# An Analysis of the Impact of Action Order on Future Performance: the Fine-Grain Action Model

Eric Van Inwegen Seth Adjei Yan Wang Neil Heffernan
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA, 01609-2280
+1-508-831-5569
{egvaninwegen, saadjei, ywang14, nth} @wpi.edu

## ABSTRACT
To better model students' learning, user modelling should be able to use the detailed sequence of student actions to model student knowledge, not just their right/wrong scores. Our goal is to analyze the question: "Does it matter *when* a hint is used?". We look at students who use identical attempt counts to get the right answer and look for the impact of help use and action order on future performance. We conclude that students who use hints too early do worse than students who use hints later. However, students who use hints, at times, may perform as well as students who do not use hints. This paper makes a novel contribution showing for the first time that paying attention to the precise sequence of hints and attempts allows better prediction of students' performance, as well as to definitively show that, when we control for the number of attempts and hints, students that attempt problems before asking for hints show higher performance on the next question. This analysis shows that the pattern of hints and attempts, not just their numbers, is important.

## General Terms
Algorithms, Measurement, Performance, Reliability

## Keywords
Action order, Hint use, Tabling, Binning, Prediction of future success, Data Mining

## 1. INTRODUCTION
Intelligent Tutoring Systems usually offer help in the form of messages, scaffolding, etc. to students who cannot (or believe they cannot) solve a problem on their own. Previous work has analyzed the apparent effect of using hints on learning. Beck, et.al. [2] have studied the question of the assumption that help is, in fact, helpful. Hawkins et. al. [6] and Wang and Heffernan [13] have studied the likelihood of future success of students using particular combinations of hits and attempts (in the Assistance Model - AM). Zhu et. al. [14] and Duong et. al. [4] have looked at clickstream data to be able to make predictions (in the Sequence of Actions model – SOA); they were able to show that students who use hints first do not perform as well in the future when compared to students who use hints later in their action order. What was insufficient about these studies was that the bins were not well controlled. In AM, action order is ignored; in SOA, temporal spacing and whether the hint gives the answer were not taken into account. User modelling algorithms, such as Knowledge Tracing (KT) [3] and Performance Factor Analysis (PFA) [9], don't use the order of actions in their computations; at least intuitively, this seems to be an omission.

This paper seeks to better understand the effects of action order and help requests. Our basic answer to "Does it matter when a hint is used?" is: "Yes; students who make at least one attempt before using a hint outperform their peers who use a hint right away, even controlling for the number of attempts made."

### 1.1 Background
The role of predicting future success, based on answers given in Intelligent Tutoring Systems (ITS's) is useful both to the system (e.g. attempting to modify work stream based on student's needs of extra work), and has been used to predict future outcomes in high-stakes paper-based testing [5]. Well-known examples include Knowledge Tracing (KT) and Performance Factors Analysis. Plenty of work has been done to try to improve KT by using additional information. A quick look at the literature gives some examples of trying to improve KT by: adding individualizations [8], using parameter learning [10] incorporating time between instances to model forgetting [11], and examining the role of first response time [12]. Although these methods have had some success, neither do they use the order of actions, nor do they examine hint use.

Beck et. al. [2] used KT to explore the impact of help on student performance. Through the use of three distinct methods, they show that help has an impact on learning, but do not address the question when is help helpful? In this present study we present an investigation that shows that there are indeed times when help improves future success.

Problems analyzed had between one and seven hints. When a problem has multiple hints, successive hints usually give more information to solve the problem. See Figure 1 for an example. In Figure 1, you can see that the first hint gives the appropriate equations to use in this problem. The second hint shows a student exactly how to plug the values into the equations. The third hint – the "bottom-out-hint" simply gives the problem's answer. While not all problems in the ASSISTments system use this exact pattern, the general trends are that more hints give successively more help and that the last hint gives the answer.

Duong et. al. in the Sequence of Action (SOA) model [4] binned students into only five categories: one (correct) attempt, only attempts (student used multiple attempts, and no hints), all hints (only hints before a single attempt), Alternative Attempt First (a



*LAK '15,* March 16 - 20, 2015, Poughkeepsie, NY, USA
Copyright 2015 ACM 978-1-4503-3417-4/15/03…$15.00
http://dx.doi.org/10.1145/2723576.2723616

mix of hints and attempts, starting with an attempt), and Alternative Hint First (a mix of hints and attempts, starting with a hint).

