn} = 0.08, P_{slip} = 0.12, P_{guess} = 0.3$. When data are collected without attrition bias, the fitted parameters correspond well to the ground truth (e.g., for trace of fixed length 20 we get $P_{init} = 0.27, P_{learn} = 0.08, P_{slip} = 0.1, P_{guess} = 0.27$). But when data are collected using the "3 correct in a row" condition, the fitted parameters are significantly different: $P_{init} = 0.72, P_{learn} = 0.23, P_{slip} = 0.52, P_{guess} = 0.15$.

These illustrations show that when researchers attempt to interpret or further use model parameters, e.g., when doing "discovery with models" [1], they should carefully investigate whether fitted parameters are dependent on details of data collection (trace length, attrition bias). At least researchers should report properties of the used data set (which is not the current practice).

3.3 Evaluation of Models

A standard approach for evaluation and comparison of models is to use historical data, split them into training and testing set, train model parameters on the training set, and evaluate the performance of models on testing set using metrics [35] like RMSE, AUC, or log-likelihood.

Results of such comparison are typically interpreted as ability of models to fit "student behaviour". However, these results can be influenced not just by student behaviour, but also by the way data were collected, specifically by the length of trace and the stopping condition (mastery learning). Models may differ in they ability to model "initial phase of learning" and "plateau of performance". For example if we generate data using the logistic function ($\sigma(\theta + k \cdot 0.1)$, where $\theta \sim \mathcal{N}(-0.4, 2)$) and compare the fit of PFA and BKT models (using RMSE), we get that PFA has better performance if we use constant trace length, whereas BKT has better performance if data are collected using mastery learning ("k correct in a row").

Another aspect of evaluation, which should be treated with caution, is the division of data between train and test set. In the context of student modeling there are multiple possibilities how to approach this division (e.g., student stratified, item stratified). One approach researchers have used (e.g., [17]) is to put the last 20% of attempts of some students in test set. This approach to evaluation is disputable because it evaluates ability of models to fit only part of student behaviour. It can be problematic particularly if the used data set was collected using mastery learning – in that case the last attempts in each sequence would be biased towards correct answer (whatever mastery learning criterion is used, it is based on seeing correct answers). This can bias results of model comparison. As a model situation, consider students who answer completely randomly and two student models: model A predicting probability of success 0.5, model B predicting probability of success 0.7. Since students answer randomly, the unbiased model A is better. However, if we perform evaluation on the last attempts of data collected by simple mastery learning condition "3 correct in a row", model B will achieve better performance as correct answer will dominate in this test set, even though their occurrence is just due to the data collection condition, not due to some inherent aspect of student behaviour.

4. ITEM ORDERING AND SELECTION

Another approach to personalization is to adaptively choose items to suit needs of a particular student. This selection is often done with respect to difficulty, i.e., stronger students get more difficult items quickly, weaker students keep practicing easier items.

Similarly to mastery learning, adaptive choice of items complicates evaluation and can bias results if untreated. As a specific example consider models build using tabulating "success rate for the next problem" as used in several ASSISTment papers (e.g., [41, 42]). These models would not work in the case of an adaptive choice of items, where the educational system actively tries to achieve a given target success rate. The application of such models is thus limited to (implicitly assumed) properties of a particular data set.

4.1 Item Ordering

In educational systems, it is quite natural that the order of items is related to their difficulty – students solve easier items first and then proceed to more difficult ones. If all students solve items in similar order, it may be impossible to disentangle increase in problem difficulty and student learning. This confounding effects has been in different forms noted in several recent works [12, 15, 17, 36].

To make this effect clear, let us consider the following model scenario. Assume that students' skill linearly increases ($\theta_k = \theta_0 + k \cdot \Delta$, where k is the order of an attempt), items are ordered in a fixed order with linearly increasing difficulty ($d_k = d_0 + k \cdot \Delta'$), and probability of correct answer is given by a logistic function with respect to the difference of skill and difficulty: $P(\mathit{correct}|k) = 1/(1 + e^{-(\theta_k - d_k)})$. In this situation parameters are non-identifiable, we get identical probabilities of correct answers for completely different situations, e.g., no learning and fixed difficulty of items ($\Delta = \Delta' = 0$) and significant learning and large increase in difficulty of items ($\Delta = \Delta' = 0.5$).

