Computers in Human Behavior 61 (2016) 198–204

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Is there a gender difference in interacting with intelligent tutoring system? Can Bayesian Knowledge Tracing and Learning Curve Analysis Models answer this question?

Leyla Zhuhadar [a], Scarlett Marklin [b], Evelyn Thrasher [a], Miltiadis D. Lytras [c, *]

[a] Western Kentucky University, Gordon Ford College of Business, Bowling Green, KY 42101, USA
[b] Tallahassee, Florida Area, Florida State University, USA
[c] 6 Gravias Street GR-153 42, Aghia Paraskevi, Athens, The American College of Greece, Greece

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form
17 February 2016
Accepted 18 February 2016
Available online 17 March 2016

Keywords:
Intelligent tutoring systems
Bayesian knowledge tracing
Learning curve analysis

ABSTRACT

Multiple studies have been conducted on Project LISTEN, an intelligent tutoring system (ITS) used to analyze educational learning through case analysis of students' interactions with ITS. Studies have defined the phenomenon by exploring 'what happens when/if' questions and analyzing these in the context of the specified phenomenon occurrence. While ITS often focus on student decisions regarding when and how to use the system's resources, we suggest further analysis and monitoring are needed to get the best results from these systems. In this study, we argue that boys interact differently with ITS than girls. This finding is evident in results from both the *Bayesian Knowledge Tracing* and *Learning Curve Analysis* models.

1. Introduction and related works

For almost a decade, faculty members from Western Kentucky University (WKU[1]) have been designing a variety of intelligent systems, such as *CaseGrader* (Crews & Murphy, 2007) and *HyperManyMedia* (Zhuhadar & Nasraoui, 2008). In *CaseGrader*, Crews & Murphy (2007) used intelligent methods to provide personalized automated scoring to students based on their performance in solving mathematical or business problems within Microsoft Excel. On the other hand, the *HyperManyMedia*[2] platform provided recommendations to students based on their previous browsing activities; these recommendations were based on artificial intelligence algorithms where ontology is defined and semantic web is utilized to provide the most accurate recommendations to students based on their level within the course. Prior research describes this process in more detail (Zhuhadar, 2015; Zhuhadar, Carson, Daday, & Nasraoui, 2015; Zhuhadar & Nasraoui, ; Zhuhadar, Nasraoui, & Wyatt, 2007; Zhuhadar, Nasraoui, & Wyatt, 2009a, b; Zhuhadar, Nasraoui, Wyatt, & Romero, 2009; Zhuhadar, Nasraoui, Wyatt, & Yang, 2010; Zhuhadar & Yang, 2012).

However, in this research we utilize a platform (Project LISTEN[3]) designed and developed by researchers at Carnegie Mellon University.[4] Many studies have been conducted on Project LISTEN (Huang & Mostow, 2015a, b; Mostow & Prieditis, 2014; Yuan, Chang, Taylor, & Mostow, 2014), an intelligent tutoring system (ITS) used to analyze educational learning through case analysis of students' interactions. Many of these studies required specific details such as student choices, timing intervals, student outcomes (predictions), classifiers, and types of help given. The complexity of these details can make them laborious to browse and gather. (Beck, Chang, Mostow, & Corbett, 2008) address three simple techniques to make data mining easier to interpret, stating researchers can directly store interactions and index them into a database, which allows ease of access without the need to browse log files.

Beck et al. (2008) also addressed a method to identify a tutorial event by linking the student, computer, and time interval together, as well as, restraining time intervals to define tutorial interactions

* Corresponding author.
E-mail addresses: leyla.zhuhadar@wku.edu (L. Zhuhadar), sm14as@my.fsu.edu (S. Marklin), evelyn.thrasher@wku.edu (E. Thrasher), mlytras@acg.edu (M.D. Lytras).
URL: http://www.acg.edu

[1] http://www.wku.edu.

[2] http://hmm.wku.edu/.

[3] https://www.cs.cmu.edu/~listen/.

[4] http://www.cmu.edu/.

http://dx.doi.org/10.1016/j.chb.2016.02.073
0747-5632/© 2016 Elsevier Ltd. All rights reserved.

within a hierarchical structure. The efficiency, generality, usability, and utility of a session browser aids in the successful facilitation of data mining efforts and will continue to be used in future research.

Cen, Koedinger, & Junker (2006) looked at improving the analysis of intelligent learning systems by focusing on the cognitive model, a set of predefined rules that influence the process of student problem solving tailored to provide helpful feedback and hints while increasing difficulty to improve student learning and knowledge. In their research, three questions were asked regarding ITS: 1) how can researchers describe learning behavior in existing cognitive models; 2) can a learning rate be established for the student; and 3) how can the cognitive model be improved inexpensively by defining measures of complexity to improve the curriculum for various learning styles (Cen et al., 2006). The researchers proposed the Learning Factor Analysis, a semi-automated method used in Java that combines statistics, human experience, and a combinatorial search (heuristic guidance) to add to tutor development, giving better insight into the analysis of data and log files through a knowledge-tracing algorithm (Cen et al., 2006).