One complaint against Duong et. al.'s method is that they do not take into account the distinctly different kinds of hints in the ASSISTments system; these differences have an impact on learning that is not accounted for in their analysis. Almost universally, the final hint in a series (called the "bottom out") simply gives the answer to that problem; this makes the use of that hint significantly different than the use of any other hint. (In the ASSISTments system, students cannot proceed to the next question until they have entered the correct answer; a question that does not have the bottom out hint will trap students and not allow them to finish an assignment.) In Duong's method, all hints were treated identically, even though some are instructional aides, while others simply give the answer with little to no instruction.

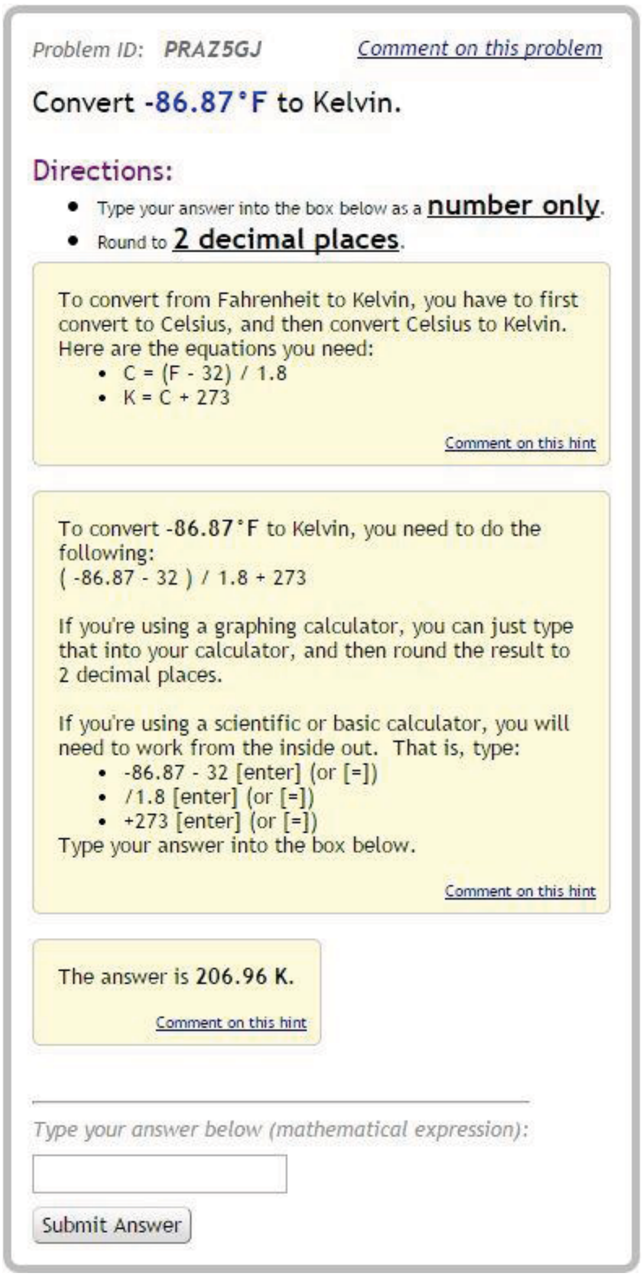


**Figure 1: A sample problem from ASSISTments showing three hints giving varying degrees of "help"**

In the Assistance Model (AM), Wang studied groups of students binned into 12 categories created from the concatenation of divisions along two axes (attempts and hint use). Hint use was divided into four categories: 0, (0-50%], (50%-100%), and 100%. Attempts were divided into three categories: 1, 2-5, 6-infinity. The Hawkins et al study takes into account the bottom-out hint, but fails to look at the action order. Also, any number of attempts between 2 and 5 inclusive are treated identically. This means that a student who makes four mistakes before getting the right answer is treated the same as a student who makes only one mistake. We test this assumption in our analysis.

In order to explore the effect of requesting help on future performance, ideally, we would want to know that students are comparable. In a randomized, controlled trial setting, we would want children that are as similar as possible to be given hints randomly throughout their attempts. Clearly, this is nearly impossible to create. Some notable attempts have been made.

One example of a non-data-mining case would be Attali and Powers [1]. In their tests, they used randomized controlled trials before GRE's to test the effects of help. Although RCT's would be perhaps the most ideal way to test the effects of help, they may also be the least practical, especially when trying to determine the effects of help when used throughout attempts. The number of students that would be needed quickly becomes prohibitive.