The ordering of items may influence collected data also in other ways. There may be a "local transfer" between consecutive items, e.g., when math problems with very similar structure are asked in sequence or due to short term memory in factual knowledge learning. Such effects can have large impact on correctness of answers, but provide little evidence about long term learning. Items in the beginning of a sequence may have lower success rate (or higher response times) just due to user interface issues (students have to get used to peculiarities of a particular system). Last items may have lower success rate due to fatigue.

It may be hard to overcome item ordering effects. We may try to incorporate some variability into item ordering for individual students, but typically the basic "easier to difficult" progression is desirable. We should nevertheless take this issue into account and at least analyze the collected data to understand its properties and potential limitations. Useful descriptive statistics for this purpose include analysis of mean presentation order of each item (used in [17]), analysis of correlation between orderings of different students (used in [36]), or analysis of transitions between items (used in [23]).

4.2 Adaptive Choice of Items

The goal of many adaptive educational systems is to select items of suitable difficulty (neither too difficult, nor too easy) to keep students in the flow state [8]. This adaptive choice of items can have important impact on evaluation of models. We illustrate this impact on data from our widely used application for learning geography [30].

We utilize data from an experiment which compares four question construction algorithms with different degree of adaptivity [31]. Questions used by the system are of the form "What is the name of the highlighted place?" and "Where is X?". The question construction process has two phases: at first, selection of the question stem (e.g., Rwanda), at second, decision how many options to use for the multiple-choice question and what distractors to use (e.g., Burundi and Tanzania). For each step two choices were considered: random and adaptive. The decisions of the adaptive algorithm are done using predictions of a student model based on the Elo rating system [30, 28], aiming at a specific target success rate [29] and using the most confusing distractors. Users of the system were allocated randomly into one of four versions of the question construction algorithm: adaptive-adaptive, adaptive-random, random-adaptive, and random-random. During this experiment we collected between 200 000 and 250 000 answers for each of the four groups. Every data set was split into a train set (20%) and a test set (80%) in a student-stratified manner. Since all models work online and update their parameters during evaluations, the choice of the size of the train set does not have big influence on reported results.

Figure 3 shows comparison of several models over these data sets. As our point is not to study models, but to show impact of data collection, we have chosen simple mod-

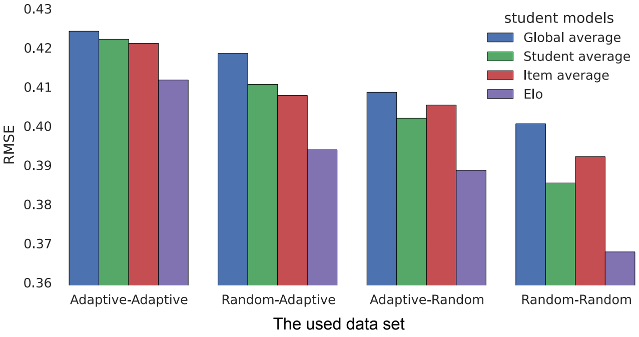

Figure 3: RMSE comparison over data from slepemapy.cz collected using different item selection algorithm.

els (predictors of student performance): constant predictor (given by global average), student average, item average, and the Elo-based model used in the actual application [30]. The figure show several interesting results. With higher adaptivity in question construction, the RMSE values are higher (prediction is more difficult) and closer together (naive predictors have similar performance as better models since questions are constructed to be off a specified difficulty). We also see a swap in ordering of models – the relative performance of two models (item average, student average) depends on what data are used to compare them. These results clearly illustrate the impact of the method that is used to collect data on model comparisons done using the data set.

