# Modeling Exploration Strategies to Predict Student Performance within a Learning Environment and Beyond

Tanja Käser
AAA Lab
Graduate School of Education
Stanford University
tkaeser@stanford.edu

Nicole R. Hallinen
Psychology Department and
College of Education
Temple University
nicole.hallinen@temple.edu

Daniel L. Schwartz
AAA Lab
Graduate School of Education
Stanford University
daniel.schwartz@stanford.edu

## ABSTRACT

Modeling and predicting student learning is an important task in computer-based education. A large body of work has focused on representing and predicting student knowledge accurately. Existing techniques are mostly based on students' performance and on timing features. However, research in education, psychology and educational data mining has demonstrated that students' choices and strategies substantially influence learning. In this paper, we investigate the impact of students' exploration strategies on learning and propose the use of a probabilistic model jointly representing student knowledge and strategies. Our analyses are based on data collected from an interactive computer-based game. Our results show that exploration strategies are a significant predictor of the learning outcome. Furthermore, the joint models of performance and knowledge significantly improve the prediction accuracy within the game as well as on external post-test data, indicating that this combined representation provides a better proxy for learning.

## CCS Concepts

**•Applied computing → Computer-assisted instruction; Interactive learning environments; •Computing methodologies** → *Knowledge representation and reasoning; Bayesian network models;*

## Keywords

probabilistic student models, learning, strategies, prediction, simulations

## 1. INTRODUCTION

A major question for the design of computerized learning environments is whether success within a learning environment translates to success outside of the environment. Many data mining efforts have primarily focused on modeling and predicting performance within the trajectory of the learning environment.



*LAK '17, March 13-17, 2017, Vancouver, BC, Canada*



DOI: http://dx.doi.org/10.1145/3027385.3027422

One of the most popular approaches to representing and predicting student knowledge accurately is Bayesian Knowledge Tracing (BKT) [13]. Predictive performance of the original BKT model has been improved by applying clustering [27] and individualization techniques [26, 39, 40, 42]. Other widely used student modeling approaches include latent factors models [7, 8, 29] or dynamic Bayesian networks (DBN) [12, 19, 20, 23]. Most of these models represent student knowledge based on the students' past performance within the computerized learning environment, i.e., the students' answers to tasks are assessed and serve as observations for the respective method. When the (predicted) student knowledge within the learning environment does not fully predict success outside of the environment, it may be necessary to consider additional features such as engagement, affect or learning behavior for student modeling.

It has been shown that features such as strategies or choices influence the learning outcome. The strategies students' applied in an educational game influenced their implicit science learning [32, 15]. Furthermore, students inevitably have to make choices when they learn, such as for example the decision about what and how to learn. Choice-based assessments interpret students' choices as an outcome of instruction and use them as a proxy for students' future learning [34]. By integrating choice-based assessments in short interactive computer games, the influence of critical thinking [10], consultation of literature [11], and feedback seeking behavior [14] on the success outside the game was demonstrated.

Furthermore, there has been an increase in the use of open-ended simulations [41] over the last decade. Ideally, students explore different configurations of parameters to infer the underlying principles. Under the best of circumstances, students learn the principles more deeply through exploration than if they are simply told the principles and asked to practice applying them [35]. Moreover, learning how to explore a simulation or empirical phenomenon is a major goal of science education in its own right. A significant technical challenge involves evaluating exploration choices to help predict student learning, perhaps with the intent of intervening, characterizing students, or simply to understand the exploratory behaviors worth teaching. If exploratory behaviors are relevant to learning, then we should be able to detect exploration patterns that are associated with learning outcomes and integrate these patterns into our student models.

However, research on integrated models of performance and strategies is sparse. While other additional features influencing the learning outcome such as help-seeking [4,

30, 31] and off-task behavior [2, 3] have been integrated or added to existing student modeling approaches, research on students' strategies in learning environments mainly focused on detecting [17, 32] player strategies in an educational game, classifying students' problem solving strategies [5, 24] or modeling strategies using interaction networks [15, 16]. FAST [18] is a technique for integrating general features into BKT. Dynamic mixture models [22] and DBNs [33] have been used to trace student engagement and knowledge in parallel. However, none of these existing models combine the representation of performance and learning strategies.

In this paper, we demonstrate that including an analysis of students' exploration strategies within the game increases our ability to predict out-of-game performance compared to an analysis that only considers student success within the game. We present a first-of-kind model for integrating exploratory behaviors and problem-solving success to predict both in-game and out-of-game performance. Our work is based on data collected with a short interactive computer-based game assessing students' exploration choices. The game is centered around a tug-of-war topic and gives students the possibility of simulating their own tug-of-war setups and testing their knowledge about the (hidden) rules (i.e., the forces) governing the tug-of-war. By extensively analyzing the collected log-file data, we demonstrate that students' exploration choices and strategies significantly influence the learning outcome. Furthermore, we build a set of simple probabilistic student models jointly representing student knowledge and strategies and evaluate their prediction accuracy within the computer-based game as well as on an external post-test. Our results demonstrate that modeling the influence of learner strategies on student knowledge significantly improves predictive performance and therefore constitutes a better representation of learning.

## 2. BACKGROUND

Probabilistic graphical models are widely used for representing and predicting student knowledge and learning. One of the most popular approaches is Bayesian Knowledge Tracing (BKT). BKT represents student knowledge by employing one Hidden Markov Model (HMM) per skill. The latent variable of the network represents (binary) student knowledge. The observed variable models the binary answers (correct or wrong) of students to questions associated with the respective skill. The model can be specified using five parameters. The transmission probabilities are described by $p_L$, the probability that a student learns a previously unknown skill and $p_F$, the probability of forgetting an already learned skill. In traditional BKT, we assume $p_F = 0$. The emission probabilities of the model are specified using $p_G$, the probability of correctly applying an unknown skill and $p_S$, the probability of incorrectly answering a question associated with an already learned skill. Finally, $p_0$ describes the initial probability of knowing a skill a-priori. Given a sequence of observations $O_1 = o_1, O_2 = o_2, \ldots, O_T = o_T$ the learning task amounts to estimating the five parameters by maximizing the likelihood function

$$\sum_L p(O_1, \ldots, O_T, L_1, \ldots, L_T | p_0, p_T, p_S, p_G), \qquad (1)$$

where we marginalize over all the hidden states $L$. Maximization of the likelihood is relatively simple and is commonly performed using expectation maximization [9], brute-force grid search [1] or gradient descent [42].

## 3. EXPERIMENTAL SETUP

All evaluations of this paper were conducted using data from an interactive computer-game. In the following, we describe the training environment, the associated post-test as well as the collected data.