Lastly, user modelling algorithms such as KT and PFA attempt to learn four or three parameters per skill. Using a model similar to Duong et. al.'s [4], we attempt to learn one parameter per skill (in a logistic regression) and subdivide students into bins based on attempts and hint use. This gives us a single parameter per skill and a small number (~20) to learn based on our bins.

## 1.2 Questions we are seeking to answer:

1. Does help correlate to an improvement in future performance, and does it matter how many attempts are made compared to help requested (and vice versa)?
2. Does the order of requesting help (interspersed within attempts) make a difference to learning?
3. Can action order be used to predict future performance?

# 2. METHODS

The dataset we use[1] has "click-stream" data; by which we mean the order of attempts and hints used to complete a problem. Having this additional data allows for new analyses of student performance. Using this data enables the use of fewer parameters in a user model than KT and PFA.

The original goal of our analysis was to determine the impact of when hints are used. We came to the conclusion that this analysis could be used to create a model that predicts future performance. We explore the impact of the different factors in our final model

[1] The dataset analyzed comes from ASSISTments (an online learning system developed and maintained at Worcester Polytechnic Institute). In this system, hints are available on demand, but must be requested, while scaffolding is automatically given to students when they submit an incorrect answer. Scaffolded questions (<10%) were ignored in this analysis. The dataset included ~ 400K problem-instances, ~14K students, and 165 skills. Multiple choice questions were ignored; all questions have fill-in style answers. Questions covered middle and high school mathematics topics.

individually to explain their importance; we then explain how the final model was created. In short, we use these factors to differentiate students into “bins”, calculate the probability of next problem correctness (NPC) for 80% of students in each bin, and test our model on the remaining 20%.

## 2.1 Attempts and help

### *2.1.1 Using multiple attempts*

This feature in our model hardly needs explaining (and barely deserves a section). The more attempts a student uses, the worse his or her chance of success on the next problem. The only question we need to answer is how to use attempt count to categorize students.

Similar to the work done by Hawkins et. al. [6] and Duong et. al. [4], we grid students by attempts and hints. However, instead of grouping attempts as [1, (2-5) and 6+], we leave attempts as individual bins for attempts of 1-4, but only for instances where no hints are used. Other instances are regrouped (in an effort to keep the number of divisions within our model reasonable).

### *2.1.2 Using multiple hints*

Much as attempt use, we would expect that students who use more hints have a lower chance of success on the next question. However, based on the notion that these students are receiving help (additional information), this might not be true. Well-written hints could give a barely struggling student information that helps them in the future.

In our analysis, we initially start by treating different numbers of hint use individually. When a student uses only one attempt, there are differences between the outcomes for different hint uses. However, as students use more attempts, the impact of hint use becomes smaller; attempt/hint use combinations are regrouped.

## 2.2 Bottom-out hint use

In ASSISTments, in order to advance to the next problem, a student must type in the correct answer. In the vast majority of cases, this means that the last hint gives the final answer. (See Figure 1 above for an example.) This means that this hint does something different than give instruction. A logical question would be to examine the difference between using the bottom out hint and compare the outcome to using the penultimate hint.

## 2.3 Action order

In order to evaluate the effect of action order on student knowledge, a few combinations of attempts (*a*), hints (*h*) and correct attempt (*A*) are subdivided according to action order. By comparing the results of groups where the only difference is the order of actions, we hope to tease out the effect that action order has on future performance. E.g. instances of two attempts and one hint can either occur as a-h-A (wrong attempt, hint, right attempt) or h-a-A (hint, wrong attempt, right attempt). If action order does not have an effect on future performance, there should be no difference between the groups. It is also useful to compare slightly different groups. E.g. comparing a-A with a-h-A and h-a-A lets us compare the effect of hints use to no hint use with a constant number of attempts. Due to sample sizes, only a few combinations are explored.

# 3. RESULTS AND ANALYSIS

In the analyses that follow, we primarily focus on the probability of the student getting the next problem correct (NPC) within the same skill as the measure of future success.

## 3.1 Attempts and help

There is an interplay between the effect of attempts used and hints used on probability of next problem correctness. When there are large numbers of both hints and attempts used, keeping individual combinations no longer makes sense. Our model regroups predictions based on similarity of outcomes, and similarity of combinations.