4.3 Feedback Loop between Student Model and Data Collection

The evaluation described in the previous section is a specific illustration of a feedback loop between student model and data collection – a model influences which data are collected, the collected data are used to evaluate the model. To explore this feedback loop in more detail we performed experiments with a simulation of a simplified version of an adaptive question answering systems described in the previous section. Details of the simulation are described in [26], the basic idea of the experiment is that we choose one student model and use it as an input for the adaptive choice of items. At the same time we let other models do predictions as well and log answers together with all predictions. Since we are using simulated data, we know the ground truth and we can compare models with optimal predictions.

Figure 4 shows the resulting RMSE for each model in individual runs (data collected using specific model). The figure shows the same basic results that we have seen for real data. When the data are collected using the optimal model, the RMSE values are largest (at least for more sophisticated models) and closest together; even the ordering of models is different from other cases. In this case even the constant model provides comparable performance to other models – but it would be very wrong to conclude that "predictive accuracy of models is so similar that the choice of model does not matter", since in the simulated system different models lead to different choice of items and consequently to different student experience. The reason for small differences in RMSE is not similarity between models, but characteristics of data ("good choice of suitable items"), which make predictions difficult and even a naive predictor comparatively good.

In real systems, the content is often organized in knowledge components (KCs, also called concepts or skills), and this domain model is also used by student models. The relation between items and knowledge components can be generally described by a Q-matrix [2] or – in case of the strict division of items to components – by item-KC mapping. The quality of a Q-matrix (or a item-KC mapping) can have strong impact on adaptive behaviour of the system [18], but this quality is difficult to measure. The difference in performance of models with different Q-matrices can be – as demonstrated by following hypothetical scenario – relatively small. Consider two Q-matrices, one of them is 'correct' and has 10 knowledge components, the other one is 'incorrect' and merges two of the skills together. The difference between performance metrics of these two models will be necessarily small, since in most cases their predictions will be identical. The difference is, however, of practical significance, because if a system uses the incorrect model, students may miss practice of one of the concepts. A realistic scenario of this type is reported in [21].

Our simulated experiment suggests that this naturally small but important difference in performance metrics is also influenced by the used data set (resp. method used to collect the data). To highlight the point we use a simple setup: two knowledge components, every item is assigned to exactly one of these KCs and every student has two independent skills corresponding to these KCs. The basic "Elo" model does not consider item division and assumes only one KC, the "Elo concepts" model considers correct item-KC mapping, and the "Elo wrong concepts" model contains random mistakes in its item-KC mapping. Figure 4 shows that models with concepts achieve nearly the same performance (i.e., models seem to be of the same quality), when data are collected by the most sophisticated models (models with concepts and optimal one). But over other data sets, the difference between the two models becomes more apparent and the results clearly show that "Elo concepts" model is better. Note that the data are able to better distinguish between models if they are actually collected by worse models.

Similarly, it can be difficult to detect wrongly specified prerequisites if we are using them to collect our data. To illustrate this point we consider the following simple scenario. We wrongly assume that a concept A is prerequisite for a concept B. It would be possible to detect this error by observing students who are able to solve problems from the concept B but have difficulty solving problems from the concept A. However, if our adaptive system uses the wrongly specified prerequisite, it offers problems from the concept B only to students who have mastered the concept A and thus is not able to collect data which would provide evidence that the specified prerequisite is wrong.

4.4 Parameter Estimation

In Section 3.2 we have shown how the length of trace and attrition bias can influence estimated parameter values. Adaptive choice of items can also have impact on parameter estimates. In previous work [34] we have shown this effect using simulated data, here we illustrate the effect using real data.