### 3.1 Training Environment

Learners need to make choices based on their prior knowledge and the (imperfect) information available to them. Students for example need to decide what and how to learn. Choice-based assessments 'measure' students' choices to get a proxy for their future learning. These assessments give students explicit opportunities to engage in learning behaviors, such as seeking feedback, creating visualizations, or consulting references. `TugLet` is a short, interactive computer-based game assessing students' exploration choices. The topic of the game is a tug-of-war, modeled with respect to forces and motion simulation. Each tug-of-war team consists of a maximum of four team members. There are small (weight $w = 1$), medium ($w = 2$), and large ($w = 3$) characters. To determine the winning side, the strength of each party needs to be computed, i.e., the weights need to be summed up. The position of the weights does not matter. The students are not told the relationships between the different weights, they must be discovered by interacting with the game. In the game, players explore by interacting with a simulation: They can set up opposing tug-of-war teams and see how they fare against each other. The player's goal is to figure out how a team's size/weight corresponds with the strength of its pull, so that they will be able to accurately predict which team will win when presented with alternative scenarios.

Students have the choice between two different activities: *Explore* and *Challenge*. In the *Explore* mode (illustrated in Figure 1 (left)), different characters can be set up and the results can be viewed to induce and test hypotheses. The *Challenge* mode tests the student's knowledge about the weights, i.e., the outcome of tug-of-war questions needs to be predicted (see Figure 1 (right)). This mode consists of eight questions ordered by increasing complexity. If a question gets answered incorrectly, the student is put back into *Explore* mode. The student is free to choose the *Challenge* mode at any point in time. The game is over after correctly answering eight *Challenge* questions in a row.

The interactive computer-game `TugLet` comes with an associated post-test, which assesses the students' knowledge about the rules (i.e., the weights and relationships of the different characters) governing the tug-of-wars. The post-test is a paper-and-pencil test consisting of ten questions. Children are presented a fixed tug-of-war team for the left side as well as ten different tug-of-war teams for the right side. The task is to select all the cases resulting in a tie. A summary sketch of the post-test is provided in Figure 2, where 'L' denotes a large character, 'M' a medium character, and 'S' stands for a small character.

### 3.2 Data Set

The data set used consists of 127 students (68 male, 59 female) in the 8-th grade of a middle school. The students had no prior experience with the topic from the science curricu-

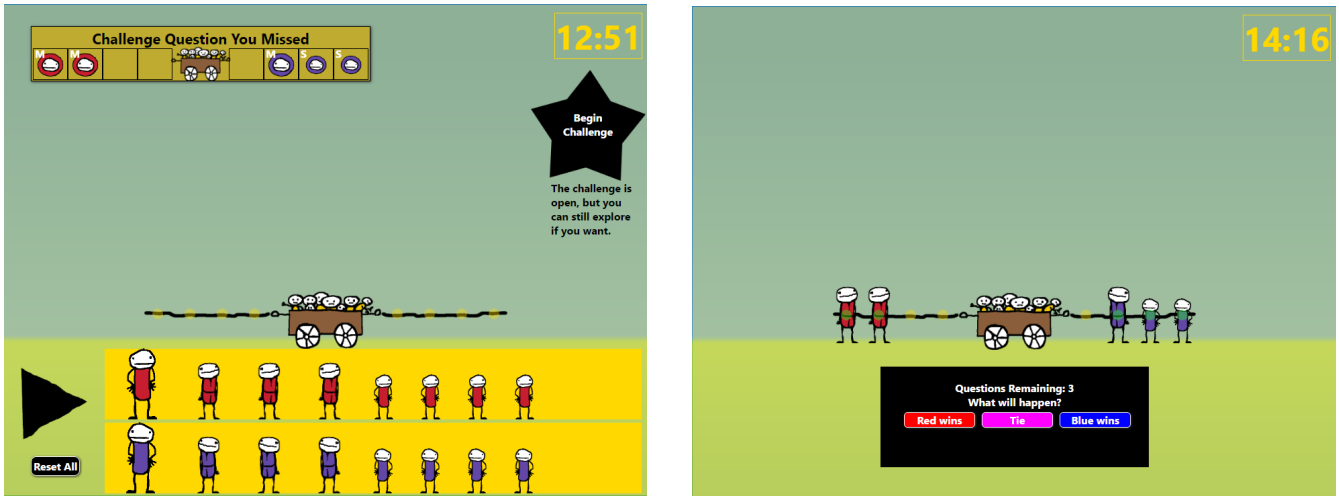


**Figure 1: *Explore* (left) and *Challenge* (right) activities for TugLet. Students are free to enter *Challenge* mode at any point in time by clicking on the *Challenge* button (left).**

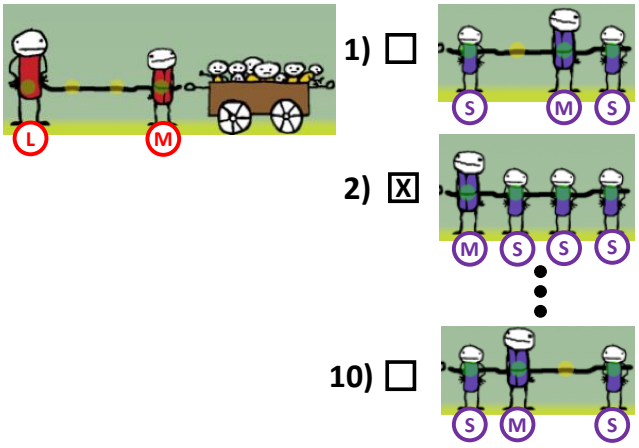


**Figure 2: In the post-test, children have to select the cases resulting in a tie. The second case for example results in a tie, since the weight of three small (S) characters is equal to the weight of one large (L) character.**

lum. Students played TugLet for a maximum time of 15 minutes, followed by a short paper-and-pencil post-test. During game play, all the prompts were recorded in log files. Children solved on average 44.2 challenge questions ($\sigma = 33.3$). They spent 42% of their time in the *Explore* mode. Most of the students ($n = 111$) successfully completed all challenge questions. The average accuracy in the post-test was 0.76 ($\sigma = 0.20$), $n = 31$ students had a perfect post-test.