While ITS often focus on learner control, the power of use is given to the student. Therefore, the student decides when and how to use the system's resources, essentially self-monitoring and judging when/if they can benefit from the help provided. (Aleven & Koedinger, 2000), using the PACT geometry tutor, suggested that students often times do not have the required cognitive skills to take advantage of the resources available through the tutor. They argue for a meta-cognitive help-seeking model that can monitor the student's strategies in using resources provided from the tutor to lend the most support for on-request help or glossary access (Aleven & Koedinger, 2000). Results indicated that students used the tutor's intelligent help facilities (hints) more frequently than the non-intelligent resources (glossary).

Aleven & Koedinger (2000) argue that the meta-cognitive help-seeking model could be implemented as a production rule model and could be used for model tracing, taking into account the student model information and whether a student might be over-using or under-using resources. With the meta-cognitive strategy, the tutor would make greater use of the glossary, a low cost resource, to find relevant information and apply it to the current problem. The tutor would also initiate intelligent help after two errors, thereby reducing the overall number of errors and the time required. They state that “in order for an intelligent tutor system to be adaptive”, meta-cognitive skills must be taken into consideration to produce better learners (Aleven & Koedinger, 2000).

Regarding response intervals, Joseph (2005) further addresses the issue of time as an indicator and predictor of how much a student learns. Previous research conducted by Aleven and Koedinger (2000) indicated that students do not always try their best in solving problems; therefore, Joseph (2005) proposed an engagement tracing model that would better model student engagement by primarily focusing on disengagement. This approach would analyze the response times and correctness of responses to model overall engagement while using an intelligent tutor. The method is inexpensive and sensitive enough to detect temporal changes during the student's interactions with the tutor (Joseph, 2005). By modeling a student's engagement, research can predict how much an individual will benefit from using intelligent tutors, while allowing for modifications that adapt to student interactions, for greater learning efficiency.

Lallé, Mostow, Luengo, & Guin (2013) noted the challenge in evaluating student models by their impact on the success of an intelligent tutor's decision about which type of help to offer students. Individualized help can have a strong impact on learning; therefore, the better the tutor can adapt its help to the student and situation, the more likely the student will learn from it. Using logs of randomized tutorial decisions and ensuing student performance, Lallé et al. (2013) trained a classifier to predict tutor decision outcomes (success or failure) based on situational features, such as student and task. Using historical data to simulate a policy by extrapolating its effects from the subset of randomized decisions that happened to follow the policy, the authors tested the method on data logged by Project LISTEN's reading tutor, which randomly chooses what type of help to give (Lallé et al., 2013). They also compared the impact of student models (knowledge-tracing model, constraint based model, and control based approach) on the expected success of tutorial decisions (greatest probability) to offer help. Using the learner policies to pick which type of help yields the greatest probability of success, taking into consideration the types of help available, student features, domain features and the student model, the measure has greater utility for measuring student learning. (Lallé et al., 2013) found that all learned policies tested improved the reading tutor's expected success compared to its original randomized decisions. Yet, this only applies to tutors that make decisions based on multiple types of available help. Furthermore, (Beck & Mostow, 2008) assessed learning decomposition to examine how much students learn from instruction. Learning decomposition determines the relative adequacy of different types of learning opportunities using a generalization of the learning curve analysis with non-linear regression. The authors suggested that students learn words better when they read a wide selection of stories rather than reading the same story multiple times (Beck & Mostow, 2008). Reading new stories, thereby expanding the exposure to words, is good for long-term learning.

Beck and Mostow's model further indicated that when students reread words, the effectiveness of learning that word decreased, supporting the argument that students benefit less from mass practice (2008). Individuals who benefited from mass practice and repeated reading were older, less proficient readers who were tagged as requiring learning support. As (Beck & Mostow, 2008) indicated, learning factor analysis, as noted earlier by Cen et al. (2006), is used to create better fitting learning curves; and learning decomposition (focused on determining impact of practice) is concerned with greater understanding of student learning potential. González-Brenes & Mostow (2011) assessed the prediction value of models by using a regularized logistic regression, arguing that conventional classifier learners require large amounts of data to avoid over-fitting and do not generalize well to unseen examples (predictions). Using regularized logistic regression makes it feasible to classify dialogues in a high dimensional space and to demonstrate on real data from Project LISTEN's Reading Tutor (González-Brenes & Mostow, 2011). One classifier predicts task completion to characterize differences in the behavior of children when they choose the story they read (71% accuracy), while another classifier (73.6% accuracy) infers who chose the story based on dialogue. Their approach solved two problems in classifying children's oral reading dialogue, predicting which stories they would finish and characterizing the student behavior according to who chose the stories. They achieved a 71.1% and a 73% cross validated classification accuracy on a balanced set of data from unseen students, indicating that regularized logistic regression is the best for assessing these problems in prediction (González-Brenes & Mostow, 2011).