### *3.1.1 Using multiple attempts*

With attempt count, values range from one to nearly 1000. However, the use of one to four attempts accounts for well over 90% of all problems. For this reason, the domain of attempt uses is: [1, 2, 3, 4, 5+]. However, when more hints are used, there is less of an impact of the number of attempts used. Thus, when many attempts and hints are used, the granularity of attempts used is reduced.

### *3.1.2 Using multiple hints*

If we look at how *NPC* changes with respect to hint use (leaving out bottom-out hint use), we must decide how fine-grain our bins should be. Non-bottom-out-hint use ranges from zero uses to six. However, the number of problems that use more than three hints is relatively small. The impact of using multiple hints changes based on how many attempts are made. Especially after the action order analysis, it makes sense to regroup hints based on the number of attempts made.

## 3.2 Bottom-out hint use

To analyze the impact of using the last hint, we compare it to students who use the penultimate hint. We find that, across three student knowledge groups (low, medium and high[2]), there is a reliable difference in future performance between penultimate hint use and bottom-out hint use.

**Table 1: The difference between using the penultimate hint and bottom-out-hint.**

| Penultimate hint | avg NPC | std NPC | n |
|---|---|---|---|
| Low knowledge students | 0.4089 | 0.4917 | 3,061 |
| Medium knowledge students | 0.6543 | 0.4756 | 6,143 |
| High knowledge students | 0.8430 | 0.3640 | 1,594 |
| Bottom-out-hint | avg NPC | std NPC | n |
| Low knowledge students | 0.2939 | 0.4556 | 30,757 |
| Medium knowledge students | 0.5439 | 0.4981 | 42,658 |
| High knowledge students | 0.7616 | 0.4262 | 4,654 |

Table 1 suggests that making use of the bottom hint (compared to using the penultimate hint) has roughly 10% (absolute) difference in future performance. This is almost the same effect as dropping a student by one knowledge category. From this analysis, it is clear that bottom-out-hints must be treated differently from any other hint.

As for grouping students who use the bottom-out-hint, they can be broken into two distinct groups: those who go through the bottom out hint and then type in the correct answer, and those who try at least once before they type in the answer. The former category had a future success rate of only 32%, while the latter had a rate of nearly 50% (fairly consistently across all attempts counts). This gives us 22 bins; this initial version (not shown) differs from

[2] Knowledge groups are: [0 to $\mu$-$\sigma$/2), [$\mu$-$\sigma$/2 to $\mu$+$\sigma$/2), [$\mu$+$\sigma$/2 to 1], where $\mu$ = mean knowledge score and $\sigma$ = standard deviation.

the Assistance Model only slightly. This 22-bin version will be modified by action order after analysis of the impact action-order.

## 3.3 Action order

When examining the impact of action order on future performance, there are only a few combinations of action and hint use that have large enough numbers of instances that subdivision is warranted. For this dataset, the groupings worth investigating are (2a,0h) & (2a,1h) and (3a, 0h) & (3a,1h). The zero hint groups are a useful comparison to the hint-used groups, and, this comparison gives surprising results.

The slight difference between a-A and a-h-A may not be surprising; however, the large difference between a-h-A and h-a-A demonstrates the impact that action order has on future success. The p-value for comparing a-h-A and h-a-A is 0.0002. (P-values for a-A to a-h-A and a-A to h-a-A are 0.01 and <0.0001, respectively.) A model that uses only action combination to generate its prediction array is leaving out information that can improve a prediction.

(Key: “a” = wrong attempt; “h” = hint used; “A” = right answer.)

**Table 2a: 2 attempts, 0 or 1 hint, no bottom out hints**

| order | avg NPC | std NPC | n |
|---|---|---|---|
| a-A | 0.732 | 0.443 | 30,402 |
| a-h-A | 0.713 | 0.452 | 3,123 |
| h-a-A | 0.624 | 0.485 | 414 |

**Table 2b: 3 attempts, 0, 1 hint, no bottom out hint**

| order | avg NPC | std NPC | n |
|---|---|---|---|
| a-a-A | 0.670 | 0.470 | 7,810 |
| a-a-h-A | 0.676 | 0.468 | 1,001 |
| a-h-a-A | 0.687 | 0.464 | 500 |
| h-a-a-A | 0.583 | 0.495 | 153 |

The implication here is that the order of action makes a difference.[3] If the first action taken is a hint, the likelihood of future success is reduced. If we are to use action combinations to predict student outcome, the first action becomes important.