## 4. KNOWLEDGE REPRESENTATION

We represent the knowledge of the students as a set of rules describing the relationships between the weights of the different characters. The winning side of a specific tug-of-war configuration is then determined by iteratively applying the available rules. The complete TugLet rule set consists of $n = 12$ rules $\mathcal{R} = \{R_i\}$ with $i \in \{1, \ldots, n\}$ and is listed in Table 1. Remember that a large character has a weight of $w = 3$, a medium character weighs $w = 2$ and the weight of a small character is $w = 1$. The rule set $\mathcal{R}$ consists of nine rules describing inequality and equality relationships between the different characters (weights). Furthermore, three meta-rules define basic tug-of-war concepts. Rule $R_{10}$ states that if the left and right team have the exact same number of characters (and weights), the configuration will result in a tie. In rule $R_{12}$ the fact that more characters of the same weight are stronger (i.e., three small characters will win against two small characters) is recorded. Rule $R_{11}$ finally allows for canceling out characters with the same weight on both sides. If the left team for example consists of a large and a medium character and the right side contains a medium and a small character, $R_{11}$ can be applied to cancel out the medium characters. The rule set in $\mathcal{R}$ contains all the rules necessary to solve all possible configurations in the game as well as in the post-test. Note that already a subset of the rules would (theoretically) be enough to derive the relationships between the weights of all characters. The rules $R_1, ..., R_4$ and $R_7, ...R_9$ can for example be derived from rules $R_5$ and $R_6$. This hierarchy of the rule set is necessary, since the students tend to learn in smaller steps, i.e., they test simpler hypotheses first (e.g., 'Large > Small'), and since the questions in *Challenge* mode are ordered by complexity. The final rule set $\mathcal{R}$ therefore is the subset of all possible correct rules necessary to determine the winning side of all tug-of-war set-ups encountered in TugLet and in the associated post-test.

Each tug-of-war configuration is associated with a subset $\mathcal{R}_N \subseteq R$ of rules necessary to determine the winning side. The calculation of $\mathcal{R}_N$ is performed as follows: each rule $R_i \in \mathcal{R}$ has a set of conditions attached under which this specific rule can be applied. Rule $R_6$ for example requires the presence of at least one medium character on the left (or right) side, respectively and a minimum of two small characters placed on the right (or left) side, respectively. To

Table 1: **Rule set $\mathcal{R}$ representing the domain knowledge in TugLet.**

| Rule | Description |
|---|---|
| $R_1$ | Large > Small |
| $R_2$ | Large > Medium |
| $R_3$ | Medium > Small |
| $R_4$ | Large > 2·Small |
| $R_5$ | Large = 3·Small |
| $R_6$ | Medium = 2·Small |
| $R_7$ | Large = Medium+Small |
| $R_8$ | 2·Medium > Large |
| $R_9$ | Small+Large = 2·Medium |
| $R_{10}$ | Equality |
| $R_{11}$ | Cancellation |
| $R_{12}$ | More is better |

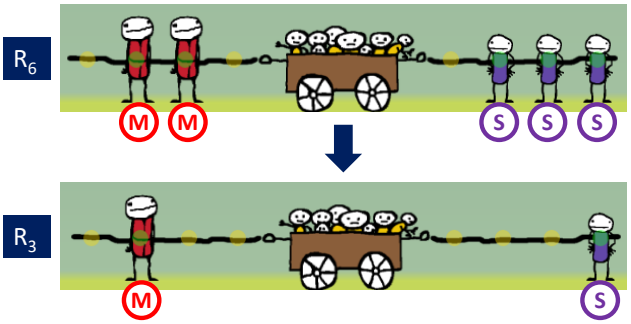


**Figure 3: Example tug-of-war configuration with two medium (M) and three small (S) characters. The winning side can be determined by applying the rule set $\mathcal{R}_N = \{R_3, R_6\}$.**

build $\mathcal{R}_N$, the system iterates through the rules $R_i \in \mathcal{R}$ and applies them, until no more rule can be applied and hence the winning side is determined. During this process, simpler rules describing basic relationships between characters (e.g., $R_1$ or $R_2$) are prioritized. The resulting rule set $\mathcal{R}_N$ consists of all the applied rules. Figure 3 shows the rule set $\mathcal{R}_N$ for an example configuration, where 'L' denotes a large weight, 'M' a medium weight and 'S' stands for a small weight.

During game play, the students are exposed to the rules when testing out tug-of-war configurations in the *Explore* mode and when answering questions in *Challenge* mode. We assume that each tug-of-war configuration encountered provides an opportunity for learning. The rules, which can be acquired (or strengthened) from a specific configuration are exactly the rules $R_i \in \mathcal{R}_N$ associated with the given configuration.

## 5. EXPLORATION STRATEGIES

To analyze the influence of students' exploration choices and behavior, we mined the log file data collected with TugLet as well as the external post-test data (see Section 3.2).

While 87% of the students passed the TugLet game, i.e., managed to answer all eight challenge questions at the end of the training, post-test performance is mixed. While the top 24% of the students have a perfect post-test, the bottom 20% reach an accuracy (ratio of correct answers) less or equal than 0.5. Therefore, the students' training performance measured by there answers in *Challenge* mode seems to describe the learning and knowledge of the students insufficiently.

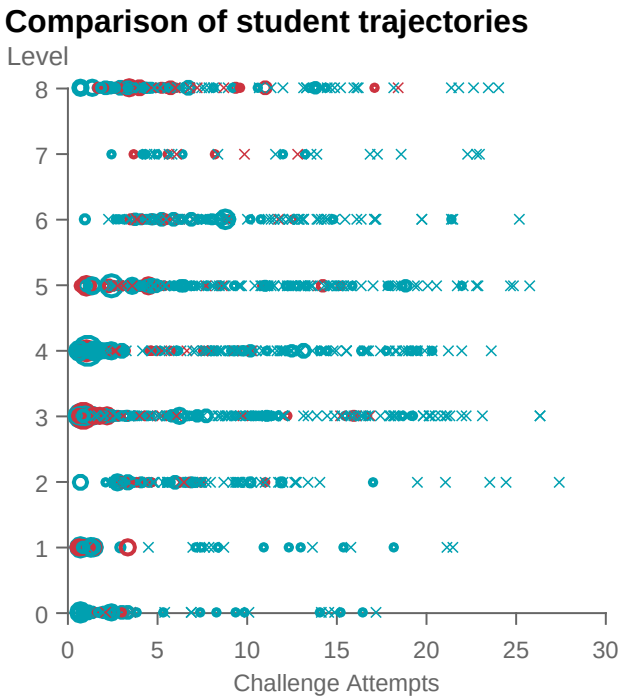


**Figure 4: Comparison of student trajectories. Each circle (or cross) denotes exactly one challenge attempt of one student, i.e., a circle at $(2, 5)$ means that the student answered five (out of eight) questions correctly at his $2^{nd}$ attempt to pass the *Challenge* mode. The size of the circle denotes the number of explored tug-of-war set-ups right before challenging, a cross means that zero set-ups were simulated. Challenge attempts of students with a perfect post-test are colored in red.**

Therefore, we investigated students' exploration behavior by analyzing their trajectories through the game as well as by examining students' specific hypotheses. Figure 4 illustrates the trajectories of the students within the game. The x-axis denotes the number of attempts so far in passing the *Challenge* mode, the y-axis denotes the level (number of correctly answered questions) reached: each circle or cross in Figure 4 denotes a challenge attempt of a student. A circle (or cross) at $(4, 5)$ means that the student answered five (out of eight) questions correctly at his $4^{th}$ attempt to pass the *Challenge* mode. The size of the circles denotes the number of tug-of-war set-ups simulated in the *Explore* mode right before changing to *Challenge* mode. A cross signifies that no set-ups were simulated, i.e., the student changed right back to *Challenge* mode. Attempts of students with a perfect post-test are marked in red, the attempts of all other students are colored in blue. Figure 4 shows that while the