As previous research has shown, several models have been tested to assess how and when students learn and whether or not tutor help is effective in increasing student learning. In assessing the findings, some researchers have called for a unified framework that simultaneously allows both the skills and impact of practice to vary (Beck & Mostow, 2008). Some researchers have addressed the cost of analyzing ITS data (Yuan et al., 2014), suggesting an

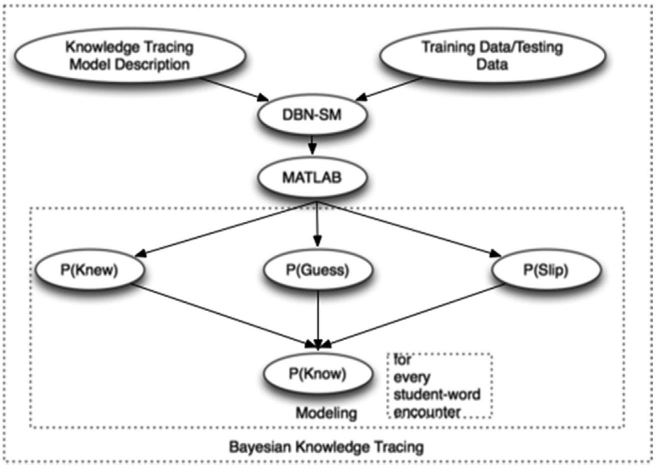

Knowledge Tracing Model Description
Training Data/Testing Data
DBN-SM
MATLAB
P(Knew)
P(Guess)
P(Slip)
P(Know)
for every student-word encounter
Modeling
Bayesian Knowledge Tracing

Fig. 1. Bayesian knowledge tracing.

inexpensive EEG model that can assess information about comprehension, but noting that future work needs to increase prediction accuracy by assessing other dimensions of knowledge and applying those assessments to improve learning outcomes. Other research has examined modeling dialogue, arguing that high dimensional space opens the door from small manually generated sets of features to richer, automatically generated sets of features, thus improving the ability to assess student learning outcomes with regard to ITS (González-Brenes & Mostow, 2011).

Lallé et al. (2013) compared the impact of student models (knowledge tracing model, constraint based model, and control based approach) on the expected success of tutorial decisions to offer help through learner policies. These policies are vulnerable to under-covering and over-fitting; therefore, more accurate student models such as LR-DBN and stronger classifiers such as Support Vector Machines (SVM) or Random Forests should be used to improve the prediction accuracy of successful help in ITS. While there is still debate on what method works best, there seems to be consensus on the Bayesian Evaluation and Assessment approach (Beck et al., 2008), which assesses both student and tutorial interventions, allowing students to transfer knowledge gained to later problems as a model for predicting learner outcomes and learner factor analysis (Cen et al., 2006). Furthermore, Beck et al.'s research on the two effects of help, scaffolding immediate performance and boosting actual learning, argues that evaluations of synthetic data are the most promising (2008).

During *The 10th Annual LearnLab Summer Research School,*[5] *at Pittsburg Science of Learning Center*, we acquired a large dataset from Project LISTEN.[6] This dataset consisted of students' knowledge tracing while interacting with the system. This data provided information about 90,000 student-word encounters. Each “word” is considered a skill that can be known, but that knowledge is hidden from us. We used Bayes Net Toolbox for Student Modeling (BNT-SM[7]) to facilitate the use of dynamic Bayes nets in the student modeling community. In this study, we argue that boys interact with Project LISTEN's Intelligent Tutoring System differently than girls. This finding was not only evident by using *Bayesian Knowledge Tracing* but also by using *Learning Curve Analysis*. In the following paragraphs, we discuss our findings in detail.

2. Comparative study

A challenge to evaluating ITS is to evaluate student models by their impact on the success of help given. Which type of help do students really need? Is the help better addressed on an individualized or mass practice basis? Beck & Mostow (2008) looked at three approaches for evaluating this issue, including experimental trials, learning decomposition, and Bayesian Evaluation and Assessment (using Bayesian networks). The model controlled for student knowledge while estimating the intervention's effectiveness. This is similar to item response theory, which enables better comparisons of students across groups by estimating student proficiency and question difficulty. Findings indicated that the only method that helped students with long-term learning was the Bayesian Evaluation and Assessment approach, which assesses both student and tutorial interventions, allowing students to transfer knowledge gained to later problems (Beck et al., 2008). Interestingly, the other two actually hurt student learning – a negative consequence. Furthermore, the authors indicated that if a student does not know the skill, they are more likely to generate a correct response with help than without. This supports the idea that tutor help has a scaffolding effect on assisting immediate performance; but the teaching effect is more beneficial in the long run than the scaffolding effect, further showing that student knowledge and student performance are indeed affected by tutor help.

Prior research was conducted on Project LISTEN (Yuan et al., 2014), showing that reading comprehension assessment can be costly and obtrusive. To address these issues, the research analyzed the efficacy of EEG devices for assessing reading comprehension (the correctness of responses) in the classroom, taking into consideration that these devices are unobtrusive and low cost. Yuan et al. (2014) used EEG signals to produce above-chance predictors of student performance. Their findings indicated that the classifier achieved significantly above-chance accuracy trained on only the reading portion; but there were no above-chance predictions made by the model. Therefore, no conclusions could be drawn about different binary distinctions (Yuan et al., 2014). While the EEG model could not successfully make above-chance predictions, it does suggest that some information regarding student comprehension can be teased out by using EEG devices (Yuan et al., 2014).

In our study, we use the same methods proposed by González-Brenes & Mostow (2011) to assess student learning outcomes with regard to ITS. We also look at three approaches for evaluating this issue, including experimental trials, learning decomposition, and Bayesian Evaluation and Assessment; however, our goal is to examine if there is a significant difference in the way students interact with the tutoring system. More specifically, is there a gender difference in these interactions? If so, does this support the idea of gender-specific tutoring system designs?