## 3.4 Using actions to predict future success

The bins described in 3.2 can be subdivided by action order, much as we did in section 3.3. However, even if we merely subdivide all possible bins (nine have only one order) by first action, we arrive at 35 different bins. Intuitively, this seems like too many. Upon examination of the numbers of questions that fall into the bins, it quickly becomes apparent that conclusions made on 35 bins will be statistically unreliable due to small differences in predictions and small numbers of instances in each bin. With that, we regroup bins based on similarity of prediction and actions. We call this the Fine-Grain-Action model (FGA).

Table 3 gives the values based on student actions in the FGA. (Number in each bin is given below each bin’s value.) To differentiate between the two possible first actions, the table has been sub-divided into a (attempt) and b (hint). It is impossible to have only one attempt come before a hint use. Likewise, it is impossible to have the first action be a hint if no hints are used. It may help to think of Table 3 a / b as a third dimension and “overlap” the two sub-tables.

**Table 3a: The Fine-Grain-Action model**
**1st action = attempt**

<table>
<tr><th></th><th>1 att.</th><th>2 att.</th><th>3 att.</th><th>4 att.</th><th>5 + att.</th></tr>
<tr><td>0 hint</td><td>0.8156<br>215,870</td><td>0.7380<br>22,229</td><td>0.6771<br>5,616</td><td>0.6380<br>2,326</td><td>0.6211<br>2,518</td></tr>
<tr><td>1 hint</td><td>-----</td><td colspan="2">0.7012<br>3,414</td><td colspan="2">0.6321<br>1,408</td></tr>
<tr><td>2 hint</td><td>-----</td><td colspan="4" rowspan="2">0.5812<br>4,011</td></tr>
<tr><td>3+ hint</td><td>-----</td></tr>
<tr><td>BOH</td><td colspan="5">0.5099<br>40,652</td></tr>
</table>

**Table 3b: The Fine-Grain-Action model**
**1st action = hint**

<table>
<tr><th></th><th>1 att.</th><th>2 att.</th><th>3 att.</th><th>4 att.</th><th>5 + att.</th></tr>
<tr><td>0 hint</td><td>-----</td><td>-----</td><td>-----</td><td>-----</td><td>-----</td></tr>
<tr><td>1 hint</td><td>0.7083<br>1,958</td><td colspan="2">0.6192<br>407</td><td colspan="2">0.5702<br>114</td></tr>
<tr><td>2 hint</td><td>0.5250<br>541</td><td colspan="4" rowspan="2">0.4688<br>465</td></tr>
<tr><td>3+ hint</td><td>0.4118<br>289</td></tr>
<tr><td>BOH</td><td colspan="5">0.3396<br>13,989</td></tr>
</table>

Once the bins were created, predicting future success rate was as simple as using 80% of students to create a training subset, and testing the results on the other 20%. Bin values were calculated; in addition, a multivariate logistic regression was run for each skill (much like a very simple PFA). This gives us a chance to compare the results of KT (fit in MATLAB using [7]), as well as two earlier attempts at binning students based on attempts and hint use (SOA by Duong et al and AM by Wang et al.). The comparisons of the methods are found in Table 4.

The RMSE’s presented in Table 4 represent RMSE by student. We felt that this is a better comparison between the tabling models as it removes the effect that a student who has a larger number of instances in the set would thus have a larger weight in the final calculation of the error. R-squared values were found by using the Excel (2010) function applied to all data points; AUC was calculated using SPSS.

**Table 2: RMSE results comparing Knowledge Tracing (KT), Sequence of Action model (SOA), Assistance Model (AM), and the Fine Grain Action Model (FGA)**

| Model | RMSE | $R^2$ | AUC |
|---|---|---|---|
| KT | 0.4069 | 0.1147 | 0.704 |
| SOA | 0.4036 | 0.1155 | 0.708 |
| AM | 0.4002 | 0.1268 | 0.714 |
| FGA | 0.3996 | 0.1282 | 0.715 |

Although the differences in RMSE between the tabling models are slight, the analyses from section 3 demonstrate that incorporating action order into bins derived from action combinations is helpful in making more robust predictions for students who have not

[3] P-values for a-a-A to h-a-a-A and (a-a-h-A & a-h-a-A) to h-a-a-A are 0.016 and 0.011, respectively

gotten a question right on the first attempt. (It's pretty easy to predict the future success of students who are getting questions right.) However, examining the RMSE's of most of the individual bins in FGA demonstrates that this model has room for improvement; bins that predict close to 0.500 have RMSE's close to 0.500 (which is essentially the predictive ability of a coin toss). Two bins even have RMSE's larger than 0.5! Needless to say, a prediction worse than a coin toss needs improvement.