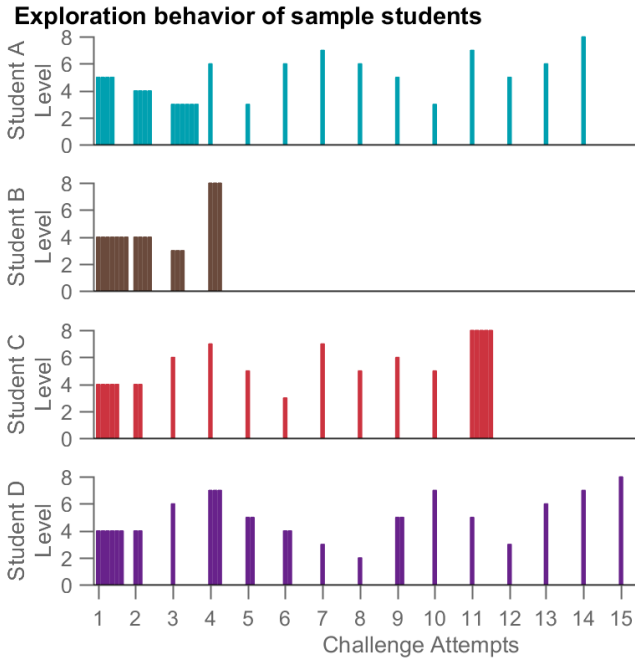


**Figure 5: Example student trajectories. The x-axis denotes the number of challenge attempts, the y-axis the level (number of correctly answered questions) reached in the actual attempt. The width of the bar shows the number of exploration set-ups tested right before the actual challenge attempt.**

students with a perfect post-test pass the game soon, other students need a lot of attempts in *Challenge* mode before passing. Indeed, there is a significant negative correlation ($\rho = -0.28, p = .001$) between the number of challenge attempts and the achieved post-test accuracy. Figure 4 also demonstrates that the better performing students exhibit a different behavior regarding exploration than those students with lower post-test accuracies. At the beginning, all the children test many tug-of-war set-ups, this number decreases over time (as visible from the decreasing circle sizes as well as the many crosses). Therefore, there is no significant correlation between the post-test performance and the number of tug-of-war set-ups tested before the first challenge attempt ($p = .203$). However, while the students with perfect post-test tend to test (few) tug-of-war set-ups in-between two challenge attempts, students with lower post-test accuracy stop exploring completely as can be seen from the many blue crosses in Figure 4. Indeed, the average number of tug-of-war set ups tested in *Explore* mode in-between two attempts to pass the *Challenge* mode is positively correlated to post-test accuracy ($\rho = 0.18, p = .048$).

Figure 5 illustrates the trajectories of four example students. The x-axis again shows the number of attempts in passing the *Challenge* mode, the y-axis shows the achieved level (number of correctly answered questions). The width of the bar denotes the number of tug-of-war set-ups tested before changing to *Challenge* mode. Student B and Student C had a perfect post-test, while the post-test accuracy of Student A and Student D was below 0.5. The sample trajectories confirm that students with low performance need more challenge attempts to pass the game. It seems that, while all students spend much time in *Explore* mode in the beginning, students performing badly in the post-test give up exploration much earlier. Student C is an exception: this student does not explore in the beginning, but realizes later that he will not pass without doing so. Student D persists, but does not seem to profit from the investigated tug-of-war set-ups.

We hypothesize that the reason behind these observations might be the fact that the conclusions drawn in the *Explore* mode are of higher value for the good performers, i.e., that the good performers test more informative tug-of-war set-ups. Figure 6 illustrates this behavior. The initial part of the trajectories of Student B (left) and Student A (right) are displayed. In the initial explore phase, both students exhibit similar exploration strategies. They test hypotheses such as equality (tug-of-war set-ups 1 and 4 of Student A), position independence (tug-of-war set-ups 2 and 5 of Student B), and relationships between the weights of the characters (tug-of-war set-up 3 of Student A). Note that during the first exploration phase, the characters available for simulation are limited: only one large and one small character are provided for each team. After the first wrong answer in *Challenge* mode, the students are put back into *Explore* mode and have the full set of characters available for hypothesis testing. Now the wheat is separated from the chaff. Student B (see Figure 6 (left)) systematically tests tug-of-war set-ups exploring the relationships between the different characters leading to the (possible) derivation of rules $R_6$, $R_4$, and $R_5$ (see Table 1). Student A on the other hand seems overcharged with the many characters available for testing. As we can see from Figure 5, Student A explores once more before the third challenge attempt, but completely quits exploring later on.

Given these observations, we divide all tug-of-war set-ups tested in the *Explore* mode into three categories: 'strong', 'medium', 'weak'. This categorization is computed automatically based on the set of rules $\mathcal{R}_N$ necessary to determine the winner of the given tug-of-war configuration. We found that a good exploration strategy focuses on isolating one underlying principle at a time. Therefore, a set-up is considered as 'strong', if $|\mathcal{R}_N| = 1$ and $R_i \in \mathcal{R}_N$ is seen for the first time, i.e., the student tests exactly one new rule. If the rule $R_i$ has been tested or seen previously, the set-up is categorized as being 'medium'. If the set-up tests two rules, i.e., $|\mathcal{R}_N| = 2$ and $R_{11} \in \mathcal{R}_N$ the tested configuration is labeled as a 'medium' hypothesis. We assume that the student could still draw conclusions (i.e., find a new rule $R_i$) by first applying the cancellation rule $R_{11}$ (see Section 4 and Table 1) and thus reducing the configuration to a set-up testing exactly one rule. If $|\mathcal{R}_N| = 2 \wedge R_{11} \notin \mathcal{R}_N$, the tested set-up is put into the 'weak' category. We also categorize tug-of-war set-ups as being 'weak' hypotheses if they require more than two rules to determine the winning side, i.e., if $|\mathcal{R}_N| > 2$. A set-up testing too many principles at the same time does not allow to draw conclusions on relationships between single characters. An analysis of the training data reveals, that better performers indeed seem to have superior exploration strategies: there is a significant positive correlation between the number of 'strong' tug-of-