The sections that follow provide details of our framework, our research questions, our findings, and our future research.

3. Proposed research framework and data analysis

Project LISTEN[8] logs its interactions with children directly into a database. These interactions are archived at multiple grain sizes ranging from sessions to stories, from sentences to utterances, from individual words to mouse clicks and key presses. The Reading Tutor administers within-subject randomized controlled trials by selecting randomly among alternative tutorial actions; the

[5] http://www.learnlab.org/opportunities/summer/.
[6] http://www.cs.cmu.edu/~listen/pubs.html.
[7] http://www.cs.cmu.edu/~listen/BNT-SM/.
[8] http://www.cs.cmu.edu/~listen/pubs.html.

Table 1
Excerpt from ***Evidence.xls***.

Id	Utterance start time	Utterance_sms	Target_word_number	Skill
fBS7-7-1990-02-03	5/12/00 17:52	640	8	WORLD
…				
fDL7-5-1993-11-28	10/13/04 13:56	421	13	WORLD
Help	**Knowledge**	**asr_accept**	**Confidence score**	**asr_confidence**
1	1	2	.0668763	1
…				
2	2	2	.0430581	1

Table 2
Excerpt from ***Param_table.xls***.

Skill	Num of users	Num of cases	ll	L1
skill_HELLO	14	23	−3.609	.744
…				
skill_WORLD	46	218	−90.177	.695
Skill	**Guess**	**Slip**	**t**	**Forget**
skill_HELLO	.721	.000005	.982517	.000001
…				
skill_WORLD	.634	.113612	.256071	.000001

Table 4
Effect size between genders.

	Male	Female	Effect size (d')	*P_value*
Knew	.41 (.03)	.44 (.04)	.580	<.0001
Learn	.19 (.03)	.20 (.04)	.300	<.0001
Guess	.70 (.04)	.71 (.04)	.507	<.0001
Slip	.07 (.03)	.06 (.03)	−.032	<.0001

resulting experiments have had as many as 180,000 trials. In addition, the logged data includes text and speech, with some gaze and EEG as well.

We used the Session Browser to explore individual interactions and MySQL queries to aggregate them. Finally, we used tools such as Matlab and SAS to run statistical analysis and machine learning algorithms on the resulting data. We used various approaches to predict whether a child would finish a story or not. For instance, we analyzed the relative value of different types of reading practice for oral reading fluency; and we compared the efficacy of different types of help being used by each child on hard words.

Dynamic Bayes Nets (DBNs) provide a powerful way to represent and reason uncertainty in time series data and are, therefore, well-suited to model a student's changing knowledge state during skill acquisition (Murphy, 2001).

Table 3
Excerpt from ***inference_result.xls***.

	Utterance start time	Utterance_sms	Target_word_number	Skill
fCA8	2004-11-09/11:25:24	218	1	HELLO
…				
fCA7	2005-01-24/11:33:12	468	1	HELLO
Help	**Knowledge**	**asr_accept**	**Confidence score**	**asr_confidence**
1	.801846	2	.124751	1
…				
1	.997498	2	.0785152	1