We use the same data as in Section 4.2 – experiment with

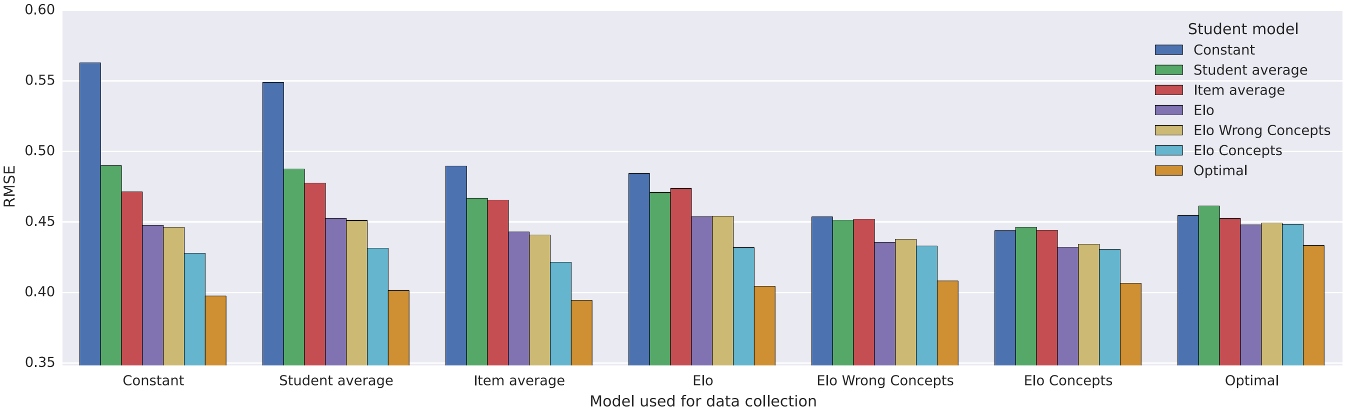

Figure 4: Comparison of student model performance (measured by RMSE) over data collected using different models.

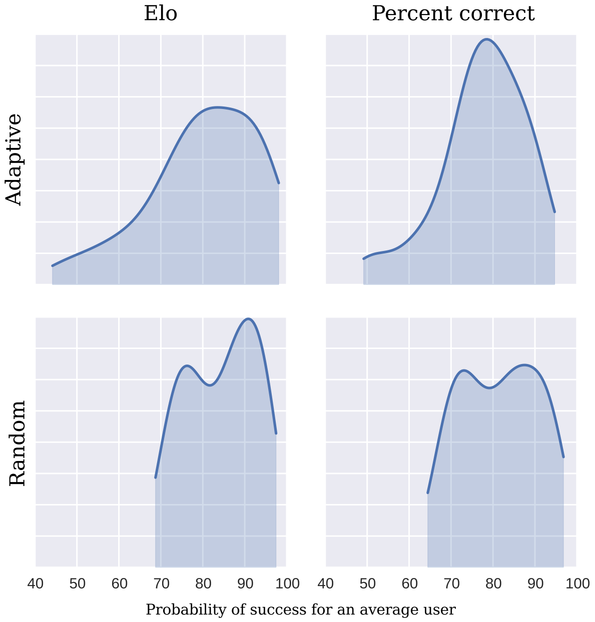

Figure 5: Distributions of estimates of difficulty for European countries across different data sets and models.

adaptive and random question selection in geography practice system. We compare two techniques for estimating difficulty of individual items – a naive "percent correct" technique and a variant of the Elo rating system that is used in the actual implementation [30]. Figure 5 shows distributions of the estimated difficulties (expressed as probability of correct answer for an average student) for European countries (which have most answers in the used data set). The figure shows differences between the used techniques, particularly the ability of the Elo rating system is to better differentiate the difficulty of individual items in the case of adaptive data collection. But the impact of the data set is much larger than the impact of the used model.

This experiment clearly illustrates that the way data are collected may have much larger influence on the fitted parameters than a choice of the model. This is important particularly for "discovery with models" [1], in which model parameters are further utilized and interpreted. Such analysis needs to take the data collection mechanism into account.

5. CONSEQUENCES FOR PRACTICE

We have described different ways how data collection mechanism influences interpretation and evaluation of student models. These issues have direct consequences not just for the realization of evaluation of models, but also for publication of data sets and the way we collect data in our systems in the first place.