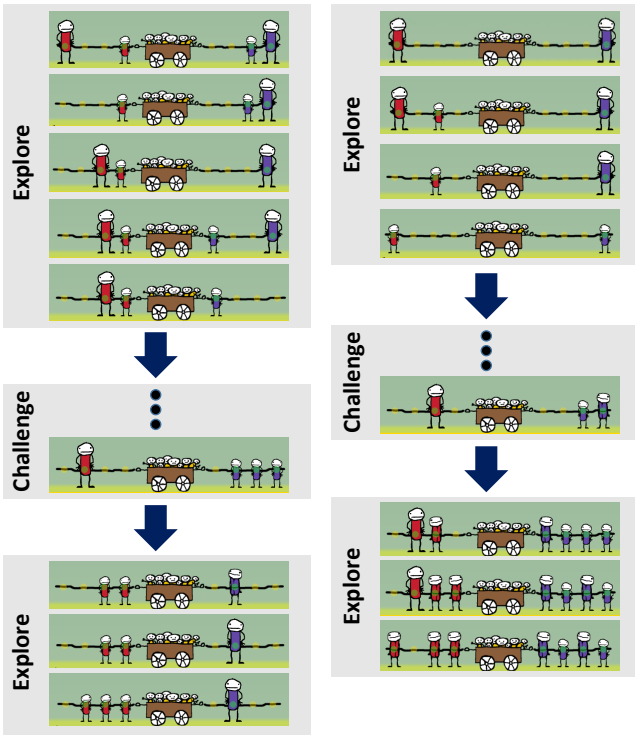


**Figure 6: Comparison of initial tug-of-war set-ups for Student B (left) and Student A (right). While they exhibit similar exploration strategies in the beginning, strategies start to differ considerably with increasing difficulty.**

war set-ups tested and the achieved accuracy in the post-test ($\rho = 0.21, p = .019$).

# 6. PROBABILISTIC MODELS OF STRATEGIES

To investigate the benefits of modeling performance and strategies jointly, we constructed probabilistic graphical models representing student knowledge and exploration behavior in one network and evaluated their predictive performance within the `TugLet` environment as well as in the post-test.

## 6.1 Simple Probabilistic Models

To model the learning process of the students and to make predictions about their performance in the game as well as in the post-test, we build probabilistic graphical models based on the representation of domain knowledge as a set of rules (see Section 4).

**Pure Challenge Model**. The pure challenge model (PCM) is a HMM, employing one model per rule. Figure 7 illustrates the structure of the graphical model. The binary latent variable $K_{R_i,t}$ represents, whether the student has mastered rule $R_i$ at time $t$. The observed variable $O_{R_i,t}$ is also binary and indicates, whether a student has correctly applied $R_i$ at time $t$. Correctness is encoded as follows: If a student answers a challenge question at time $t$ correctly, we assume that all

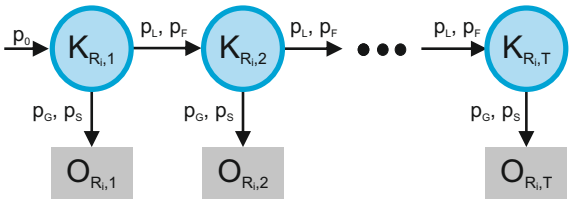


**Figure 7: Structure of the graphical model over $T$ time steps for the PCM, the CHM and the WHM.**

rules $R_i \in \mathcal{R}_N$ have been applied correctly, i.e., $o_{R_i,t} = 1$, $\forall R_i \in \mathcal{R}_N$. If the student gives an incorrect answer, we assume the all rules $R_i \in \mathcal{R}_N$ have been applied incorrectly, i.e., $o_{R_i,t} = 0$, $\forall R_i \in \mathcal{R}_N$. This encoding method also influences prediction: the predicted probability $\hat{p}_{C,t}$ that the student will correctly determine the winning team of a tug-of-war configuration $C$ at time $t$ depends on the predicted probabilities $\hat{p}(O_{R_i,t} = 1)$ of the rules $R_i \in \mathcal{R}_{N_C}$:

$$\hat{p}_{C,t} = \prod_{R_i} \hat{p}(O_{R_i,t} = 1),\ R_i \in \mathcal{R}_{N_C}. \qquad (2)$$

While this model is based on BKT, we allow a small amount of forgetting ($p_F > 0$). Note that in the PCM, we do not represent actions performed in the *Explore* mode.

**Correct Hypotheses Model**. The correct hypotheses model (CHM) is an extension of the PCM. It again employs one HMM per rule (see Figure 7) and the interpretations of the latent and observed variables are accordingly. We encode the answers to the challenge questions in the same way as for the PCM. However, in contrast to the challenge model, the CHM also incorporates the actions performed in *Explore* mode. For each tug-of-war set-up $H$ tested in the *Explore* mode, the rule set $\mathcal{R}_{N_H}$ necessary to find the winning side of the simulated set-up are computed. We then assume that all rules in $\mathcal{R}_{N_H}$ have been applied correctly, i.e., $o_{R_i,t} = 1$, $\forall R_i \in \mathcal{R}_{N_H}$.

## 6.2 Modeling Strategies

Both the PCM and the CHM are variations of BKT models and therefore allow for efficient parameter learning and predictions. However, the two models do not (in case of the PCM) or only in a limited way (in case of the CHM) take the exploration behavior of the students into account. Yet, our data analysis has shown that students' exploration choices and strategies are significantly correlated to the learning outcome (see Section 5).

**Weighted Hypotheses Model**. The weighted hypotheses model (WHM) is based on the observation that exploration behavior significantly influences post-test performance. It again employs one HMM per rule and uses the graphical structure illustrated in Figure 7. The binary latent variables $K_{R_i,t}$ again denote, whether the student has mastered rule $R_i$. The observed variables $O_{R_i,t}$ are also binary and denote an application of rule $R_i$ when answering a challenge question $C$ or the testing of a rule $R_i$ in a tug-of-war set-up $H$ in *Explore* mode. We encode answers in *Challenge* mode as described in the PCM (see Section 6.1) and rules encoun-

tered in *Explore* mode as explained in the CHM (see Section 6.1). However, the WHM introduces a weighting of the different observations. Observations associated with a tested tug-of-war set-up are weighted according to the three categories 'strong', 'medium', 'weak' as defined in Section 5. Challenge answers are weighted differently based on their correctness. The sequence of $T$ observations $\mathbf{o_{R_i}}$ for a rule $R_i$ is therefore given by

$$\mathbf{o_{R_i}} = (o_{R_{i,1}}^{w_1}, o_{R_{i,2}}^{w_2}, ..., o_{R_{i,T}}^{w_t}), \tag{3}$$

with weights $w_j$, $j \in 1, ..., T$ specified as follows:

$$w_j = \begin{cases} w_{hs} & o_{R_{i,j}} \text{ is a strong hypothesis.} \\ w_{hm} & o_{R_{i,j}} \text{ is a medium hypothesis.} \\ w_{hw} & o_{R_{i,j}} \text{ is a weak hypothesis.} \\ w_{cw} & o_{R_{i,j}} \text{ is a wrong challenge answer.} \\ w_{cs} & o_{R_{i,j}} \text{ is a correct challenge answer.} \end{cases} \tag{4}$$

The weights $w = (w_{hs}, w_{hm}, w_{hw}, w_{cw}, w_{cs})$ are positive integers and can be learned from the collected data using cross validation.