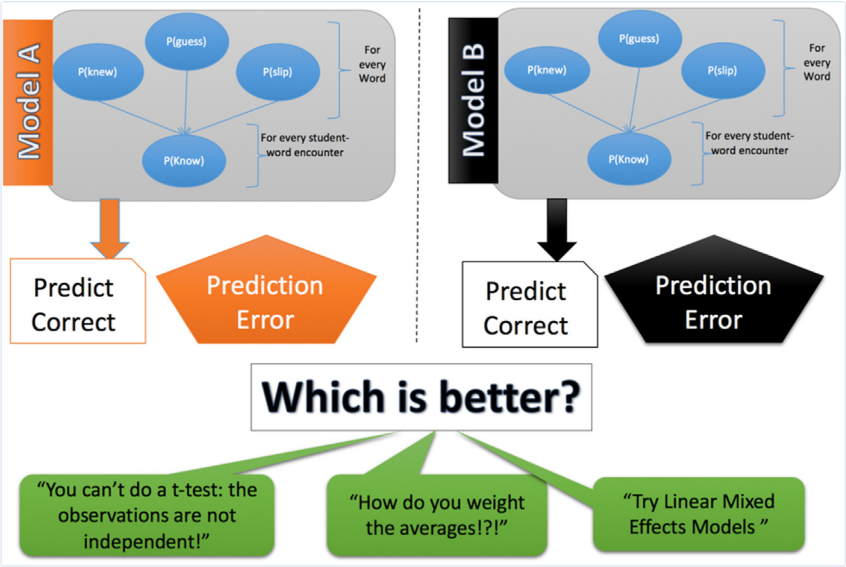
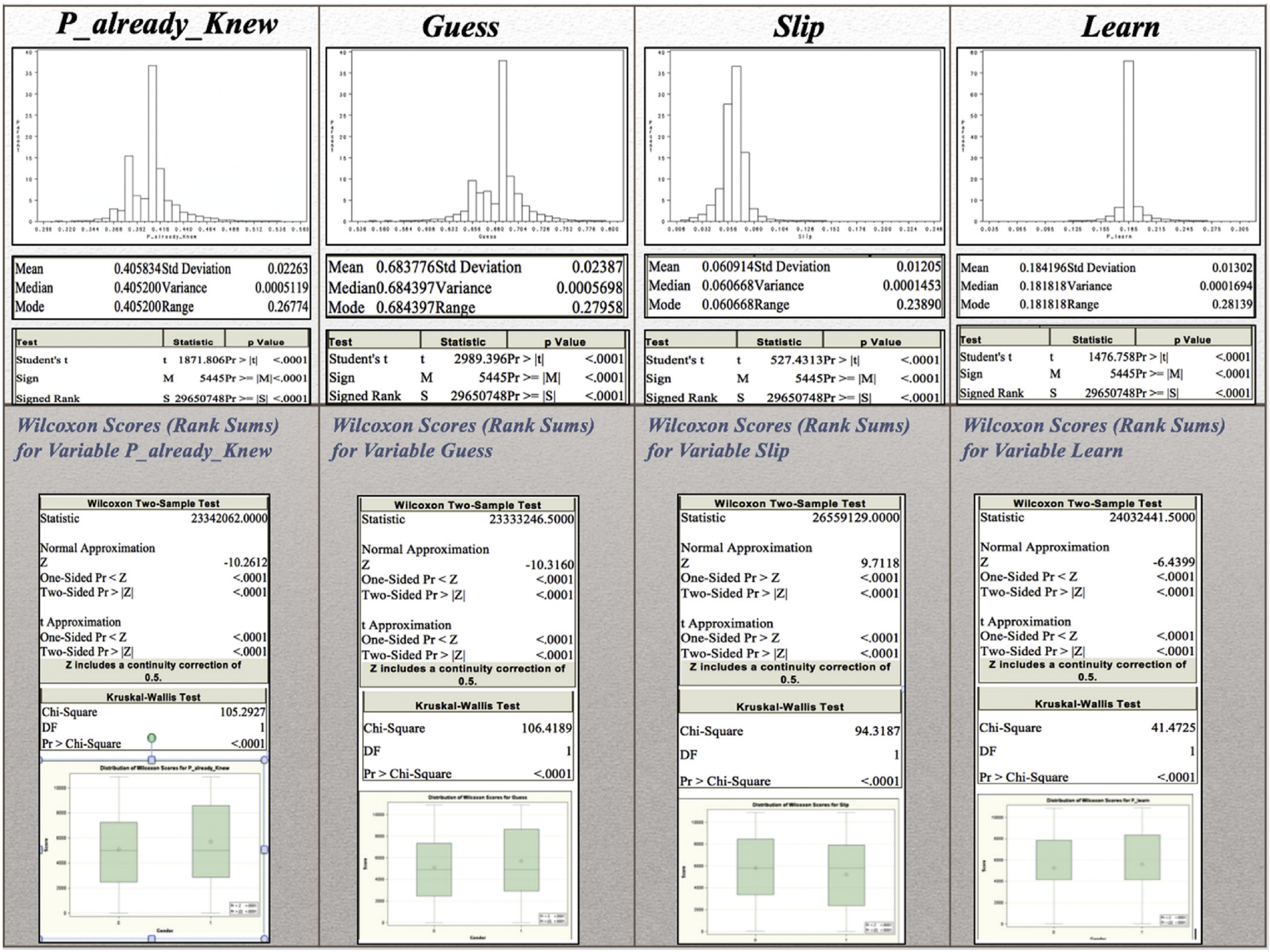

Model A
P(guess)
P(knew)
P(slip)
For every Word
P(Know)
For every student-word encounter
Predict Correct
Prediction Error
Model B
P(guess)
P(knew)
P(slip)
For every Word
P(Know)
For every student-word encounter
Predict Correct
Prediction Error
Which is better?
"You can't do a t-test: the observations are not independent!"
"How do you weight the averages!?!"
"Try Linear Mixed Effects Models "

Fig. 2. Proposed research framework.

Table 5
Wilcoxon Scores.


P_already_Knew
Guess
Slip
Learn
P_already_Knew
Guess
Slip
P_learn
Distribution of Wilcoxon Scores for P_already_Knew
Distribution of Wilcoxon Scores for Guess
Distribution of Wilcoxon Scores for Slip
Distribution of Wilcoxon Scores for P_learn
Score
Gender

P_already_Knew

Mean	0.405834	Std Deviation	0.02263
Median	0.405200	Variance	0.0005119
Mode	0.405200	Range	0.26774

Test	Statistic		p Value	
Student's t	t	1871.806	Pr > |t|	<.0001
Sign	M	5445	Pr >= |M|	<.0001
Signed Rank	S	29650748	Pr >= |S|	<.0001

Wilcoxon Scores (Rank Sums) for Variable P_already_Knew

Wilcoxon Two-Sample Test	
Statistic	23342062.0000
Normal Approximation	
Z	-10.2612
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	105.2927
DF	1
Pr > Chi-Square	<.0001

Guess

Mean	0.683776	Std Deviation	0.02387
Median	0.684397	Variance	0.0005698
Mode	0.684397	Range	0.27958

Test	Statistic		p Value	
Student's t	t	2989.396	Pr > |t|	<.0001
Sign	M	5445	Pr >= |M|	<.0001
Signed Rank	S	29650748	Pr >= |S|	<.0001