5.1 Publication of Data Sets

Currently, most published data set document only the data itself, but not the way in which the data were collected. As the data collection can have important impact on the interpretation of the data, it is necessary to document data collection mechanism as well.

The behaviour of adaptive systems is quite complicated (e.g., many parameters often influence the exact choice of items) and it may not be feasible to document the data collection mechanism up to all details. But authors of data sets should explicitly discuss all major issues and potential limitations due to the data collection mechanism, particularly attrition bias and algorithms used for item ordering and selection.

5.2 Evaluation and Interpretation of Models

Researchers who use published data that they did not collect themselves should inquire into details of the used data collection mechanism. It is useful to perform exploration of data set properties to get understanding of the data and its potential biases.

When doing evaluation and interpretation of student models (and their parameters), special attention should be paid to the influence of properties of the used data set. We should make sure that our results are not superficially created by the data collection mechanism. Careful attention should be paid particularly to division of data between train and test set as data collection mechanism can easily cause bias, which may lead to poor generalization of results.

In order to avoid biases caused by data collection, it is useful to probe stability of achieved results (comparison of model performance, fitted parameter values). What happens when we use artificially shorter trace lengths? Do results stay the same (similar)? In some cases it may be impossible to perform such probes – for example if we have access only to offline data and we care about ordering of items for different student, we cannot perform experiments with different orderings. In such cases it is important to explicitly discuss limitations and future work should try to replicate results with newly collected data overcoming stated limitations.

5.3 Data Collection

Our results also have consequences for the data collection itself. We have repeatedly illustrated how the use of adaptive techniques leads to data set, which make it difficult or even impossible to compare student models and find mistakes in their specification (e.g., in knowledge components or prerequisites). The adaptive behaviour is the purpose of student modeling and is beneficial for students, it is, however, detrimental for evaluation purposes.

It should be possible to find a reasonable compromise between our different goals. We can modify behaviour of our educational systems in a way that would enable easier evaluation without hampering their main goal (i.e., student learning). Specifically, we may employ controlled use of randomization. If some items are chosen randomly (from a reasonably defined set of items), the impact on user experience may be negligible and the collected data can be used for evaluation in much more straightforward manner than adaptively chosen items.

6. ACKNOWLEDGMENTS

This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

7. REFERENCES

[1] R. S. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009.

[2] T. Barnes. The q-matrix method: Mining student response data for knowledge. In *Educational Data Mining*, 2005.

[3] L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[4] C. Brooks, O. Chavez, J. Tritz, and S. Teasley. Reducing selection bias in quasi-experimental educational studies. In *Learning Analytics And Knowledge*, pages 295–299. ACM, 2015.

[5] H. Cen, K. Koedinger, and B. Junker. Comparing two irt models for conjunctive skills. In *Intelligent Tutoring Systems*, pages 796–798. Springer, 2008.

[6] H. Cen, K. R. Koedinger, and B. Junker. Is over practice necessary?-improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in Artificial Intelligence and Applications*, 158:511, 2007.

[7] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[8] M. Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.

[9] R. De Ayala. *The theory and practice of item response theory*. The Guilford Press, 2008.

[10] A. E. Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[11] S. E. Fancsali, T. Nixon, A. Vuong, and S. Ritter. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*, 2013.

[12] J. González-Brenes, Y. Huang, and P. Brusilovsky. General features in knowledge tracing: applications to multiple subskills, temporal item response theory, and expert knowledge. *Educational Data Mining*, 2014.

[13] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *The Journal of Machine Learning Research*, 10:2935–2962, 2009.

[14] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[15] P. Jarušek, M. Klusáček, and R. Pelánek. Modeling students' learning and variability of performance in problem solving. In *Educational Data Mining*, pages 256–259, 2013.

[16] T. Käser, K. R. Koedinger, and M. Gross. Different parameters-same prediction: An analysis of learning curves. In *Educational Data Mining*, 2014.