## 6.3 Experimental Evaluation

We evaluated the predictive accuracy of our models within the `TugLet` environment as well as on the post-test using the data set described in Section 3.2. We used a train-test setting, i.e., parameters were fit on the training data set and model performance was evaluated on the test set. All the models were fit using a Nelder-Mead (NM) optimization [25]. The NM algorithm is often used for optimization problems due to its simplicity and fast convergence rate. Predictive performance was evaluated using the root mean squared error (RMSE) as well as the area under the ROC-curve (AUC). The RMSE is widely used for the evaluation of student models, e.g., [26, 39, 40, 42]. The AUC is a useful additional measure to assess the resolution of a model.

**Within-Game Prediction**. The prediction accuracy of the PCM and the CHM models on the log files collected from *TugLet* was evaluated using student-stratified (i.e. dividing the folds by students) 10-fold cross validation. Since the estimation of model performance during parameter tuning leads to a potential bias [6, 38], we use a **nested** 10-fold student-stratified cross validation to estimate the predictive performance of the WHM and to at the same time learn the optimal weights $w_{opt}$ for this model. We used $r = 50$ random re-starts for the NM algorithm for all models, since the NM algorithm is known for being trapped into local optima and to be sensitive to the initial starting values [25, 28]. We used the same parameter constraints for all models: $p_i \leq 0.5$, if $i \in \{L, F, G, S\}$. The prior probability $p_0$ remained unconstrained. Figure 8 displays the RMSE and the AUC for the PCM, CHM, and WHM models.

The WHM demonstrates the highest prediction accuracy within the game ($RMSE_{WHM} = 0.3328$). The inclusion of exploration choices into the model led to an improvement in RMSE by 2.6% ($RMSE_{PCM} = 0.3574$, $RMSE_{CHM} = 0.3480$), the representation of strategies further reduced the RMSE by 4.4% ($RMSE_{CHM} = 0.3480$, $RMSE_{WHM} = 0.3328$). A one-way analysis of variance performed on the

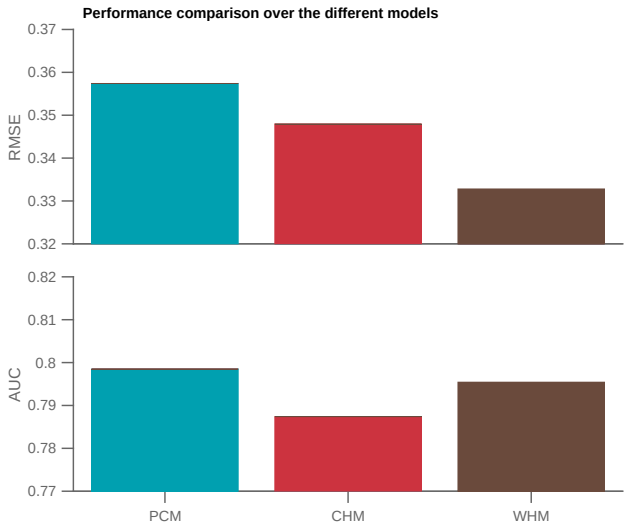


**Figure 8: Comparison of within-game prediction accuracy of the PCM (modeling knowledge only), CHM (modeling knowledge and exploration) and WHM (modeling knowledge and exploration strategies).**

**Table 2: Mean pair-wise differences $\mu$ in RMSE between the models along with confidence intervals $ci$ and significance values $p$ for the within-game prediction.**

| | Mean $\mu$ | 95% ci of $\mu$ | p |
|---|---|---|---|
| $\mathbf{d_{PCM,CHM}}$ | 0.0092 | [-0.0055,0.0239] | .309 |
| $\mathbf{d_{PCM,WHM}}$ | 0.0240 | [ 0.0093,0.0387] | <.001 |
| $\mathbf{d_{CHM,WHM}}$ | 0.0148 | [ 0.0001,0.0296] | <.05 |

per-student RMSE of the different models shows that there are indeed significant differences between the mean RMSEs of the different models ($F = 7.45, p < .001$). The results of multiple comparisons (using a Bonferroni-Holm correction) between the different models are listed in Table 2. There is no significant difference in performance between the PCM and the CHM models. However, the WHM significantly outperforms the PCM and CHM models. All three models are performing well in discriminating challenges from failures ($AUC_{PCM} = 0.7985$, $AUC_{CHM} = 0.7874$, $AUC_{WHM} = 0.7954$), there are no significant differences in AUC between the models.

The optimal weights found for the WHM are $w_{opt} = \{3, 1, 1, 1, 2\}$. Tug-of-war set-ups classified as 'strong' hypotheses have a higher impact than set-ups falling in the 'medium' or 'weak' categories ($w_{hs} = 3, w_{hm} = 1, w_{hw} = 1$). 'Strong' hypotheses are also assigned more weight than correct answers to challenge questions ($w_{hs} = 3, w_{cs} = 2$).

**Post-Test Prediction**. To evaluate the predictive performance of the different models on the post-test, we used all within-game observations (i.e., actions performed within the `TugLet` environment) for training and predicted the outcome of the external post-test. We again used $r = 50$ random re-

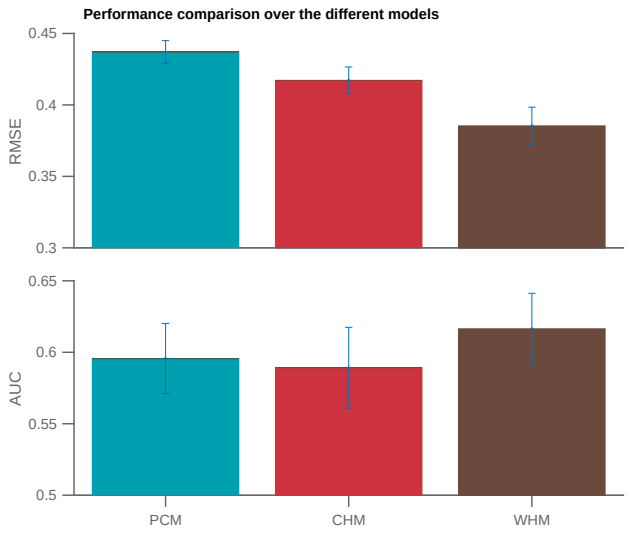


**Figure 9: Comparison of post-test prediction accuracy along with standard deviations for the PCM (modeling knowledge only), CHM (modeling knowledge and exploration) and WHM (modeling knowledge and exploration strategies).**

starts for the NM algorithm. We constrained the parameters of all models as described for the within-game prediction: $p_i \leq 0.5$, if $i \in \{L, F, G, S\}$. The prior probability $p_0$ remained unconstrained. For the WHM, we can safely use the optimal weights $w_{opt} = \{3, 1, 1, 1, 2\}$ found in the nested cross validation, since this optimization was performed on within-game data only. Prediction accuracy in terms of the RMSE and the AUC was computed using bootstrap aggregation with re-sampling ($b = 100$). Figure 9 displays the error measures (with standard deviations) for the PCM, CHM, and WHM models.