Wilcoxon Scores (Rank Sums) for Variable Guess

Wilcoxon Two-Sample Test	
Statistic	23333246.5000
Normal Approximation	
Z	-10.3160
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	106.4189
DF	1
Pr > Chi-Square	<.0001

Slip

Mean	0.060914	Std Deviation	0.01205
Median	0.060668	Variance	0.0001453
Mode	0.060668	Range	0.23890

Test	Statistic		p Value	
Student's t	t	527.4313	Pr > |t|	<.0001
Sign	M	5445	Pr >= |M|	<.0001
Signed Rank	S	29650748	Pr >= |S|	<.0001

Wilcoxon Scores (Rank Sums) for Variable Slip

Wilcoxon Two-Sample Test	
Statistic	26559129.0000
Normal Approximation	
Z	9.7118
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr > Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	94.3187
DF	1
Pr > Chi-Square	<.0001

Learn

Mean	0.184196	Std Deviation	0.01302
Median	0.181818	Variance	0.0001694
Mode	0.181818	Range	0.28139

Test	Statistic		p Value	
Student's t	t	1476.758	Pr > |t|	<.0001
Sign	M	5445	Pr >= |M|	<.0001
Signed Rank	S	29650748	Pr >= |S|	<.0001

Wilcoxon Scores (Rank Sums) for Variable Learn

Wilcoxon Two-Sample Test	
Statistic	24032441.5000
Normal Approximation	
Z	-6.4399
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	41.4725
DF	1
Pr > Chi-Square	<.0001

Table 6
Comparison between models.

Model	Student N	Word N	Prior knowledge	Transfer	Guess	Slip	p (Know)	p (Cn)
Female specific	33	4811	.41 (.03)	.19 (.01)	.69 (.03)	.06 (.01)	.61 (.23)	.85 (.07)
Male specific	30	3322	.40 (.02)	.18 (.01)	.68 (.02)	.06 (.01)	.50 (.21)	.81 (.06)
Aggregated female	33	5199	.41 (.03)	.19 (.02)	.69 (.03)	.06 (.01)	.61 (.23)	.85 (.07)
Aggregated male	30	4395	.41 (.03)	.19 (.02)	.69 (.03)	.06 (.02)	.60 (.22)	.85 (.06)
Aggregated	63	5829	.41 (.05)	.19 (.05)	.69 (.22)	.06 (.03)	.60 (.22)	.85 (.06)

Table 7
Weighted average per word encounter.

	Accepted	Fluency	Help seeking
Male	.88	.8	.03
Female	.88	.8	.03

Accordingly, we used Knowledge Tracing (KT) for our student modeling toolbox.[9] The goal was to assess the student's knowledge from his or her observed actions at each successive opportunity to apply a skill by using the updated *Knowledge Tracing* estimated probability that a student knows the skill. These updates are associated with skill-specific learning, performance parameters, and the observed student performance (evidence).

Fig. 1 shows a diagram of DBN-SM workflow. Three parameters are used to estimate the probability that a student knows the encountered word, represented in the diagram as *P(Know/Learn)*. These parameters are: *P(already Knew)*, *P(Guess)*, *and P(Slip)*. In the next sections, we present BNT-SM outputs and our research question.

[9] https://code.google.com/p/bnt/.

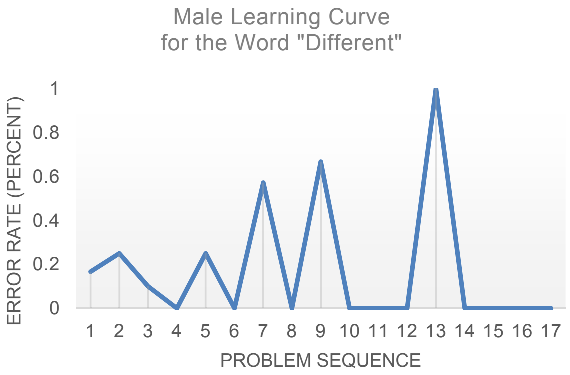


Fig. 3. Male learning curve.

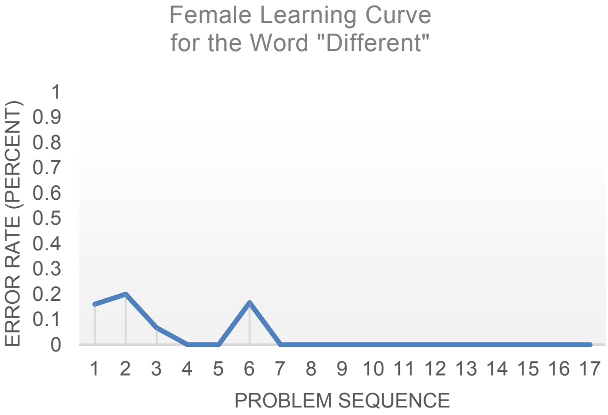


Fig. 4. Female learning curve.

3.1. Generated output

Evidence.xls consists of Bayes Net data for each skill from all children. Evidence data are comprehensive. Hidden variables and missing observations are marked with NULL. For discrete variables, the values cannot be 0 because Matlab uses array subscripting (starting with 1). Therefore, we often increment the discrete variables by 1. In the case of a binary variable, 1 is used for 0 or false values, whereas 2 is used for 1 or true values, as shown in Table 1.