Even though some bins are clearly using the wrong algorithm, the basic premise that we should use both action combination and action order to improve the prediction of future success is valid.

# 4. CONCLUSION

From these analyses, we can conclude that help does, in certain circumstances, help. We feel that we can answer our questions as:

1. Can generalizations of help use be made?
We can conclude that there are times when help use leads to a better chance of future success. However, help use must be combined with student attempts. Students who need too much help or too many attempts are at a disadvantage for their future performance; students who use all available help (especially when they only make one attempt) are among the least likely to succeed in the future.

2. Does the order of hint use matter?
It is better for a student to try a problem before seeking help. However, after a couple of attempts, students may be more likely to do better with help (or are at least not at a disadvantage).

3. Can action order be used to predict student performance?
This knowledge can be used to create a better prediction of future student success. However, when analyzing the error of the prediction, we find that there is still plenty of room for improvement in categories of high hint use and large numbers of attempts. Even with a data set of ~400K attempts, we are encountering small bin sizes and thus statistical insignificance when trying to examine the effects of action order with larger combinations. In addition, there are some circumstances where our methods clearly fail and a new algorithm is needed.

## 4.1 Contributions

Although other authors have examined the role of hints and attempt numbers [5, 7, 13, 14], no one has been able to examine the action combination and order in the detail presented herein. What we find is that the order of hint use and attempts (and not just the combinations) plays a role in the future performance of that student. This analysis also gives further insight into understanding the circumstances under which hints enable learning. The detail of analysis within the action shows that this method gives new insight to user modelling. To our knowledge, no author has shown the detail of analysis to describe how action order impacts student performance.

## 4.2 Future work

The Fine-Grain Action model demonstrates that action-order analysis enhances prediction algorithms on a dataset from ASSISTments; the obvious next question is the applicability to other datasets. Additional future work could include re-analyzing the bins to simplify the model (without loss of predictive power) and using a student's history to see if they improve over time.

# 5. ACKNOWLEDGEMENTS

We acknowledge funding from NSF (# 1440753, 1316736, 1252297, 1109483, 1031398, 0742503), ONR's 'STEM Grand Challenges' and IES (# R305A120125 & R305C100024) grant for ASSISTments and support of the author.

# 6. CITATIONS

[1] Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement*, *70*(1), 22-35.

[2] Beck, J. E., Chang, K., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the Bayesian Evaluation and Assessment methodology. *Intelligent Tutoring Systems*. Springer Berlin Heidelberg.

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

[4] Duong, H. D., Zhu, L., Wang, Y., & Heffernan, N. T. (2013). A Prediction Model Uses the Sequence of Attempts and Hints to Better Predict Knowledge: Better to Attempt the Problem First, Rather Than Ask for a Hint. *Paper Submitted to EDM.*

[5] Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. Intelligent Tutoring Systems. Springer Berlin Heidelberg.

[6] Hawkins, W., Heffernan, N., Wang, Y., & Baker, R. S. Extending the Assistance Model: Analyzing the Use of Assistance over Time.

[7] Murphy, K.: Bayes Net Toolbox for Matlab. < https://code.google.com/p/bnt/ > Accessed 4 September, 2014

[8] Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. *User Modeling, Adaptation, and Personalization*, 255-266.

[9] Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis--A New Alternative to Knowledge Tracing. *Online Submission*.

[10] Qiu, Y., Pardos, Z. A., & Heffernan, N. T. (2012). Towards Data Driven Model Improvement. *FLAIRS Conference*.

[11] Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., & Heffernan, N. T. (2011). Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing. EDM.

[12] Wang, Y., & Heffernan, N. T. (2012). Leveraging First Response Time into the Knowledge Tracing Model. International Educational Data Mining Society.

[13] Wang, Y., & Heffernan, N. T. (2011). The" Assistance" Model: Leveraging How Many Hints and Attempts a Student Needs. *FLAIRS Conference*.

[14] Zhu, L., Wang, Y., & Heffernan, N. T. The Sequence of Action Model: Leveraging the Sequence of Attempts and Hints.