The WHM shows the best performance for both error measures. Modeling exploration behavior even in a simplistic way leads to an improvement in RMSE of 4.55% ($RMSE_{PCM} = 0.4370$, $RMSE_{CHM} = 0.4171$), categorization of the different explored set-ups along with the introduction of weighted observations decreases the RMSE by another 7.6% ($RMSE_{CHM} = 0.4171$, $RMSE_{WHM} = 0.3854$).

The low standard deviations in RMSE ($\sigma_{PCM} = 0.0079$, $\sigma_{CHM} = 0.0094$, $\sigma_{WHM} = 0.0131$) indicate significant differences between the different models. A one-way analysis of variance confirms that there are indeed significant differences between the mean RMSEs of the different models ($F = 633.46$, $p < .001$). Multiple comparisons (using a Bonferroni-Holm correction) between the mean RMSEs of the different models demonstrate that all model means are significantly different from each other. Table 3 illustrates this fact: The 95% confidence intervals for the differences in RMSE between the models do not include zero.

The WHM also exhibits a higher AUC than the PCM and the CHM ($AUC_{PCM} = 0.5956$, $AUC_{CHM} = 0.5891$, $AUC_{WHM} = 0.6164$). Although the standard deviations ($\sigma_{PCM} = 0.0246$, $\sigma_{CHM} = 0.0283$, $\sigma_{WHM} = 0.0248$) are higher than for the RMSE, a one-way analysis of variance suggests that the mean AUCs of the different models are not the same ($F = 30.22$, $p < .001$). The multiple comparisons (employing a Bonferroni-Holm correction) between the mean AUCs demonstrate that while the differences between the $PCM$ and the $CHM$ are not significant, the $WHM$ significantly outperforms the other two models. Table 4 lists the mean values $\mu$ for the differences between the models' average AUCs along with 95% confidence intervals and significance values.

**Table 3: Mean pair-wise differences $\mu$ in RMSE between the models along with confidence intervals $ci$ and significance values $p$.**

| | Mean $\mu$ | 95% ci of $\mu$ | p |
|---|---|---|---|
| $\mathbf{d_{PCM,CHM}}$ | 0.0199 | [0.0165,0.0233] | <.001 |
| $\mathbf{d_{PCM,WHM}}$ | 0.0517 | [0.0483,0.0551] | <.001 |
| $\mathbf{d_{CHM,WHM}}$ | 0.0318 | [0.0284,0.0352] | <.001 |

**Table 4: Mean pair-wise differences $\mu$ in AUC between the models along with confidence intervals $ci$ and significance values $p$.**

| | Mean $\mu$ | 95% ci of $\mu$ | p |
|---|---|---|---|
| $\mathbf{d_{PCM,CHM}}$ | 0.0065 | [-0.0021,0.0151] | 0.184 |
| $\mathbf{d_{PCM,WHM}}$ | -0.0209 | [-0.0295,-0.0123] | <.001 |
| $\mathbf{d_{CHM,WHM}}$ | -0.0273 | [-0.0359,-0.0187] | <.001 |

## 7. DISCUSSION AND CONCLUSION

The strategies and choices of students in a learning environment have a significant influence on their learning outcome. Previous work has shown that strategies used vary considerably across students [36, 37]. Furthermore, students' abilities in critical thinking [10], their literature inquiries [11], and their feedback seeking behavior [14] have a significant impact on the learning outcome.

Recent research in educational data mining has investigated the strategic behavior of children in games. However, most of this work has focused on the data mining part, i.e., measuring implicit science learning based on player moves in an educational game [17, 32] or the classification of problem solving strategies [5, 24]. Research on the modeling part has focused on representing the problem solving behavior only [15].

In contrast to previous work, we represent student knowledge and exploration strategies jointly in one model. Our work is comparable to research on engagement modeling, where student knowledge and engagement are simultaneously traced [33]. FAST [18] also allows for the integration of additional features into a BKT model, however, these additional features influence prediction of the observed state only. In contrast to this approach, in our joint model of knowledge and strategy, the strategies directly influence the (hidden) knowledge state. This technique allows us to predict performance on an external post-test, where information about strategies is not available.

Our results demonstrate that even simple probabilistic models of strategies offer a better representation of learning than a pure performance model. Modeling the strength of student hypotheses leads to a small, but significant im-

provement of 6.9% of the RMSE ($RMSE_{PCM} = 0.3574$, $RMSE_{WHM} = 0.3328$), when predicting students' answers to challenge questions within the learning environment. Improvements are much larger for the post-test: the joint representation of performance and strategies improves the RMSE by 11.8% ($RMSE_{PCM} = 0.4370$, $RMSE_{WHM} = 0.3854$). Modeling strategies also improves the AUC in the post-test, i.e., the WHM is better at discriminating failures (incorrectly answered challenge questions) from successes than the PCM. The increased prediction accuracy on the post-test demonstrates that 1) using probabilistic models of strategies, we are able to improve the detection of 'shallow' learning [21]: From the 111 students passing the game (measured by an assessment of their performance), 21 students achieved an accuracy less or equal than 0.5 on the post test. The better predictive performance on the post-test also shows that 2) simple probabilistic models representing performance and knowledge jointly are superior at identifying understanding. The post-test required a higher level of rule understanding and also a transfer, since tasks were asked in a different way than in the game (selecting tug-of-war set-ups resulting in a tie vs. determining the outcome of a given tug-of-war set-up).

The improved predictive performance of our joint representation of strategies and performance as well as the significant correlations found between exploration choices, strength of hypotheses and the learning outcome confirm the findings of previous work: Students' choices [10, 11, 14] and learning strategies [15, 32] have a significant impact on the learning outcome. These findings give important directions for assessment: not only performance data, but also students' strategies and choices need to be measured to reliably predict future learning.

The strategies represented in our models are of course specific to the presented educational game. They can, however, be generalized to the inquiry strategies of simplification and testing one principle at a time. In future work, we would therefore like to model these inquiry strategies for different educational games and simulations in order to analyze and demonstrate the generalizability of our models.

To conclude, we have proposed the use of probabilistic graphical models jointly representing student knowledge and strategies. Our results demonstrate that simple probabilistic models of strategies are sufficient to significantly improve prediction accuracy. Furthermore, we have shown that students' strategies significantly influence the learning outcome and therefore, augmented models are a better predictor for learning than pure performance models.