Param_table.xls consists of **BNT-SM** estimates for skill-specific models where a Bayes Net is trained for each skill in the training dataset, as shown in Table 2.

Inference_result.xls has a format identical to that of **evidence.xls**, except that BNT-SM performs inferences on the hidden variables and estimates their values as follows:

- For binary hidden variable X, the estimated value will be the probability of X = 1 (represented X = = 2 in Matlab).
- For discrete hidden variables Y with values greater than 2, the probability of all values will be output in the form of [p_1; p_2; … p_n;]
- By default, BNT-SM infers posterior probability (after observing the evidence)
- If instead, you want to infer prior probability (before observing the evidence, e.g. classic Knowledge Tracing), you can make a switch in ***RunBnet.m*** when calling ***inference_bnet.m***, as shown in Table 3.

4. Proposed research framework

Fig. 2 illustrates our research framework using two models: **Model A** (boys) and **Model B** (girls). Our null hypothesis states that there is no difference between these two models. Therefore, if we compare the knowledge tracing parameters, *P(Know/Learn)*, *P(already Knew)*, *P(Guess)*, *and P(Slip)*, between theses two models, we should not find a significant difference.

4.1. Research question?

- *Is there a difference between girls (female) and boys (male) in their ways of interacting with the intelligent tutoring system?*

5. Experiments

5.1. Data set

We queried a dataset of **student-word encounters** (a balanced dataset of easy words and hard words). For complete results, refer to these links (male results,[10] female results[11]).

5.2. Building models

We used BNT-SM to generate these four training parameters for each student-word encounter: *P(Know/Learn), P(already Knew), P(Guess), and P(Slip).*

Table 4 shows how our population holds a significant effect size between genders for these four training parameters (please refer to the *P_value* associated with each parameter).

We used the ***Wilcoxon signed-rank test*** instead of *t-test* since our observations are not independent. The Wilcoxon signed-rank test is a nonparametric test used to compare two sets of scores that come from the same participants. This can occur when we wish to investigate any change in scores from one point in time to another, or when individuals are subjected to more than one condition. For more information, refer to the resource below.[12] The Wilcoxon Scores of *P(Know/Learn), P(already Knew), P(Guess), and P(Slip)* are reported in Tables 5 and 6.

Table 6 shows the number of encountered student-words for each model. In addition, we listed the average value of each parameter. For more information about this dataset, refer to the resource below.[13] In addition, Table 7 provides the weighted average per word encounters for boys (male) and girls (female).

6. Findings

6.1. Bayesian Knowledge Tracing

Despite the observation of females encountering more words, we found that the means and standard deviations do not appear to differ between the gender-specific model and the aggregated model. However, a difference is seen across genders within the gender-specific model, as shown in Table 6. When comparing the percentage of a correctly spoken word (accepted), the rate of familiarity the student has with a word (fluency), and the percentage

[10] https://www.dropbox.com/home/learnlab2014/bnt-sm/results/comparison_results/male_model.

[11] https://www.dropbox.com/home/learnlab2014/bnt-sm/results/comparison_results/female_model.

[12] https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php.

[13] https://www.dropbox.com/sh/0hc030ct4u2p08j/AAB6WiLh-1AOE7NFYeNblZ5za?dl=0.

of help seeking for a word, it appears there are no differences in the weighted averages between genders, as shown in Table 7.

On the other hand, by looking at Wilcoxon Scores, seen in Table 5, we see that our population holds a significant effect size between gender for these four training parameters (Z-score[14]). More specifically, we can report the following findings:

1. *The P_already Knew, P_Guess, P_Know/Learn for girls (female) are significantly higher than boys (male), and*;
2. *The P_Slip for boys (male) is significantly higher than girls (female).*

6.2. Learning curve analysis

In addition to investigating *Bayesian Knowledge Tracing*, we compared the *Learning Curve Analysis* for a specific word "**Different**" between male and female, as shown in Figs. 3 and 4. Comparing these figures, we see a significant difference between boys and girls in language acquisition. They learn the same words in a different way. Boys tend to have a high error rate percent. In addition, boys tend to need more time (more problems to solve) to excel. Of course, this analysis needs further research, such as observing these patterns on different types of words (easy vs. hard) and bigger sets of words. In addition, we should also consider the context in which these words appear. In general, we conclude that a difference exists between genders in the way children interact with the tutoring systems. Therefore, we suggest a gender specific tutoring system where, for example, the reading materials are correlated with gender type.

We would assume that a boy would be more interested in reading a story about car races, whereas a girl would be more interested in reading princess stories. Knowing there is a difference does not mean that girls are better than boys regarding reading comprehension. Rather, our findings suggest that there is a difference in their way of learning; and we should embrace this difference to effectively promote greater reading comprehension by providing reading materials that would be of interest to them.