[17] M. M. Khajah, Y. Huang, J. P. González-Brenes, M. C. Mozer, and P. Brusilovsky. Integrating knowledge tracing and item response theory: A tale of two frameworks. *Personalization Approaches in Learning Environments*, page 7, 2014.

[18] K. R. Koedinger, J. C. Stamper, E. A. McLaughlin, and T. Nixon. Using data-driven discovery of better student models to improve student learning. In *Artificial Intelligence in Education*, pages 421–430. Springer, 2013.

[19] J. Langford, A. Strehl, and J. Wortman. Exploration scavenging. In *Proceedings of the 25th international conference on Machine learning*, pages 528–535. ACM, 2008.

[20] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Web search and data mining*, pages 297–306. ACM, 2011.

[21] R. Liu, K. R. Koedinger, and E. A. McLaughlin. Interpreting model discovery and testing generalization to a new dataset. In *Educational Data Mining*, pages 107–113, 2014.

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

[23] M. Lopes, B. Clement, D. Roy, and P.-Y. Oudeyer. Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, 7(2):20–48, 2015.

[24] B. Martin, A. Mitrovic, K. R. Koedinger, and S. Mathan. Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3):249–283, 2011.

[25] R. C. Murray, S. Ritter, T. Nixon, R. Schwiebert, R. G. Hausmann, B. Towle, S. E. Fancsali, and A. Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

[26] J. Nižnan, R. Pelánek, and J. Papoušek. Exploring the role of small differences in predictive accuracy using simulated data. In *AIED Workshop on Simulated Learners*, 2015.

[27] T. Nixon, S. Fancsali, and S. Ritter. The complex dynamics of aggregate learning curves. In *Educational Data Mining*, 2013.

[28] J. Niznan, R. Pelánek, and J. Rihák. Student models for prior knowledge estimation. In *Educational Data Mining*, pages 109–116, 2015.

[29] J. Papoušek and R. Pelánek. Impact of adaptive educational system behaviour on student motivation. In *Artificial Intelligence in Education*, volume 9112, pages 348–357, 2015.

[30] J. Papoušek, R. Pelánek, and V. Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining*, pages 6–13, 2014.

[31] J. Papoušek, V. Stanislav, and R. Pelánek. Evaluation of an adaptive practice system for learning geography facts, 2015. Submitted.

[32] A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, 2010.

[33] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.

[34] R. Pelánek. Application of time decay functions and Elo system in student modeling. In *Proc. of Educational Data Mining*, pages 21–27, 2014.

[35] R. Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2), 2015.

[36] R. Pelánek and P. Jarušek. Student modeling based on problem solving times. *International Journal of Artificial Intelligence in Education*, pages 1–27, 2015.

[37] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[38] A. C. Smith, L. M. Frank, S. Wirth, M. Yanike, D. Hu, Y. Kubota, A. M. Graybiel, W. A. Suzuki, and E. N. Brown. Dynamic analysis of learning in behavioral experiments. *The journal of neuroscience*, 24(2):447–461, 2004.

[39] M. Streeter. Mixture modeling of individual learning curves. In *Educational Data Mining*, 2015.

[40] B. van de Sande. Properties of the bayesian knowledge tracing model. *Journal of Educational Data Mining*, 5(2):1, 2013.

[41] E. Van Inwegen, S. Adjei, Y. Wang, and N. Heffernan. An analysis of the impact of action order on future performance: the fine-grain action model. In *Learning Analytics And Knowledge*, pages 320–324. ACM, 2015.

[42] E. G. Van Inwegen, S. A. Adjei, Y. Wang, and N. T. Heffernan. Using partial credit and response history to model user knowledge. In *Educational Data Mining*, 2015.

[43] S. Wager, N. Chamandy, O. Muralidharan, and A. Najmi. Feedback detection for live predictors. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3428–3436. Curran Associates, Inc., 2014.

[44] M. Yudelson. Tool for fitting bayesian knowledge tracing models, 2014. https://github.com/IEDMS/standard-bkt.