# 8. REFERENCES


[1] R. S. Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere. Contextual Slip and Prediction of Student Performance after Use of an Intelligent Tutor. In *Proc. UMAP*, pages 52–63, 2010.

[2] R. S. Baker, A. T. Corbett, and K. R. Koedinger. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proc. ITS*, pages 531–540, 2004.

[3] R. S. J. d. Baker, A. T. Corbett, I. Roll, and K. R. Koedinger. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3):287–314, 2008.

[4] J. E. Beck, K.-m. Chang, J. Mostow, and A. Corbett. Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. In *Proc. ITS*, pages 383–394, 2008.

[5] P. Blikstein. Using Learning Analytics to Assess Students' Behavior in Open-ended Programming Tasks. In *Proc. LAK*, pages 110–116, 2011.

[6] A.-L. Boulesteix and C. Strobl. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Medical Research Methodology*, 9(1):85+, 2009.

[7] H. Cen, K. R. Koedinger, and B. Junker. Is Over Practice Necessary? -Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. In *Proc. AIED*, pages 511–518, 2007.

[8] H. Cen, K. R. Koedinger, and B. Junker. Comparing Two IRT Models for Conjunctive Skills. In *Proc. ITS*, pages 796–798, 2008.

[9] K.-M. Chang, J. Beck, J. Mostow, and A. Corbett. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proc. ITS*, pages 104–113, 2006.

[10] M. Chi, D. L. Schwartz, K. P. Blair, and D. B. Chin. Choice-based Assessment: Can Choices Made in Digital Games Predict 6th-Grade Students' Math Test Scores? In *Proc. EDM*, pages 36–43, 2014.

[11] D. B. Chin, K. P. Blair, and D. L. Schwartz. Got game? A choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge, and Learning*, 21:195–210, 2016.

[12] C. Conati, A. Gertner, and K. VanLehn. Using Bayesian Networks to Manage Uncertainty in Student Modeling. *UMUAI*, 12(4):371–417, 2002.

[13] A. T. Corbett and J. R. Anderson. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *UMUAI*, 4(4):253–278, 1994.

[14] M. Cutumisu, K. P. Blair, D. B. Chin, and D. L. Schwartz. Posterlet: A Game-Based Assessment of Children's Choices to Seek Feedback and to Revise. *Journal of Learning Analytics*, 2(1):49–71, 2015.

[15] M. Eagle and T. Barnes. Exploring Differences in Problem Solving with Data-Driven Approach Maps. In *Proc. EDM*, pages 76–83, 2014.

[16] M. Eagle, D. Hicks, B. Peddycord, III, and T. Barnes. Exploring Networks of Problem-Solving Interactions. In *Proc. LAK*, pages 21–30, 2015.

[17] M. Eagle, E. Rowe, D. Hicks, R. Brown, T. Barnes, J. Asbell-Clarke, and T. Edwards. Measuring Implicit Science Learning with Networks of Player-Game Interactions. In *Proc. CHI in Play*, pages 499–504, 2015.

[18] J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge. In *Proc. EDM*, pages 84–91, 2014.

[19] J. P. González-Brenes and J. Mostow. Dynamic Cognitive Tracing: Towards Unified Discovery of Student and Cognitive Models. In *Proc. EDM*, 2012.

[20] J. P. González-Brenes and J. Mostow. Topical Hidden Markov Models for Skill Discovery in Tutorial Data. *NIPS - Workshop on Personalizing Education With Machine Learning*, 2012.

[21] S. M. Gowda, R. S. Baker, A. T. Corbett, and L. M. Rossi. Towards Automatically Detecting Whether Student Learning is Shallow. *IJAIED*, 23(1):50–70, 2013.

[22] J. Johns and B. Woolf. A dynamic mixture model to detect student motivation and proficiency. In *Proc. AAAI*, pages 163–168, 2006.

[23] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks. In *Proc. ITS*, pages 188–198, 2014.

[24] L. Malkievich, R. S. Baker, V. Shute, S. Kai, and L. Paquette. Classifying behavior to elucidate elegant problem solving in an educational game. In *Proc. EDM*, pages 448–453, 2016.

[25] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[26] Z. A. Pardos and N. T. Heffernan. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In *Proc. UMAP*, pages 255–266, 2010.

[27] Z. A. Pardos, S. Trivedi, N. T. Heffernan, and G. N. Sárközy. Clustered knowledge tracing. In *Proc. ITS*, pages 405–410, 2012.

[28] J. M. Parkinson and D. Hutchinson. An Investigation into the Efficiency of Variants on the Simplex Method. In *Numerical Methods for Non-linear Optimization*, pages 115–135. Academic Press, 1972.

[29] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance Factors Analysis - A New Alternative to Knowledge Tracing. In *Proc. AIED*, pages 531–538, 2009.

[30] I. Roll, V. Aleven, B. McLaren, and K. Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21:267–280, 2011.

[31] I. Roll, R. Baker, V. Aleven, and K. R. Koedinger. On the benefits of seeking (and avoiding) help in online problem solving environment. *Journal of the Learning Sciences*, 23(4):537–560, 2014.

[32] E. Rowe, R. Baker, J. Asbell-Clarke, E. Kasman, and W. Hawkins. Building Automated Detectors of Gameplay Strategies to Measure Implicit Science Learning. In *Proc. EDM*, pages 337–338, 2014.

[33] S. E. Schultz and I. Arroyo. Tracing Knowledge and Engagement in Parallel in an Intelligent Tutoring System. In *Proc. EDM*, pages 312–315, 2014.

[34] D. L. Schwartz and D. Arena. Measuring what matters most: Choice-based assessments for the digital age. *The MIT Press*, 2013.

[35] D. L. Schwartz, C. C. Chase, M. A. Oppezzo, and D. B. Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4):759–775, 2011.

[36] R. S. Siegler. Strategy choice and strategy discovery. *Learning and Instruction*, 1(1):89–102, 1991.

[37] R. S. Siegler and Z. Chen. Developmental Differences in Rule Learning: A Microgenetic Analysis. *Cognitive Psychology*, 36:273–310, 1998.

[38] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006.

[39] Y. Wang and J. Beck. Class vs. Student in a Bayesian Network Student Model. In *Proc. AIED*, pages 151–160, 2013.

[40] Y. Wang and N. T. Heffernan. The student skill model. In *Proc. ITS*, pages 399–404, 2012.

[41] C. E. Wieman, W. K. Adams, and K. K. Perkins. PhET: Simulations That Enhance Learning. *Science*, 322(5902):682–683, 2008.

[42] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized Bayesian Knowledge Tracing Models. In *Proc. AIED*, pages 171–180, 2013.