7. Conclusion and future works

The LearnLab Summer Research School[15] at the *Pittsburgh Science of Learning Center* provides researchers with opportunities to access large datasets from various projects dealing with, but limited to: cognitive psychology or educational psychology, computer-supported collaborative learning, development of technology-enhanced course content, and analysis of student data or educational data mining. In this research, we used a large dataset generated by students using *Project LISTEN*.[16] This dataset consisted of students' knowledge tracing while interacting with the system. This data provided information about 90,000 student-word encounters. We found that boys (males) interact with *Project LISTEN's* ITS differently than girls (females). This finding was not only evident by using the *Bayesian Knowledge Tracing* but also by using *Learning Curve Analysis*. While ITS often focus on the student decisions regarding when and how to use the system's resources, we suggest further analysis and monitoring are required to get the best results from these systems. Considering gender differences in the way students learn, in addition to the context of the reading materials presented, is essential, especially for boys. In this study, we argue that boys learn vocabulary differently than girls. This finding was evident in both *Bayesian Knowledge Tracing* and in *Learning Curve Analysis*. Our future work will include an extension of the learning curve analysis considering *Linear Mixed Effects Models*.

[14] If you have a large number of participants, you can convert Wilcoxon into a z-score.

[15] http://www.learnlab.org/opportunities/summer/.

[16] http://www.cs.cmu.edu/~listen/pubs.html.

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: do students know when they need help?. In *Paper presented at the intelligent tutoring systems*.

Beck, J. E., Chang, K.-m., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the Bayesian evaluation and assessment methodology. In *Paper presented at the intelligent tutoring systems*.

Beck, J. E., & Mostow, J. (2008). How who should practice: using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Paper presented at the intelligent tutoring systems*.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis—a general method for cognitive model evaluation and improvement. In *Paper presented at the intelligent tutoring systems*.

Crews, T., & Murphy, C. (2007). *CaseGrader: Microsoft office Excel 2007 casebook with autograding technology*. Course Technology Press.

González-Brenes, J. P., & Mostow, J. (2011). Classifying dialogue in high-dimensional space. *ACM Transactions on Speech and Language Processing (TSLP), 7*(3), 8.

Huang, Y.-T., & Mostow, J. (2015a). Evaluating human and automated generation of distractors for diagnostic multiple-Choice cloze questions to assess children's reading comprehension. In *Paper presented at the artificial intelligence in education: 17th International Conference, AIED 2015*. Madrid, Spain, June 22-26, 2015. Proceedings.

Huang, Y.-T., & Mostow, J. (2015b). Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension. In *Paper presented at the artificial intelligence in education*.

Joseph, E. (2005). Engagement tracing: Using response times to model student disengagement. In *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology* (Vol. 125, p. 88).

Lallé, S., Mostow, J., Luengo, V., & Guin, N. (2013). Comparing student models in different formalisms by predicting their impact on help success. In *Paper presented at the artificial intelligence in education*.

Mostow, J., & Prieditis, A. E. (2014). Discovering admissible search heuristics by abstracting and optimizing. *Machine Learning Proceedings, 1989*, 240.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing Science and Statistics, 33*(2), 1024–1034.

Yuan, Y., Chang, K.-m., Taylor, J. N., & Mostow, J. (2014). Toward unobtrusive measurement of reading comprehension using low-cost EEG. In *Paper presented at the Proceedings of the Fourth International Conference on learning analytics and knowledge*.

Zhuhadar, L. (2015). A synergistic strategy for combining thesaurus-based and corpus-based approaches in building ontology for multilingual search engines. *Computers in Human Behavior, 51*, 1107–1115.

Zhuhadar, L., Carson, B., Daday, J., & Nasraoui, O. (2015). A universal design infrastructure for multimodal presentation of materials in STEM programs: universal design. In *Paper presented at the Proceedings of the 24th International Conference on World Wide Web, Florence, Italy*.

Zhuhadar, L., & Nasraoui, O.. Personalized cluster-based semantically enriched web search for E-learning. Paper presented at the CIKM 08: ONISW the 2nd International workshop on ontologies and information systems for the semantic web.

Zhuhadar, L., & Nasraoui, O. (2008). Semantic information retrieval for personalized E-learning. In , *20th IEEE International Conference on ICTAI '08: Vol. 1. Tools with artificial intelligence, 2008* (pp. 364–368).

Zhuhadar, L., Nasraoui, O., & Wyatt, R. (2007). *Knowledge mining for adaptive multimedia web-based educational platform*.

Zhuhadar, L., Nasraoui, O., & Wyatt, R. (2009a). Automated discovery, categorization and retrieval of personalized semantically enriched E-learning resources. In *International Conference on semantic computing, 0* pp. 414–419).

Zhuhadar, L., Nasraoui, O., & Wyatt, R. (2009b). Dual representation of the semantic user profile for personalized web search in an evolving domain. In *Paper presented at the Proceedings of the AAAI 2009 Spring Symposium on social semantic web, where Web 2.0 meets Web 3.0*.

Zhuhadar, L., Nasraoui, O., Wyatt, R., & Romero, E. (2009). Multi-model ontology-based hybrid recommender system in e-learning domain. In *Paper presented at the Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on web intelligence and intelligent agent technology* (Vol. 03).

Zhuhadar, L., Nasraoui, O., Wyatt, R., & Yang, R. (2010). Visual knowledge representation of conceptual semantic networks. *Social Network Analysis and Mining*, 1–11.

Zhuhadar, L., & Yang, R. (2012). The impact of social multimedia systems on cyberlearners. *Computer in Human Behaviors*.